A Northern City Going Elsewhere:

Apparent and Real-Time Sound Change in Ogdensburg, New York

A dissertation
submitted to the Faculty of Humanities at the University of Bern
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Anja Thiel
May 2019

Supervisors:

Prof. Dr. David Britain,
University of Bern

Prof. Dr. Aaron J. Dinkin,
San Diego State University

Abstract

The *Atlas of North American English* (Labov, Ash, & Boberg, 2006) found that dialect diversity in North America was increasing, via the continuing advancement of regional sound changes such as the Northern Cities Shift (NCS). In the decade since the *Atlas*'s publication, however, indications have emerged that that conclusion was premature, with multiple studies finding retreat from the NCS in communities where it was expected to be stable or advancing (e.g. Wagner, Mason, Nesbitt, Pevan, & Savage, 2016; Driscoll & Lape 2015). This dissertation reports a real-time study demonstrating that the loss of NCS can be rapid indeed.

This study examines the loss of the NCS in Ogdensburg, a small city in rural Northern New York, on the Canadian border. On the basis of nine speakers interviewed there in 2008, Dinkin (2009, 2013) described Ogdensburg as the northeasternmost limit of the NCS, with some evidence that the NCS was advancing in apparent time. Furthermore, the data suggested an incipient merger of the low back vowels LOT and THOUGHT in the community, a feature that has been believed to be incompatible with the NCS (Labov et al., 2006). In this study, I compare those nine speakers interviewed in 2008 with a new sample of 39 speakers from the same city that I interviewed in 2016, and supplement speech production data with social perception data.

The results suggest that, in the eight years between 2008 and 2016, the NCS apparently disappeared from Ogdensburg, a change that is visible in nearly all phonemes of the NCS. It appears that the community is orienting toward a new system, the Elsewhere Shift, including the merger of LOT and THOUGHT, a development that has been reported in other traditional NCS communities as well (e.g. Wagner et al., 2016). It seems that increasing negative evaluation of at least two NCS features may have been the driving force behind the abandonment of the NCS in Ogdensburg, and the points in apparent-time at which the changes emerge in the data suggest that these evaluations and the consequential restructuring of the community's vowel system might be a response to a myriad of social changes in the community in the second half of the 20th century (Coupland, e.g. 2009).

Acknowledgements

Although my name graces the front page of this thesis, it was, in many ways, a collaborative effort, and I would like to take this opportunity to express my sincere gratitude to the individuals and institutions that have helped me to get to this point.

This project was funded partially by the Bern University Research Foundation, the Faculty of Humanities, the GSH and the CSLS at the University of Bern, and I'm grateful for their generous support over the years.

I would like to express the utmost appreciation to my supervisor Dave Britain for bringing me to beautiful Bern as his PhD student and assistant, thereby giving me the opportunity to pursue a research project that had been lingering in the back of my mind for years; for expanding my sociolinguistic horizon with his insightful and constructive feedback on my chapter drafts; and for his reassuring and encouraging comments from beginning to end.

I owe my deepest gratitude to my second supervisor Aaron Dinkin for *his* PhD dissertation that got me thinking about mine 10 years ago; for his endless patience in explaining statistics and discussing our data (and, of course, for letting me use his data); for his invaluable advice, comments and suggestions on my ideas, without which this dissertation would not have turned out the way it did; for getting me to Toronto; for getting me through NWAV; and (in advance) for getting up early for my defense.

Special thanks also to Anne Fabricius, my wonderful MA supervisor, for allowing me to explore sociophonetics in a program that didn't really have room for it and for spending countless hours guiding me in the process; for her faith in my academic potential when I was lacking it; for encouraging me to pursue a PhD; and for forwarding that faithful call for applications for my current position in Bern.

I want to thank Kathryn Campbell-Kibler and Adrian Leemann for helping me make the best of my matched guise experiment when I had no idea what I was doing, and Jo Fruehwald for his speedy assistance with FAVE when the website crashed in the midst of data processing.

My deepest heartfelt appreciation goes to everyone in New York who's helped me before, during and after my fieldwork in the North Country: to the silly college kids at SUNY Potsdam who made fun of the sound of my name /ɑnjə/, constantly changing it to /ʌndər jə/ (which only works in the North), and who are responsible for my ensuing obsession with dialects in New York; to Krista Chisholm LaVack, Sara Trimm, Allonah Ezro-Christy and the Lawrence Family for their companionship and for helping me find additional informants; and particularly to my participants in Ogdensburg, without whom this project would never have been possible. Special thanks to the participants identified in this dissertation by the pseudonyms Richard, Amber, Sarah, and Monica for all their efforts that went way beyond just being interviewed, as well as to the Ogdensburg Public Library for allowing me to use their Local History Room for my interviews.

I would like to express my gratitude to my friends, fellow grad students and coworkers for having contributed to this project in one way or another throughout the past 3 ½ years. At UniBe, Gwynne, Domi, Sara and Tobi, and in Toronto, Melanie, Katharina, Savannah, Lex, Radu and Jessica. I would particularly like to thank Eva and Hannah for all our enlightening discussions, their warm support and encouragement, and their much-appreciated feedback on my (often shitty) drafts. Special thanks to Eva also for being my partner in crime during project finish line, and, very importantly, for being Dr. Pepper's *favorite* dog sitter when I needed to clear my head on the slopes. And to Hannah, a very special thank you for taking on the role of my trusted human thesaurus and pun generator, and, well… for being *Hanja* with me and everything it entails. My time at UniBe certainly wouldn't have been the same without my office-mate (for which we have yet to find a better word).

Of course, I would never have made it to this point without the continuous support of my family, who has always helped me in any way possible, even though they still have no clue what it is I actually do and don't understand why I constantly have to move so far away from home in the pursuit of it.

And last, but certainly not least, I am and will forever be immeasurably grateful to the one and only Dr. Pepper for being the shining star in my life, always.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Chain Shifts and Mergers

Variationist linguistic research has established that linguistic changes tend to either be changes from below or from above (e.g. Labov, 1966), referring primarily to the level of conscious social awareness, and secondarily to the position of the change in the socioeconomic hierarchy. In this sense, change from above refers to change that is consciously adopted with full social cognizance, i.e. above the level of conscious awareness, such as lexical borrowings of prestige forms, and frequently affects more formal speech styles first. Often, this type of change is driven by the dominant, i.e. higher social classes (Labov, 2001, 2006). Changes from below, on the other hand, are argued to be internally motivated by many in the variationist tradition, such as Labov (e.g. 2001). There are, however, examples of changes from below that, at least in some cases, have been externally triggered by dialect contact, as is likely the case with the changes of interest in the present study: the Northern Cities Shift (NCS) and the merger of the low back vowels LOT and THOUGHT (COT-CAUGHT merger, low back merger). Regardless of their motivation, changes from below often appear first in certain phonological environments and in less formal speech styles, generally in younger speakers of the centrally located social classes. These types of changes occur below the level of conscious awareness, at least in their initial stages. Once they near completion, however, they may become a matter of social commentary (Labov, 2001, p. 196), and show a greater degree of social stratification (see Chapter 1.4 for more detail).

This shift from an unconscious change from below to one that has attracted a certain degree of social awareness can currently be observed in the two sound changes of interest in the present study. The NCS and the COT-CAUGHT merger have generally been described as changes from below that took place below the level of conscious awareness (Labov, 2001). More recently, however, these changes, or at least elements of them, appear to have reached this level of awareness, and have therefore attracted social commentary and indeed, become more socially stratified. For the NCS, this appears to be the case both in places where it originated, i.e. developed organically, and in places of diffusion[1] (e.g. Driscoll & Lape, 2015; Nesbitt & Mason, 2016; Savage & Mason, 2018; Savage, Mason, Nesbitt, Pevan, & Wagner, 2016; Thiel & Dinkin, under review). Social awareness of the COT-CAUGHT merger, on the other hand, has, as of yet, not been reported. However, the results of the present study suggest that the elements involved in this merger, i.e. changes in the realization of LOT and THOUGHT, may have reached this level of awareness in a place of diffusion (see Chapter 6.5.3).

Both vowel chain shifts and vowel mergers involve interrelated changes in two or more vowels, i.e. changes in one vowel phoneme affect one or more neighboring vowels in the vowel system. However, the two processes differ in the affected vowels' reactions to the changes in the first vowel: In chain shifts, the involved phonemic categories maintain their phonemic distinction as one vowel moves in adjustment to the movement of another, either by following the first vowel in a pull or drag chain, or by moving out of the way to make room for the first vowel in a push chain (see Chapter 1.2.1); in mergers, on the other hand, phonemic distinction collapses owing to a lack of such adjustment, resulting in homonymy between word pairs that, without the merger, are distinct.

Which factors determine whether changes in one vowel that affect another lead to a chain shift or a merger are, as of yet, not particularly well understood. It has been suggested that mergers are a result of language-internal pressures such as an overcrowded phonetic space or the attempt to achieve symmetry in the phonological system (Martinet, 1955), which suggests that chain shifts might be more common in systems that are less crowded and more symmetrical. A second contributing factor might be the functional load of the phonemes involved in the changes, i.e. the number of words that are distinguished through the contrast between the respective vowels (minimal pairs) (Labov, 2001). If this is the case, a chain shift would be a more likely reaction to changes in vowels with greater functional loads, and a merger would be more likely for vowels with smaller functional loads. Both of these accounts assume that chain shifts are driven by system-internal pressures to maintain phonological distinction between the phonemes involved in the shift. In other words, subsequent shifts occur in order to preserve contrast and avoid potential mergers. However, based on the observation that, in some alleged chain shifts, certain changes occur without others, and that in some of these cases margins of security are endangered, the ontological status of chain shifts has

[1] Where the NCS originated and where it potentially diffused to will be the focus of Chapter 1.3.1.

been questioned, and some consider co-occurring vowel shifts to be independent, unrelated phenomena (e.g. Stockwell & Minkova, 1988). It has also been suggested that the motivation for chain shifts and mergers may be external to the language system, and that linguistic contact is the driving force for these kinds of changes, which appears to be the case with the NCS (see Chapter 1.2.1) and COT-CAUGHT merger (see Chapter 1.3.2), at least in some instances.

Vowel chain shifts and vowel mergers are common occurrences in languages and have been characteristic of the diachronic development of English.[2] One of the best-known examples of a chain shift is the Great Vowel Shift of the 15th and 16th century, which "is often taken to define the boundary between Middle English and Early Modern English" (Dinkin, 2012). Examples of more recent chain shifts are the Short Front Vowel Shift in e.g. Australia and New Zealand (e.g. Cox & Palethorpe, 2008; Maclagan & Hay, 2007), the Canadian or California Shift, and the Southern Shift (e.g. *ANAE*). Similarly, mergers have occurred throughout the history of the English language, such as for example the 16th/17th century FACE[3] merger of [a:] in words like *cake* and *safe* with [ɛi] or [æi] in words like *faith* and *way,* which eventually became [eɪ], and the three-way merger of historical /eɪ/ with /æ/ and /e/ in pre-r position in most US varieties of English, known as the MARY-MARRY-MERRY merger (Wells, 1982). More current examples of vowel mergers are the NEAR-SQUARE merger in New Zealand English (e.g. Hay, Warren, et al., 2006) and (to a certain extent) Norwich (Trudgill, 1988), and the PIN-PEN merger in southern US varieties of English (e.g. Baranowski, 2013).

While many of these shifts and mergers have been researched thoroughly, the NCS and the COT-CAUGHT merger are probably two of the best studied examples of contemporary chain shifts and mergers in US varieties of English, potentially owing to their sheer magnitude: The NCS has been described as "a massive change that bears no resemblance to any chain shift previously recorded in the history of the language" (Labov, 1994, p. 10); likewise, the COT-CAUGHT merger has been referred to as one of the major phonological changes in progress in North American English (*ANAE*; Labov, 1994), and is

---

[2] In addition to vowels, mergers can also occur in consonants, as e.g. in the case of /w/ and /v/ in 18th and 19th century Southeastern England (Trudgill, Schreier, Long, & Williams, 2009), which will be explored further in Chapter 1.6.2.

[3] Throughout this dissertation, I use the lexical sets developed by Wells (1982) to refer to vowel classes, with the exceptions of descriptions of phonological environments, quotes that include other denotations, and mergers, which will be identified by their respective names.

arguably the most wide-spread present-day vowel merger, found not only in North America, but also in e.g. Northern Ireland (e.g. Harris, 1985; J. Milroy et al., 1983), Scotland (Lass, 1987; Wells, 1982), and potentially Singapore (Deterding & Hvitfeldt, 1994). Wells (1982) also reported the merger in parts of Ireland, while Hickey (2016) claims that LOT and THOUGHT remain securely distinct in Irish English. But not only the magnitude of these two changes make them particularly interesting variables to examine in more detail. Interestingly, the NCS and the COT-CAUGHT merger have long been considered incompatible, as LOT as part of the NCS is fronted out of its low back position away from THOUGHT (see Chapter 1.2.1), thus preventing the merger from taking place. Given the vigorous geographic expansion of this merger across US varieties of English (see e.g. Katz, 2016 and Chapter 1.3.2 in this thesis), an in-depth analysis of how it is or is not affecting the NCS promises to be a particularly interesting undertaking, especially in light of recent research which has found that LOT in NCS communities is retracting and potentially merging with THOUGHT (e.g. Dinkin, 2009, 2013; D'Onofrio & Benheim, 2018; Driscoll & Lape, 2015; Durian & Cameron, 2018; Fox, 2014, 2016; King, 2017; McCarthy, 2010; Milholland, 2018; Morgan, DeGuise, Acton, Benson, & Shvetsova, 2017; Wagner et al., 2016). This appears to be part of the expanding Elsewhere Shift, a chain shift that very closely resembles, or may in fact be identical with the California and Canadian Shift (discussed further in Chapters 1.6.1 and 7.1).

Both the NCS and the COT-CAUGHT merger involve complex phonetic processes subject not only to certain linguistic but social constraints too, often varying across communities in which they have occurred. Thus, in the following subchapters, I will present both the NCS and the COT-CAUGHT merger in more detail based on previous research. In doing so, I will provide an overview of various aspects that have been found to trigger and drive (or prevent) chain shifts and mergers, including both internal factors, i.e. those concerning system-internal, structural aspects, and external factors, i.e. social considerations, also drawing on other chain shifts and mergers to exemplify various elements involved in these types of changes. Along the way, it should become clear that, while both the NCS and the COT-CAUGHT merger have been studied extensively, certain gaps remain in the literature – some of which motivate the present study.

1.2 The NCS and COT-CAUGHT Merger as Changes from Below

1.2.1 The Conceptualization of the NCS as a Chain Shift

Labov (e.g. 1994) claims that vowel chain shifts are generally triggered by system-driven changes in the vowel space, such as monophthongization or vowel mergers[4]; the NCS, however, he believes to have been initiated by koineization resulting from language contact. Starting in 1825, the population of cities in Central and Western New York, such as Syracuse, Rochester and Buffalo (see Figure 1), grew rapidly, with increasing immigration from the UK, Ireland and Germany owing to the completion of the Erie Canal (Labov, 1966).

Figure 1: Map of New York State. The Erie Canal runs from near Buffalo in the western part of New York State toward Albany near the eastern state border.

[4] The Southern Shift, for example, is believed to have been initiated by the monophthongization of PRICE (e.g. Labov, Ash, & Boberg, 2006, p. 244) and the Elsewhere Shift (also referred to as the California or Canadian Shift) by the merger of LOT and THOUGHT (see Chapters 1.6.1 and 7.1).

In combination with earlier settlers from primarily Southwestern New England in each of these cities, this created a contact situation of different speaker groups with at least four different TRAP systems (see Chapter 3.1), which involved the raising of TRAP, or lack thereof, to varying degrees. This mixing of different TRAP configurations, in addition to an already rather turbulent history of the vowel class (see Chapter 3.1), is believed to have let to the formation of a koiné, a simplified system of unconditioned TRAP raising, and Labov (1994) argues that this koineization constituted the initiation of the NCS in the middle of the 19th century. Thus, Central and Western New York is the most likely birthplace of the NCS. It is therefore not surprising that some of the earliest evidence of NCS-like developments can be found in descriptions of English spoken in this part of the state, where elements of the NCS were observed as early as the late 19th and early 20th century (e.g. Emerson, 1891; Monroe, 1896; C. K. Thomas, 1935-1937).

Labov (1994) believes that the general raising of TRAP led to the ensuing changes in five to six vowels in a chain shift that became known as the NCS. These changes, illustrated in Figure 2 below, include the fronting of LOT, lowering and fronting of THOUGHT, backing and/or lowering of DRESS and KIT, and backing of STRUT.[5] In its traditional conceptualization, the NCS is argued to be a combination of multiple pull (or drag) and push chains. The initiating changes, the fronting and raising of TRAP and the fronting of LOT, caused THOUGHT to front and lower toward the space formerly occupied by LOT in a pull chain, often in combination with unrounding of the vowel. In a next step, the backing of DRESS toward STRUT may have been motivated by TRAP raising, and created a push chain that caused STRUT to retract toward THOUGHT (Labov, Ash, & Boberg, 2006; henceforth *ANAE*), as well as a potential pull chain leading to the lowering and backing of KIT. This chronology was based on the time of first observations as well as the relative advancement of each shift. The raising of TRAP, fronting of LOT and lowering and fronting of THOUGHT were first reported by Fasold in 1969. Labov, Yaeger and Steiner (1972; henceforth *LYS*) added the backing and/or lowering of DRESS and KIT to this shift a few years later in 1972[6], and STRUT was added in the 1980s by Eckert (e.g. 1988). However,

[5] A detailed description of the phonetic outcome of each of these shifts will be provided in the introductions to the respective chapters for each vowel (Chapters 3.1, 4.1, 5.1, 6.1, 7.1).

[6] Although centralization of KIT was observed alongside DRESS by *LYS* in 1972, *ANAE* labels the shifting of KIT as the last stage of the NCS (p. 190), without providing an explanation for this decision other than that Eckert (1988) proposed the same overall order based on her Detroit data. It is possible that this is related to *ANAE's* general disregard for KIT shifting, as they do not include KIT in their discussion of the NCS, and

the exact order of these changes has been the subject of an ongoing debate (e.g. Durian & Cameron, 2018; Gordon & Strelluf, 2016), and it is possible that the order of changes was dependent on the individual speech community in which the shift occurred, for example depending on whether the shift arose organically or was adopted through diffusion (Dinkin, 2012), which will be discussed further in Chapter 1.3.1.

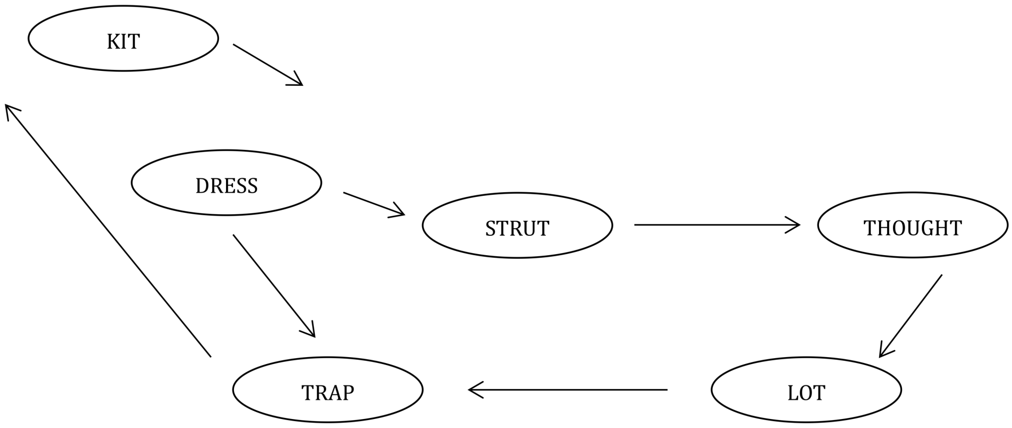


Figure 2: The Northern Cities Shift.

Whether the changes commonly observed in NCS communities are in fact systematically dependent changes operating as a chain shift has been questioned as well. Gordon (2001), for example, found that the NCS does not necessarily follow the criterion of contrast preservation. While representations of the NCS such as in Figure 2 above are common and clearly illustrate how the movement of one vowel is an adjustment to a shift by another vowel, Gordon argues that this representation is an extremely simplified abstraction of the phonetic processes involved in the NCS, where some vowel distinctions may in fact be endangered. *ANAE*, for example, acknowledges that DRESS and STRUT backing is commonly observed in originating NCS communities without accompanying lowering of THOUGHT (p. 191). As an alternative to the notion of the NCS as a coherent event, Durian and Cameron (2018) argue that the NCS is more likely to be a combination of two independent, co-occurring chain shifts: that of TRAP and LOT on the one hand, and DRESS and STRUT on the other. THOUGHT and KIT, under this scenario, are not linked to these shifts, THOUGHT being an optional shift and KIT sporadically shifting as a parallel development to DRESS backing.

instead describe it as a change that is "closely associated with the Northern Cities Shift", but "not an essential part of it" (p. 79).

However, DRESS has also been observed to be backing while STRUT remained stable (*ANAE*, p. 191), which questions the systematical dependencies of these two shifts.

In general, chain shifts can occur on a phonemic level (e.g. Gordon, 2013) or be conditioned by phonological environments (e.g. Langstrof, 2006). The NCS has been described as being subject to only fine-grained phonetic conditioning (*ANAE*, p. 120), in that some phonological environments were found to favor certain shifts more than others. The strongest effect on the height of TRAP, for example, appears to have been the manner of the following segment, especially nasals (Labov, 1994, pp. 362–363). Because the present study is more concerned with the social conditioning of the NCS rather than with potential effects of different phonological environments on NCS variables[7], a detailed account of such effects is beyond the scope of this study, but see *LYS*, Callary (1975), Gordon (2001), *ANAE* and Benson, Fox, and Balkman (2011) for more detail.

1.2.2 The Inner Workings of the COT-CAUGHT Merger

Mergers result from the collapse of contrast between two or more vowels in both production and perception. In the case of the COT-CAUGHT merger, the two phonemes involved are the low back vowels LOT and THOUGHT.

The reasons for common loss of phonemic distinction between LOT and THOUGHT are, at this point, not fully understood. It has been suggested that it is related to a rather small functional load of both phonemes, as minimal pairs for LOT and THOUGHT are relatively rare (Labov, 2001). Additionally, the tumultuous historic developments of LOT and THOUGHT may have contributed to the frequent merger of these two vowel classes, as "a long series of historical accidents … led to the creation of the highly skewed and unstable long open-*o* class – a back rounded vowel, distinguished from short open *o*[8] only by length" (Labov, 2010, p. 173), which will be discussed in more detail in Chapter 6.1. Language contact is a third reason that has been suggested as a cause for the COT-CAUGHT merger, as mergers may be the outcome of contact with speakers without the distinction (e.g. Maguire, Clark, & Watson, 2013). Under this scenario, initial instances of the merger in US English may have been triggered by the immigration of Scots and Scotch-Irish (Lass,

---

[7] Phonological environments are factored into the statistical analyses (see Chapter 2.6.1), and their estimated effects are provided in Appendix H, but they are not commented on in the analysis, with the exception of TRAP, for which pre-nasal tokens are analyzed separately.

[8] Long open-*o* (roughly) refers to THOUGHT, short open *o* to LOT.

1987; Milroy, 1995) and/or by immigrants from Slavic-speaking countries (Herold, 1990), as neither of these two groups distinguished LOT from THOUGHT (see Chapter 1.3.2).

While the initial triggers for mergers remain somewhat of a question mark, the mechanisms that drive mergers once they are set in motion are better understood. The literature suggests four main mechanisms through which mergers are advanced in production (Dinkin, 2009; Herold, 1990; Trudgill & Foxcroft, 1978), three of which have been found to apply to the COT-CAUGHT merger, either independently or in combination:

- Merger by approximation (or drift)
- Merger by expansion
- Merger by lexical transfer
- Merger by phonological transfer

A merger achieved by the gradual approximation of LOT toward THOUGHT, vice versa, or in simultaneous movements, has been reported in Charleston (Baranowski, 2013), Southern Illinois (Bigham, 2010)[9], Western Massachusetts and California (D'Onofrio, Eckert, Podesva, Pratt, & Van Hofwegen, 2016; Hall-Lew, 2013). In this process, LOT and THOUGHT decrease their phonetic distance and consequently increase their phonetic overlap. This process is complete once the members of the merger are sufficiently close so as to not be distinguishable, which can take up to four generations (Labov, 1994). There is, however, no consensus on *how* close the members have to be in order to be judged as merged (see Chapter 2.6.3), though it has been found that a minimum distance of 200 Hz on the front-back dimension is necessary in order to maintain stable distinction (*LYS*; Labov, 1994). It is also possible for this process of approximation to never be completed. In this case, the members of the putative merger are merely approximating each other in phonetic space, but their phonemic contrast never fully collapses (see below).

Merger by expansion was found to be the driving mechanism behind the COT-CAUGHT merger in Eastern Pennsylvania (Herold, 1990). In this process, the involved phonemes expand their phonetic range in the vowel space, thereby increasing the overlap with the other member. This is, arguably, the fastest of all mechanisms, and the collapse

[9] Bigham (2010) reports disagreement on the perceived qualities of the "new" phoneme between him and his research assistant in the process of auditory coding, one consistently hearing LOT-like realizations, the other hearing something in between LOT and THOUGHT that does not quite resemble either of the two phonemes.

of contrast between the involved phonemes can be completed in a single generation (Herold, 1990; Labov, 1994, p. 322). Thus, there are generally no intermediate stages to be observed. It can, however, divide a community into merged and distinct speakers, as speakers who have already acquired a secure distinction, generally older generations, are less likely to adopt the merger (Johnson, 2007).

In a merger by transfer, elements of one vowel class are transferred to a different vowel class, one lexical item (lexical transfer) or one phonological environment (phonological transfer) at a time in quite abrupt processes (Milroy & Harris, 1980). Transfer is the slowest of the three merger mechanisms (Labov, 1994), and merger by transfer is complete when one of the members of the merger is realized as the second member in all relevant lexical items or phonological environments. However, it is also possible for the merger to remain partial. While lexical transfer has not yet been reported for the COT-CAUGHT merger, phonological transfer of LOT to THOUGHT appears to have initiated the merger in New York State, followed by merger by approximation (Dinkin, 2009, 2016). Phonological transfer, in this case, appears to have remained partial, as LOT was transferred to THOUGHT only when preceding /lv/ or /lf/.

A different combination of mechanisms has been observed in Southeastern New England. Here, the merger appears to have started as a merger by approximation, and continued as a merger by sudden expansion at a later stage (Johnson, 2007).

The outcome a of merger, i.e. the phonetic realization of the resulting "new" phoneme, depends on the underlying mechanism and the direction of change. Traditionally merged areas like Canada, Eastern New England and Western Pennsylvania merged LOT and THOUGHT in low back or low mid-back position to something like [ɒ], though rounding seems to be variable (Boberg, 2000, 2001; Clarke, Elms, & Youssef, 1995; E. R. Thomas, 2001). One exception to the Canadian merger outcome is St. John's, Newfoundland, where speakers of the Anglo-Irish dialect have been found to be fully merged in a position approaching low central (*ANAE*). A low central merger has also been reported in the West (e.g. Drager & Hay, 2011; Hall-Lew, 2013; Kennedy & Grama, 2012), in Erie, Pennsylvania (Evanini, Isard, & Liberman, 2009; Labov, 2010), and in Southwestern Vermont, while in Northwestern Vermont, the merger took place in low back position (Boberg, 2001). Similar in-state differences have been found in Eastern Pennsylvania and California: In Eastern Pennsylvania, realizations of the merger range from [ɑ] to [ɔ] depending on the speaker and the community (Herold, 1990). In California, speakers in the central valley have merged on THOUGHT (D'Onofrio et al., 2016), while speakers in San Francisco have been found to flip-flop LOT and THOUGHT, i.e. LOT is realized with THOUGHT-like qualities, while THOUGHT is realized with LOT-like qualities. In other words, the two vowels are moving past each other, resulting in switched but still opposing qualities (Hall-Lew, 2013). Evidence of a potential flip-flop has also been reported in Utah (Di Paolo, 1992). In the South, e.g. Oklahoma and Texas, the merger has been reported to take place mainly through the unrounding of THOUGHT (Guy Bailey, Wikle, Tillery, & Sand, 1993). Because the outcome of the COT-CAUGHT merger, i.e. the phonetic realization of the resulting "new" phoneme, can vary, it has been suggested that this merger exists in multiple, rather than just a single version (Irons, 2007).

So far, I have described mergers mostly in terms of their spectral qualities in the vowel space, and typically, these are the parameters by which mergers are defined, i.e. the relative distance between the involved phonemes on the height and front-back dimension, and the amount of overlap between them. However, vowel contrast can be sustained through acoustic features other than those referring to space. In these cases, the vowels may have merged phonetically, but are kept distinct phonemically through secondary acoustic cues. The nature of the distinctive feature can vary from merger to merger, but also from speaker to speaker (Paolo & Faber, 1990). In the South, for example, spectrally similar LOT and THOUGHT are differentiated by the presence of a back upglide on THOUGHT (*ANAE,* p. 256; Irons, 2007). Alternatively, speakers can distinguish spectrally identical LOT and THOUGHT through lip rounding (Havenhill, 2018). Differences in phonation, such as creakiness and breathiness, have been found to serve the same purpose (e.g. Di Paolo, 1992). Di Paolo also found duration to be a differentiating feature[10] for spectrally identical LOT and THOUGHT, though more recent studies found few significant duration differences between these two vowels, even among the most merged speakers (e.g. Benson, Fox, & Balkman, 2011; Irons, 2007; Majors, 2005). Likewise, formant trajectories for F1 and F2 of merged LOT and THOUGHT have been found to be *more* similar than those of distinct LOT and THOUGHT (Majors, 2005), i.e. changes in height and frontness during the articulation of LOT and THOUGHT are more similar for merged speakers than they are for distinct speakers, and are therefore unlikely to introduce phonetic contrast to spectrally similar LOT and THOUGHT.

[10] Klatt (1976) defines the perceptual threshold for differentiation through length as 25 milliseconds.

Like chain shifts, mergers can occur either on an allophonic or phonemic level, i.e. be conditioned or unconditioned. Mergers that occur only in certain environments, i.e. conditioned mergers, only affect certain allophones of their respective phonemes, and therefore do not alter the vowel inventory of the speaker. For example, the PIN-PEN merger, found predominantly in the southern states in the US (e.g. Baranowski, 2013) and parts of Ireland (Wells, 1982), only occurs in pre-nasal position (e.g. *bin, Ben*); the merger of FOOT and GOOSE in western Pennsylvania, Albuquerque and Salt Lake City is limited to positions preceding /l/ (e.g. *pull*, *pool*), as is the merger of KIT and FLEECE in some parts of the South (e.g. *fill*, *feel*). The COT-CAUGHT merger has been found to be an unconditioned merger, i.e. it is generally not limited to certain phonological environments. The contrast between LOT and THOUGHT is therefore typically lost completely, reducing the vowel inventory by one phoneme. This merger does, however, seem to *favor* certain environments, and it has been reported to be more advanced in positions preceding /n/ as in *Don* and *dawn* (e.g. *ANAE*, Boberg & Strassel, 1995; Gordon, 2013) and /l/ as in *collar* and *caller* (Boberg, 2001; Dinkin, 2009, 2016). Following /k/ as in *stock* and *stalk*, on the other hand, appears to disfavor the merger (e.g. *ANAE*, Boberg & Strassel, 1995; Boberg, 2001; Dinkin, 2009, 2016). However, only few studies have reported an incomplete merger of LOT and THOUGHT: In Utah, both vowels are neutralized only before nasals (*Don, dawn*) (*ANAE*), while in Southeastern Ohio and Northern West Virginia, the merger occurred only preceding /t/ (e.g. *tot, taught*) (E. R. Thomas, 2001).

As mentioned above, a merger involves the loss of distinction between two sounds in production *and* perception, i.e. a diminishing ability to produce a distinction as well as to perceive it. While these two processes are related, they are frequently found to progress at a different pace. In areas that generally are merged in production, speakers with fully merged production may maintain their ability to distinguish the respective phonemes in perception (e.g. Baranowski, 2013; Hay, Drager, & Thomas, 2013). In distinct or transitional areas, on the other hand, merger in perception is most commonly found to *precede* merger in production, a pattern that has also been observed for the COT-CAUGHT merger (e.g. *ANAE*, Boberg & Strassel, 1995), especially in older generations (Labov, 1994, pp. 362–363).

Depending on the ratio of merger in perception and production, the outcome might be a special kind of merger, termed near-merger. Labov (e.g. 1994) defines near-mergers as being characterized by an asymmetric relationship between merger in production and merger in perception, with the latter preceding the former. Thus, speakers are merged in minimal pair judgments and categorization tasks, but maintain a distinction in their own speech – though sometimes reduced in more careful speech styles, with or without overlap of the vowel classes.[11] The distinction, according to Labov, is generally maintained through differences on the front-back dimension rather than in height. This pattern is particularly common for older speakers in the community, while younger speakers tend to be significantly more advanced in the merger in production. As a result, there tends to be notable variation regarding the status of the merger in the community, as has been found to be the case with the COT-CAUGHT merger in Pennsylvania (Labov, 1994) and Utah (Di Paolo, 1992). Other examples of reported near-mergers are GOOSE and FOOT in pre-/l/ position (e.g. *pool*, *pull*) in Albuquerque (Labov, 1994) and Salt Lake City (Di Paolo, 1988), as well as GOOSE and GOAT, and BEER and BEAR in Norwich (Trudgill, 1974). Near-mergers may, or may not, be responsible for the reversal of mergers, i.e. the reintroduction of phonemic contrast in the past, as will be discussed in Chapter 1.6.2.

1.3 The Spread of Sound Change

1.3.1 Transmission and Diffusion of the NCS

As described in Chapter 1.2.1, Central and Western New York is the likely birthplace of the NCS. From there, Labov (2001) believes the NCS was transmitted to the Great Lakes region, which was largely settled by a westward movement from New York State to cities such as Detroit and Chicago, where the shift continued to develop concurrently in each community. Transmission in this sense is defined by Labov (2007) as a process through which linguistic innovations, including the underlying structures that link these innovations, are carried over from one generation to the next via first language acquisition in children. The outcome, Labov argues, is a faithfully reproduced pattern, i.e. the NCS as a unitary chain shift, which continued to be advanced in the same direction with each successive generation (incrementation) throughout the Great Lakes region. Labov believes that this process of transmission was made possible by the particular

[11] In extreme cases, it is also possible for speakers to merge both members completely in minimal pairs while they are otherwise distinct in their speech, an effect Labov (1994) refers to as the "Bill Peters effect", named after a speaker in Pennsylvania and his production and perception of LOT and THOUGHT.

settlement pattern of westward migrants, who moved as entire, intact communities, with continuous contact among children speaking the same dialect, which is assumed to account for the uniformity of the Inland North dialect.

Transmission of the NCS throughout the Great Lakes region appears to have been a rather speedy process. TRAP raising, for example, seems to have first appeared in Lansing and Grand Rapids in Michigan, more than 300 miles from its birthplace in New York, at the turn of the 20th century (Gordon & Strelluf, 2016). In Detroit, raising of TRAP, and fronting of LOT and THOUGHT were first observed in 1969 (Fasold, 1969), and backing of STRUT was first noticed in the community at the end of the 20th century (Eckert, 1988). In Chicago, the NCS is estimated to have started to develop around 1920 (McCarthy, 2010), but was not observed in its entirety until *LYS* (1972), who reported a systematic shift of TRAP, LOT, THOUGHT, KIT and DRESS in their data from Chicago as well as Detroit and the Buffalo-Rochester area. In fact, they were the first to describe these observations as a structured chain shift. By 2006, *ANAE* was able to identify the majority of NCS elements in speakers from not only Western New York, Detroit and Chicago, but also Ohio (Cleveland, Akron, Toledo), Michigan (Grand Rapids, Flint, Kalamazoo), Illinois (Rockford, Joliet), Indiana (Gary), and Wisconsin (Kenosha, Milwaukee, Madison, Green Bay). Thus, the NCS became the defining feature of the Inland North dialect, a subsection of the Northern dialect, identified in Figure 3 below.

The Inland North has been described as a remarkably uniform dialect, distinguished from surrounding dialects, particularly Canada and the Midlands, by unusually sharp isoglosses. In fact, Labov (2007, p. 373) argues that "the linguistic boundary separating the Inland North from Midland vowel patterns is the sharpest and deepest division in North American phonology". This divide appears to be the result of different settlement patterns in the periods of westward migration, and opposing lifestyles of the Yankees in the North and the Southerners in the Midland (Labov, 2010). The details of this opposition will be discussed in more detail in Chapter 1.5.1. In addition, Labov attributes some of the reasons for the North/Midland boundary to early communication patterns. While there was plenty of east-west infrastructure facilitating communication in this direction in both the North and the Midland, no such connections existed in north-south direction between the Midland and the North on a larger scale, contributing to the separation of these two regions, also in terms of linguistic features.

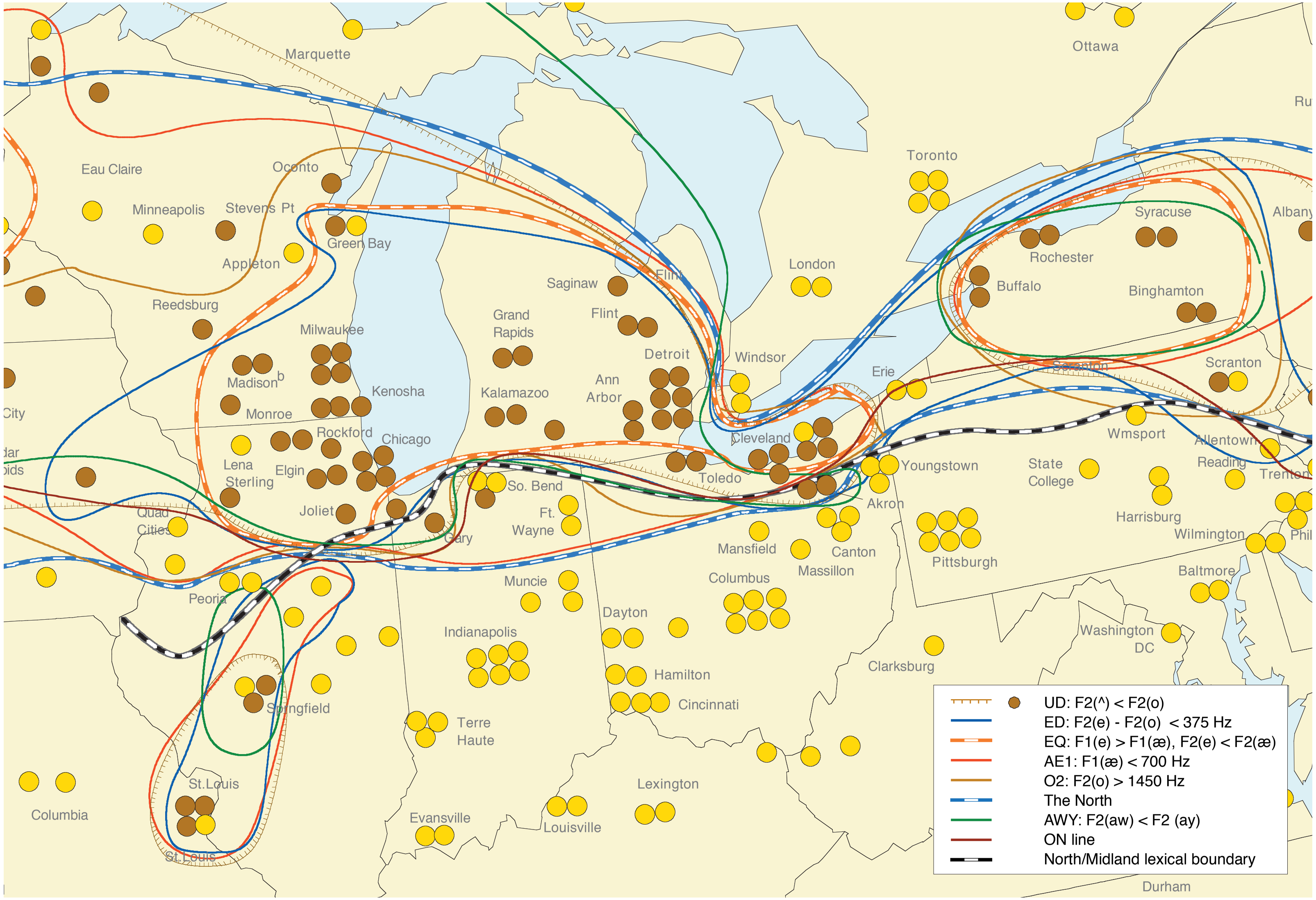


Figure 3: The Inland North dialect area based on elements of the Northern Cities Shift[12] (*ANAE*, p. 207).

The NCS - An Urban Phenomenon?

As the name suggests, the Northern *Cities* Shift has generally been conceptualized as an urban phenomenon. Thus, it is no coincidence that the majority of studies mentioned above focused on urban centers as "the most important place to draw data" (*LYS,* p. 13), an approach that has long been predominant in variationist sociolinguistic research. However, the effects of dialect contact, which was found to be the driving force behind the transmission of the NCS to the cities of the Great Lakes region, have been argued to be largely identical in urban and rural areas (e.g. Britain, 2009, 2012). Nevertheless, only few studies have examined the NCS in smaller towns (e.g. Dinkin, 2009, 2013; Fox, 2014; Gordon, 2001; Ito, 2001; Ito & Preston, 1998), and while there appears to be agreement on how the NCS developed in urban centers, the conditions for its presence in smaller communities are less established.

The most common assumption about the NCS in smaller and more remote communities is that, rather than having developed there through transmission and incrementation, it diffused along networks of communication from urban centers according to the cascade model of diffusion (also known as hierarchical diffusion) proposed by C. J. Bailey (1970). According to this model, linguistic innovations arise in urban centers, and from there spread to the next lager cities, temporarily bypassing intermediate and smaller communities, to which the innovation will diffuse at a later stage on its downward slope. Evidence for the applicability of this model to the NCS stems from Callary's (1975) study of Northern Illinois, where TRAP raising strongly correlated with community size: While participants from various urban centers seemed to participate in this feature, those from more rural areas did so only sporadically, and only in certain phonological environments. Unfortunately, the loss of structural integrity in the process of diffusion of chain shifts outlined above cannot be tested in Callary's study, as his analysis focused on TRAP raising only. Additionally, the sample of speakers was restricted to only 18 college-aged female speakers, so that apparent-time trends, or the lack thereof, would not be observable. Labov (2001) initially supported Callary's suggestion that the NCS follows the cascade model of diffusion. Later, however, he found that this model may only account for TRAP raising in Illinois, as city size did not correlate

---

[12] For an overview and explanation of the criteria used to map NCS features, see Chapter 2.6.2.

in any significant way with the advancement of the NCS in the thirty Telsur[13] cities in the Inland North (Labov, 2010, p. 205). Only for DRESS backing did city size seem to be a determining factor (*ANAE*, p. 197).

While Callary rejected the idea that settlement patterns determine receptiveness to the NCS through diffusion[14], patterns observed in New York State and along the St. Louis Corridor[15] suggest the opposite. In Central and Northern New York, the main determiner of whether or not smaller communities were subject to the NCS seems to be their early settlement: While communities that derived their early settlement primarily from Southwestern New England did adopt the NCS, though with some variability, those that were settled mostly by descendants of the Dutch colonists of New Netherland[16] did not (Dinkin, 2009, 2013). Similarly, Labov (2007) suggests that the NCS diffused to St. Louis and cities along the corridor from Chicago owing to the mixed settlement history of this area, which was originally settled from the South, but attracted more people from the North starting with the second half of 19th century.

That diffusion cannot be the sole relevant mechanism in the spread of the NCS was demonstrated in a study of small-town participation in the NCS in Michigan. Gordon (2001) investigated NCS participation in speakers from two towns, Paw Paw and Chelsea, with population sizes between 3.000 and 4.000.[17] Neither of the two communities is remote or isolated; instead, both are located in between the two urban centers of Chicago and Detroit, to which they are well connected by an interstate highway. Paw Paw is geographically closer to Chicago as well as to the medium-sized city of Kalamazoo. Chelsea is nearer Detroit, and neighbors Ann Arbor, which is slightly bigger than Kalamazoo. All of these cities, Detroit, Chicago, Ann Arbor and Kalamazoo have been affected by the NCS according to *ANAE*. Gordon, among his participants, identified frequent shifting of TRAP, LOT and THOUGHT, as well as some rare instances of shifted DRESS, KIT and STRUT. Overall, however, shifting was much more common in Paw Paw than it was in Chelsea. Since

---

[13] "Telsur" refers to a telephone survey of linguistic change, conducted in 1995 and 1996 in North America (see *ANAE* (p. 8) for further detail).
[14] Callary (1975) found that Illinois' settlement history as identified by Shuy (1962) did not correlate in any way with the observed stratification of TRAP raising.
[15] The St. Louis Corridor runs along the I-55 (an interstate highway in central US, running from Illinois in the north to Louisiana in the south) from Chicago to St. Louis, and includes the cities of Fairbury, Springfield and Bloomington. Geographically, this area belongs to the Midland rather than the North (*ANAE*).
[16] New Netherland was a 17th century Dutch colony that comprised parts of what today is New York State (including New York City), New Jersey, Connecticut and Delaware.
[17] These numbers refer to the estimate in the 1990 US census as cited by Gordon (2001).

Chelsea is closer to Detroit than Paw Paw is to Chicago, and because Ann Arbor is a larger city than Kalamazoo, this was an unexpected finding that contradicts the expected outcome of hierarchical diffusion. Gordon suggested that the difference in NCS participation between the two towns may be related to attitudinal differences, which may have led Paw Paw speakers to adopt the NCS, while Chelsea speakers rejected it. Similar effects were reported by Ito and Preston (1998) and Ito (2001), and will be explored in more detail later on in Chapter 1.5.1.

If diffusion was the underlying mechanism driving the spread of the NCS to smaller communities, it is likely to have affected the systematic make-up of the shift in the recipient communities. Diffusion, as defined by Labov (e.g. 2001, 2007), is driven by adults (as opposed to transmission, which relies on language acquisition in children), who, owing to limited learning abilities are not able to acquire abstract structural elements of language[18], e.g. that raised TRAP leads to backing of DRESS. Instead, they acquire surface-level elements, i.e. those features that can be observed, such as raised TRAP and/or backed DRESS. This results in the loss of the structural links that characterize, for example, chain shifts. Thus, whereas elements of an actual chain shift are not expected to occur in isolation because they are assumed to be interdependent and structurally linked, in locations of diffusion, single elements of a chain shift can be adopted without others. Additionally, if smaller communities have in fact been subject to a diffusing NCS rather than a transmitted and incremented NCS, this would have consequences for the synchronic profile of the shift, with no clear apparent-time trends toward advancing NCS in these communities. Instead, the most advanced speakers would be those who are in regular contact with NCS speakers, i.e. adults (e.g. Dinkin, 2012).

1.3.2 Expansion of the COT-CAUGHT Merger

Regarding their expansive tendency, mergers have been described as being governed by Herzog's Principle (Labov, 1994), which states that mergers expand geographically at the expense of distinction. The merger of LOT and THOUGHT in North America was first identified in Eastern New England (1930s), Eastern Pennsylvania (1940s), Canada

---

[18] Stanford and Kenny (2013) argue that learning abilities are secondary in the difference between transmission and diffusion of the NCS. The primary factor, they argue, is simply the difference in the density of interactions of agents.

(1960s) and the West (1970s). Currently, it covers more than half of the North American territory, as can be seen in Figure 4 below, and Labov (2010) predicts that it will, at some point, dominate most of the continent. In addition to the regions marked in Figure 4 below, the COT-CAUGHT merger, or progress toward this merger, has been reported for much of the South (e.g. *ANAE*; Baranowski, 2007, 2013; Irons, 2007), the Midlands (*ANAE*; Bigham, 2010), Southwestern and Southeastern New England (*ANAE*; Johnson, 2007), Eastern Pennsylvania (Herold, 1990), Washington DC (Lee, 2018), San Francisco (e.g. Hall-Lew, 2013), and Hawaii (e.g. Drager & Hay, 2011). Unexpectedly (for reasons detailed below), the merger has also expanded into communities affected by the NCS, including Michigan (Wagner et al., 2016), Missouri (Gordon, 2006; Majors, 2005), Northwestern Wisconsin (e.g. Benson et al., 2011) and New York (Dinkin, 2009, 2010, 2016).

Although the expansion of mergers is a common assumption and observation, a detailed description of the mechanisms behind their geographical advancement seems to be missing from the literature. Labov (2007) explains that, while it is often found that children of non-merged parents are fully merged, the mechanism behind this process is not particularly well understood. The two mechanisms that are most commonly mentioned in the literature are independent developments, and contact-driven, contagious diffusion from one affected area to an adjacent one.

In a few cases, the COT-CAUGHT merger has been found to have originated independently. In Western Pennsylvania and the South, for example, the merger is likely to have developed on its own, as both are geographically separated from the nearest merged areas by unmerged areas (*ANAE*; Irons, 2007). While it is possible that this separation may be the result of hierarchical diffusion of the merger, this process appears to be an unlikely driving force behind the spread of this merger, as will be discussed in the following paragraph. The development of the merger in Eastern Pennsylvania, while geographically independent, appears to have been contact-induced, either by Scotch-Irish immigrants (Lass, 1987; Milroy, 1995) or by foreign miners from Slavic-speaking countries (Herold, 1990), both of which failed to make the distinction between LOT and THOUGHT.

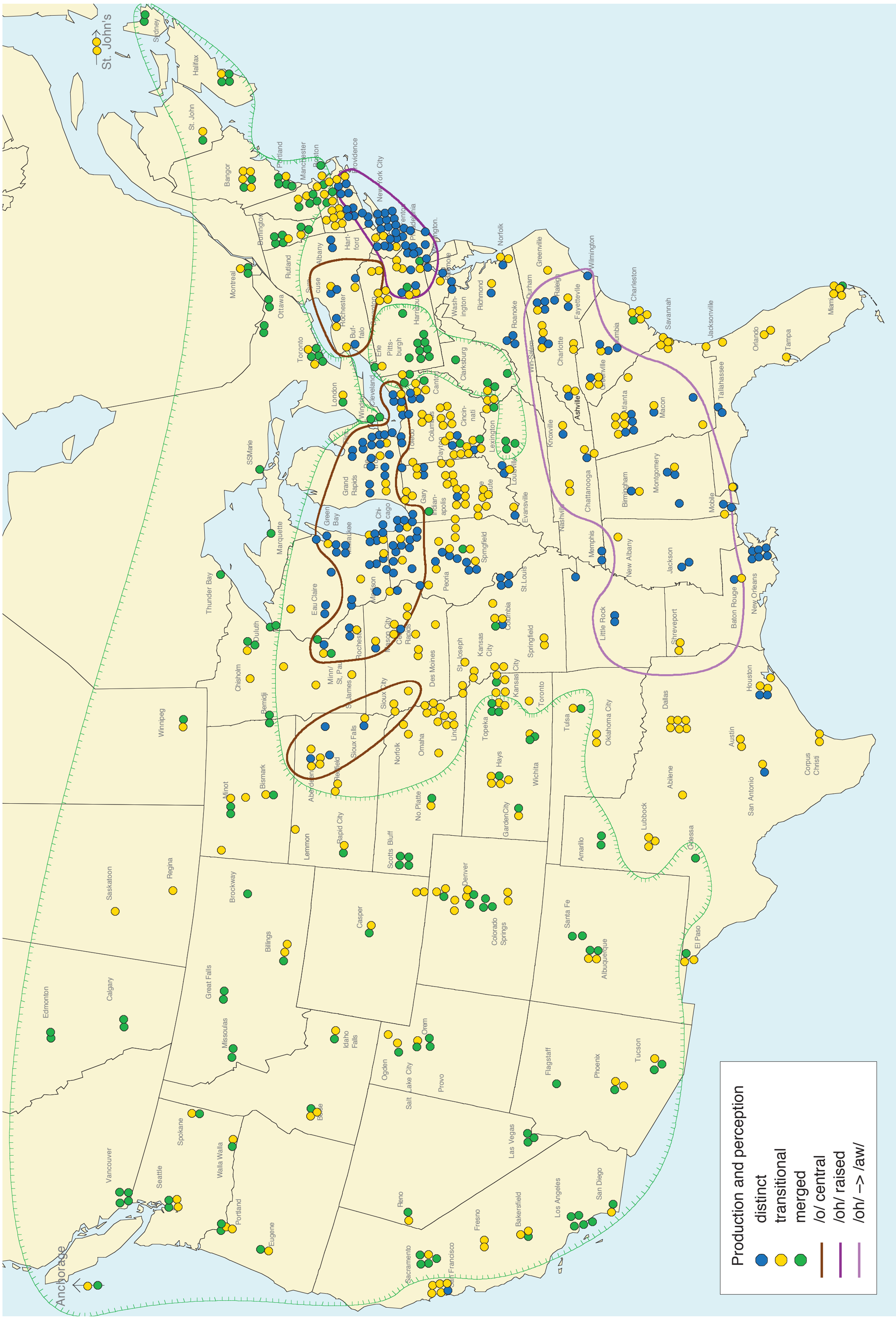


Figure 4: The COT-CAUGHT merger in North America, and areas of (presumed) resistance to the merger *ANAE*, p. 60).

For the most part, it seems that mergers are assumed to be expanding through the process of contagious diffusion, the magnitude of which becomes clear by comments such as "it is well known that phonemic mergers are very easily diffused" (Labov, 2001, p. 27). It has been argued that, in contact situations between merged and distinct speakers, phonemic contrast becomes redundant, and that mergers are preferred over distinction (e.g. Herold, 1990). Rather than redundancy, Trudgill (e.g. 1986, 1996) believes that the reason for the preference for mergers over distinction are simplification processes involved in adult language and dialect acquisition, as a merger is easier to acquire than a distinction. Labov (1994), on the other hand, argues that listeners cannot distinguish between merged and distinct speakers. Because Labov (e.g. 1994, 2001) has also argued that only linguistic features that are observable can be subject to diffusion, the claim that listeners cannot perceive phonological contrast, or the lack thereof, questions whether mergers are available for diffusion at all. A solution to this contradiction was offered by Dinkin (2009, p. 415), who suggests that, in the diffusion of a merger, the diffusing element is the phonetic surface-level change leading to the merger, rather than the lack of contrast between the members of a merger itself. Thus, in the case of the COT-CAUGHT merger, the diffusing element(s) might be the changes in the realizations of LOT and/or THOUGHT that lead to a weakened contrast between the two elements and, eventually, to a full merger.

Regardless of the underlying mechanism, diffusion is by far the most frequently reported cause for the expansion of the COT-CAUGHT merger. A contact-induced spread of this kind has been reported, for example, in Eastern Pennsylvania (Herold, 1990) and Northwestern Wisconsin (Benson et al., 2011). Dinkin (2019) found contagious diffusion to apply to the spread of the COT-CAUGHT merger in Northern New York as well. In a study comparing Northern New York communities along or close to the Canadian border, he found no significant north-south difference, but a striking east-west progression of the merger. In other words, communities that are closer to Canada did not show a higher degree of merger than those further away from the border, but communities further east were found to be more merged than communities to the west. This was interpreted as the merger diffusing from the fully merged Northeastern New England through the transitional and nearly merged (dialectological) North Country[19] to the, as of yet distinct,

[19] "North Country" generally refers to a certain geographic part of Northern New York, but the same term is used by Dinkin to refer to the dialect region bordering the Inland North to the east in New York.

Inland North. The operation of contagious diffusion, rather than hierarchical diffusion, in the case of the COT-CAUGHT merger is supported by *ANAE* on a broader level, as it found no significant effect of population size on the spread of this merger across multiple communities in the US.

However, there are a few counterexamples to this generalization. Guy Bailey et al. (1993) reported that diffusion of the COT-CAUGHT merger in Oklahoma was proceeding hierarchically, bypassing areas that were further away from urban centers. They did, however, point out that the difference between urban and rural areas leveled out among younger generations. Similarly, Irons (2007) found that the diffusion of the COT-CAUGHT merger in certain areas of Kentucky may have proceeded hierarchically. More generally, however, she found that none of the traditional models for geographic diffusion applied to the geographic distribution of the COT-CAUGHT merger in Kentucky, as those areas with the most advanced merger were those that were characterized by very low population densities and were located far away from major infrastructure that would have connected them to merged areas. In Southeastern New England, the COT-CAUGHT merger is argued to not have developed through contagious diffusion from a neighboring, traditionally merged dialect, but through in-migration of merged children (Johnson, 2007).

Although the COT-CAUGHT merger has been found to expand beyond regional boundaries and can be found in numerous dialects with varying vowel systems, certain phonological and political conditions have been believed to prevent the merger, at least for a while. *ANAE* argues that among the phonological preventative mechanisms are the raising of THOUGHT away from LOT, commonly found in the Mid-Atlantic States and New York City, and the presence of a back upglide on THOUGHT in the South. Likewise, the NCS has been described as offering "stable resistance" to the merger, as LOT is substantially fronted out of its low back position, away from THOUGHT. In addition, raised TRAP has been associated with the maintenance of distinction between LOT and THOUGHT, while unraised TRAP appears to have promoted the merger (*ANAE*; E. R. Thomas, 2001). However, as outlined above, the merger *has* been found to be spreading to these presumably resistant areas in recent years. In terms of politics, the national border between the US and Canada has been found to prevent the merger from spreading from Canada to the US in some places. US and Canadian communities located on Lake Erie were found contrast quite strongly with regard to the COT-CAUGHT merger, with Canadians being fully merged, while communities on the US side of the lake maintain the distinction (Boberg, 2000).

One contradiction that has been ignored in previous research is the clear lead of younger speakers in the adoption of the COT-CAUGHT merger and the assumption that adults lead linguistic change adopted through diffusion, which is the presumed mechanism behind the spread of the merger. Only Johnson (2007) addresses this issue, suggesting that the main agents behind the spread of the COT-CAUGHT merger to Rhode Island are not adults, but children of in-migrating merged speakers from Massachusetts, as children *with* the distinction lose this distinction in continuous interaction with merged children. This hypothesis is supported by Yang (2009), who was able to prove that even small percentages of in-migrated merged children can lead to the acquisition of the merger by children in dialect areas that otherwise do not have the merger, and questions the assumed adult lead in changes acquired through the process of diffusion.

1.4 The Social Distribution of the NCS and COT-CAUGHT Merger

Especially in their early stages, changes from below such as the NCS and COT-CAUGHT merger are somewhat socially stratified and tend to concentrate among younger speakers of the upper working and lower middle classes. Thus, sociodemographic characteristics such as social class and age are central factors in the development of changes from below[20], and need to be considered in the life cycle of a sound change. Each of these contributing factors, and their role in the development and spread of the NCS and COT-CAUGHT merger, will be explained in the following subchapters.

1.4.1 Social Class

Linguistic change from below often follows the curvilinear social pattern of linguistic change, i.e. it originates in the interior social classes and spreads to members of different classes. While Labov (e.g. 2001) has argued that these types of changes tend to be internally motivated, others (e.g. J. Milroy & Milroy, 1985) believe that the common lead of the medial social classes in change from below is a result of a higher degree of face-to-face contact between speakers of different dialects owing to increased social mobility and

[20] Ethnicity is another factor that has been shown to affect change from below, including the Northern Cities Shift (e.g. King, 2017) and the COT-CAUGHT merger (e.g. Guy Bailey et al., 1993). However, ethnicity will not be considered in this study, as the speaker sample used for analysis is homogeneously white, with the exception of one speaker (see Chapter 2.4.1).

weaker social networks in these classes. Likewise, once an innovation has become stigmatized, speakers of the middle classes, especially those orienting upward, tend to be the first to reject this newly stigmatized feature (Chambers, 2008). Members of these medial classes have also been found to show the greatest amount of shifting in the usage of linguistic features depending on how much attention they pay to their speech (henceforth referred to as "speech styles" for convenience' sake, following Labov (1996)[21]) (Chambers, 2008).

Findings regarding the correlation between the NCS and social class have been contradictory in earlier studies. Although some findings support the expected curvilinear hypothesis for the NCS, this appears to have been the case for only some of the elements of the NCS. For example, Fasold (1969) reported a lead of the upper working class and lower middle class in the raising of TRAP and fronting/lowering of THOUGHT in Detroit; however, this pattern did not apply to LOT fronting in the same community. Eckert (e.g. 1989), on the other hand, found the opposite effect of social class in Detroit, i.e. a lead of working-class-oriented participants in the shifts of THOUGHT, STRUT and DRESS, while TRAP and LOT shifting showed no effect of social class. In Chicago, middle-class speakers were initially found to be leading the fronting of LOT as well as raising of TRAP (Herndobler, 1993), though a follow-up study of Chicago did not support this pattern for any of the NCS variables (McCarthy, 2007). In small-town Michigan, Gordon's (2001) results suggested that strongly middle-class oriented Michigan girls shift more than less middle-class oriented girls.[22] Ito (2001), on the other hand, found no class differences in the raising of TRAP in Michigan. The only correlation between social class, as indicated by education, and the NCS reported by *ANAE* pertains to DRESS, which was found to be shifted more commonly among speakers with lower levels of education. Based on the same data, Labov (2010) also reported a strong negative correlation between education and the backing of STRUT.

Despite having been researched extensively in the US, very few studies have found a strong link between social class and the degree of merger of LOT and THOUGHT. Only Baranowski (2013) found that a curvilinear pattern emerges among speakers in Charleston, South Carolina, with the middle class leading over the working and upper

---

[21] See Chapter 2.1 for a brief discussion.
[22] Gordon (2001) attributes these differences to gender affiliation rather than social class.

class in the acquisition of the merger. However, these findings contradict the *lack* of a significant social class effect on the COT-CAUGHT merger in an earlier study of the same community (Baranowski, 2007). Labov (2010) attributes this general lack of correlation to the fact that the COT-CAUGHT merger is not a salient sociolinguistic variable, and operates below the level of social awareness. However, mergers are not inherently resistant to social effects. The PIN-PEN merger, for example, appears to be led by speakers of lower classes in Charleston (Baranowski, 2013), and has been reported to correlate with the speakers' educational backgrounds (Labov, 2010).

1.4.2 Gender

Differences in the usage of phonological variants between male and female speakers are very common and well documented. In change from below, women generally use innovative forms more frequently than men, and in change from above, women are commonly found to prefer overtly prestigious forms while rejecting those that are stigmatized (Labov, 1994, 2001). This rejection of negatively evaluated variants in change from above is attributed to women being more aware of, or sensitive toward, stigmatization of non-standard features (e.g. Wolfram, 1969) and to a greater degree of social mobility among women (Labov, 2001). Men, on the other hand, tend to use non-standard features more commonly, as they frequently carry covert prestige. Eckert (e.g. 1989) takes a different approach to explaining the common female lead in the usage of linguistic change, arguing that choosing one variant over another may be a way of expressing group memberships of different kinds, a social practice that (especially adolescent) females tend to be more concerned with than males. Thus, she argues, gender is not directly linked to linguistic behavior, but rather to differences in social practices, which manifest themselves in linguistic behavior.

The general pattern of females leading over males in linguistic change has been found to apply to elements of the NCS in the Inland North. In the Detroit area, women led men in TRAP raising and THOUGHT fronting/lowering (Fasold, 1969) as well as LOT fronting (Eckert, 1989). Likewise, in Chicago, women were found to lead over men in the raising of TRAP (Herndobler, 1993; McCarthy, 2007) as well as in the backing/lowering of DRESS and fronting/lowering of THOUGHT (McCarthy, 2007). In small-town Michigan, TRAP raising was more advanced among female speakers as well (Gordon, 2001; Ito, 2001), and the same appeared to be the case for fronting of LOT and THOUGHT and backing of STRUT (Gordon, 2001). In the shifting of KIT and DRESS, females were also in the overall lead in Michigan; however, for both vowels, this lead depended on the direction of shifting, with males preferring the backed variants of both vowels, while females employed a broader range, including backing, lowering, or a combination of both. In New York, backing of DRESS was observed *only* for women, while STRUT backing was found in males as well, though still led by women (Clopper, Pisoni, & de Jong, 2005). For the Inland North as a whole, *ANAE* reported overall gender differences for the raising of TRAP[23], the backing of DRESS, and the lowering of THOUGHT, all of which were found to be led by women.

Evidence for the role of gender in mergers is considerably weaker than it is for vowel shifts. For the COT-CAUGHT merger, only a few studies have reported a (slight) female lead (Baranowski, 2013; Benson et al., 2011; Clopper et al., 2005; Gordon, 2006; Jasewicz, Fox, & Salmons, 2011), suggesting that, overall, gender is not a particularly relevant factor in the adoption of the COT-CAUGHT merger.

1.4.3 Age

One of the most commonly considered social variable in linguistic change is age, and linguistic differences between speakers of different age groups in a community are often taken to signal a generational linguistic change in progress, referred to as apparent-time change. The apparent-time paradigm is based on the hypothesis that people of different ages can be taken as representatives of different times because speakers will not change their language after having reached a certain age (Labov 1966, 1972). Relying on the validity of this hypothesis, the apparent-time paradigm allows for the observation of diachronic developments of linguistic change in a synchronic comparison of speakers of different ages (L. Milroy & Gordon, 2003). In change from below, the frequency of occurrence of innovative forms tends to increase gradually from the oldest to the youngest speakers. However, this correlation is not always a linear one; instead, sound change progresses slowly at the outset and the end, peaking at a maximum rate somewhere in between. Change from above, on the other hand, tends to be led by adults

[23] The female lead was found to be maintained despite the observation that TRAP raising was in recession in the Inland North.

who orient toward a new perceived standard in contact situations (Labov, 2007), of which children are, presumably, unaware.

Rather than advancing with every succeeding generation, linguistic change can also occur on a communal level, typically affecting lexical and syntactic features, but also phonetics. Communal change occurs when both the individual and the community are linguistically unstable, and is the result of a change affecting all members of the entire community simultaneous at the same rate, either by altering their frequency of use or by acquiring a new form (Labov, 1994). Although communal change affects all members of the community regardless of age, apparent-time trends generally signal this upcoming change, i.e. apparent- and real-time trends point in the same direction of change in the community (e.g. Hollett, 2006; Sankoff & Blondeau, 2007).

The NCS has been described as a change in progress as late as 2010 (Labov, 2010), however, the last source to report robust apparent-time progress toward the NCS in all variables was *ANAE* in 2006, with the exception of TRAP raising, which was found to be in recession throughout the Inland North (pp. 197, 211). Gordon (2001, p. 21) points out that *ANAE*'s findings contradict earlier reports by *LYS*, who found that shifting of KIT and DRESS, two of the later changes in the NCS, were observed only in older, but not in younger speakers, even in cities where the NCS was, presumably, present as a result of transmission. A lead in older speakers was also observed in individual communities throughout the Inland North, e.g. for TRAP raising and LOT fronting in Chicago (Herndobler, 1993). In places of (presumed) diffusion, TRAP raising was more advanced in older speakers in small-town Michigan (Gordon, 2001; Ito, 2001), and the NCS as a whole was more advanced in Sidney and Cooperstown in Northern New York, while in the rest of this part of New York, older and younger speakers overall did not differ significantly in the frequency with which they used NCS features. Only in one community, Ogdensburg, did the NCS seem to still be in progress (Dinkin, 2009, 2013). Similarly, in Michigan, adults and younger speakers were tied in the fronting of LOT, while THOUGHT fronting/lowering was more advanced among younger speakers. Additionally, Murray (2002), in a real-time comparison of St. Louis speakers in 1982 and 2001, found that the dialect pattern in the city was moving toward the NCS. St. Louis, however, is a special case and not necessarily part of the Inland North.

There are multiple potential reasons for the lack of apparent-time progress toward the NCS in these studies. One possible explanation might be that the presence of the NCS in the communities studied is the outcome of diffusion, i.e. change driven by adults, and thus would be expected to be more advanced among older speakers than in the younger generations (Chapter 1.3). However, diffusion is unlikely to account for the absence of apparent-time trends toward the NCS in *LYS* and *ANAE*, as their analyses focused on cities that were affected by the NCS through transmission rather than diffusion. *ANAE* explains the lack of apparent-time progress in NCS features with linguistic change in adults as follows: If adults participate in NCS features to the same extent as younger speakers, contrary to what would be expected, their linguistic behavior would not differ considerably from that of the younger generation, producing a flat-seeming apparent-time profile in the community (p. 211). Judging by the latest research on the NCS, a more likely explanation for the absence of any significant apparent-time trends toward NCS patterns in these studies is that the NCS was already in recession, with younger speakers leading the change away from NCS variants. This will be discussed in more detail in Chapter 1.6.1.

Contrary to the NCS, the COT-CAUGHT merger appears to be progressing rather rapidly in apparent time. In the adoption of the merger, *ANAE* found an overall advantage for younger speakers, a lead that was particularly strong in Eastern New England and the West, where the merger was assumed to be progressing toward completion, as well as in the South and the Midland, where it was assumed to be in gradual transition. Similar results have been reported in virtually all studies examining the progress of the merger in North American communities (e.g. Boberg & Strassel, 1995; Gordon, 2006; Irons, 2007; Jasewicz et al., 2011; Johnson, 2007). Only a few regions showed little to no progress toward the merger in *ANAE*. One of these was Canada, where the merger is "well enough established to show no correlation with age" (*ANAE,* p. 65). In other words, because Canada was already fully merged at the time of the study, no apparent-time difference between older and younger generations in their degree of merger could be observed in the data. Another area that showed no progress toward the merger of LOT and THOUGHT was the Inland North. This was argued to be related to the presence of the NCS in this area, which supposedly offered "stable resistance" to the merger owing to the fronting of LOT. However, three out of nine *ANAE* speakers in the Western New York portion of the Inland North were found to be transitional, mostly in production, indicating an incipient COT-CAUGHT merger. For the Inland North as a whole, the percentage was much lower, and the few speakers who were transitional in the rest of the Inland North were transitional in *perception*, not production (*ANAE*, p. 63). Further support for the emergence of the COT-CAUGHT merger in New York State has been provided by Dinkin (2009, 2010, 2019), as he found the COT-CAUGHT merger to be in progress throughout most of Upstate New York. Similarly, Benson et al. (2011) found significant apparent-time progress toward the merger in the northwesternmost corner of the Inland North, with younger speakers being fully merged or showing only insignificant qualitative differences between LOT and THOUGHT, while older speakers were clearly more distinct. Additionally, as pointed out above, Gordon (2001) found that the lowering of THOUGHT, a shift he observed regularly in his data (presumably NCS-related), was the only change that was led by younger speakers, while LOT was the only shift for which younger and older speakers were tied. Although Gordon does not attribute these findings to a potential merger, this pattern might be indicative of an incipient merger in the young Michiganders in Gordon's data as well.

1.5 The Social Perception of the NCS and COT-CAUGHT Merger

In changes from above and changes from below that have reached a certain level of social awareness, a hearer's social perception and evaluation of a linguistic feature is central to the fate of this feature, as the social meanings attached to linguistic change can enhance or hinder its further development, or even reverse it, as will be discussed in Chapter 1.6. Usually, these social meanings are reflections of the attitudes toward the speakers who use them (Preston, 2013), which, in turn, often depend on the socio-economic context in a particular community or dialect region (Coupland, 2009). In general, all superficial, i.e. observable, elements of language might be subject to social evaluation (Labov, 2001), and regarding sound change, this pertains to the concrete phonetic realization of a speech sound, while the underlying phonological structure this change occurs in, such as a chain shift or a merger, are argued to *not* be subject of social perception (Eckert & Labov, 2017). However, not all observable features seem to attract social meaning to the same degree, which leads Labov (e.g. 1994, 2001) to differentiate between "indicators", "markers" and "stereotypes":

- An indicator is described as a linguistic feature that correlates with sociodemographic categories, such as region, or social class, and thus signals group membership. However, these kinds of features are only observable by cultural outsiders, generally linguists; speakers are completely unaware of these indicators, which are therefore generally characterized by a lack of intra-speaker variation. Typical examples of indicators are vowel shifts in their initial stages and vowel mergers.
- A marker is a linguistic feature that has become associated with a particular style of speech and is thus socially meaningful. Although speakers are not necessarily aware of it, markers can carry some social evaluation in the form of prestige or stigma. Therefore, substantial intra-speaker variation is common for markers, typically depending on the level of formality. If a particular form has come to be negatively evaluated, this is often reflected in negative reactions on subjective reaction tests. Some chain shifts, or rather, elements thereof, have been found to carry social evaluation in the form of a marker, e.g. the Southern Shift, which has been reported as being negatively evaluated (e.g. Baranowski, 2013).
- A stereotype, like a marker, is associated with a particular speech style and/or with a particular social group. In contrast to markers, however, speakers tend to be fully aware of stereotypes. The lack rhoticity in New York City, for example, is a highly stigmatized feature, which is often associated with uneducated speakers of the lower classes (e.g. Becker, 2014).

If listeners establish a relation between a certain phonetic form and particular characteristics of the speakers who use them, this form will undergo two or all of these phases in its life cycle. This idea has been expanded on by Eckert (e.g. 2008), who argues that the social meanings of a linguistic form can be reinterpreted infinitely, creating a multifaceted field of potential, related meaning, which she calls an indexical field. For example, the two variants for the realization of the English ending –*ing* have been found to be indexed with qualities relating to the speakers' level of education, the formality of the situation, articulateness and the amount of effort put into the speech act (e.g. Campbell-Kibler, 2006, 2007; Eckert, 2008). These indexical links, in turn, allow speakers to actively use linguistic forms in processes of identity construction or performance.

1.5.1 The NCS – From Indicator to Marker

Traditionally, the NCS has been argued to be unavailable for social evaluation, and NCS variables have been assumed to be indicators rather than markers, based on the observation that there was "little style shifting[24] associated with their social distribution" (Labov, 2010, p. 194); however, over time, some elements of the NCS do appear to have attracted social meaning. A number of earlier studies reported that raised TRAP was perceived as an emerging prestige pronunciation, a conclusion that was based on increased usage of the raised variant in more careful speech (Ash, 1999 as cited by Labov, 2010, p. 59; Herndobler, 1993; McCarthy, 2007). On a broader level, Preston (1996) was able to show that Michiganders perceive the speech of (most of) the Inland North[25] as one of the most correct in comparison with other US dialects. Likewise, outsiders' perspectives on the NCS suggested a certain level of prestige. For example, in St. Louis, the Inland North dialect was perceived as having a high standard of correctness (Murray, 2002).

Additionally, NCS variables were found to be indexed with meanings relating to gender and urbanity in earlier research. In the suburbs of Detroit, the backing of DRESS and STRUT were found to index urban toughness, while raised TRAP, fronted LOT and lowered THOUGHT appeared to be associated with femininity (Eckert, e.g. 1988). Likewise, in Chicago, raised TRAP was found to be associated with femininity as well as sophistication and cosmopolitanism, while fronted LOT signified masculinity and independence, though without the connotation of being a nonstandard feature (Herndobler, 1993). Similarly, in small-town Michigan, raised TRAP was perceived as signifying the urban and fashionable lifestyle of the city (Ito, 2001; Ito & Preston, 1998), and fronted LOT as indexing femininity (Gordon, 2001). Additionally, Gordon found that the NCS may have been associated with city people in one of the small towns he analyzed. He suggests that, similar to the native islanders in Labov's Martha's Vineyard study (1963), the influx of people from the city into small-town Chelsea may have created feelings of resentment among Chelsea natives, leading them to reject the NCS and retain the conservative vowel forms to symbolize their local Chelsea identity.

---

[24] See Chapter 2.1.

[25] New York State was rated as less "correct" than other Inland North states, and Washington State and Colorado were rated as similarly "correct" as the Inland North.

In addition to locally defined social meanings, Labov (2010) argued that cultural oppositions between the North and upland Southerners in the Midland may have been a contributing factor to the abrupt dialect boundary between these two areas, suggesting that social perceptions of the NCS also operated on a much larger scale, and were embedded in sociohistorical as well as political aspects that affected the Inland North as a whole. As mentioned in Chapter 1.3.1, Yankee settlers in the North differed quite drastically from upland Southerners who settled in the Midland. Not only did they contrast in terms of migration style, preferred community types, housing location, and time spent in the community, they have also differed in cultural and political attitudes. Labov (2010, p. 216) cites Power (1953, p. 6) as describing Midlanders of southern origin to perceive Yankees as busybodies and meddlers with an "inclination to regulate the morals of the whole society". Labov (2010) suggests that the Yankees' general vision of being superior and their interference with Midland culture also affected Midlanders' linguistic practices, which Yankees disapproved of. The perception of the Yankee way of speaking being the "better" one is reflected in the fact that the dialect of the North was selected as the broadcasting standard in the US and became the baseline for "General American English", of which Midlanders were rather critical.

Labov (2010) points out that these cultural differences have manifested themselves in several ways. Based on Elazar's (1972) definition of political cultures in the US, Labov (p. 218) explains that the North is generally described as following Yankee "moralistic" tradition, understanding the government as an institution of public service that should serve the community. On the other hand, Midlanders are "individualists" who reject the Yankee's inclination of government interference in individuals' lives. In the early 2000s, these differences were also reflected in voting patterns, with the Inland North (among others) voting predominantly Democrat, and the Midland mainly Republican. A similar, though arguably weaker link was established between the Inland North and Midland dialect areas and the history of the death penalty in the respective states, as the Northern dialect region largely coincides with states where, as of 2004, the death penalty was no longer authorized. One notable exception to this is the State of New York. Labov argued that, because most of New York's population was outside of the Inland North, this was not a considerable exception; however, as Dinkin's (2009, 2013) research has shown, the part of New York State that belongs to the Inland North is considerably greater than originally assumed. Overall, however, these patterns suggest that the NCS might have been associated with particular political ideologies. In an experiment that tested this hypothesis, Labov (2010) found this to be the case at least for some of the participants in the study.

One exception to the North-Midland opposition is the corridor from Chicago to St. Louis. As was mentioned in Chapter 1.3.1, communities in this area have adopted many of the NCS features despite being surrounded by the Midland dialect. Labov (2007) references Frazer (1979) in pointing out that ideological factors very likely contributed to the possibility of these communities adopting phonological features of the North. While most Midlanders rejected the ideologies of the North, speakers along the corridor, particularly those with German background, favored the developments in the North, thus increasing their receptivity to Northern influence, including their speech patterns. Along the same lines, Bigham (2007) found that a speaker's regional orientation correlated with their NCS participation: Speakers orienting toward Chicago were found to adopt NCS features more commonly than those who orient more toward Southern Illinois.

The findings presented in the previous paragraphs suggest that NCS variables may have been socially evaluated, though presumably unconsciously, from very early on. More recent research has found that at least some of the NCS variables have now become available for *overt* social evaluation, generally in a negative sense. The first indication of a potential negative perception of an NCS variable was observed in Chicago, where raised TRAP was found to be corrected in more formal speech, seemingly owing to social pressure from above (Callary, 1975). More recently, accounts from focus-group interviews and laymen's interpretations of raised TRAP that were posted on social media have shown that raised realizations of TRAP have attracted overt comments (Savage et al., 2016). Anecdotal evidence suggests an increasing stigma around raised TRAP in Syracuse (Driscoll & Lape, 2015), and in Lansing, style-shifting patterns in the usage of raised TRAP among speakers born in the 1990s reveal that the unraised variant is preferred in more careful speech (Nesbitt & Mason, 2016). Additionally, fronted variants of LOT are evaluated negatively by participants in Lansing, while lowering of DRESS is associated with positive characteristics such as intelligence, confidence, articulateness, and friendliness. Especially among women, these positive evaluations of lowered DRESS are reflected in their own speech production, as they tend to lower DRESS more in careful speech to a greater extent than in spontaneous speech (Nesbitt & Mason, 2016; Savage & Mason, 2018; Savage et al., 2016).

1.5.2 The COT-CAUGHT Merger as the Perceived Standard?

Mergers have generally been described as unobservable elements of language, because they are abstract structural phonological processes or categories (e.g. Eckert & Labov, 2017); thus, similar to chain shifts in their initial stages, they have been argued to be more or less invisible to social evaluation (Labov, 2001). Unlike chain shifts, Labov (1994, p. 342) argues, "remain at this level, without social prestige or stigma, long after they have gone to completion". It is, however, possible for the observable changes in vowel quality that lead to the merger to raise to the level of social awareness and carry social meaning. Under this scenario, listeners would not be able to evaluate the fact that someone does or does not make a distinction between two sounds, however, they could react to deviating qualities in certain sounds involved in the merger. Eckert and Labov (2017) argue that this is only the case for mergers that occur in high frequency words, with one exception: the COT-CAUGHT merger.

Because mergers tend to remain below the level of conscious awareness, style shifting is rarely observed. Nevertheless, differences in the degree of merger across speech styles have been reported in the literature, though the reasons for these differences are the subject of an ongoing debate. On the one hand, these style differences are believed to be due to the fact that mergers are commonly found to be stylistically gradual types of change (Chambers, 2008), i.e. they occur in one style before they occur in the other, with merger in spontaneous speech frequently preceding merger in more careful speech (e.g. Boberg & Strassel, 1995; Johnson, 2007). In this case, style differences may merely be a result of a change in progress. In other cases, however, mergers have been found to be more advanced in more careful speech, especially minimal pairs, while in spontaneous speech the distinction is maintained (D'Onofrio et al., 2016; Di Paolo, 1992; *LYS*). This style pattern, which is particularly common for near-mergers (Labov, 1994, p. 402), has been interpreted as an orientation toward a perceived norm, which the speaker strives to achieve when paying more attention to speech, but not in casual conversations. The validity of this assumption has been challenged, however, and it has been argued that merged production in careful speech is an unreliable indicator of the degree of merger, because it might be an "artifact that only comes about through formal testing" (Labov, 1994, p. 402).

For some mergers, there is clear evidence for a certain degree of social awareness and evaluation attached to them. The perception of the NEAR-SQUARE merger in New

Zealand, for example, has been found to depend on social characteristics of the speakers, e.g. age and social class, indicating that social expectations play a role in the perception of mergers (Hay, Warren, & Drager, 2006). Similar conclusions have been drawn for the PIN-PEN merger, which appears to be (negatively) associated with a certain dialect and age group in the South (Baranowski, 2013; Koops, Gentry, & Pantos, 2008; E. R. Thomas, 2004). Similarly, the NURSE-NORTH merger in Tyneside English (Geordie) has been reported to be highly stigmatized (Maguire et al., 2013).

Studies on the COT-CAUGHT merger, on the other hand, have rarely reported evidence of social awareness or evaluation. To the best of my knowledge, only one study, conducted by Di Paolo (1992), has tested the social perception of the COT-CAUGHT merger in an experimental setting. The results suggest that merged production is perceived to be more standard than distinct production. Additionally, merged speakers were perceived as more successful and as having a more favorable personality. All of these judgments held for participants who originated from both merged and distinct dialect areas.

1.6 Reversal of Sound Change

As was pointed out above, negative social evaluation of a sound change can not only hinder its spread, but even reverse it. For example, Becker's (2014) study of New York City suggests that lowering of raised THOUGHT among certain communities in the city is motivated by contemporary social meanings of the raised variant, which is associated with persona that are negatively evaluated. Becker also points out that these negative associations seem to have spread far beyond city limits, as other scholars have found evidence that negative evaluations of raised THOUGHT in New York City have led to a withdrawal from this feature in Philadelphia, almost 100 miles away.

Commonly, increasing negative evaluations of a linguistic feature stem from drastic social changes in the community, so that the rejection or reversal of a sound change is due to social pressures (e.g. Hall-Lew, 2017; Labov, 1994). In other words, "given the right cultural configuration, there are very few general patterns that cannot be reversed" (Labov, 1994, p. 120). If this is the case, it seems possible that growing negative social evaluation of NCS variants might trigger their reversal, which is, in fact, what the growing body of research on the NCS in recent years has found. For the COT-CAUGHT merger, on the other hand, no such indications exist.

1.6.1 Recession of the NCS

The most recent studies examining the status of the NCS have indicated an abandonment of NCS features. The first explicit descriptions of NCS recession were presented by McCarthy (2010), who found that, in Chicago, LOT was backing away from its NCS fronted position in apparent time, and Dinkin (2009, 2013), who found that, in New York State, LOT was retracting and TRAP starting to lower. The retreat from NCS variants of LOT and TRAP have since been observed in several other NCS communities, including, Rochester (King, 2017; Morgan et al., 2017), Buffalo (Milholland, 2018), Syracuse (Driscoll & Lape, 2015), Lansing (Wagner et al., 2016), and Northwestern Wisconsin (Fox, 2014, 2016). In two new studies on Chicago, ongoing reversals of both sounds confirmed McCarthy's early indications of NCS recession (D'Onofrio & Benheim, 2018; Durian & Cameron, 2018).

However, a potential recession of the NCS was already evident in earlier studies. Herndobler (1993) found that middle aged speakers, rather than the youngest ones, had the highest degree of TRAP raising. She interpreted this finding to be related to identity formation processes among adults rather than caused by a recession, based on a female advantage in the raising of TRAP. Likewise, Gordon (2001) reported that the middle generation was more advanced than the oldest or youngest generations in most NCS variables. Although he considered the possibility of an NCS recession, he did not pursue this thought further owing to a lack of data that would allow for a real-time comparison. In his case, the observed age-patterns might also have been a result of diffusion. As his study focused on small-town communities, to which the NCS may have diffused, the age patterns he observed might have been expected, as adults would be more likely to pick up the changes first. This does, however, not explain why the younger generation was more advanced in the lowering of THOUGHT, and why the younger and the middle generation were tied in the fronting of LOT. On a broader level, as mentioned earlier, *ANAE* (pp. 197, 211) found that TRAP raising was receding all over the Inland North.

The loss of the NCS appears to involve both reversals and continuations of trajectories, as well as reconfigurations of allophonic variation. In combination, these developments indicate an orientation of NCS communities toward the supra-local system of the Elsewhere Shift, characterized by the COT-CAUGHT merger, the retraction of pre-oral TRAP resulting in a low nasal or continuous TRAP system, and the lowering and/or backing of DRESS and KIT.[26] Commonly, these changes are accompanied by the fronting of GOOSE, FOOT, STRUT and GOAT (see Chapter 7.1). The Elsewhere Shift is also known as California or Canadian Shift, and as these two latter names suggest, this shift has traditionally been associated with California and Canada (e.g. Boberg, 2005; Kennedy & Grama, 2012). However, it has also been found, for example, in Vermont (Boberg, 2011), Ohio (Durian, 2012), Southern Illinois (Bigham, 2009), Kansas (Kohn & Stithem, 2015), South Carolina (Baranowski, 2013) and Alaska (Bowie et al., 2012).

In the Inland North, a potential formation of the Elsewhere system has been observed in various communities as well. It was first reported in Lansing (Wagner et al., 2016), where TRAP is developing an allophonic alternation in which pre-nasal TRAP remains in raised position while pre-oral TRAP retracts, resulting in either a low nasal or low continuous TRAP system. Additionally, Nesbitt and Mason (2016) reported GOOSE and GOAT fronting, as well as retraction of LOT in the community. The only ongoing change reminiscent of the NCS in Lansing is the lowering and backing of DRESS; however, this change is also part of the Elsewhere Shift, and in the absence of any other observable NCS features, it is likely that DRESS is following this trajectory as part of the Elsewhere Shift, rather than the NCS. Morgan et al. (2017) found very similar patterns in Detroit, including pre-oral TRAP lowering, DRESS and KIT lowering/backing, LOT backing and GOOSE fronting.

These developments have also been identified in the New York portion of the Inland North. In Syracuse, TRAP seems to be lowering and backing in apparent time, while LOT has been found to be retracting, and STRUT was fronting (Driscoll & Lape, 2015). Dinkin (2009, 2013) and King (2017) reported the development of a low nasal or continuous TRAP system in Rochester and Northern New York in addition to the retraction of LOT, so that it seems likely that the Elsewhere Shift is starting to take shape in New York as well.[27] King also found THOUGHT to be backing in apparent time in Rochester, which is consistent with NCS reversal, but not with the Elsewhere Shift, as it prevents THOUGHT from merging with LOT. In fact, King reported no progress toward a merger of these two vowels in Rochester, while Dinkin did find evidence for an incipient merger of LOT and THOUGHT in Northern New York.

---

[26] A known exception to this seems to be the St. Louis Corridor, which has been found to orient toward Midland patterns instead (Friedman, 2014).

[27] However, Driscoll and Lape (2015) also report fronting of DRESS and fronting and raising of KIT, which runs counter to the Elsewhere Shift.

While the reversal of the NCS seems to be a relatively robust trend across the Inland North, it appears that not all speakers participate in this reversal to the same extent. Recent research on the loss of the NCS in Buffalo, Chicago, and Lansing has shown that TRAP-raising persists longer in speakers with lower levels of education (Milholland, 2018) or those that belong to the working class (Durian & Cameron, 2018; Nesbitt, 2018). Middle-class speakers, on the other hand, have been found to lead the change away from the NCS, as exemplified by pre-oral TRAP lowering in Lansing (Wagner et al., 2016) and Detroit (Morgan et al., 2017). Additionally, research in Chicago has shown that local identity concerns may play a role in the rejection or maintenance of the NCS (D'Onofrio & Benheim, 2018).

It seems likely that the recession of the NCS is a direct result of the negative social evaluations of some of the elements involved in the shift, which in turn may result from economic changes in each of the NCS communities. As explained in Chapter 1.5, at least some of the NCS features have become stigmatized to some degree in Syracuse and Lansing, including raised TRAP and fronted LOT. Consequently, TRAP is being reorganized, and LOT reversed its trajectory. Nesbitt (2018) found that the reorganization of the TRAP system in Lansing started with speakers of the Baby Boomer generation (born 1946–1955), which, in turn, coincides with the industrial decline in the community: Lansing, in the early 20th century, was a manufacturing powerhouse, and working-class identity was closely tied to this industry. Because of that, Nesbitt argues, local speech patterns such as the NCS were beneficial in the community at this time. In the late 20th century, Lansing's manufacturing industry collapsed, leading to a shift to a service industry, as well as to population decline, unemployment and an increase in crime rate. These sociocultural changes and the ensuing decline in prestige for the city, Nesbitt suggests, may have led to the markedness of all things local, including speech patterns such as raised TRAP. This, in turn, may have led to a rejection of these features, particularly among middle-class and upwardly mobile speakers. In Chicago, the timing of NCS reversal was found to be somewhat later, starting with speakers born after 1975; however, it has not been discussed whether this change is linked to industrial decline (Durian & Cameron, 2018). D'Onofrio and Benheim (2018) suggest that reversals of sound change are not necessarily a result of orientation toward a supra-local norm, but rather are driven by shifts in the way local identity can be indexed.

1.6.2 Once a Merger, Always a Merger?

"Garde's Principle" of mergers generalizes that true mergers have unidirectional character, i.e. are irreversible by linguistic means (Labov, 1994, p. 311). This principle is based on the logic that a speaker who does not have a distinction between two phonemes in production and perception cannot restore the contrast, because the required linguistic information does not exist in their phonology. Attempts at doing so would either be unsuccessful or, if somewhat successful, would entail a certain degree of hyper-correction (i.e. error) (Maguire et al., 2013), which, in turn, can attract social commentary (e.g. Baranowski, 2007).

There are, however, a few, albeit rare, cases of full restoration of distinction. Examples of this are the historic mergers, and un-mergers, of MEAT and MATE in Belfast (J. Milroy & Harris, 1980) and of /w/ and /v/ in Southeastern England (Trudgill et al., 2009). More recent mergers that have become or are currently becoming undone are for example the merger of /ahr/ and /ɔhr/ in St. Louis (*ANAE,* p. 277) and the PIN-PEN merger in Houston (Koops et al., 2008). Two explanations for the undoing of these mergers have been proposed: Labov (1994) suggests that the reason for mergers to unmerge is that phonemic contrast between the phonemes involved in these mergers was never fully lost in production, i.e. they were merely near-mergers in which the respective vowel pairs were produced with a small but consistent distinction. Thus, according to this hypothesis, merger reversals are not the manifestation of a speaker's ability to reintroduce a distinction, as phonological information is never lost completely. The second hypothesis that attempts to account for the undoing of mergers assumes that the driving force behind it is dialect contact (e.g. Trudgill et al., 2009). According to this hypothesis, distinction is reintroduced through contact with distinct speakers, and evidence for this is drawn from the un-merging of MEAT and MATE (J. Milroy, 1992) and /w/ and /v/ (Trudgill et al., 2009). The reversal of the PIN-PEN merger in Houston is also believed to be due to increased contact with distinct speakers (as well as negative evaluation of merged production) (Koops et al., 2008).

For the merger of LOT and THOUGHT, no *true* reversals have yet been reported to the best of my knowledge. Kurath (1939) described speakers in Eastern New England as producing distinct LOT and THOUGHT, which would indicate a reversal of their very early merger. Later research, however, found a complete merger in some parts of Eastern New England, while in other parts, the merger is said to never have been present (Johnson, 2007), and thus not reversed. Another instance of a potential reversal of the COT-CAUGHT merger was reported by *ANAE* for the Upper Midwest. *ANAE* explains that, in 1966, the merger was present throughout South Dakota and Nebraska, as well as most of Minnesota. In the *ANAE* data, collected about 30 years later, on the other hand, the authors did not find more than a few transitional cities in these states. The authors do not offer an explanation for this incongruity, and the question of whether or not the merger actually reversed in these areas or was never really present to begin with remains unanswered. On a smaller scale, Johnson (2007) described children of merged parents acquiring the distinction if frequently exposed to it by their peer group.

1.7 Purpose of the Current Study

From the descriptions of the NCS and the COT-CAUGHT merger presented above, it is apparent that both changes have been studied extensively in the past. Nevertheless, a number of gaps have remained in the research, some of which I would like to address in the present study.

A first concern is the trigger of a potential incipient Elsewhere Shift in NCS communities. While recent research reports that the NCS is receding and Inland North communities are likely orienting toward the supra-local Elsewhere system, few of these studies have analyzed or reported on the status of the merger of LOT and THOUGHT, which is assumed to be the trigger of this particular shift.[28] Although most studies report LOT retraction, its relation to THOUGHT is largely under-researched.

Potential social evaluation of the variables involved in the changes currently observed in the Inland North is another concern that will be addressed in the current project. While a number of studies indicate that some elements of the NCS have reached the level of conscious awareness, most of them say little about the nature of the social evaluations attached to them. Additionally, few of these recent studies take style shifting into account in their analysis. The only exception seems to be a recent study by Savage and Mason (2018), who have been able to show that *positive* evaluations of lowered DRESS seem to lead to DRESS lowering in the community. However, the question of how potential *negative* evaluations might affect production patterns remains unanswered.

[28] It has, however, also been argued that TRAP retraction might be the triggering event of the Elsewhere Shift in traditional NCS communities (Mason, 2018).

Furthermore, most of the studies only take a few selected NCS variables into account, leaving others unexplored. The COT-CAUGHT merger, on the other hand, has never been tested for social evaluation in this area. As a result, the underlying motivations for the changes currently progressing in the Inland North remain largely unknown.

The third objective of the study is to contribute to the existing literature of NCS reversal with a focus on a small-town community instead of urban centers. It has been established that rural and urban communities have been treating the NCS somewhat differently owing to issues related to transmission and diffusion (Chapter 1.3.1). Likewise, recent research has suggested that the reversal of the NCS might be progressing at different rates depending on community size. For example, Driscoll and Lape (2015), in their analysis of speakers in and around Syracuse, found that TRAP retraction seems to be progressing significantly faster among speakers in the suburbs than among urban speakers. However, very few recent studies have included smaller and rural communities, and even fewer studies have carried out an in-depth sociolinguistic analysis of these communities (e.g. Benson et al., 2011; Fox, 2014). Likewise, the advancement of the COT-CAUGHT merger in small-town communities remains relatively unknown. As a result, developments in these and other sound changes in smaller communities are largely left unexplored. If this primary focus on urban centers continues, Britain (2012) argues, we run the risk of uncovering only a limited part of the picture of language variation and change.

In the current project, these gaps will be addressed by analyzing changes in the perception and production of variables involved in the NCS, the COT-CAUGHT merger and the Elsewhere Shift in a rural, small-town community in Northern New York, named Ogdensburg. Based on the objectives outlined above, the following research questions have been formulated for this project:

I. To what extent is there evidence that the NCS, the COT-CAUGHT merger, the Elsewhere Shift, and STRUT, GOOSE, FOOT and GOAT fronting are (still) in progress or have gone to completion in Ogdensburg?
    a. How far and in which direction have they progressed?
    b. How do different demographic factors correlate to the changes in each of the involved variables?

II. To what extent do the variables of interest carry social meaning in Ogdensburg?
   a. To what extent are speakers aware of phonetic and phonological features in their region compared to other regions?
   b. How does potential social evaluation of the variables of interest relate to speech production patterns?

III. To what extent does phonetic change in Ogdensburg resemble developments in urban Inland North communities?

1.8 The Speech Community: Ogdensburg, New York

Ogdensburg, also referred to as The Maple City or, casually, the Burg, is a small community of about 11,000 people in New York's rural North Country. The city stretches for about 7 miles along the southern shores of the St. Lawrence River, the national border between the US and Canada, in the sparsely populated St. Lawrence County (see Figure 5 below). Most of St. Lawrence County is made up of farmland, interspersed with several villages. Ogdensburg, despite its small size, is one of the most populous settlements in the county, and the only community with the legal status of a city.

Ogdensburg was founded in 1749, and the first settlers under American flag arrived in 1796 (Durant & Peirce, 1878). Because the majority of Ogdensburg's town records were destroyed in a fire in 1839 (Hough, 1853), information about the origins of these early settlers is sparse, and that provided in the available sources are quite vague. The majority of these sources indicate the rather broad region of New England as the early settlers' origin (e.g. Kenrick, 1846; Merriam, 1907), though some seem to point more specifically to Vermont (e.g. Durant & Peirce, 1878; Landon, 1932), and Hough (1853, p. 398) explains that early Ogdensburgers made an effort to draw in people from Vermont, particularly "that part of Vermont from whence the greatest emigration to this country comes … beyond the mountains, near the borders of New Hampshire". Garand (1927), on the other hand, states that only a few emigrants from that part of New England found their way to Ogdensburg. In addition to New England, Northern New Jersey and Southern and Central New York are mentioned as the early settlers' points of origin (Durant & Peirce, 1878). A genealogy search on ancestry.com produced equally ambiguous results as these reports. Of the nine identified early settlers of Ogdensburg that resulted in a match on the search engine, I found that three originated from Southern New York, two from Northern New Jersey, two from Vermont and two from Connecticut.

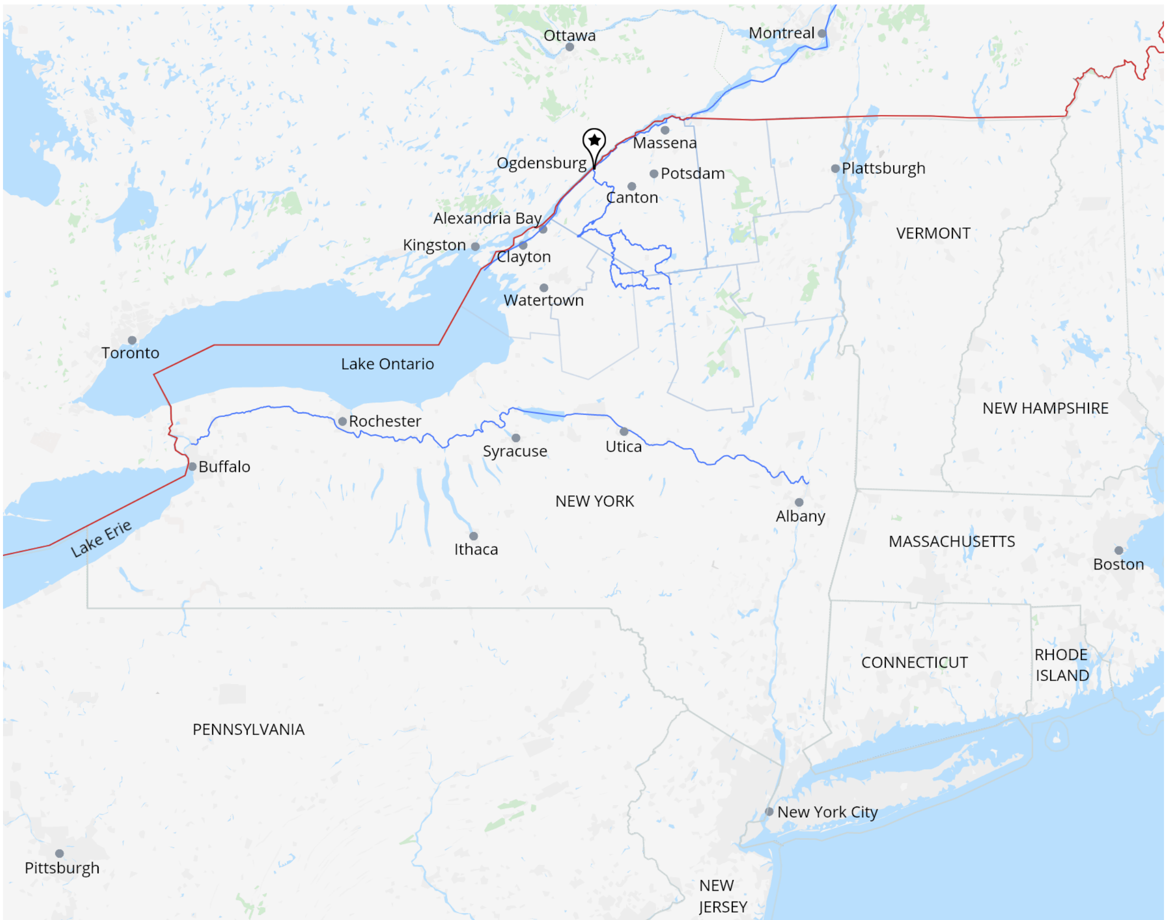


Figure 5: Ogdensburg's location in New York State. The counties outlined in the northern part of the state constitute the area referred to as the (geographic) North Country. The red line indicates the national border with Canada. The St. Lawrence River flows from Lake Ontario through Clayton, Alexandria Bay, Ogdensburg and Massena on to Montreal and the Atlantic Ocean, and is joined by the Oswegatchie River in Ogdensburg. Running through Western and Central New York toward Albany is the Erie Canal.

If the settlement patterns of Ogdensburg's neighboring communities are any indication of the roots of the city's own early settlers, Vermont does appear to be the primary origin. Overall, sources seem to agree that St. Lawrence County was primarily settled from Vermont (Durant & Peirce, 1878; Landon, 1932). The same specification is provided for the neighboring towns of Potsdam, Canton, Lisbon and Massena (Hough, 1853; Landon 1853), which were established around the same time as Ogdensburg. It therefore seems likely that a good number of Ogdensburg's early settlers originated from this northern part of New England as well. Western Vermont, the part of the state most relevant to the settlement of St. Lawrence County, was predominantly settled from Connecticut during the 18th century (Kurath, Bloch, & Hansen, 1939, p. 104), and Connecticut in turn was settled during the 17th century by the Dutch, as well as by migrants from earlier New England communities.

During the 19th and early 20th century, Ogdensburg flourished, capitalizing and relying on its location at the confluence of the St. Lawrence and Oswegatchie rivers, and on its proximity to Canada. Because Ogdensburg was the eastern head of river navigation on the St. Lawrence, ships coming from the Great Lakes had no choice but to stop in Ogdensburg, where freight was transferred to trains (the railroad system in and around Ogdensburg was completed in 1850 (Hough, 1853)) or smaller boats, and forwarded to New England or Montreal (Kenrick, 1846). By 1820, Ogdensburg had developed into an important port with about 4,000 inhabitants (Landon, 1932) [29], and although the opening of the Erie Canal in 1825 and other canals took away some of Ogdensburg's trade by offering alternative routes from the Great Lakes to the sea (Kenrick, 1846), Ogdensburg continued to grow. Soon, major manufacturers moved into the city and established their plants.[30] Along with the St. Lawrence State Hospital and the railroad, they became the biggest employers in Ogdensburg, and maintained this position well into the city's thriving era in the 1950s (Sandburg, 2015). Naturally, the city's economic growth affected its population, which continued to increase[31] and peaked at almost 17,000 in the early and mid 1900s ("Ogdensburg, NY Population," n.d.). Information about the origin of the migrants who made their way to Ogdensburg during this time is scarce, but it seems that, during the 1840s and 1850s, the largest group of migrants were the Famine Irish, followed by French Canadians in the 1880s, who continued to migrate to Ogdensburg throughout the 1900s alongside English speaking Canadians (Madlin, p.c., based on Taylor, 2017).

In the second half of the 20th century, a series of events caused a sharp economic decline in Ogdensburg. The completion of the St. Lawrence Seaway in 1959 opened the entire length of the river to large vessels and ships were now able to bypass Ogdensburg,

[29] The website of the Ogdensburg Chamber of Commerce states that by 1820, Ogdensburg's population had reached 16,610. However, this seems unlikely, because it was just under 16,000 in the census from 1910, though population was still increasing then ("Ogdensburg, NY Population," 2018).
[30] Standard Shade Roller Co., Diamond Paper Mill, and Acco Products are examples of the best-known companies in Ogdensburg.
[31] With the exception of 1920, when it saw a minor decline ("Ogdensburg, NY Population," 2018).

which damaged much of the city's economy (EIS, 1979, p. 10). In the 1960s and 70s, Ogdensburg underwent an Urban Renewal program in an attempt to modernize the city's downtown area and to encourage the opening of new business (EIS, 1973, p. 10). The project failed miserably, and "gutted much of the downtown area" (Vita Nuova, 2012, p. 23) by removing a great many of the city's historic buildings. Additionally, a bypass that routed traffic around instead of through the city was constructed alongside a number of stretch-malls outside the city. As a result, Ogdensburg lost numerous local businesses in the following years, and many of the city's manufacturing plants shut down operations owing to a declining market demand and the discontinued railway service in and out of the city (EIS, 1982). Only the St. Lawrence State Hospital was left as one of the main employers, but was drastically downsized and renamed St. Lawrence Psychiatric Center in 1972 (Sandburg, 2010). While a few of the hospital buildings are still in use, most of the campus, as well as the production sites of manufacturers, have been abandoned and left to deteriorate.

Since then, the city has been trying to stimulate economic development. Focus has been put mostly on redeveloping the downtown area and the waterfront for recreational and commercial use (Robinson, 2017) in order to increase the city's potential to attract tourism. Although Ogdensburg's port does currently accommodate about 12 river cruise boats per season, many of them merely clear US customs in Ogdensburg, without passengers disembarking to visit the city (Vita Nuova, 2012). The main challenges in the attempt to revitalize the city are the lingering effects of the industrial decline and Urban Renewal. Because of the stretch-mall projects outside the city, shopping in Ogdensburg is limited, and there are few restaurants and entertainment resources in the downtown area (Vita Nuova, 2012). Due to disinvestment and decay of buildings and vacant sites, which are surrounded by chain link fences (Robinson, 2017), the city is described as having a "depressed appearance" and "depressed physical environment" (DADRAS Architects, 2011, p. 10). But it was not only appearance that suffered. Some of the former production sites were left deeply contaminated with pollutants, including tainted soil and water. So far, the city has been able to acquire a number of these deserted sites and has removed some of the abandoned buildings. Polluted plant sites are still in the process of being cleaned up (Robinson, 2017) in preparation for development. One of the biggest successes so far has been the expansion of Ogdensburg's airport, which, starting in 2016, has been able to offer cheap direct connections to Florida (Robinson, 2018), one of the area's main travel destinations.

As a result of the economic decline and slow improvements, Ogdensburg currently performs below average in terms of socioeconomic success compared to the rest of St. Lawrence County and New York State[32]. The city's population has declined drastically and continuously since the 1960/70[33] to its current 11,000, as people have moved away for job opportunities and in-migration has been limited: In 2016, only about 21% of the city's residents had relocated to Ogdensburg within the last year, the majority of them from within St. Lawrence County or another county within New York State. Only 2% moved in from outside the state. In-migration from other countries is even more limited: In 2016, only about 500 of the 11,000 residents were foreign-born, Latin and North America being the two most frequent places of origin. This limited in-migration is also reflected in the city's rather homogeneous population: 88.3% of Ogdensburg's inhabitants are White, and 7.3% are Black or African American. The remaining percentages are distributed mostly between American Indian and Alaska Natives, Asians and "unidentified other races" (US Census Bureau, n.d.). Of the 8,900 people over the age of 16 currently living in Ogdensburg, about half are in the city's labor force, the majority being employed in the educational, health and social services (US Census Bureau, n.d.), constituting what locals call the "grey collar" class[34] of Ogdensburg. Because of limited employment opportunities in the city itself, most residents commute to work, many to the surrounding towns of Canton and Potsdam, located in the dialectological North Country (see Figure 6 below). The main employers in the city itself are now two correctional facilities, which opened in 1982 and 1988. Since 2007, they also include a sex offender treatment facility, which, like the prisons, is surrounded by barbed wire (Sandburg, 2016). Although this contributes "to what is currently a negative image for the urban character of Ogdensburg" (DADRAS Architects, 2011, p. 16), the prisons are a source of pride for the residents of Ogdensburg and are considered integral parts of the community. Both prisons were voluntarily brought into the city in hopes of creating new jobs after the closing of the plants (Sandburg, 2016), and Ogdensburgers are grateful for the employment opportunities

[32] The demographic estimates in this paragraph refer to the estimates for 2016.
[33] With the exceptions of 1990, 2012, 2014, when the city's population saw minor inclines, each below 1%.
[34] This refers to a middle ground between white and blue-collar workers, described as semi-professional workers.

these facilities have brought along. Nevertheless, unemployment is relatively high in Ogdensburg[35], and the city has one of the lowest median household incomes in St. Lawrence County ($36,832), third only to Massena and Clifton. This places Ogdensburg well below the county average of $46,000, which in turn is among the 10 lowest household incomes of the 62 counties in the State of New York. Although there are several institutions of higher education in the vicinity[36], only about 16% of Ogdensburg's population holds a college degree, while the state average is more than twice as high (US Census Bureau, n.d.).

One aspect that has not changed over the years is the city's contact with Canada. From very early on, Ogdensburg had very strong ties to its Canadian neighbor, the town of Prescott. The two communities were connected by a ferry as early as 1812 in the summer and by the ice on the St. Lawrence River in the winter (Kenrick, 1846). To the dismay of many, the ferry was replaced in 1960 by the Ogdensburg-Prescott International Bridge that now spans the St. Lawrence River. Nevertheless, Canadians frequently visit Ogdensburg, mainly for grocery shopping and cheaper gas (Burke, 2018), and many Ogdensburgers frequently visit Prescott and Brockville, or travel to Montreal, Ottawa or Kingston for shopping and entertainment purposes. In fact, Ottawa and Kingston are the nearest cities with populations greater than 50,000, to which Ogdensburg is, in terms of infrastructure, better connected than any US population centers. Additional destinations on a daily basis are the surrounding towns of Canton and Potsdam.

So far, few studies have looked at speech patterns in this rural part of New York State, and none have focused in-depth on a single speech community. Dinkin (2009, 2013) sampled Ogdensburg as part of his survey of the dialect geography of Upstate New York. Based on data collected in 2008, he described the city as the northeasternmost limit of the NCS, with clear, but variable and less consistent participation in the NCS compared to traditional NCS cities like Buffalo, Rochester and Syracuse. This description was corroborated by a later study that included data from additional communities in the vicinity (Dinkin, 2019). Thus, Ogdensburg was grouped, along with other smaller New York communities, into the Inland North Fringe dialect (see Figure 6 below), which is characterized by a slightly less advanced NCS than the Inland North Core.

---

[35] Ogdensburg's unemployment rate was estimated at 9.2% in 2016, which compares to 7.5% and 7.4% in New York State and the US, respectively.
[36] SUNY and Clarkson University in Potsdam, SUNY and St. Lawrence University in Canton.

How the NCS got to Ogdensburg is, at this point, not very clear. As was pointed out in Chapter 1.3.1, Dinkin (2009, 2013) found that the origin of the first settlers seems to be an important aspect to consider in answering the question of which communities in New York State the NCS diffused to, as only communities that were settled primarily from Southwestern New England seem to have been subject to the NCS. NCS patterns found along the St. Louis Corridor supported the assumption that settlement history is a determining factor in the diffusion of the NCS. However, as was detailed above, Ogdensburg was most likely settled primarily from Vermont rather than Southwestern New England, and thus should not have been able to adopt the shift, or features thereof.

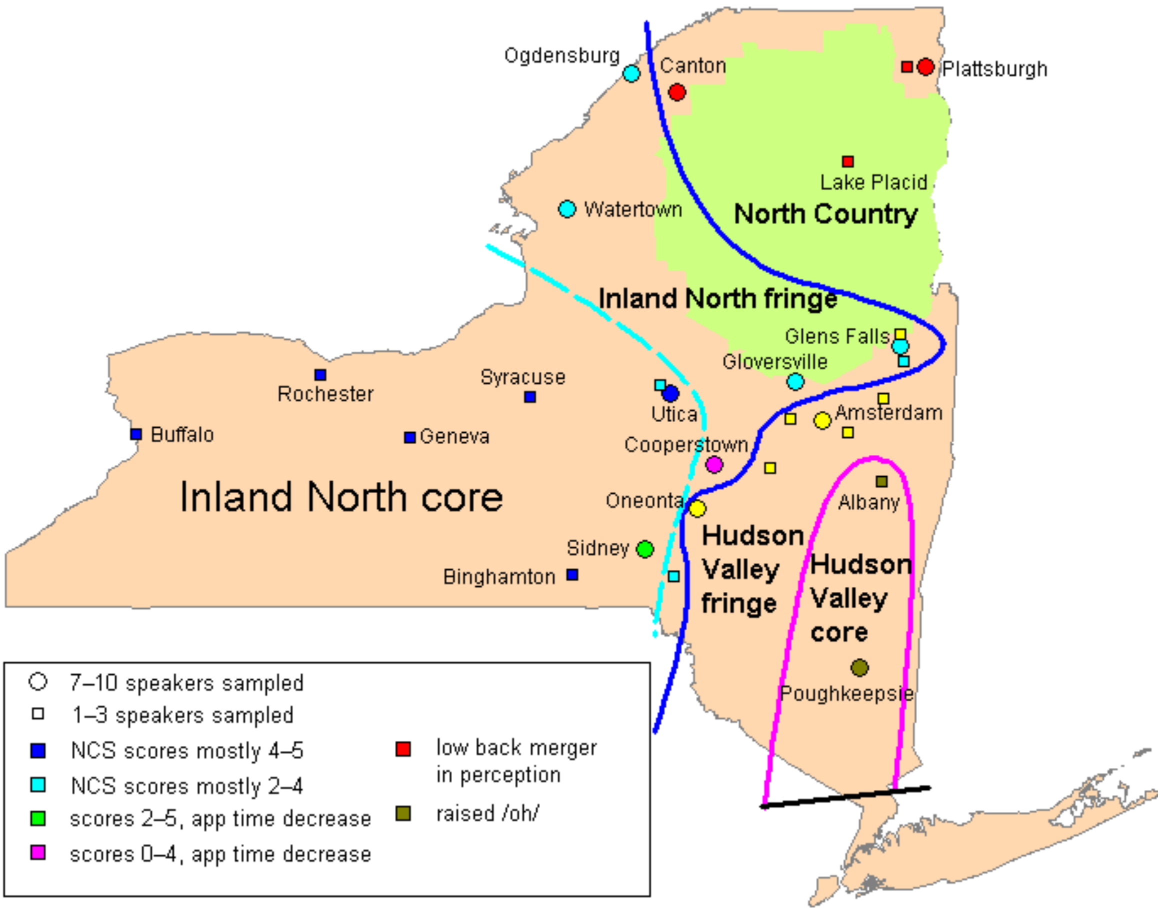


Figure 6: Dialect regions of Upstate New York as defined by Dinkin (2009, 2013); used with permission. An explanation of the scoring system can be found in Chapter 2.6.2.

Although Dinkin (2009) argued that his results show that the spreading of the NCS is not merely related to channels of communication, communication may in fact be of vital importance in the case of Ogdensburg. As was described above, Ogdensburg was in intensive trade contact with both Central and Western New York as well as the Great Lakes region during the time when the NCS was in its initial stages. Before the completion of the St. Lawrence Seaway in 1959, ships coming from the Great Lakes could not go any further down the river than Ogdensburg, making this the easternmost point of contact. This would explain why communities further east along the river, such as Waddington and Massena, do not have the NCS and are thus not part of the Inland North. If the NCS really did spread to Ogdensburg via diffusion, however, we would not expect to see apparent-time trends toward NCS patterns, because adults supposedly are the driving agents in this type of spread. However, Dinkin (2009, 2013) did observe that at least two of the changes, the backing of DRESS and lowering of THOUGHT, were in progress in apparent time in the community. The youngest speakers in his sample, however, were born in the 1980s, and by the time of the interviews in 2008 were, in fact, adults.

Dinkin's (2009, 2013) findings suggest that, in addition to the NCS, the COT-CAUGHT merger was in the process of being adopted in Ogdensburg. This seems to have been the result of diffusion from Northeastern New England through the neighboring North Country (Dinkin, 2019). Although the presence of the NCS and early evidence of a merger in progress seemed to apply to communities throughout Upstate New York, findings in Ogdensburg were surprising in two ways:

I. Compared to similar communities, speakers lagged behind in adopting features of the NCS, making Ogdensburg the only Inland North Fringe community in which the NCS appeared to still be in progress in apparent time.
II. Compared to similar communities, speakers were more advanced in adopting the COT-CAUGHT merger, in both perception and production.

While Dinkin (2009) pointed to this anomaly in his data from Ogdensburg, he was not able to explore the contradiction further owing to a lack of sufficient data, which gave reason for an in-depth follow-up study of the community. There is no obvious reason why Ogdensburg should have been behind in adopting the NCS. It *was* one of the smallest, most remote and least wealthy communities sampled in Dinkin's study, however, communities of similar sizes, and with comparable geographical distances to urban NCS communities and median household income, had completed the adoption of the NCS features despite these shared characteristics: Gloversville, for example, is only slightly bigger, and has a median household income that is even lower than that of Ogdensburg. Glens Falls, while somewhat bigger and wealthier, is just as remote from urban NCS communities as Ogdensburg. In neither of these two communities was there evidence for an ongoing process toward the adoption of NCS features.

Because of Ogdensburg's ties to Canada, and its early contact with Central and Western New York and cities around the Great Lakes, it is an ideal location to study the current development of the NCS and the COT-CAUGHT merger. Central and Western New York, as was established in Chapter 1.2.1, was the likely birthplace of the NCS, and it was transmitted to the Great Lakes area soon after. Today, however, Ogdensburg's only geographical connection to other known NCS communities lies to the southwest, toward the city of Watertown, since trade contact with the Great Lakes area has mainly ceased. To the east, the city borders the dialectological North Country, where the NCS is absent but the COT-CAUGHT merger transitional (*ANAE*; Dinkin 2009, 2013), and to the north of Ogdensburg lies Canada, also without the NCS, but where merged LOT and THOUGHT were observed as early as the mid-19th century. Thus, Ogdensburg has been surrounded by, and been in extensive contact with both linguistic features of interest in this study.

It seems likely that Ogdensburg's turbulent social history may have affected speech patterns in the city, especially considering that it is located on two dialect borders (Canada to the north, and the North Country to the east). Coupland (2009), for example, found that a shift from a manufacturing industry to a service industry reshaped language ideologies and thus linguistic practices in Britain. He therefore argues that social change and language change are "necessarily inter-related" and "mutually constitutive entities and indeed processes" (p. 27). In the case of Ogdensburg, it seems likely that the economic collapse in the second half of the 20th century and the ensuing disconnection from the Inland North may have resulted in stigmatization and recession of the NCS, which has been characteristic of this area, as well as in the adoption of linguistic features from a neighboring dialect that may now be perceived as more prestigious.

Chapter 2: Methodology

2.1 Methodological Goals

In the previous chapter, I outlined the motivation for and the aim of the present study, which is to examine patterns of social variation in the production and perception of multiple target vowel phonemes in Ogdensburg. In this chapter, I will describe what kind of data and tools were necessary in an attempt to answer the research questions that were outlined in Chapter 1.7, how I went about collecting this data in the community, and how the data was processed and analyzed. This also includes a review of the theoretical foundations that motivated my methodological choices.

For an analysis of the extent and direction of vowel movement, the social distribution of these potential changes and their potential social meanings in Ogdensburg, wide-ranging speech production and perception data from a wide variety of speakers in the community as well as access to the speakers' demographic information were fundamental requirements.

Because numerous vowel phonemes are involved in the changes of interest in this study, speech production data needed to be exhaustive, in order to ensure that every phoneme occurs in the data, preferably numerous times. This required recordings of free, spontaneous speech of a certain length, as well as of speech produced in experimental elicitation tasks such as the reading of a wordlist and minimal pairs which focused specifically on the target vowels. Thus, data collection required a method that allowed for inclusion all of these speech elicitation techniques. One such method is the sociolinguistic interview, which was the method of choice in the present study and which will be discussed in detail in Chapter 2.2.

In addition to ensuring that each target vowel is produced at least once by every speaker, speech elicited in reading tasks can provide great insight into the mechanisms of language change: As Labov (e.g. 1966) has demonstrated, intra-speaker variation across different elicitation tasks with varying amounts of attention paid to speech (i.e. speech styles) can help to find answers to questions of ideologies, standard and prestige. It can also function as an indication of change in progress, as phonetic changes such as vowel shifts and mergers tend to be stylistically gradual: For changes from above, it has been found that speakers have a tendency to exhibit an ongoing change more drastically and/or frequently when they pay closer attention to their speech, i.e. in careful speech elicited in reading tasks, than in their normal conversational style, often because the incoming change is perceived as the new standard. Changes from below, on the other hand, tend to affect less monitored, spontaneous speech earlier than more careful speech. Including tasks to elicit controlled speech enabled me to test for such intra-speaker variation (i.e. style shifting) in the present study, which is concerned with changes in progress as well as potential social evaluation attached to these changes. The tools I used to elicit controlled speech from my participants are a wordlist reading, and additionally for LOT and THOUGHT the reading of minimal pairs and the reading of a repetitive *cot-caught* line-up, all of which will be described in detail in Chapter 2.3.1. According to Labov (1966), speech elicited in minimal pairs, wordlist readings and casual conversations form a continuum from very closely monitored to less monitored speech, and thus allow for comparison across two to three different speech styles.

Because perceptual data can provide great insight into language variation, I did not limit data collection to speech production data in this study but collected speech perception data from the community of interest as well. Perception here refers to two different processes, both of which are of relevance to the present study: the perception of linguistic categories, and the perception of social categories.

Insight into the perception of linguistic categories provides answers to the questions of whether or not speakers perceive speech sounds as different, similar or the same, and which acoustic qualities influence this perception. Thus, testing the perception of linguistic categories is particularly vital in the study of phonemic mergers, which in this project entails testing for the ability to distinguish and correctly identify LOT and THOUGHT. As Di Paolo and Yaeger-Dror (2011b) argue, it is "important to delve into the mental structure of the speakers' phonological system to better understand the underpinnings of the speakers' spontaneous speech", which is why, in addition to tracking merger in production, I included two experiments to test for the perception of the two linguistic categories LOT and THOUGHT, which are assumed to be undergoing merger in the community. The experiments of choice were the commutation test (see Chapter 2.3.2.2) and self-judgments of minimal pair production (see Chapter 2.4.4).

The second aspect that perception data can address is the perception of social categories. This refers to "the process engaged when people are exposed to … linguistic material, and extract information from it" (Campbell-Kibler, 2010, p. 378). Since people are exposed to external linguistic stimuli in nearly every communicative situation, it follows that they are constantly engaged in this process of extracting information from speech material. Consequently, listeners make judgments about their interlocutor and their social background, e.g. age, gender, ethnicity, education, etc. Variation in how the speaker conveys their message leads to different perceptions and judgments in the listener, and this in turn may ultimately influence the linguistic behavior of the listener. Thus, in order to understand sociolinguistic variation in speech production, it is essential to unveil the social meaning that it carries for the listener (Campbell-Kibler, 2006a). In the present study, which is concerned with potential social evaluation of changes in progress, this was accomplished by including a matched guise experiment in the study (see Chapter 2.3.2.1).

Each of the techniques and tools mentioned above will be discussed and described in detail in the following subchapters. The order of these subchapters reflects the timeline of the project: They start out with a discussion regarding the choice of data collection methods and the design of the experiments included in these methods prior to data collection, then go on to describe how these various techniques were implemented in the field, and conclude with an overview of how the data was processed and analyzed.

2.2 The Sociolinguistic Interview

There are a number of data collection methods that have been used successfully in variationist studies. These include, for example, using data from publicly available sources, (e.g. written text or media broadcast), written or fieldworker-administered surveys, as well as more ethnographic approaches such as participant observation (L. Milroy & Gordon, 2003, p. 51). Although all of these methods have their advantages and have been used successfully in previous research, none of them were appropriate for the present study for various reasons. The aim of this study is to examine patterns of social distribution regarding the production and perception of the target vowel phonemes, most of which are assumed to behave differently across various speech styles. The data collection method therefore needed to allow me to capture speech perception and speech production in different speech styles for multiple phonemes. Another requirement was that I had access to the participants' demographic information. Written texts fulfill none of these requirements; media broadcasts do not provide perception data or background information about the speaker; surveys, both in written and fieldworker-administered form rely either on self-reported speech production data or on auditory coding, both of which are not particularly adequate when the variables of interest entail fine-grained phonetic detail as is the case with vowel shifts and mergers (Di Paolo, Yaeger-Dror, & Beckford Wassink, 2011); and ethnographic participant observation requires long-term intense social engagement with and involvement in the target community, which requires a) an exceptionally sociable and outgoing researcher, which are qualities that do not align with my temperament, and b) a tremendous amount of time, which I did not have. Since none of these methodologies for data collection seemed appropriate, I made use of the sociolinguistic interview instead.

The sociolinguistic interview is one of the standard tools for data collection used by variationists and fulfills all of the requirements outlined above. It allows the researcher to record the participants' speech production in conversational speech style, and by implementing e.g. reading tasks of various kinds, different speech styles can be captured during the interview. Additionally, the format allows for the inclusion of perception experiments for the gathering of perception data, which can be designed in a way that elicits attitudes indirectly (see Chapter 2.3.2.1). Since sociolinguistic interviews are generally conducted face-to-face, the interaction is not anonymous, and the researcher is able to ask for the participants' demographic information. Another advantage of the sociolinguistic interview is that all of the above mentioned can be achieved relatively quickly, with interviews lasting anywhere from 10 to 30 minutes for a Short Sociolinguistic Encounter (Ash, 2002) to a few hours.

However, despite those benefits, there are several shortcomings with this technique. I had limited contacts in Ogdensburg prior to my fieldwork; thus, all of my informants were unknown to me and I had no prior personal relationship with any of them. The main drawback of this, and the sociolinguistic interview as a whole, is the potential loss of casualness in the interview (Tagliamonte, 2006, p. 18), which may lead speakers to shift away from their casual language use, i.e. vernacular (L. Milroy & Gordon, 2003, p. 49). This is referred to as the "observer's paradox" and the "interviewer effect". The observer's paradox describes the discrepancy between wanting to *observe* language as it is being used while speakers are *not being observed*. Adding a microphone and voice recorder to the scene makes the element of observation particularly prominent in a sociolinguistic interview (L. Milroy & Gordon, 2003, p. 49) and increases the likelihood of the interviewer effect, a potential outcome of the observer's paradox. The interviewer effect may (or may not) arise when speakers are being observed and recorded by a stranger (L. Milroy & Gordon, 2003) and may (or may not) entail a shift in linguistic behavior. Depending on the interviewer, the speaker's impression of the interviewer and the context in general, this shift can take the form of convergence to or divergence from the language of the interviewer or a perceived standard. This can either be a standard in which speakers believe they are supposed to be talking in an interview situation (i.e. matching the perceived expectations of the interviewer), or a standard that they believe portrays them in the best possible light, regardless of the interviewer's expectations. These kinds of effects were, for example, observed by Hay, Drager and Warren (2009) in their study of the NEAR-SQUARE merger in New Zealand, where they found that an interviewer without the merger might trigger speakers to produce a greater distinction between the two phonemes than they would with a merged interlocutor. My recruiting and interview procedures, which will be described below and in Chapter 2.4, involved steps and techniques that arguably mitigate the observer's paradox and the interviewer effect as much as possible. While the observer's paradox "can never be entirely resolved" (L. Milroy & Gordon, 2003, p. 49), employing these techniques should have eliminated most of the speakers' potential focus on their speech resulting from observation (Di Paolo & Yaeger-Dror, 2011a) so that the recordings can be analyzed under the assumption that the recorded speech closely resembles the speakers' vernacular, despite a somewhat heightened attention to speech.

The conversational part of the interview (henceforth "interview proper") was semi-structured. Before starting the process of interviewing participants, I compiled an interview schedule which included a set of questions that were arranged into "conversational modules" (Labov, 1972), i.e. general topics to talk about during the interview, all of which pertained to the general topic of life and experiences in Ogdensburg. Following Tagliamonte's (2006) advice to start the interview with questions about demography, I generally started with questions about the participant's personal background, e.g. where they were born and grew up, where their parents are from, and whether they have always lived in Ogdensburg or moved away at one point or another. The second module included questions about living in Ogdensburg, e.g. whether or not they like living in Ogdensburg, if they intend to leave at some point, what they like to do in Ogdensburg, and travelling in and around New York and Canada. For school-aged participants I also included a module on school, asking questions about favorite and least favorite classes, plans for college etc. Questions in the last module revolved around language, and included questions about the participants' experience with their own accent, and perceived language differences within New York and between New York and Canada. Tagliamonte (2006) recommends saving the language module until the end of the interview, and although I intended to do so, it was not always possible to stick to this order, as language was often brought up by the informants themselves at any given point during the interview proper. Since "a sociolinguistic interview should have no rigid insistence upon a pre-set order of topics" (Tagliamonte, 2006, p. 39), I avoided sticking to my pre-determined schedule during the interview proper. Instead, I generally "allow[ed] the subject's interest in any particular set of topics to guide transition … from module to module" (L. Milroy & Gordon, 2003, p. 60). The aim of this was to let the informants talk freely and without interruptions about things they were interested in. This made it easier for them to switch into storytelling mode, talking about personal memories and telling me stories about their lives, without needing much encouragement from my side. As Labov (1984, p. 34) argues, once informants are "engaged in this type of discussion speakers tend to produce vivid recollections rich in vernacular features", the capturing of which was the main purpose behind the collection of spontaneous speech during the interview proper. This flexibility also meant that not all participants provided answers to all the scheduled questions. Three participants in particular seemed to have their own idea of which topics the interview was going to cover: the history and culture of Ogdensburg and their relation to Canada. One of them, like me, had prepared a list of topics to talk about. Another one even remarked during the interview that I, of course, would not be interested in his personal stories. However, all those informants produced enough speech material in talking about the things they were interested in, and still gave me the demographic and attitudinal information I needed.

2.3 Designing Speech Production and Perception Experiments

Once I had decided on the sociolinguistic interview as my main method of data collection, which would allow me to include structured elicitation of speech production and perception, I chose and designed experiments to accomplish these tasks, which will be described and discussed in detail in the following subchapters.

2.3.1 Speech Production: Structured Elicitation

As outlined above, spontaneous speech is relatively easy to target in conversation; the elicitation of controlled speech, on the other hand, requires more planning and preparation. In the next paragraphs, I am going to outline how I designed the tools I used to elicit controlled speech from my participants: a wordlist reading, the reading of minimal pairs, and the reading of a repetitive *cot-caught* line-up.

2.3.1.1 Wordlist Reading

The wordlist that participants were asked to read contained words for five NCS vowels: TRAP, LOT, THOUGHT, STRUT and DRESS, as well as one word of the NORTH/FORCE class. TRAP, DRESS and STRUT were represented by 10 items each. LOT and THOUGHT were represented by a total of 13 and 10 items respectively. This resulted in a wordlist of 54 words in total. All target vowels on this wordlist occurred in a CVC environment, with all neighboring consonants being obstruents. This generally allows for easier acoustic analysis (see Chapter 2.5.1), as the transition from vowel to obstruents and vice versa is generally clean and therefor easy to detect on a spectrogram. Even though for this study vowels were extracted automatically using the FAVE software package (Rosenfelder, Fruehwald, Evanini, & Yuan, 2011), cleaner transitions between segments arguably made it easier for the software to accurately align the segments and measure them more reliably. Exceptions to this rule of placing each vowel in between obstruents were three of the LOT items (*revolve, golf, sorry)* as well as *barn* (START[37]), and *born* (NORTH/FORCE). *Revolve* and *golf* were added based on Dinkin's (2009, 2016) finding that words in which LOT is followed by an /lf/ or /lv/ cluster are likely to be transferred into the THOUGHT class, even by speakers who otherwise do not have the merger (see Chapter 1.2.2). Including these words in the wordlist provided an opportunity for a more detailed analysis of this phenomenon. *Sorry* was added a few interviews into my fieldwork, as it turned out to be a salient lexical item that is subject to considerable variation in the community. *Barn* and *born* were added as a control pair, which will be explained further in the following section on minimal pairs (Chapter 2.3.1.2). In the interest of keeping the wordlist reading task as short as possible, I did not add any filler words to the wordlist. Including five target

[37] Although *barn* belongs to the START class, it is likely to be identified with LOT in most US varieties of English (see Chapter 6.1 for detail).

vowels, which were presented in random order (as detailed in Appendix A), arguably provided enough variation so as to not raise suspicion among the participants (Di Paolo & Yaeger-Dror, 2011a, p. 15). Table 1 contains all wordlist items for each of the target vowels.

	TRAP[38]	DRESS	STRUT	LOT	THOUGHT[39]	NORTH/FORCE
1	tab	sketch	bus	deposit	gawk	born
2	badge	bed	hut	box	jaw	
3	path	peck	hush	pot	dog	
4	bad	kept	bubble	cop	cause	
5	pass	best	hub	pop	toss	
6	bash	bet	cut	hockey	fought	
7	bat	feather	hug	chop	coffee	
8	cab	keg	huddle	top	pause	
9	bag	pep	hutch	pod	bought	
10	back	beg	but	revolve	cough	
11				golf		
12				sorry		
13				(barn)		
	10	10	10	12(+1)	10	1

Table 1: Wordlist items. They were presented to the participants in a randomized order.

2.3.1.2 Minimal Pair Reading

The minimal pair reading was included to test for the suspected merger of LOT and THOUGHT in production. The list consisted of 14 minimal pairs for LOT and THOUGHT, all of which were real words, as well as a START-NORTH/FORCE (false) pair, which served as a filler and control pair. START was represented by *barn,* and NORTH/FORCE by *born,* which are always pronounced differently by the majority of native US English speakers. Thus, participants who judge the two words to sound the same and produce them with the same vowel in the reading of minimal pairs might be unlikely to accurately recognize and

[38] Note that *path* and *pass* are BATH words in some varieties of English; however, this class is not distinct from TRAP in most varieties of American English, as they never underwent the TRAP-BATH split (see Chapter 3.1).
[39] Note that, owing to the complicated diachronic development of the THOUGHT class (see Chapter 6.1), not all of the words listed in this category belong to the THOUGHT class in all varieties of English, including some US varieties, such as e.g. *toss*, *coffee* and *cough* (CLOTH) and e.g. *dog*, which are often realized with LOT instead. In Ogdensburg, however, all of these appear to be commonly realized with THOUGHT, as the results of the analysis in Chapter 6.2.1 will show.

realize the more subtle distinction between the genuine LOT-THOUGHT pairs, and might therefore not be a reliable indication of an advancing merger (Johnson, 2007). However, none of the subjects in this study produced *barn* and *born* with the same vowel.

As the set of potential minimal pairs for LOT and THOUGHT is relatively limited, some studies opt to include nonsense words in these minimal pair reading tasks (see Hay, Drager, & Thomas, 2013). For my purposes, however, 14 pairs of real words provided enough material to work with, and I opted not to include any nonsense words in the list. Table 2 lists the minimal pairs that were included in this study. The items in blue represent the LOT words, items in wine-red belong to the THOUGHT class.

odd - awed	hauler - holler	collar - caller	tot - taught	hock - hawk
fond - fawned	knotty - naughty	sod - sawed	cot - caught	pond - pawned
stock - stalk	sought - sot	gnawed - nod	cawed - cod	born - barn

Table 2: Sets of minimal pairs included in the reading task.[40]

Because the set of possible pairs is limited for this merger, I was not able to control the phonological environments as much as I could with the wordlist. While the majority of LOTs and THOUGHTs do occur in between consonants (and word initial for one pair), not all of the surrounding consonants belong to the obstruent class. In two of the pairs, the target vowels are followed by a nasal, and in another two of the pairs they are preceded by a nasal. For two pairs, the target vowels are followed by /l/. Although varying phonological environments affect comparability across pairs to a certain degree, including them allows for the possibility to investigate whether the merger takes place in some phonological environments earlier than others (see Chapter 6.2.3.3).

For the same reason, i.e. limitations of available minimal pairs for LOT and THOUGHT, the absence or presence of morpheme boundaries was not taken into consideration in the selection of the pairs. In the majority of the pairs, the THOUGHT word contains a morpheme boundary while the LOT word does not; in one pair, the LOT word has a morpheme boundary while the THOUGHT word does not; and in another pair the THOUGHT word is an irregular past tense, which may or may not constitute a morpheme boundary (Dinkin, p.c.). Table 3 below lists the minimal pairs with respect to morpheme boundaries.

[40] They are listed in the order in which they were presented to the participants when read top to bottom and left to right.

It is possible that these differences have affected the degree of merger in the participants perception and production. Whether or not it affected perception cannot be assessed in this study, as minimal-pair judgments were not elicited for individual pairs (see Chapter 2.4.4). In production, the presence or absence of morpheme boundaries does not appear to have significantly influenced the degree of phonemic contrast between LOT and THOUGHT. This is illustrated in Figures 7 and 8 below, based on the examples of two of the most merged speakers in the sample.

Morpheme boundary in THOUGHT	Morpheme boundary in LOT	Irregular past tense in THOUGHT	No morpheme boundary
odd - awe/d	knot/ty - naughty	tot - taught	hock - hawk
fond - fawn/ed		cot - caught	stock - stalk
haul/er - holler		sought - sot	
collar - call/er			
sod - saw/ed			
gnaw/ed - nod			
caw/ed - cod			
pond - pawn/ed			

Table 3: Morpheme boundaries in minimal pairs.

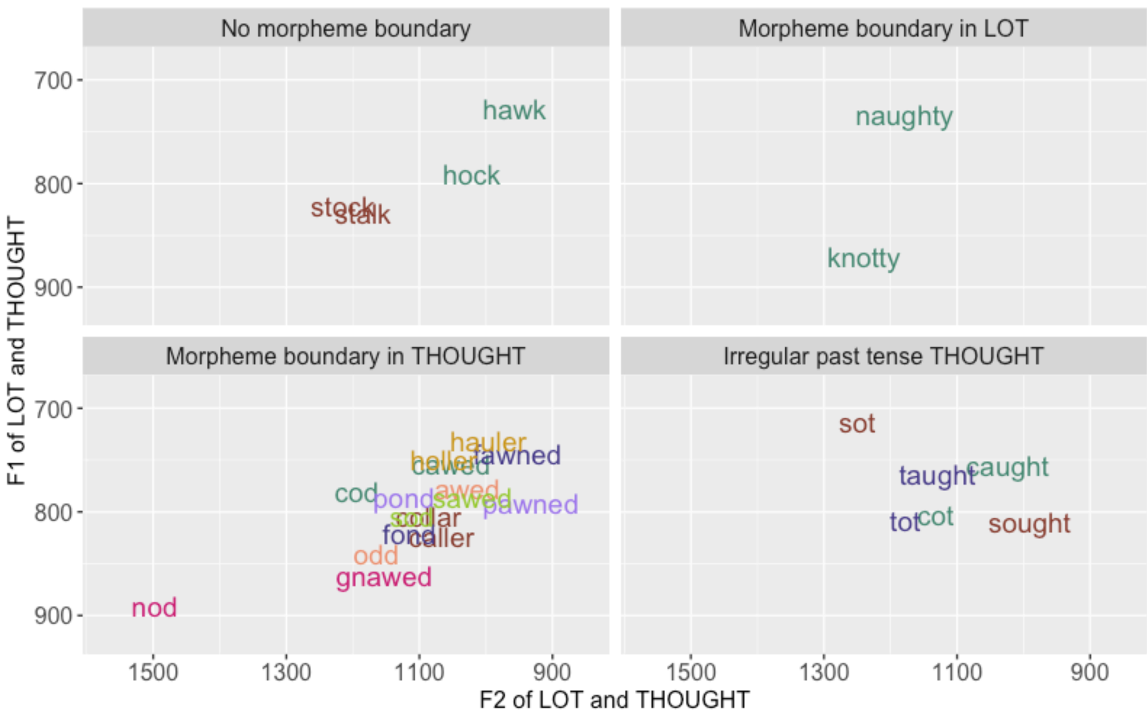


Figure 7: Effect of morpheme boundaries on the degree of contrast between LOT and THOUGHT for Jason, born in 1998.

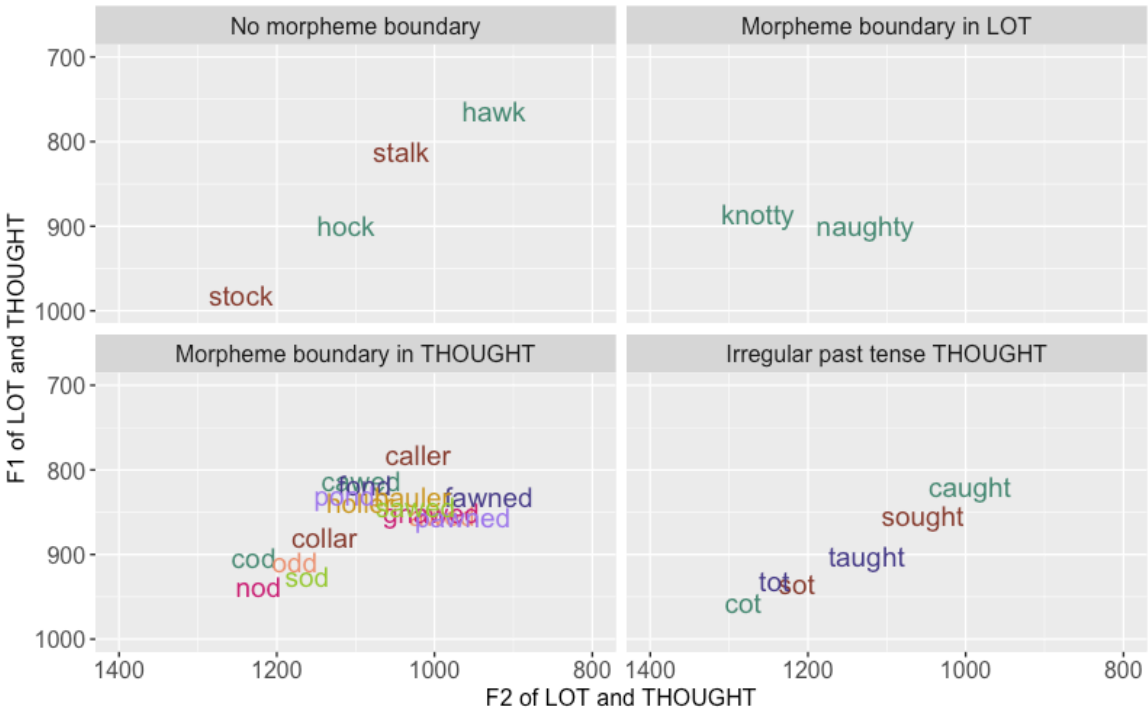


Figure 8: Effect of morpheme boundaries on the degree of contrast between LOT and THOUGHT for Allison, born in 1993.

For neither of the two speakers does there seem to be a particular pattern in the degree of phonemic contrast depending on morpheme boundaries. The same was found for other speakers with a comparable degree of merger in production, so that it can be assumed that morpheme boundaries do not impact phonemic contrast between LOT and THOUGHT to a significant extent in the present sample. They will therefore not be taken into consideration in the analysis.

2.3.1.3 Reading of a cot-caught line-up

The third part of the reading task involved a list of 10 repeating instances of *cot* and *caught*, i.e. five *cot* and five *caught* in random order[41], which later served as the stimuli for the commutation test. What exactly the commutation test entails and how it was designed will be discussed in Chapter 2.3.2.2.

[41] The order in which they were presented to the participants can be found in Appendix B.

2.3.2 Speech Perception Experiments

As outlined above, the speech perception experiments for this study target both the perception of social and the perception of linguistic categories. The techniques used to accomplish these tasks will be discussed in the following subchapters, starting with the matched guise experiment which tests for potential social meanings of sound change, followed by the commutation test that tests the ability to correctly identify minimal pairs.

2.3.2.1 Matched Guise Experiment – Perception of Social Categories

As mentioned above, one of the main questions asked in this study is whether or not the target vowels are subject to social evaluation in Ogdensburg. While style shifting as discussed above can and does provide first indications of this, it was assumed that the data gained through perception experiments can provide "a deeper understanding of how sociolinguistic variation is transmitted between talker and listener, and of how listeners deal with variability for both linguistic and social purposes" (Clopper, Hay, & Plichta, 2011, p. 149). However, speech perception in the sense of feelings toward and beliefs of a particular variety are difficult to examine because they cannot be directly observed (Kircher, 2015, p. 197). Thus, designing experiments to capture those attitudes and beliefs requires thorough preparation and careful attention to detail.

Experiments designed to elicit listeners' perception of social information in speech "draw primarily on research in cognitive psychology examining the perception of talker variability and in social psychology examining social attitude judgments" (Clopper et al., 2011, p. 149). Based on those strands of research, linguists have developed three main types of methods to elicit and assess language attitudes (Kircher, 2015, p. 197). These include the analysis of the societal treatment of language varieties through participant observation, *direct* elicitation methods such as surveys, interviews, and categorization tasks. For the present study, I decided to make us of an *indirect* elicitation method, the Matched Guise Technique (MGT).

The MGT is one of the best known and most sophisticated (partially) indirect methods of language attitude elicitation. It was developed in an attempt to avoid the problem of social desirability biases often found when using direct methods, i.e. the participant's wish to present themselves in a desirable light, thus giving dishonest responses regarding their language attitudes (Kircher, 2015). This is achieved by obscuring the purpose of the experiment, i.e. assessing potential unconscious evaluation of linguistic features, as participants are asked to evaluate the *speaker* rather than the speaker's *language,* as is the case in more direct elicitation methods (Preston, 2009, p. 207 as cited in Kircher, 2015, p. 206). Furthermore, the MGT elicits *spontaneous* attitudes, allowing the participant less time for reflection than is the case with direct methods, thus avoiding, or at least minimizing social desirability biases (Ryan, Giles, & Hewstone, 1987).

The MGT was devised by William Lambert and colleagues in the 1960s and has since been used in various studies covering a range of linguistic situations, such as comparing attitudes toward multiple languages, e.g. French and English in Canada (Genesee & Holobow, 1989), different varieties of the same language, e.g. regional varieties of British English (Dixon, Mahoney, & Cocks, 2002), as well as single phonological features such as the –*ing* variable (e.g. Campbell-Kibler, 2006).

In a matched guise experiment, participants are presented with pairs of auditory guises which are completely identical, with the exception of the sociolinguistic variable that is being studied. The variable occurs with its first variant in the first guise, and with its second variant in the second guise. Participants listen to the recordings and are asked to evaluate the speaker based on what they hear. Campbell-Kibler (2006a, p. 74) explains the MGT as follows:

> Because listeners are not told that the alternate recordings have been produced by the same person, they evaluate each guise (language or accent performance) as an individual speaker. However, because the recordings have been produced by the same person, many of the paralinguistic cues are (hopefully) held constant, for example speech rate, pitch contours, and various aspects of voice quality ... so that any differences between the evaluations can be (in theory) assigned to different perceptions of the languages or varieties under study.

There are several techniques through which matched guises can be evaluated by the participants. The options range from open-ended questions to survey-style testing, where the participant evaluates the speaker on a range of pre-defined qualities. Using open-ended questions has the advantage of giving the participant room to freely express whatever thoughts about the speaker come to mind. Thus, this data can "provide more nuanced insights into the participants' ideologies and conscious thought processes" (Campbell-Kibler, 2006a, p. 76). In using pre-defined categories, on the other hand, the researcher runs the risk of ignoring categories that might be important to the participant who is evaluating the guises, and the survey responses may lack nuance and insights into the reasoning behind the judgments. The main benefit of using survey responses over open-ended questions is that they are easier to code, quantify and generalize, and thus easier to correlate to observed patterns in speech production.

In my study, I tested both approaches. I used the first few interviews to pilot-test the experiment with open-ended questions, trying to elicit general reactions to the guises and to see which words or categories the participants might use to describe the speakers in the guises. However, it proved difficult to elicit any concrete evaluations from the participants. This is a common problem when the salience of the variables tested is relatively low, as tends to be the case with vowel shifts and mergers (Labov, 1994). Thus, I switched to pre-defined categories with future participants, asking the participants to rate the voices in the recordings in five categories: age, friendliness, level of education, how local to Ogdensburg and how Canadian they sounded. I decided on these five categories for two main reasons. Firstly, I wanted to include categories that would pertain to both the status dimension (e.g. education) as well as the solidarity dimension (e.g. friendliness). As Kircher (2015, p. 201) points out

> Empirical research from numerous parts of the world has revealed that status and solidarity are independent dimensions of language attitudes, and that it is indeed in terms of these two primary dimensions that the identities of speakers of different varieties tend to be evaluated.

Those two dimensions “are considered to have ‘a universal importance’ for the understanding of language attitudes” (Kircher, 2015, p. 201), which is why I included categories that pertained to both dimensions. A second reason for choosing those particular categories was that some of them have been used in previous studies on the social evaluation of the NCS in different areas within the same dialect region (e.g. Wagner et al., 2016). Using the same categories will allow for later comparison of the results. I added the “Canadian-ness” category out of interest in potential border-effects, because Ogdensburg is located on the Canadian border, and it seemed likely that US American vs.

Canadian identity might be a salient sociolinguistic category for locals, especially in the COT-CAUGHT merger. In the interest of time, I opted to not add any more categories to the experiment, although there are certainly many more that could have provided rich information about the social evaluation of the features of interest, both on the status as well as the solidarity dimension. Not all of the pre-defined categories I included turned out to be relevant for all of the variables involved in Ogdensburg. Overall however, the survey responses provided material to work with, whereas the initial open-ended questions did not.

Like most social evaluation elicitation techniques, the matched guise format does have its drawbacks as well. The main concern with this experiment design is that speaker evaluations are based solely on language. This is a rather artificial situation, as evaluations are generally based on a combination of various characteristics, e.g. language, physical appearance, clothing, etc. In a matched guise experiment, however, none of these additional speaker characteristics are provided. This is done intentionally, because it allows the researcher to conclude that evaluations are in fact based solely on language and not influenced by external factors. However, it has been argued that this creates a setting that is "a bit far removed from real-life contexts" (Fasold, 1984, p. 154), and that the naturalness of the judgments might therefore be questionable. A related issue with the MGT is that it presents a case of one-way communication that does not involve any interaction between the voice (i.e. the speaker in the recording) and the participant. Ryan et al. (1987, p. 1076) argue that this is "different from normal interactions in which both parties would actively participate in a speech exchange". Because this one-way channel is unnatural for a communicative situation, it contributes to the artificialness of the situation and might influence the ratings (Ryan et al., 1987). The process of rating itself is another point of criticism with the MGT. Fasold (1984, p. 76) argues that explicitly asking participants to rate speakers might highlight their prejudices and lead them to make "evaluative judgments in a way that doesn't happen in ordinary interactive settings". However, Chambers (2008, p. 224) argues that "the consistency with which subjects make their decisions indicates that the prejudices are not merely individual but are communal … they are not random or arbitrary".

Despite those drawbacks, the MGT is arguably one of the "most fruitful experimental measures of subjective reactions to linguistic variation" (Labov, 2001, p. 194) and has, according to Fasold (1984) "become virtually standard in language attitudes research" (as cited in Kircher, 2015, p. 206). Labov also notes that the matched guise format is successful in eliciting reactions to changes "that are not available for overt recognition in self-reported tests" (Labov, 2001, p. 212), such as chain shifts and mergers. Therefore, the MGT seemed the most appropriate experiment in the attempt to elicit language attitudes toward the NCS and the COT-CAUGHT merger for this study. The MGT has been used for a similar purpose by Campbell-Kibler (e.g. 2005, 2006b, 2007, 2008, 2011). In her work on *-ing*, she was able to document "a connection between the use of a given variant of *-ing* in a specific situation and a change in the rating of a listener on a list of labeled scales", and to establish "information about the relative attitudes of the listener(s) to the two linguistic styles presented" (Campbell-Kibler, 2006a, p. 74). Thus, it seemed likely that the MGT would provide ample insight into attitudes toward the NCS and the COT-CAUGHT merger as well.

In the following sections I will explain how I created the guises used for this experiment. I will start out by going over the collection of guise material, then move on to illustrating the integration of NCS and merger stimuli (splicing), and finally discuss how I selected and lined up the guises.

Guise Material

In order to create auditory materials for the matched guise experiment I recorded four native speakers of US American English (henceforth "voices"), using a Zoom H5 recorder paired with an AKG C 417 PP lavalier microphone. One of the voices was female, the other three were male. However, since creating the guises (see below) proved to be considerably more difficult for the female voice than for male voices, I dropped the female voice and created the matched guises from the three male voices only. This also allowed for easier analysis, as "there are systematic differences in the manner in which female and male speakers of the same variety are evaluated" (Kircher, 2015, p. 199). By eliminating the female voice, the gender of the voice was no longer a potential factor in the listeners' judgments. Since previous studies have shown that perceived social information about the voice can affect social perception (e.g. Hay, Warren, & Drager, 2006), I tried to keep the relevant demographic characteristics across voices as consistent as possible. All three voices were graduate students at a Swiss university, of similar age and white. While all three voices came from different dialect areas of the US[42], none of them exhibited signs of the Northern Cities Shift, but all of them had fully merged low back vowels. Since instances of their original target vowels were removed in the splicing process (see below), their origins are unlikely to have affected the participants' ratings. Target vowels without primary stress and non-target vowels (also generally without primary stress) were not altered, so that some authentic vowels, which might be representative of the regions the voices were from, were heard by the participants. However, as the majority of these occurred in unstressed positions, they were likely realized with reduced qualities and little local or regional color. To me, none of them stood out in any particular way.

For this experiment, I used read material for the carrier phrases of the guises. The main advantage of using read material for matched guise experiments is that it allows the researcher to control both the content of the materials (including factors such as word choice and sentence structure) as well as the phonological environments in which the target vowels occur. Holding content stable across all guises and all voices is important for the experiment, as the content of what a speaker says can obviously impact the judgments others make about them. The content of all sentences was as neutral and trivial as possible, as suggested by Giles and Coupland (1991), so as to avoid content influencing the participants' rating of what they are hearing (Drager, Hay, & Walker, 2010; Hay, Warren, et al., 2006). This meant avoiding topics relating to e.g. ideologies, politics, and language in the guises (Kircher, 2015). I also excluded topics that are (stereo-) typically associated with either the US or Canada, so as to avoid giving the participant the impression that a speaker might be of one or the other nationality based on what they were saying. Studies have shown that speech perception often depends on who the participant thinks the speaker is, including their nationality and the associated dialect area (e.g. Hay & Drager, 2010; Hay, Nolan, & Drager, 2006; Niedzielski, 1999). As discussed above, Canadian-ness was one of the categories I tested for in the experiment, so I made sure that the design of the experiment, including the content of the guise material, did not contain primes that would shift the responses of the participants in any particular direction.

[42] Voice R is from San Diego, California; voice J from Montana City, Montana; and voice T from the White Mountains of New Hampshire.

Controlling for phonological environment by using read material for the guises was crucial for this experiment as well, because rather than manipulating the formants of the original vowels produced by the voices, I spliced alternative stimuli into the carrier phrases (see below), which can be problematic with certain environments. In general, it is easier to replace vowels that occur before and after obstruents than sonorants. Thus, all of the target vowels in my guises occur between two obstruents. This level of control is not possible when using spontaneous speech material as carrier phrases, as speakers will inevitably say different things, so that the researcher is limited to whatever tokens happen to appear and whatever environments they happen to occur in. The major drawback of read material, on the other hand, is that listener evaluations can differ between read and spontaneous speech (Giles, Smith, Browne, Whiteman, & Williams, 1980; Smith & Bailey, 1980), which makes it "problematic to assume the results for read speech will reflect percepts in real, spontaneous interactions" (Campbell-Kibler, 2006a, p. 75). It is therefore necessary to exercise caution when generalizing the perception results from the experiment in this study.

To create the carrier frames for the experiment, the voices were asked to read a number of sentences. One of the three voices was asked to read a total of 25 sentences, each of them containing two or three instances of one of the target NCS vowels (TRAP, LOT, THOUGHT, STRUT, or DRESS) in stressed position. None of the sentences, however, contained more than one NCS vowel phoneme. For each target vowel, there were five sentences. The other two voices were asked to read a total of 28 sentences. These included the same 25 sentences for the NCS vowels as described above. In addition, they included three sentences that contained one instance of LOT and one instance of THOUGHT each. These were designed to create guises that represented merged or distinct production of LOT and THOUGHT. Table 4 below lists all sentences the voices were asked to read.

TRAP	DRESS
Her d**a**d's so s**a**d about it.	They were ass**e**ssing the ch**ef.**
That's ex**a**ctly what h**a**ppened.	They need good ch**e**ddar for their g**ue**sts.
The c**a**t p**a**ssed **a**fter the att**a**ck.	He did b**e**tter in the s**e**cond t**e**st.
She has a p**a**ssion for b**a**bbling.	They conf**e**ssed to the th**e**ft.
I had my j**a**cket in my b**a**g.	They sugg**e**sted g**e**tting a p**e**t.
LOT	STRUT
I need to sh**o**p for s**o**cks.	They made s**u**ch a f**u**ss about this st**u**ff.
The d**o**ctor was in sh**o**ck about it.	They c**u**ddled with the p**u**ppy.
The j**o**b is imp**o**ssible with a t**o**ddler.	She made d**ou**ble f**u**dge brownies for me.
Those sc**o**tch b**o**ttles weren't in st**o**ck.	My b**u**ddy took the b**u**s a c**ou**ple times.
He didn't b**o**ther st**o**pping for them.	Two c**u**ps were like five b**u**cks.
THOUGHT	Merger
Their b**o**ss made this **a**wesome s**au**ce.	His j**a**w p**o**pped out of place.
Their d**o**g has huge p**a**ws.	We got a good sh**o**t of the ch**a**lk cliffs.
Her d**au**ghter knew this **au**thor.	They t**o**ssed it off the d**o**ck.
He th**ou**ght they were in his **o**ffice.	

Table 4: Carrier phrases for the matched guises. Only unreduced vowels carrying primary stress, indicated in bold, were altered in the experiment.

Splicing

There are various techniques to adjust the vowel quality in carrier frames for a matched guise experiment, the three most common being the performance of differing versions by the voices who provide the material, digital manipulation of the formants as produced by the voices, and splicing in different material. For this study, I opted for the splicing method, which entails cutting out the target variable as it was produced by the voice of the carrier phrase and replacing it with an alternate. The advantages of splicing (and digital manipulation) over the performance approach is that it gives the researcher precise control over the quality of the target sounds and ensures that only one variable in the carrier, the target vowel, is affected, i.e. the rest of the guise is exactly the same for each pair. It also makes recruiting speakers much easier, as the researcher is free to use lay people and monodialectal speakers who do not need to be able to perform two differing versions of the same variable at a native-like level (Campbell-Kibler, 2006a). The splicing technique has been used in studies by, for example, Campbell-Kibler (2006) and Labov, Ash, Baranowski, et al. (2006).

For this study, I used two alternates for each target vowel to create NCS guises: one NCS-shifted, and one non-NCS vowel. For example, I spliced out the voice's original TRAP using Praat acoustic analysis software (Boersma & Weenink, 2017) and replaced it with a raised TRAP stimulus for the NCS-shifted guise, and an unraised stimulus of TRAP for the non-NCS guise. These alternates that served as the actual stimuli were pulled from two of Dinkin's 2008 participants, Dan L. (born in 1959) and Myke U. (born in 1992). Like the voices I recorded for the carrier frames, these two speakers were white males. I selected stimuli from these particular speakers because, in agreement with Dinkin (p.c.), I deemed the qualities of their vowels to be sufficiently different for differentiation through auditory perception. Most of the stimuli I extracted from their wordlist reading, except for the unshifted stimulus for STRUT, which was taken from spontaneous speech owing to a lack of appropriate wordlist tokens.[43] I used alternates from read material, as the carrier phrases were also read by the voices, which meant that pitch and intensity would be relatively stable across the carrier phrases and the alternates. Dan, whose vowels were used to create the NCS shifted guises, was a speaker from Ogdensburg who showed moderate participation in the NCS[44] and maintained a clear distinction between LOT and THOUGHT. Myke, the speaker whose vowels were used to create the non-NCS guises was from Canton. Canton, although very close in location to Ogdensburg, is not part of the Inland North dialect area (Dinkin, 2009, 2013) and therefore does not participate in the NCS, but has LOT and THOUGHT almost fully merged. Both of these characteristics apply to Myke: He did not participate in the shift, while his LOT and THOUGHT were close to merged. Below is an example of how splicing was accomplished for TRAP.

Carrier:	Her d**a**d's so s**a**d about it. – Original TRAP was removed from the frame
Shifted guise:	Spliced in alternate: raised TRAP (Dan)
Unshifted guise:	Spliced in alternate: unraised TRAP (Myke)

In both guises, the two instances of original TRAP as produced by the voices reading the phrase were spliced out. To generate the NCS-shifted guise, their original TRAP was

---

[43] STRUT was realized notably backer in spontaneous speech than it was in wordlist style by speakers in 2008 (see Chapter 5.2.3)
[44] Moderate NCS participation seemed to be the norm in Ogdensburg according to Dinkin's data.

replaced by the raised TRAP alternate as produced by Dan. Likewise, the non-NCS version was created by replacing the voice's original TRAP with Myke's unraised production. I spliced the unshifted guises as well rather than just keeping the originals produced by the (unshifted) voices, so that potential perceptual differences cannot merely be due to the fact that one of the guises in a pair was spliced, and thus tampered with, while the other was not. Even though efforts were made to make the tampering as unnoticeable as possible (see below), splicing both versions seemed to be the safest way of ensuring that the two guises within a pair did not differ in anything but the actual spliced-in vowel stimuli (Campbell-Kibler, 2006a). Guises for LOT, STRUT and DRESS were produced by this same procedure. Since I was not able to find appropriate stimuli for THOUGHT in the data (Dan did not seem to participate in that particular NCS feature, and Myke was nearly merged), I excluded NCS THOUGHT from the experiment.

The guises for the merger were created in the same way as the guises for the NCS as described above. The same two speakers whose NCS vowels I spliced into the NCS guises were also used to create both the merged and distinct merger guises. Myke's LOT was used to replace the original LOT in both the merged and unmerged guises. It also served to replace the original THOUGHT in the merged guises. Thus, the merger here was constructed by transferring THOUGHT words to the LOT category. Since Myke did not participate in the NCS, it was assumed that his LOT (normalized F1 829 and F2 1375) would be roughly in the phonetic space of where the two phonemes would merge in Ogdensburg. Later comparison to nearly merged speakers from my sample confirmed this assumption: The most merged speakers in my sample produce LOT and THOUGHT within an F1 range of 700 to 850 Hz, and within an F2 range of 1200 to 1400 Hz (normalized) (see Chapter 6.2). In the unmerged guises, Dan's THOUGHT was used as the alternate for the original THOUGHT in the carrier phrases. Although Dan generally participated in the NCS, his THOUGHT was not significantly lowered or fronted, so that it provided a highly representative realization of a traditional THOUGHT vowel. Thus, the unmerged guises were created with Myke's LOT vowel, and Dan's THOUGHT vowel. I chose to use two different speakers here because Dan had fronted LOT, which would have been unsuitable for creating a merged guise. At the same time, Myke had THOUGHT very close to (yet still distinct from) LOT so that it could not be used for unmerged guises. Table 5 below details the qualities of all spliced in stimuli as well as their source.

	Target	Version	F1	F2	F1 difference	F2 difference	Stimulus speaker	Source word
Lobanov normalized	TRAP	Unraised	880	1964	159	32	Myke	*bat*
		Raised	721	1996			Dan	
	LOT	Unfronted	789	1164	10	497	Myke	*socks*
		Fronted	779	1661			Dan	
		Unmerged/ merged	829	1375	0	0	Myke	*cot*
	THOUGHT	Unmerged	789	1147	40	228	Dan	*caught*
		Merged	829	1375			Myke	*cot*
	STRUT	Not backed	701	1454	12	136	Myke	*subs*
		Backed	689	1318			Dan	*dusk*
	DRESS	Not backed	735	1948	51	183	Myke	*gender*
		Backed	684	1765			Dan	*desk*
raw formants	TRAP	Unraised	785	1657	204	76	Myke	*bat*
		Raised	582	1581			Dan	
	LOT	Unfronted	727	1096	92	232	Myke	*socks*
		Fronted	636	1328			Dan	
		Unmerged/ merged	759	1267	0	0	Myke	*cot*
	THOUGHT	Unmerged	713	1125	45	142	Dan	*caught*
		Merged	759	1267			Myke	*cot*
	STRUT	Not backed	660	1301	89	245	Myke	*subs*
		Backed	571	1055			Dan	*dusk*
	DRESS	Not backed	633	1607	48	131	Myke	*gender*
		Backed	584	1476			Dan	*desk*

Table 5: Quality and sources of the matched guise stimuli. The values in the table on top are Lobanov normalized, those in the bottom table are raw formants measurements.

I used the same method to create both the NCS and the merger guises. I spliced at zero crossings, i.e. at the points in time when there is no perceived signal when the sine wave changes from positive to negative values or vice versa (see Figure 9 below), and only spliced the nucleus of the vowel, leaving neighboring segments as untouched as possible. However, in a few cases it proved helpful to shorten or delete closure and release times from neighboring stops as well as frication from fricatives to make the guise sound more natural. After splicing, I adjusted the vowel length to approximate that of the original vowel as realized by the respective voice. This was achieved by either deleting or adding full wave cycles, i.e. at least one full sine curve as shown in Figure 9 below. This was done in equal parts at the beginning, middle and end of the vowel, always matching the beginning and end of the cycle to zero crossing as shown in the screenshot above. When I added a cycle to increase vowel length, I copied a cycle adjacent to where I was going to insert it. For example, in order to add cycles to the beginning of the vowel, I copied and pasted cycles from the first few cycles of that vowel. In doing so I was able to approximate the length of the original vowel. However, since only full cycles should be added or deleted, matching the exact duration was not possible. In order to get an exact match I could instead have used Praat to speed up or slow down the vowel to adjust its length (Styler, 2016). However, doing so resulted in very unnatural sounding vowels. Since maintaining naturalness was more important to me than matching the exact vowel length, I used the adding/deleting cycles method instead.

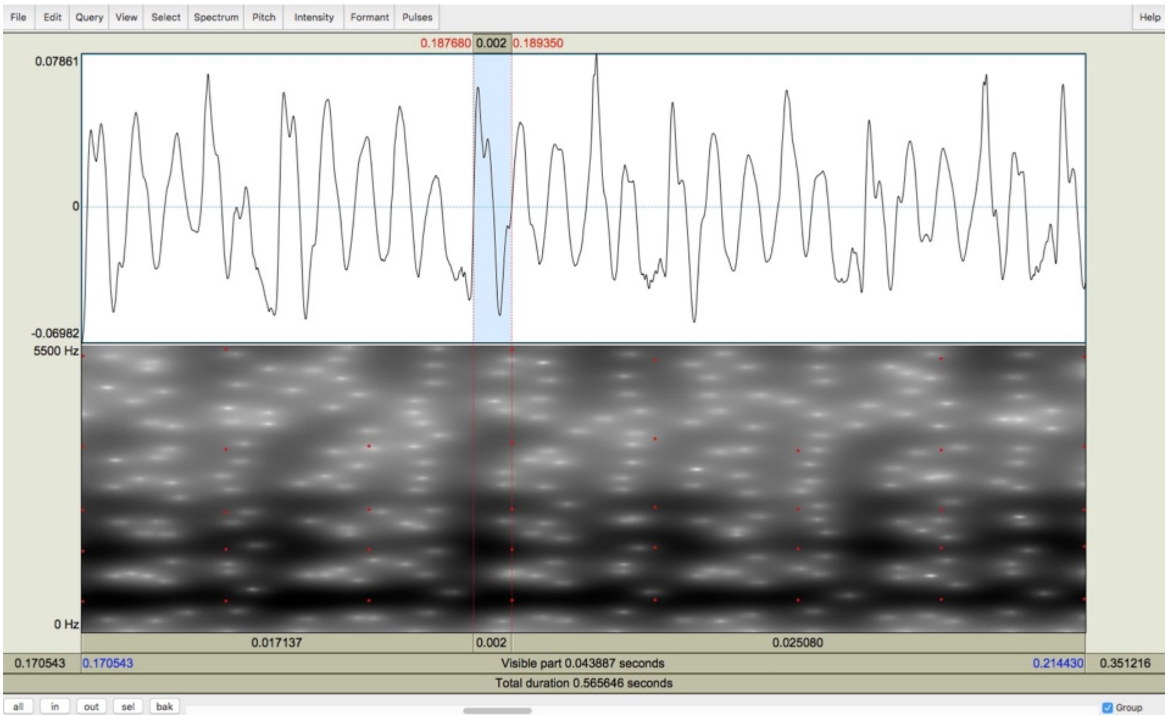


Figure 9: Adjustment of vowel length. Full wave cycles were deleted or added to the spliced in alternatives in order to approximate the length of the original vowels as realized by the voices. The points at which the sine waves cross the horizontal blue line are zero crossings.

In addition to length, I also adjusted pitch and amplitude using Praat to match the original vowel as closely as possible while still maintaining maximum naturalness in the stimulus material. In the process of matching length, pitch and amplitude I did not prioritize matching the alternates across guises to each other, so that for example TRAP in one of the raised guises might be slightly longer or shorter than TRAP in the unraised counterpart. As Campbell-Kibler (2006a, p. 78) argues, the question of how much of a match the pairs need to be is not an easy one to answer, because there might be "regular differences between the two variants with respect to, for example, length". Thus, matching e.g. the length of the variants in both guises to each other might generate a typical duration of one variant and an atypical duration of the other variant. The comparison then is not necessarily between "shifted" and "unshifted" but between "typical" and "atypical" (Campbell-Kibler, 2006a). Therefore, I prioritized matching alternate vowels to the originals to maintain naturalness, rather than matching alternate to alternate across guises. In a few exceptional cases I also adjusted parts of the guises outside of the spliced vowel, which I will explain in more detail in the next subsection.

Choosing and Lining up the Guises

Once splicing was completed, I sent the guises to fellow linguists and asked them to judge the sound quality and naturalness of the material. Based on their suggestions, I adjusted some guises in parts other than the spliced vowel in order to make them sound more natural. This affected four of the sentences produced by voice J:

I. *Her dad's so sad about it.*

This sentence was produced by the voice with an uncharacteristically long GOAT diphthong in *so*, which I shortened using the same method outlined above.

II. *He did better in the second test.*

In this sentence, the voice very clearly articulated the word-final /d/ of *second* as well as the word-initial /t/ of the following *test*, which is atypical of connected speech, where consonant clusters would be reduced. In this particular case, word-final /d/ on *second* would be omitted in connected speech, which is what I did manually by deleting the closure and release of /d/.

III. *We got a good shot of the chalk cliffs.*

This sentence stood out because of a longer-than-average pause between the words *shot* and *of* which resulted in disconnected speech. I shortened the pause to give the sentence a more natural flow.

IV. *My buddy took the bus a couple times.*

This sentence was marked by longer-than-average frication in word-final /s/ on *bus*. I shortened the frication manually to make the sentence sound more natural.

All of these adjustments were made in both the shifted as well as the unshifted guise, so that they still differed only in the target vowel. These additional adjustments were necessary to make the carriers produced by voice J sound more natural. It also helped to make J's guises stand out less in comparison to the other two voices. One of the requirements of a matched guise experiment is that the listeners are not supposed to know that they are hearing the same voices over and over again. By eliminating uncharacteristic features that would stand out to anyone, these adjustments made it less noticeable that some of the guises were in fact produced by the same person.

After all necessary adjustments were made, I selected the three best NCS guises per vowel and per voice as well as the two best guises for the merger for each voice. I included guises from all three voices in the experiment rather than using only one voice that overall produced the best stimuli, because, as Kircher (2015) argues, using multiple voices increases the reliability of potential findings. In total, this amounted to 72 guises for the NCS (3 voices x 3 sentences x 4 vowels x 2 versions each) and eight guises for the merger (2 voices x 2 sentences x 1 vowel pair x 2 versions each). The two to three sentences per vowel and voice were grouped together, so that one stimulus consisted of two or three sentences (all with the same target vowel) produced by the same voice. For example, the non-NCS guise for TRAP as produced by voice T contained the following sentences:

- That's exactly what happened.
- Her dad's so sad about it.
- She has a passion for babbling.

There are two reasons why I grouped all sentences of one guise together for each voice. Firstly, this reduced the length of the experiment and thus allowed me to include additional experiments in the interview without taking too much of the participants' time. Secondly, participants need a certain amount of auditory input before they can be asked to respond to and evaluate what they hear. In previous studies that made use of the matched guise technique, the length of the texts containing the stimuli ranged from 30 to 150 seconds, which Kircher (2015, p. 199) deemed to be "sufficient for participants to make systematic speaker evaluations". The stimuli in this study were markedly shorter, ranging from about 5 to 7 seconds. However, the target features were presented to the participants in a relatively condensed form, since each sentence contained two or three instances of the stimuli. Thus, despite the relative shortness of the guises, I presume that the guises provided enough audio input for the participants to evaluate what they heard in the guises.

In total, the participants were presented with 28 stimuli to evaluate. For the most part, I was able to select the same sentences for all three speakers, thus holding the content of the utterance 100% constant, so that content has little to no influence on the evaluations, and differences in evaluation can be assigned to different perceptions of the target phoneme (Campbell-Kibler, 2006a). However, there is some variation and not all sentences are the same for all vowels for every voice. The stimuli were presented to the participants in a randomized order during the experiment, the same order for each participant (see Appendix C). I did not add any filler sentences to this experiment, as evaluating 28 guises was a lot to ask of the participants to begin with and adding fillers to distract the participants seemed too much for this task. This may, however, have compromised the obscurity of the voices. Although I made sure that the guises were lined up pseudo-randomly and no voice (and no target feature) occurred consecutively (Kircher, 2015), so that each stimulus simultaneously served as a filler, some of the participants did mention that they thought they heard the same voices repeatedly, which may have affected their ratings. Adding fillers might have prevented this from happening, although there is no guarantee for that either. Cases in which participants noticed the repetitiveness of the guises will be discussed further in Chapter 2.5.2.

In the interest of saving time and so as to not overwhelm the participants with too much stimulus material, I also chose not to include practice voices at the beginning of the experiment. Practice voices are voices that deliver the same text as the actual voices, but are not part of the experiment, and give the participants the chance to familiarize themselves with the experimental procedure, so that when they later listen to the recordings of the actual voices, they will no longer be preoccupied with the practicalities of the evaluation process (Kircher, 2015). While this is in general a good idea and undoubtedly would have been helpful, adding more guises to an experiment that already contained 28 stimuli did not seem practical.

Avoiding potential order effects is another aspect to consider when lining up stimuli in a matched guise experiment. Order effects occur when a "previous treatment influences the participants' behavior in a subsequent treatment" (Abbuhl, Gass, & Mackey, 2014, p. 120), i.e. the rating of one guise influences the rating of the following guise. This influence can be due to practice effects or fatigue effects among the listeners. If all participants are presented with an identical line-up of stimuli, it is possible that potential patterns in the participants' ratings are not based on their actual attitudes toward the speakers and their language, but are merely due to order effects. One strategy to avoid this is counterbalancing the sample, i.e. presenting one half of the participants with the stimuli in order A and the other half of the participants with order B. The stimuli in the present study were, however, not counterbalanced, which might be one of the main flaws of the experiment. Fatigue in particular may have played a role in the ratings, as some participants were quite explicit about getting tired or bored with the experiment, some even asking to skip some of the stimuli. However, Clopper et al. (2011, p. 155) argue that counterbalancing is particularly important when consecutive guises in the line-up "represent an experimental manipulation", which does not seem to be the case for the guises in the present study, as they were randomized (by voice and target vowel). Thus, counterbalancing may not have been absolutely necessary in this case.

Designing the Evaluation Sheet

For the matched guise experiment, the participants were given an evaluation sheet that listed all 28 stimuli (see Appendix C). This allowed them to read along if they wished to do so (Kircher, 2015). The sheet also contained a rating grid on which the participants were asked to indicate their ratings of the voices they heard. Figure 10 below shows the rating grid that I used for the experiment, and which was part of the evaluation sheet.

As discussed above, I used pre-defined categories for this experiment: friendliness, education, age, localness and Canadian-ness. The participants were asked to indicate to what degree they agreed with the applicability of all five categories to each stimulus they heard, thereby rating all of the stimuli. Their choices ranged from "strongly agree" to "strongly disagree" on a 6-point forced-choice Likert scale. Using a scale with an even number meant that participants were required to either agree or disagree with each category to a certain extent, thus forcing them to pick a side. Another option would have been to use a 5 or 7-point scale with the opportunity for the participants to indicate a neutral opinion. While this can be beneficial, offering this option creates a "significant risk of respondents giving a neutral answer for neutrality's sake" (Rasinger, 2013, p. 67). Especially when it comes to an uncomfortable task like judging and evaluating people, it seems very likely that participants might have made excessive use of the option to remain neutral, which is why I decided on a forced-choice scale with an even number instead. Even-numbered scales, however, can lead to biases in the ratings. Rasinger (2013) references Ping (2005) in pointing out two issues with those scales: the acquiescence response set and extreme response styles. The former refers to the phenomenon of informants disproportionately agreeing to statements regardless of the content or the participant's actual opinion. Extreme response styles, on the other hand, are "characterized by respondents' tendency to select responses at the extreme end of the scales" (Rasinger, 2013, p. 67). In the present study, neither of these two phenomena has been observed, however.

This speaker sounds...	**Strongly agree**	**Agree**	**Weakly agree**	**Weakly disagree**	**Disagree**	**Strongly Disagree**
friendly:						
educated:						
old:						
local to Ogdensburg:						
Canadian:						

Figure 10: Evaluation sheet for the matched guise task. To rate each stimulus, participants were asked to write the number of the stimulus into the appropriate box for each of the five categories.

Triangulation

Despite all the advantages of the matched guise technique in eliciting language attitudes indirectly, I added two ways to elicit attitudes overtly and directly to complement the quantitative data collected through this experiment. In doing so, I triangulated my methods, which makes interpretations of the results more reliable by enabling comparison across three methods rather than making statements about attitudes from just one measurement (Kircher, 2015). The two direct methods I included were a comment section to the matched guise experiment, where participants had the option to write down any comments they had about any of the stimuli (see Appendix C) and open questions during the interview proper. However, none of the participants made use of the first option, suggesting that the participants did not have too many conscious attitudes to any of the stimuli. The only stimuli that received a few comments were those that contained the merged LOT-THOUGHT guises, to which participants occasionally responded with "this guy is from Boston", or "I've never heard it pronounced that way" (meaning THOUGHT realized as LOT). The implications of these comments will be discussed in Chapter 8.3.2. Open questions about language variation, on the other hand, resulted in some valuable information, as some participants did bring up certain linguistic features they are aware of, including phonetic ones. Although none of them passed any judgment on the features they mentioned, this provided insight into the participants' awareness of dialectological differences in and around New York.

The benefit of these direct attitude elicitation methods is that they can provide valuable information while being very easy to carry out. However, on the downside, overt comments do not always reflect the participant's true feelings about certain features or varieties. This is because sometimes those feelings and beliefs are unconscious, and the participant is not aware that they are making those judgments. In some cases, it is also possible that the responses about language attitudes are dishonest. As discussed above, when participants realize what the purpose of the question is, namely eliciting attitudes, they have a tendency to respond in socially desirable ways so as to present themselves in a better light.

2.3.2.2 Commutation Test – Perception of Linguistic Categories

For this study, I decided to use the commutation test to assess the participants' ability to perceive phonemic contrast between LOT and THOUGHT. Drawing on research in phonetics, cognitive psychology, and speech and hearing sciences (Clopper et al., 2011), this test provides rich insights into the participants' phonological structure while being relatively efficient, as it requires little planning and can be executed quickly compared to e.g. the Coach Test (Labov, Mark, & Miller, 1991) or the Vowel Categorization Experiment (Di Paolo, 1988). The commutation test is an identification tasks in which participants are asked to label segments of speech with labels provided by the experimenter, thereby identifying a set of stimuli (e.g. Niedzielski, 1999; Plichta, Preston, & Rakerd, 2007). This is a more advanced methods that developed out of simpler discrimination tasks that merely involved same/different evaluations of a set of stimuli, preferably the participant's own articulation of a minimal pair or minimal pairs in writing. Although same/different judgments were integrated into the study as well (see Chapter 2.4.4), a commutation test seemed to be better fitted for this project, as it generally provides more reliable and valid data (Labov, 1994, p. 356).

The initial version of the commutation test had one native speaker read a randomized list of members of minimal pairs to a native listener. The listener would then be asked to identify which word he heard. A listener with a secure distinction would be able to correctly identify all words, while the results for a merged listener would be roughly 50%. Anything in between 50% and 100% would indicate transitional perception (Labov, 1994, p. 356). However, the question that remains with this method is whether the listener's success rate in this test is due to their own perception or the production of the speaker whose minimal pairs they hear as the input stimuli. Thus, I used a modified and advanced version of this test. For each participant, I generated a line-up of five *cot* and five *caught* items of their own production that was recorded during the minimal pair reading part of the interview (see Chapters 2.3.1.3 and 2.4.4). For the identification task of the test, this line-up of their own production was played back to the participants, and they were asked to identify which word they heard. However, this self-commutation test leaves us with a similar problem as the initial version of the test: A speaker who is more merged in production is likely to have more trouble identifying their own production of *cot* and *caught* than a speaker who produces them with a clear distinction. Therefore, I extended the test and included a second line-up of *cot* and *caught*. This second line-up was created with distinct stimuli from an Ogdensburg native speaker recorded in 2008 by Dinkin.[45] From his reading of *cot* and *caught* (he produced each word once), I created a line-up of five randomly alternating instances of *cot* and *caught*[46], each separated from the next by a 2-second interval of silence (which corresponds to the intervals between each item in the participants' reading of the *cot-caught* list). As with the line-up of their

[45] This was the same speaker that I used to create the shifted guises in the matched guise task (Dan).
[46] The order in which the participants heard these items can be found in Appendix B.

own production, the participants were again asked to identify which of the two words they heard. Since the 2008 Ogdensburg speaker produced the two words with a very clear distinction in his reading of the minimal pair, the failure to correctly identify the words in his line-up could with relative certainty be attributed to the merged perception of the listener rather than to indistinguishable stimuli. On the other hand, if a participant failed to correctly identify their own *cot* and *caught*, but was able to do so for the second line-up with clearly distinct tokens, this would indicate that they are in fact still able to discriminate and identify both sounds correctly as long as they are produced with sufficient contrast. Without the second line-up, this information would be missing, which might result in an overestimation of the degree of merger in perception in those participants, based on their failure to correctly identify LOT and THOUGHT in their own production.

2.3.3 Pilot Study

Unfortunately, timing did not allow me to pilot test any of the experiments before starting to interview participants. Instead, I used the first few interviews as a way to test out the material. Based on the impressions and feedback from those first interviews I made a few adjustments to the experiments:

- Instead of open questions I opted for a forced-choice test with predetermined categories for the matched guise experiment. Open questions may be able to provide more detail in some cases, however, in this particular study, participants did not seem to be able to consciously pick up on and comment on the differences in most of the guises, so that it seemed more sensible to give them a few options on which to rate the stimuli.
- Three words were added to the wordlist: *sorry* proved to be interesting, as its pronunciation ranged from a very raised NORTH/FORCE to LOT. *Golf* and *revolve* were added following Dinkin's finding that /ɑlf/ and /ɑlv/ clusters might be subject to phonological transfer in Northern New York (Dinkin, 2009, 2016).

2.4 In the Field: Data Collection

In the previous subchapters I described how I planned the sociolinguistic interview and designed the speech production and perception experiments for this study. In this chapter, I will discuss how these techniques were implemented in the field. This includes a description of my sampling method, how I recruited participants, conducted my interviews, and how I administered the speech production and perception experiments.

2.4.1 Sampling

I initially defined the criteria for participation in my study as follows:

- Participants must have been born in Ogdensburg
- Participants must have grown up in Ogdensburg
- Participants must currently live in Ogdensburg

However, the turnout with these criteria was quite limited, so that I dropped the "currently living in Ogdensburg" requirement. Expanding the sampling universe (L. Milroy & Gordon, 2003, p. 26) in that way drastically increased the number of potential participants. I was now able to include speakers that had left Ogdensburg but stayed in the area as well as informants who had left the area but happened to be in town during my stay in the North Country.

In an attempt to ensure that the sample of speakers would be evenly stratified I used quota (or judgment) sampling, which is argued to be the most appropriate sampling technique in linguistic research (Chambers, 1995, p. 41 as cited in L. Milroy & Gordon, 2003, p. 33). Thus, before starting the process of recruiting and interviewing informants, I designed a stratification scheme (illustrated in Table 6 below), delineating certain social categories specified by demographic characteristics that have been shown to be of importance in the study of sound change (see Chapter 1.4), with the exception of ethnicity, which turned out to be a redundant category in this project, as all but one of the participants were white.[47] I intended to record at least three speakers per cell, which

[47] This participant was born in China but was adopted by a white family in Ogdensburg at the age of one. Although her adoptive sister was born in China as well, there is otherwise no prominent Asian community in Ogdensburg, and she is fully integrated into the predominantly white community. Her social network consists of predominantly white peers. Therefore, her ethnic background was not considered in this study. Since her production patterns do not deviate markedly from those of her peers, this decision seems justified.

would allow for relatively reliable statistical testing (Tagliamonte, 2006, p. 31), as each cell of the scheme is assumed to be representative of speakers with the same characteristics in the community (L. Milroy & Gordon, 2003, p. 32).

Year of birth	>1998	1997-1987	1986-1976	1975-1955	< 1954
Female					
Student	Chloe (98) Grace (02)[48]				
No college				Breanna (64) Ashley (66)	Donna (35) Stephanie (52)
College		Lindsey (87) Summer (87) Megan (90) Sophie (91) Allison (93)	Rachel (76) Aubrey (80) Mandy (84)	Amanda (58) Charlotte (58) Monica (65) Amber (67) Sarah (69) Martina (70) Melissa (72)	Nicole (32) Tracy (44) Ruth (48) Helen (49) Bethany (50) Kelly (53)
Male					
Student	Jason (98) Will (99) Ben (99) Mark (01)				
No college		Anthony (91)			Eddie (43) Brian (45) Scott (46)
College		Daniel (93)	Ryan (83)	Patrick (68)	Richard (41) Gary (41) Henry (53)

Table 6: Stratification scheme for the speaker sample. Cells without potential members are shaded in gray. The names used for individual speakers are pseudonyms.

As is apparent from Table 6 above, I was unfortunately not able to reach the goal of three speakers for every cell, some being left blank completely, lacking male speakers in general as well as speakers with a lower level of education[49] for both males and females. As a result, the sample is somewhat skewed toward highly educated women. While efforts were made to rectify this bias in order to achieve representativeness (L. Milroy & Gordon, 2003), I was more concerned with filling the top rows for both genders with school-aged speakers. Although it does not show in Table 6, finding teenaged participants proved very difficult during the summer. Once school restarted toward the end of my fieldwork in September, I was eventually able to make contact with and interview some of the students at the local high school with the help of a teacher that I had interviewed. Since age is a

[48] Grace was actually still in middle school at the time of interviewing. However, since she was the only middle schooler, I added her to the high school category in this study.
[49] Educational background is a factor frequently used to indicate social class and was used in the current project for this purpose.

more important factor in the apparent-time framework than gender and education, filling the cells for teenaged speakers was prioritized over finding more male and less educated speakers.

All in all, I was able to recruit and interview 41 participants for the study. However, I included only 39 of those speakers in the actual corpus for this project, including 25 women and 14 men, born between 1932 and 2002 (mean: 1969). Two female participants were excluded for the following reasons: One of the recordings was of particularly bad quality because the speaker did not attach the lavalier microphone (see Chapter 2.4.3) close enough to her mouth, so that the background noise overpowered her speech. The second speaker was removed because, while she was born in Ogdensburg and lived there during the time of the interview, she had moved away from Ogdensburg as a young child, grew up outside of New York State and did not return to Ogdensburg until her early teens. Thus, all speakers included in this sample were either born in Ogdensburg or moved there at a very early age. A detailed overview of all speakers and their demographic background can be found in Appendix D.

In addition to the participants I recorded in 2016, I also included the nine native Ogdensburg speakers recorded by Dinkin in 2008. This corpus includes speakers born between 1922 and 1989, with a mean year of birth of 1972, due to an old outlier born in 1922. Seven of these speakers are female, two are male, and all nine speakers are white. Three of them had a college degree, three did not, and three were students at the time of the interview. Including data collected 8 years prior to my own fieldwork allowed me to add a real-time component to the apparent-time analysis in this study. While the apparent-time approach, introduced in Chapter 1.4.3, is commonly taken in most sociolinguistic research and the inferences drawn from these studies are generally found to be reliable (Chambers, 2008), there are significant drawbacks to its underlying hypothesis, most notably the possibility of age-grading that might be misinterpreted as linguistic change in apparent time. Additionally, real-time evidence occasionally reveals that apparent-time inferences did *not* hold true. As will become clear later, this is also the case for some of the findings in this study (Chapter 8.2.2). In order to avoid these issues that apparent-time studies can entail, Labov (1972, p. 275) advises to complement apparent-time data with "at least one measurement at some contrasting point in real time" against which the current data can be interpreted (L. Milroy & Gordon, 2003, p. 36). Adding the 2008 data to the sample allowed me to do exactly that, and potential apparent-time effects can be supported or discredited in the analysis using the 2008 data as an anchor in real time.

However, an important issue to consider when combining data from different sources is the role of the interviewer and their method of data collection and analysis. While, similar to mine, the majority of Dinkin's recordings were made during face-to-face interactions, Dinkin followed Ash's (2002) protocol for Short Sociolinguistic Encounters and interviewed people he approached spontaneously on the street. The interviews lasted 15 to 30 minutes and, in addition to the conversational part of the interview, included the reading of a wordlist and minimal pairs as well as other elicitation tasks. Two of Dinkin's participants were interviewed over the phone following *ANAE* methodology, but both were found to present a reliable picture and gave "the same general impression of the status of the community as the larger sample" (Dinkin, 2009, p. 45). I, on the other hand, followed a more formal approach of recruiting and interviewing, as will be described in detail in the following two subchapters (2.4.2 and 2.4.3). In addition to differences in sampling methods, differences in social characteristics of the interviewer, such as age, gender, nationality, and language proficiency may have influenced the linguistic behavior of the informants in the present study. While Dinkin is a male native speaker of (non-NCS) US English, I am a female non-native English speaker (German being my native language). Lastly, the formant measurements in the 2008 data were extracted at points selected by hand in Praat (see Dinkin, 2009 for detailed methodology), while the 2016 data was force-aligned and formant measurements were extracted automatically by FAVE (see Chapter 2.5.1). Although it is possible that this has led to differences in formant measurements, Severance, Evanini, and Dinkin (2015) found no significant differences between the results of the two techniques in a direct comparison of Dinkin's hand measurements to FAVE's results on a subset of interviews from Dinkin's (2009) corpus, including some NCS speakers. Nevertheless, methodological differences will be taken into consideration in the analysis, and their potential influences will be discussed in Chapter 8.2.1.

2.4.2 Recruiting Participants

The data for this study was collected during a 3-month fieldwork stay in Ogdensburg, NY in the summer of 2016. Since I had limited social ties in Northern New York and none in Ogdensburg before I started my fieldwork, I had to rely on several different methods of recruiting participants for the study. These methods included print media (a poster and a feature article in local and regional newspapers) and social media (the Ogdensburg history group on Facebook).

The poster shown in Figure 11 called for volunteers to participate in my study. It provided a brief description of the project, the requirements for participation (as outlined above), an overview of what participation would entail, as well as my academic background and contact information. I put up the poster in different locations around the city, including the library, the visitor center, cafes and diners. This was, however, the least successful of my recruiting methods, and I do not believe that any of my informants participated in my study as a result of having seen this poster. In fact, the poster at the local library, which I visited frequently, remained untouched throughout the 3 months I spent in Ogdensburg.

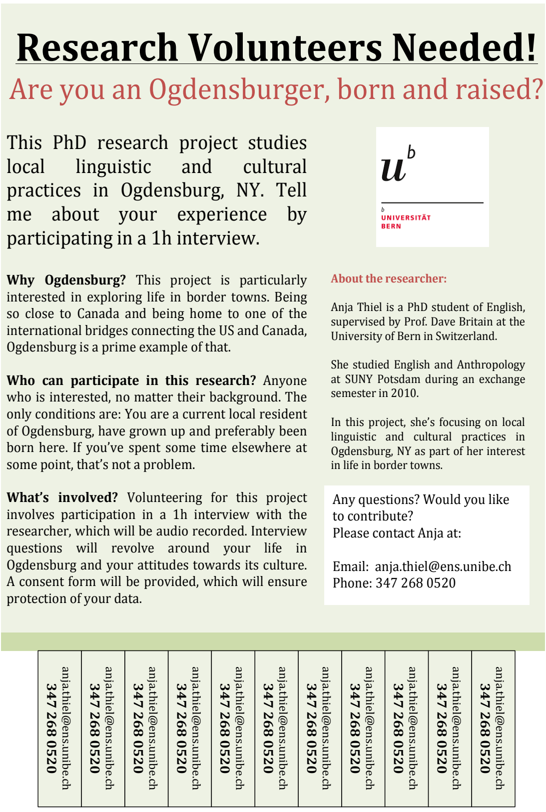


Figure 11: Poster calling for participants for my study.


The Journal
TUESDAY July 26, 2016
City's Bond Rating Downgraded
Unique Dialect
A Graduate Student Is Studying Speaking Patterns In Ogdensburg
Hillary DNA Issues Heard
Cyclist Rides Across The Country To Fight Against Cancer
Jake Kiah Wins Al O'Marah Trophy At Sandbar Classic
BRIEFS
Good Morning George Kotz Thank You for subscribing to The Journal

Figure 12: Newspaper article reporting my research.

The newspaper article was one of the more successful, if not the most successful recruiting method. The article was published on the front page of various regional newspapers, including The Journal (Figure 12 above), as well as in the form of an online article on their respective online platforms, which was shared frequently on Facebook. It included some of my personal background, my motivation for my research and, of course, a call for participation. The editor of the newspaper was very specific about the background of the picture, which he thought should be characteristic of Ogdensburg in the most positive way, and the St. Lawrence River, an immense source of pride for Ogdensburgers, is just that.

In the Ogdensburg history group on Facebook, I publicly posted a call for participation. Like the poster, this post included information regarding the project, the requirements for participation, what participation would entail, as well as my academic background and contact information. I also included the poster for visualization (the newspaper article had not been published at that point). This Facebook post immediately generated responses, and the first participants were recruited through this method. Once some of the members of this Facebook group had been interviewed, they commented on my post advocating for me and verifying the legitimacy of my research, and encouraging other members to participate as well, which helped to reduce my status as a stranger in the community.

Once a number of contacts had been established, I was also able to make use of the "snowball" technique (friend of a friend) and was referred to further potential participants by some of my informants. Generally, former participants forwarded my contact information to their friends or family members, who then contacted me to schedule an interview. In some cases, they gave me the name and contact information of people they thought might be interested in being interviewed, or put me in touch with them on Facebook. Making use of the snowball recruiting technique made me less of an outsider with many of my informants, as I had already met with and interviewed at least one of their friends (hence friend of a friend) who trusted me enough to forward me on to them, which potentially weakened the observer's paradox, as I was no longer just an observer, but someone that community members were already familiar with and who had spent at least some time in the community.

2.4.3 Interview and Recording

I followed the same procedure for every interview. Each meeting was scheduled in advance and lasted 1 to 2 hours. I generally suggested the public library as a meeting place and most of my interviews were recorded there. However, some participants chose to invite me to their office or to their house instead, and most of my school-aged participants were interviewed under supervision[50] at the local high school. The majority of the interviews were one-on-one conversations between the informant and me. In a few cases, however, I interviewed two people at a time (both recorded on separate tracks). This was generally the case with couples that lived together and invited me to their house for the interview. Before starting the actual interview, I introduced myself to the participant(s) and outlined the structure of the meeting, which would consist of a conversational part, as well as a reading and listening experiment. I asked the participants to sign the consent form[51] (see Appendix E), explained that their participation is voluntary, that they are free to withdraw at any time, and that the data will be anonymized. None of the participants were informed of the real research purpose of the study before the interview. All informants agreed to being recorded before I started the interview, and all recordings were made using the same recording equipment: a Zoom H5 recorder paired with AKG C 417 PP lavalier microphones, as well as a MacBook Air (13-inch, Mid 2013) for the reading and perception experiments.

After the interview proper, the majority of the participants agreed to also complete the listening and reading experiments. Di Paolo and Yaeger-Dror (2011a, p. 16) argue that "given the importance of the vernacular to sociolinguistic analysis, the tasks which focus on pronunciation should always be placed as late in the session as possible so that the conversation itself will be as untrammeled with self-conscious speech as possible". Thus, the experiments were always pushed toward the end of the interview. How I ran these experiments will be explained in the following section.

2.4.4 Administration of the Experiments

In addition to the interview proper, I included various production and perception experiments in my interviews. These included the matched guise experiment, the wordlist reading, the reading of minimal pairs and the commutation test. For a detailed

[50] The interviews at the high school were conducted in a side room to a larger classroom. A teacher and other students were present in the classroom, and the door connecting the two rooms was kept open at all times, as was requested by the school administration. This did, however, not interfere with the interview process.
[51] When I interviewed minors, I asked that their parents sign an adapted consent form ahead of time.

explanation of how the experiments were constructed, e.g. preparation of the guises, stimulus material and line-up etc. see Chapter 2.3. The order in which I administered the experiments was as follows:

- Social perception (matched guise)
- Production (reading)
- Linguistic perception (commutation test)

Since the reading of the wordlist and particularly the minimal pairs brings into focus the target sounds the researcher is interested in, they are usually the last tasks in sociolinguistic interviews, so that the increased awareness of those targets cannot affect any of the other tasks (Di Paolo & Yaeger-Dror, 2011a, p. 16). However, in the present study, the commutation test was conducted at the end of the interview out of logistical necessity: I needed the participants' reading of the *cot-caught* line-up in order to play it back to them, as will be explained in more detail below.

The matched guise task was the first of the experiments that followed the interview proper. The 27 (of 39[52]) participants who completed this experiment were given the evaluation sheet (Appendix C) along with the following instructions:

> *You are going to hear short recordings of voices that I have asked to read a couple of sentences. They are all reading the same sentences. I will play one voice at a time, each containing two to three sentences. After each set, I will pause the recording, and you will be asked to rate the voice in terms of friendliness, education, age, how local to Ogdensburg and how Canadian they sound to you. Try not to think about it too much, there is no right or wrong, it is really just about how you perceive the voices, you cannot fail at this task. All sentences are listed on your evaluation sheet, so you can read along if you wish.*

If the participants had no further questions, I started playing the stimuli. Each set of guises was only played once unless a participant asked to hear it again or we were interrupted, in which case I let them hear it again. The participants heard one stimulus, i.e. one set of

---

[52] Students who were interviewed at the high school were not asked to participate in the matched guise experiment owing to time constraints, but did complete the reading tasks.

three sentences, at a time. Each guise in that set contained the same target vowel, in either its "standard", shifted or merged form. After each set, I paused the recording and asked the participants to note down their ratings. This procedure was repeated until the participants gave their ratings for all 28 stimuli, and all participants were presented with and responded to all (identical) auditory stimuli. This design allows the researcher to better control for inter-individual differences (Abbuhl et al., 2014, p. 120), and is generally more powerful than the between-subject design, in which one subset of the stimuli are presented to a random subgroup of the listeners, and a different subset of the stimuli is presented to a second random subgroup of listeners. However, a potential drawback of the within-subject design are possible order effects as discussed in Chapter 2.3.2.1.

After the matched guise experiment, the participants were asked to perform the reading tasks. They were told that the reading would contain three parts in the following order: a list of words, a set of word pairs and a set of repeating words in the end (which was the *cot-caught* line-up for the commutation test. All three parts were presented to the participants in form of a PowerPoint presentation with automated and timed slide transitions. The three tasks were performed in one go without any longer breaks. Instruction slides at the beginning and in between sections separated the minimal pairs from the wordlist and the reading of the *cot-caught* line-up. Those instruction slides gave the participants simple instructions as to what to expect: "*Please read the following words/word pairs/repeating words*". For the wordlist and the set of repeating words, the participants were presented with one word at a time, i.e. one slide contained one word, with a transition time of 2 seconds per slide. In both of these tasks, words were presented in random order (the same for each participant).[53] For the minimal pairs, each slide contained both members for each pair (one on top, one at the bottom), and slides were presented at 3.2 second intervals. The members for each individual pair on each slide were arranged in pseudo-random order: For some of the pairs, the LOT word was presented on top, followed by its THOUGHT counterpart at the bottom, and vice versa for other pairs (see Chapter 2.3.1.2). This was intended as a preventative measure to avoid artificial distinction between both word groups based on the order in which they occurred. Hay, Warren, and Drager (2006, p. 467) predict that this strategy would provide

[53] The randomized wordlist and *cot-caught* line-up can be found in Appendix A and Appendix B.

"the most accurate indication of the degree to which a speaker keeps the distributions distinct in natural speech".

Every participant read the wordlist, the minimal pairs and the repeating *cot-caught* line-up one time. It is often suggested that participants should be asked to perform reading tasks, particularly the wordlist, multiple times in different random orders (Di Paolo & Yaeger-Dror, 2011a, p. 15). This way the researcher can reduce the risk of order effects, i.e. the influence the order the words are presented in, on their articulation. Because of the relatively high number of experiments in this study, some of them relatively lengthy, however, I asked the participants to perform the reading tasks only once so as to avoid fatigue among the participants.

After the participants finished the reading, they were asked whether they thought the minimal pairs sound the same or different. This provided the self-reported data that allows for an analysis of actual and perceived production (Hay et al., 2009) and was my way of integrating the minimal pair same/different evaluation method into my interviews. Some studies, e.g. Hay et al. (2009), ask for this evaluation for each minimal pair individually. I opted to ask my participants for all minimal pairs collectively, in an attempt to keep the experiment as short as possible. However, this compromised the amount of detail in the participants' responses. Some of the participants commented that some of the pairs sounded more alike than others, which may or may not be related to e.g. the absence or presence of morpheme boundaries (see Chapter 2.3.1.2). With the structure I used, I am not able to tell which of the minimal pairs they refer to in their judgment and whether or not there is agreement about this across participants. Additionally, it is possible that the participants' self-judgment might be understating the degree of merger in perception, as the reading tasks ended with the *cot-caught* pair, which has been found to be marked as distinct more frequently than other minimal pairs of LOT and THOUGHT (Gordon, 2006; Johnson, 2007).

After the participants were asked for their self-judgment of minimal pairs, I explained the concept of the commutation test and told them that they would now hear what they had just read in playback. Since I recorded the entire reading passage (wordlist, minimal pairs and the *cot-caught* line-up) on my computer using Audacity (Audacity Team, 2012) in addition to the voice recorder, I was able to administer the commutation test without interrupting the recording of the interview. I used Audacity to change the order of the repetitive *cot–caught* list on the spot, moving the first three words to the end of the list in order to avoid the possibility of the responses being based on the participants' memory of the order they read the words in. When necessary, I amplified the recording in Audacity before I played it back to the participant. I then played the reordered line-up of their own *cot-caught* production back to the participants on my computer, and participants were asked to indicate whether they heard *cot* or *caught*. To reduce error rates, both options were listed as column entries on the evaluation sheet the participants had been given (Labov, 1994, pp. 356–357), so they could check the box for whichever word they thought they heard. They also had the option to choose "I can't tell" for each item if they were not able to identify the word (see Appendix C). After they finished the commutation test for their own recording, I played them the pre-recorded line-up of five very distinct *cot-caught* stimuli (see Chapter 2.3.2.2) and asked them to again identify which of the two words they thought they heard on their evaluation sheet, again with the option to choose "I can't tell". The participants heard each line-up once unless there were interruptions or distractions that made it impossible to complete the task, in which case I played the line-up a second time. I also asked them in which of the two line-ups they found it easier to identify the words. After the completion of the commutation test, I generally stopped the recording of the interview. I then asked the participants to provide or confirm their demographic information, such as year of birth and birthplace and location of their parents. If they were interested, I also informed the participants about the actual purpose of the study.

2.4.5 Summary

Overall, the interviews were comfortable, and the experiments occurred without any major glitches that would affect the data. With the exception of the prospective participants whose names had been provided by earlier informants, I did not approach people directly to ask for participation in my study, especially not *spontaneous* participation. Thus, virtually all informants had the opportunity to contemplate whether or not they would be willing to be interviewed by me. Therefore, all participants seemed enthusiastic and eager to share their stories during the interviews, resulting in conversations that seemed very natural, and provided reliable data for the analysis of speech patterns in Ogdensburg.

2.5 Data Processing

Once fieldwork was completed, several steps were taken to process the collected recordings and prepare the data for analysis. These steps included orthographic and phonemic transcription of the recordings, automated vowel extraction, cleaning the data of unwanted tokens, renormalizing the data sets, and calculating the means. Before the data was used for analysis, I checked the reliability of the automatically extracted formant measurements. The measures that were taken to achieve this will be outlined and discussed in light of their appropriateness for this study in the following sections, starting with transcription and vowel extraction in the next subchapter.

2.5.1 Transcription and Vowel Extraction

After the interviews were recorded, I transcribed each interview orthographically using ELAN, a free software package for annotating audio files (*ELAN*, 2016). These transcriptions served as a base for FAVE, a program suite that automatically aligns and extracts vowel formant measurements from orthographically transcribed data. The suite consists of two separate programs: FAVE-align and FAVE-extract. Both FAVE-align and FAVE-extract were available as an online interface but can also be downloaded as command-line versions.[54] In the following sections I will briefly outline how each of these programs work and how they were used in this study.

FAVE-align is a program that automatically produces time-aligned phonemic transcriptions. Based on an orthographic transcription and a respective audio file provided by the researcher, FAVE-align produces phoneme-level transcriptions for each speaker in the recording which are force-aligned against the speech signal in the recording. In order for this to work through the web-based interface, I had to down sample my audio files to 44100 Hz, as uploading larger sound files was not possible. However, 44100 Hz, or 44 kHz, is, at present, “the de facto standard sampling rate” (Podesva & Zsiga, 2014, p. 171). The output of the alignment process is a time-aligned Praat text grid (Boersma & Weenink, 2017), with separate tiers for each speaker in the recording. The phonemic notations generated by FAVE-align are based on an external

[54] At the time of writing, the online interface for FAVE is only partially operational. At the time of transcribing and extracting, FAVE-align was still fully functioning so that I was able to use the online interface for forced alignment. For FACE-extract, I had to rely on the command-line version.

pronunciation dictionary, the Carnegie Mellon University dictionary (CMU Pronouncing Dictionary). The researcher is also required to manually supply transcriptions for words that cannot be identified by the dictionary. Because the CMU is based on General American English (George Bailey, 2016), FAVE-align works particularly well with varieties of English spoken in North America, as is the case in this study. Nevertheless, I manually checked the force-aligned phonemic transcriptions for potential alignment errors using Praat. This was done to make sure that time-alignment was successful, i.e. the segmentations in the annotation matched the segments in the audio file, both visually (spectrogram and waveforms) and audibly. If this was not the case, i.e. when the alignment of a transcription segment crossed into the preceding or succeeding sound segment, I manually adjusted the alignment. For the spontaneous speech section of the interviews I performed random checks at random intervals for all speakers. If these random checks did not indicate any serious alignment issues, I made no further adjustments. If, however, it seemed as though multiple sections were seriously misaligned, I went through the speaker's entire transcript more carefully, and realigned all mismatches that I could detect in the file. This was, however, only necessary for a handful of recordings, typically those for which transcription segments were too long, and those that had a good amount of background noise. All tokens in the reading passages of the interview (wordlist, minimal pairs and *cot-caught* line-up) were checked individually and carefully for every speaker and were, if necessary, corrected.

The phonemic notations produced by FAVE-align served as a basis for acoustic analysis of the data. Acoustic analysis has become the standard method of analyzing vowel variation in most sociolinguistic studies since about 1980 (E. R. Thomas, 2002) and was the method of choice for the present study, in which I made use of FAVE-extract. FAVE-extract is the second program in the FAVE suite and automatically extracts vowel formant measurements. For this purpose, the phonemic transcriptions that were produced by FAVE-align along with the down-sampled audio files were fed into the downloadable command-line version of FAVE-extract. Although FAVE allows the researcher to specify measurement settings, I kept the default settings the FAVE online interface recommends:

- Formant Prediction Method: mahalanobis
- Measurement Point Method: faav
- nSmoothing: 12
- remeasure

The mahalanobis formant prediction method uses Evanini's (2009) algorithm, which compares all measurements for a vowel to a range of expected measurements according to *ANAE*. The faav measurement point setting instructs FAVE to take measurements of the first and second formant at ⅓ of the vowels' duration, with the exceptions of PRICE and FACE, which are measured at maximum F1; GOAT and MOUTH, which are measured halfway between the vowel onset and maximum F1; and GOOSE preceded by coronal consonants, which is measured at the vowel onset. In addition, FAVE takes formant measurements at 20%, 50%, 65% and 80%; however, analysis was based primarily on the ⅓ mark measurement, with the exception of GOAT (see Chapter 7.4). The nSmoothing parameter specifies the width of the time window around which formants at a particular point in time are calculated, and the default of 12 corresponds to a 25 ms window. The remeasure setting tells FAVE to perform a second round of measurements on each speaker, which can then be compared to the speaker's individual system rather than to that of *ANAE* speakers. FAVE-extract also allows the researcher to specify in which instances vowels should be measured. I instructed FAVE to only measure vowels with a duration of at least 0.05 seconds that carry primary stress and specified stop words (i.e. words to be excluded from measurement). The list of stop words contained prepositions, pronouns, determiners, interjections, conjunctions, enumerators, auxiliaries, as well as "yes", "not", "as", including all contracted forms of these word classes. A full list of stop words can be found in Appendix F. FAVE-extract automatically normalizes formant measurements by means of the Lobanov normalization method (Lobanov, 1971), setting each speaker's overall mean F1 equal to 650 Hz (with a standard deviation of 150) and mean F2 equal to 1700 Hz (standard deviation 420), which has been shown to be one of the best performing normalization methods available (Adank, Smits, & van Hout, 2004; Fabricius, Watt, & Johnson, 2009). In addition to the normalized data, FAVE also outputs raw, unnormalized formant data as a separate file.

As Di Paolo, Yaeger-Dror and Beckford Wassink (2011, p. 87) argue, "instrumental phonetic analysis is a requisite for the sociophonetic study of vowels". This process can, however, be very time intensive if each vowel is measured individually by hand at multiple points, particularly with larger amounts of data. Automated vowel extraction makes this process more efficient and thus has become increasingly popular among sociolinguists, as it provides consistent and replicable formant measurements of the entire vowel space (measured at several points) of a large number of vowel tokens in a very short amount of time (Severance et al., 2015). In comparison with other aligners, MacKenzie and Turton (2013) found that FAVE-align offers the most accurate alignment, even when taking into account overlapping speech and background noise that distorts the recording. Additionally, Evanini et al. (2009) found that a proto version of FAVE-extract outperformed a different extractor as well as human measurements for three vowel phonemes (Severance et al., 2015). The FAVE suite as a whole, then, currently seems to be the best choice when it comes to force-alignment and automated vowel extraction and has been used in a number of recent studies, e.g. Labov, Rosenfelder and Fruehwald (2013) and MacKenzie and Turton (2013). Consequently, FAVE served as an adequate alternative to hand-measuring vowel formants for this project.

2.5.2 Data Cleaning

The data cleaning process involved the removal of all tokens in a participant's speech production that were deemed not representative of the speaker and therefore should not be included in further analysis, as well as fixing discrepancies in the phonemic transcriptions produced by FAVE. The steps that were taken to achieve this will be outlined below.

The first set of uncharacteristic tokens that I removed from the data included unfinished words, one and two letter words, and words that were produced with atypical features, such as mispronunciations that were immediately corrected or imitations of other people. Such tokens were manually removed from the output files that FAVE-extract produced. Since mispronunciations and corrections were marked separately in the transcription, those tokens were easy to identify in the FAVE output file.

In a second step, I manually checked the data for apparent outliers. "Typically, any data points that fall two or more standard deviations from the mean are considered outliers" (Clopper, 2011, p. 194). However, I did not determine outliers based on standard deviations, but rather relied on visual cues on scatter plots using Plotnik (Labov, 2017). Tokens that visually deviated significantly from the mean were remeasured manually using Praat. If necessary, formant measurements were corrected in the data set or, if formants could not be measured reliably, the token was removed from the data set. Tokens that proved to be true outliers, i.e. tokens with exceptional but reliably measurable formant frequencies, were retained in the data set. A post-hoc inspection of the data suggested that the data set still contained a number of outliers, however, they were so small in number that they are unlikely to have had a drastic influence on the results presented in this study.

The last step of cleaning speech data concerned discrepancies in the phonemic transcriptions produced by FAVE. As mentioned above, FAVE-align relies on an external pronunciation dictionary (CMU) where a lexical item with more than one possible realization can have various entries. Thus, FAVE-extract has to select the pronunciation that best fits the speech signal (George Bailey, 2016), which can lead to either incorrect transcriptions or deviating transcriptions of one and the same lexical items within the data set. Incorrect transcriptions were, for example, found for the word *kept*, which was consistently transcribed with TRAP rather than DRESS, which was corrected manually. Deviating transcriptions were, for example, spotted for variations of the word *dog*, i.e. *dog*, *dogs* and *doggy*, some of which were transcribed with LOT and others with THOUGHT. I converted those deviating transcriptions to the THOUGHT class, which, as shown in Figure 13, is a more characteristic realization of *dog* in the community in question.

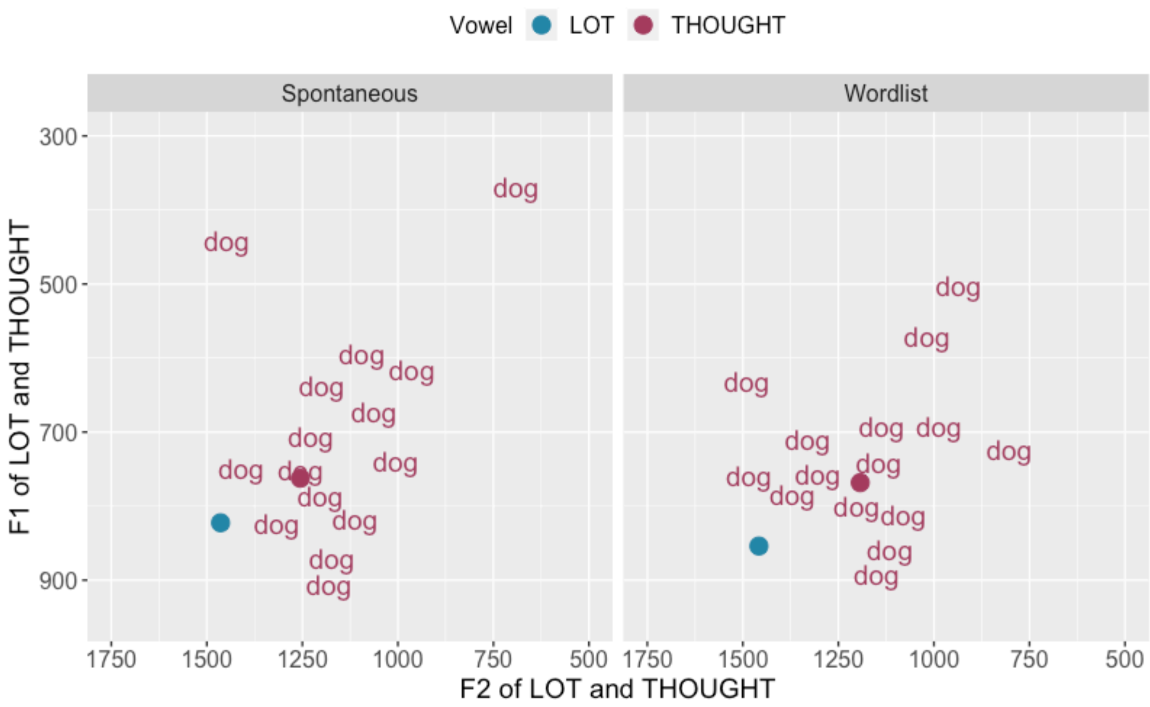


Figure 13: Distribution of *dog* tokens in relation to the community means of LOT and THOUGHT.

However, as Figure 13 above shows, there are a few speakers for whom *dog* appears to be closer to the community mean of LOT than to THOUGHT, but for whom variations of the word *dog* were nevertheless coded as THOUGHT. While in spontaneous speech, these individual tokens are unlikely to have influenced the results in the analysis of LOT, THOUGHT or the merger of these two vowels, this is more likely to be the case in wordlist style, as there are, overall, fewer LOT and THOUGHT tokens in this style. However, only for four of these speakers does *dog* appear to be realized with a vowel that is somewhat more characteristic of the speaker's LOT in this speech style: Allison (born in 1993), Ashley (born in 1966), Jason (born in 1998) and Summer (born in 1987).[55] The youngest two of them, Allison and Jason, are among the eight most merged speakers in the sample, which may be an overestimation based on this coding of *dog*. However, both of them are among the most merged speakers in the two other speech styles as well, so that, overall, coding *dog* into the THOUGHT class is unlikely to have influenced the results to a significant extent.

I opted to not exclude formant measurements from tokens that appeared early on in the interview. Although doing so is often recommended based on the argument that, once a certain amount of time has passed during the interview, participants will be more relaxed and thus more likely to use their most casual speech style (e.g. Chambers, 2008), L. Milroy and Gordon (2003, p. 58) point out that time passed in the interview is not necessarily a determining factor in the amount of style shifting:

> Recent research on style-shifting … suggests that interviewees may move in and out of styles throughout the course of an interview for a variety of reasons … Thus, researchers should be careful in assuming that speakers will adopt or maintain a particular style simply based on the fact that some period of time has elapsed in an interview.

It should be noted that some of these cleaning measures were also applied to the 2008 data set where relevant. Thus, I removed one and two letter words, as well as the stop words that FAVE was instructed to exclude from the analysis from the 2008 data. I also recoded tokens for which FAVE produced varying phonemic transcriptions as described above in this data. This was done in order to ensure comparability between the 2008 and 2016 data sets. However, this also means that results for the 2008 data might deviate

[55] The plots to illustrate this can be found in Appendix G.

slightly from those presented in Dinkin's (2009, 2013) study, as he did not follow the same steps.

Data cleaning can also be necessary with regards to perception data. For data collected in matched guise tasks, Kircher (2015) suggests that participants who figured out that they heard the same voices multiple times should be removed from the analysis, because "ignorance with regard to the methodology is a precondition to the elicitation of valid results" (p. 202). As mentioned earlier, some of my participants did comment on the repetitiveness of the guises. However, I opted not to exclude those participants from the MGT. While Kircher's point is certainly valid, it is likely that other participants noticed that they heard the same speakers in multiple guises without commenting on it, so that removing participants who did mention it would be redundant and potentially a waste of data.

2.5.3 Renormalization and Calculating Means

FAVE outputs two data sets for each speaker: one with raw formant measurements and one with (Lobanov) normalized values. However, as explained above, some tokens were remeasured or removed entirely from the output files, thus changing the set-up of the tokens on which the normalized formants are based. Thus, I renormalized the cleaned, raw data for each speaker using the same method employed by FAVE.

Based on these new values, I calculated the means for F1, F2, and duration, as well as standard deviations for F1 and F2 for all vowel classes following *ANAE* methodology, which excludes the following environments in the calculation (Labov, 2017):

- Vowels preceded by glides /j/ and /w/
- Vowels preceded by obstruent-liquid clusters (e.g. /sl, bl, gl, br, gr, dr/)
- Vowels followed by /l/
- Vowels followed by /r/
- TRAP and DRESS followed by a nasal (in addition to the above mentioned)

2.5.4 Reliability Measures

Despite the benefits of automated processes in the analysis of speech, some researchers (e.g. *ANAE*) have raised concerns about the reliability of computational approaches to acoustic analysis. Evanini et al. (2009, p. 89) argue that "the most important test for the applicability of any automatic formant prediction method from a sociolinguist's perspective is whether the predicted values … demonstrate the same type of variation as the means from the hand measurements". Thus, ensuring that the measurements obtained through FAVE are in fact reliable and comparable to those of a human analyst has been a crucial point of focus not only in this study, but in sociolinguistic circles in general. Several researchers have experimented with the reliability of FAVE, e.g. George Bailey (2016), Driscoll and Lape (2015), Evanini et al. (2009), Labov, Rosenfelder and Fruehwald (2013), and Severance et al. (2015). All of these studies have found FAVE to perform accurately and reliably. In their 2015 project, Severance et al. compared FAVE formant measurements to those of an expert measurer in order to determine the error rate, i.e. the difference between the two sets of measurements, and to identify potential patterns in the kinds of errors. The data they used for comparison came from Dinkin's (2009) study on dialect boundaries in Upstate New York. They limited the data to a subset of speakers from Utica, which is a community that participated in the NCS in Dinkin's study. This means that the reliability of FAVE has been specifically tested on NCS data, which is the kind of data this present study was presumed to be dealing with as well. Severance et al. based their comparison on seven interviews with both male and female speakers that lasted on average 13:40 minutes and contained mostly spontaneous speech as well as wordlist tokens. Dinkin (2009) measured most of his vowels manually using the maximum F1 method. FAVE, on the other hand, by default measures vowels at ⅓ of their duration (among others). Severance et al. used paired t-tests and correlations for comparison of the measurement points, average F1 and average F2. The tests showed that, for the majority of data points analyzed, there are no significant differences when comparing FAVE's to hand-measured formants (both F1 and F2). Out of 188 vowel dimensions, 27 turned out to differ significantly between analysts, mostly in terms of F1 measurements for back vowels. Because this comparison was based on data from "actual sociolinguistic interviews with typical complications that could introduce errors" (Severance et al., p. 33), the authors suggest that their study can be generalized to other data that was collected using this interview technique, as I did for the present project. Nonetheless, I conducted a similar analysis to supplement previous studies and to ensure reliability of FAVE's formant measurements in my data.

I conducted this comparative analysis on a small subset of my data. For the assessment, I selected six of my 39 participants (15%), chosen based on their NCS scores (see Chapter 2.6.2): two speakers with a relatively high NCS score, two with an average NCS score, and two with a score of zero. Each of these pairs consists of one male and one female speaker and the subsample includes both younger and older speakers. In selecting these speakers I followed Clopper (2011), who suggests that the data for reliability testing should consist of roughly 10% of the data and "should be randomly, but evenly, sampled across experimental conditions" (pp. 195-196). Using Praat, I hand-measured the first 200 tokens of all target vowels from the speakers' spontaneous speech production as well as all of their wordlist and minimal pair tokens, all at ⅓ of each vowel's duration. Tokens for which no reliable measurements could be taken were omitted from the sample. I generally kept the formant settings in Praat at default (tracking five formants at a time) but adjusted the tracker up or down depending on which setting seemed to match the spectrogram most accurately. In a few cases, I took F1 and F2 measurements from different tracker settings. This was the case when F1 or F2 did not seem to be tracked accurately using only one setting. These measurements were then compared to the raw, unnormalized data extracted by FAVE based on the absolute value of the mean differences (Clopper 2011).

The mean absolute difference is a reliability measure that describes the value of the absolute mean of the second measurement subtracted from the first measurement. In the case of this study, this refers to the difference between FAVE's formant measurements and hand-measured formants. In this test, the lower the difference, i.e. the closer to zero, the more similar the two sets of measurements are to each other. A difference from zero, on the other hand, "suggests large and/or frequent smaller differences between the two sets of measurements" (Clopper, 2011, p. 192).

According to the results of this test, FAVE seemed to have performed fairly well for this study in all three speech styles: spontaneous speech, wordlist style and minimal pairs. In spontaneous speech, the overall mean F1 across the subsample used for this comparison is 630 Hz across all vowels as measured by FAVE. On average, the hand measured formants deviate from this mean by 22 Hz (mean absolute difference), i.e. the mean of the differences between hand-measurements and FAVE measurements is 22 Hz. For F2 in spontaneous speech, the mean across vowels in the subsample according to FAVE's measurements is 1544 Hz; the mean absolute difference from hand measured formants to FAVE's data is 41 Hz. In wordlist style, hand measured formants deviate slightly more from FAVE's measurements. The mean based on FAVE's F1 measurements is 710 Hz across all vowels. The mean absolute difference to the hand measured F1 data is 37 Hz. For F2, the mean in the subsample according to FAVE is 1463 Hz in wordlist style. The mean absolute difference from hand measured formants to FAVE for F2 is 61 Hz. For minimal pairs, the mean F1 of FAVE's measurements is 765 Hz across all vowels. Compared to the hand measured formants, the mean absolute difference is 34 Hz. For F2, this difference is 28 Hz, with an overall mean F2 of 1211 Hz.

With the exception of F2 in wordlist style, the differences between FAVE's measurements and hand-measured values appear to be relatively low. However, this relativity brings up the question of how much of a difference between the data sets is acceptable. Severance et al. (2015, p. 31) point out that "there is unfortunately no consensus yet on what an acceptable absolute mean difference looks like". Clopper (2011, p. 192) corroborates this by saying that "the acceptable mean difference between measurements will vary depending on the materials and the phenomena being measured". However, she also makes the point that "mean absolute differences within approximately 1-2% of the mean over the whole data set are typically deemed reliable". The 2% marks for my data are provided in Table 7.

	F1	2%	F2	2%
Spontaneous				
Overall mean (Hz)	630	13	1544	31
Median absolute	9		16	
Mean absolute	22		41	
Wordlist				
Overall mean (Hz)	710	14	1463	29
Median absolute	17		29	
Mean absolute	37		61	
Minimal pairs				
Overall mean (Hz)	765	15	1211	24
Median absolute	15		10	
Mean absolute	34		28	

Table 7: Mean absolute differences between formant measurements: FAVE vs. hand-measured. Cells highlighted in blue indicate that the difference falls within the 2% mark of reliability. Those highlighted in red indicate that the difference falls outside of this threshold.

As Table 7 above shows, the absolute mean differences exceed this 2% mark in all styles in both F1 and F2, indicated in red. However, in most cases this seems to be due to outliers, since the absolute *median* difference falls within this 2% mark in all but one (F1 wordlist) cases – indicated in blue. This suggests that the majority of the measurements are comparable across both sets, i.e. FAVE measurements are very similar to hand measured formants. As Evanini et al. (2009, p. 1657) argue, "even when a small number of automatic formant measurements are gross errors, the sociolinguistic analysis can still be conducted successfully if they are not systematically biased in any direction". Since this does not seem to be the case with my data, it appears that the formant measurements that were extracted by FAVE can in fact be considered to be reliable.

2.6 Analytical Tools

The last step of preparing for analysis involved choosing appropriate analytical tools. This chapter will go over which tools were chosen for which purpose, and detail potential modifications to those tools, which where necessary in order to make them compatible with the data in this study, and the results comparable to those of previous research. I will start out by detailing the statistical packages that I used for data analysis, and then move on to a discussion of the tools used for the analysis of speaker participation in the NCS and the COT-CAUGHT merger.

2.6.1 Statistical Modeling

As mentioned in Chapter 1.7, the aim of this study is to investigate patterns of language use in Ogdensburg, particularly in relation to the NCS, COT-CAUGHT merger and the Elsewhere Shift along with its associated changes in the back vowels. Identifying large-scale trends like these is generally easier to accomplish using quantitative, i.e. statistical, methods, which are "interested in the patterns or trends formed by groups" (Johnson, 2014, p. 288). For this study, I made use of a combination of descriptive and inferential statistics.

Descriptive statistics, i.e. the visual representation of the data in graphs, formed the basis of all analyses. These kinds of visualizations allow for the exploration of potential patterns in the data, such as social effects and interactions between social factors, i.e. the dependency of the effect of one variable on a second variable (Hay, 2011).

To test for the statistical significance of these patterns, i.e. to test if the patterns in the sample can be inferred to be representative of real patterns in the population (i.e.

Ogdensburg) or if they are merely due to chance (Tagliamonte, 2007, p. 202), I used mixed-effects regression models, a method for doing multivariate analyses toward which most variationists have been gravitating in recent years. As Baayen (2014) explains, most data are multivariate, i.e. multiple factors affect the variable under investigation, and for “a proper understanding of the structure of the data, it is often most informative to consider the different variables simultaneously” (p. 338). This can be achieved using mixed-effects regressions, as they allow the researcher to treat “observations made on many variables simultaneously” (p. 337).

Mixed-effects models “combine fixed-effect factors with random-effect factors … [and] treat random-effect factors as sources of random variation in the data” (Baayen, 2014, p. 350). The fixed effects involved in the analyses in this study (summarized in Table 8 below) include *speaker age* (as a continuous variable), *gender* (male, female), *education* (college, no college, student), *style* (spontaneous, wordlist, minimal pairs), *sample year* (2016, 2008) and *phonological environment* as coded by FAVE (following manner, following place, following voice).[56] For LOT and THOUGHT, *vowel duration* was considered as well, though in a separate analysis.[57] Which of these factors, and which potential interactions between them, were included in the regression model for any given variable depended on which model best matched the visual representation of the data. Random effects included in the analyses were *speaker* and/or *word,* depending on which was applicable. The reason for including *speaker* as a random effect, Hay (2011, pp. 212-213) explains, is that doing so ensures that “no individual participant can dominate the significance of any reported effect. It is then possible to test whether there are overall effects which exist over and above the variation across individual speakers”. The same can be said for the second random factor: *word*. Thus, by taking random factors into account in the analysis, the researcher “can be reasonably confident that any remaining fixed effects are genuine, which increases the likelihood that the results of [the] statistical model can be generalized to the sampling population” (Walker, 2014, pp. 453–454). A series of ANOVAs (analysis of variance) was carried out to determine the statistical significance of the fixed effects predicted by the regression models. The significance

[56] FAVE codes phonological environments as five separate variables. For the purpose of statistical analysis, three of them (following manner, place and voice) were recoded as one factor, while two of them (following segment and preceding segment) were excluded from the analyses owing to rank deficiencies.
[57] *Vowel duration* is not factored into the analysis when Dinkin's (2009) data is included, because vowel duration was not measured for his data.

threshold for the analyses was set at $p \leq 0.05$. Both the mixed effects regression and ANOVAs were implemented using the R statistics package (R Core Team, 2016) for statistical analysis. Contrary to the mean-based analyses of vowel production in the graphs, statistical results are based on token-level regressions. While means were calculated by excluding certain phonological environments (see Chapter 2.5.3), the regression models include all phonological environments[58] using treatment coding, so that potential effects of these environments on the quality of the target vowel are taken into consideration in the statistical results. Their effects will not be presented in the analysis but are provided in Appendix H.

Fixed effects	Random effects
○ *Age* (continuous) ○ *Gender* (male, female) ○ *Education* (college, no college, student), ○ *Speech styles* (spontaneous, wordlist, minimal pairs for LOT and THOUGHT)[59] ○ *Sample year* (2008 and 2016) ○ *Phonological environment* ○ *(Duration)*	○ *Speaker* ○ *Word*

Table 8: Predictors in the statistical analyses of vowel production.

In the evaluation of the matched guise data, the variables that are factored into the analyses for NCS guises include *age*, *gender*, *education*, *mean F1* and *F2* of the target vowel in spontaneous speech, the amount of *style shifting* from spontaneous speech to wordlist style, *NCS scores* in spontaneous speech, the guise specific factors *guise* and *voice,* as well as *listener/rater* as a random effect. For the merger guises, the included factors are *age*, *gender*, *education*, *distance* between and *overlap* of LOT and THOUGHT in minimal pairs, and the guise specific factors *guise* and *voice,* as well as *listener/rater* as a random effect. Since matched guise data is available only for speakers from the 2016 sample, *sample year* is not considered in the analysis. Again, the starting point for analysis was data visualization.

[58] The only exception to this is TRAP, where pre-nasal tokens are analyzed separately as TRAMP.
[59] The 2008 data includes minimal pairs for phonemes other than LOT and THOUGHT, as well as a fourth speech style ("elicited speech"). Those additional tokens were excluded from the analysis to ensure comparability across the two data sets.

In the analysis chapters (3-7), not all patterns that I describe as emerging from the graphs reach the level of statistical significance (as suggested by their respective *p*-values). However, as the analyses will show, the patterns are highly repetitive for nearly all variables. Thus, taking into consideration evidence from both descriptive and inferential statistics and the nature of the sample, I decided to rely more heavily on the graphs than on *p*-values in some cases (Hay, 2011). This will be discussed further in Chapter 8.2.2.

2.6.2 Adjusted NCS Criteria

There are several ways to determine whether a vowel has shifted. The Euclidean distance can be a useful tool for this and is sometimes used to determine shifts: "In the study of a vowel shift, the Euclidean distance is taken from a stable anchor point vowel to the vowel that appears to be shifting" (Di Paolo et al., 2011, p. 104). This method has been used, for example, by Hay, Warren, et al. (2006) in their study on NEAR and SQUARE in New Zealand. For a chain shift, however, determining a stable anchor point can be difficult, as chain shifts by definition involve the shifting of multiple vowels. Thus, using the Euclidean distance is not particularly useful when analyzing variables (potentially) involved in the NCS. Labov (2007, p. 41) argues that "since the NCS is a complex rotation of its elements, the measurement of any one vowel tells us little about the progress of the shift". Instead, he developed a set of criteria for the study of the NCS that "relies on structural relations among NCS vowels" in order to map this progress (p. 41). These criteria were used for the study of the NCS by Dinkin (2009, 2012, 2013) and have been adopted for this project in a modified version as well. The five NCS criteria in their original form are defined by Labov (2007) as follows:

- AE1: F1(TRAP) < 700 Hz
- EQ: F1 (TRAP) < F1(DRESS) and F2(TRAP) > F2(DRESS)
- ED: F2(DRESS) – F2 (LOT) < 375 Hz
- O2: F2(LOT) > 1500 Hz
- UD: F2(LOT) > F2(STRUT)

The criteria describe either absolute (AE1 and O2) or relative (EQ, ED and UD) vowel qualities in terms of F1 and F2. AE1 defines shifted TRAP as having an F1 of less than 700

Hz. The EQ criterion, on the other hand, describes shifted TRAP as being higher (i.e. having a lower F1) as well as fronter (i.e. having a greater F2) than DRESS. Thus, this criterion "combines the raising and fronting of /æ/ with the backing and lowering of /e/" (*ANAE, p. 198*) and describes the reversal of the positions of both vowels in an NCS system. The AE1 and EQ criteria are both valuable for measuring participation in NCS TRAP. While AE1 measures TRAP directly, its value depends on the normalization methodology chosen and on the somewhat arbitrary value of 700 Hz. EQ is based on relative, thus more objective criteria, but cannot necessarily distinguish between the raising and fronting of TRAP and the lowering and backing of DRESS. DRESS is further described by the ED criterion, which puts the positions of DRESS and LOT into relation, in that the difference between their mean F2 values should be less than 375 Hz. Shifted (or fronted) LOT is further defined by the O2 criterion as being fronter than 1500 Hz, which puts it close to the normalized center at 1550 Hz (Labov, 2007). Another measure for NCS LOT is that it is fronter than STRUT, as defined by the UD criterion. Simultaneously, this defines shifted STRUT as being backer than LOT. The sum of fulfilled criteria is what I refer to as a participant's NCS score, which can range from zero (none of the criteria are met) to five (all criteria are met). Note that there are no criteria that formally define shifted THOUGHT or KIT, which were not treated as integral parts of the NCS by Labov (2007).

These criteria were adopted for the present study in a modified version. The modifications concerned the values that define the criteria. The thresholds that Labov (2007) used to define his criteria are based on formant values that were generated by the *ANAE* normalization method which is based on Nearey (1977). As explained in Chapter 2.5, however, the data in this present study are Lobanov normalized. Although the *ANAE* and Lobanov normalization methods are similar in that they are both vowel-extrinsic and speaker-intrinsic[60], they rely on different formulae for their calculations, so that the returned normalized vowel formants are not directly comparable. Consequently, Dinkin (p.c.) found that when trying to determine whether a speaker was shifted in his data, his *ANAE* normalized data returned a different number of shifted speakers per criterion than the same data when it was normalized to Lobanov. Since he used the exact same data for

[60] In its original form, the *ANAE* normalization was speaker-extrinsic, since it defined the overall group mean that every speaker was normalized to match in terms of the geometric mean of all speakers in the data. However, the overall geometric mean is no longer being updated, so that later speakers are normalized to match the existing group mean (i.e. speaker-intrinsic) (Dinkin, p.c.).

this comparison, the difference can only be due to the mismatch between the normalization method used on the data, and the method on which the original NCS criteria are based. To counteract this disparity, Dinkin adjusted the thresholds for Labov's NCS criteria, choosing numbers "that seem to make sense for defining the same NCS criteria - in that at least they pick out the same *number* of speakers under the new normalization, if not the same exact speakers" (Dinkin, p.c.). The revised criteria that I used to analyze participation in the NCS in this study thus look as follows:[61]

- AE1 criterion remains: F1(TRAP) < 700 Hz
- EQ criterion remains: F1 (TRAP) < F1(DRESS) and F2(TRAP) > F2(DRESS)
- ED criterion: F2(DRESS) - F2(LOT) < 391 Hz
- O2 criterion: F2(LOT) > 1668 Hz
- UD criterion remains: F2(LOT) > F2(STRUT)

2.6.3 Measuring Degree of Merger

Because determining if and to what degree two vowel classes have merged has been a significant problem in the field of sociophonetics (Di Paolo et al., 2011, p. 102), I made use of four different measures in this study. Two of these relate to speech production: the (adjusted) Euclidean distance and the Bhattacharyya score. The other two tools were used to measure the degree of merger in perception, and included commutation test scores and self-judgment. Each of these methods, as well as their implementation in the study, will be outlined in the following paragraphs.

The Euclidean (or Cartesian) distance calculates the distance between the means of two phonemes and reveals how far apart they are in a speaker's vowel space. The smaller the distance, the more merged a speaker is inferred to be in their production of the two phonemes, with a distance close to zero indicating a full merger. If a trend of decreasing distance can be observed in apparent and/or real time, it can be taken as an indication of an advancing merger in the community. As mentioned earlier, there is no consensus on *how* close in production two vowels have to be in order to be considered merged. However, there does appear to be some agreement that 200 Hz on the front-back

[61] Note that these cut-offs differ slightly from what Dinkin (2018) suggests, where AE1 is set to <702 Hz, ED to < 402 Hz and O2 to 1612 Hz.

dimension are the minimum distance that is necessary to maintain a stable distinction (*LYS*; Labov, 1994). The Euclidean distance can either be calculated based on the normalized mean values of the two phonemes in question (as was done for some of the NCS criteria described above), or it can be estimated using a mixed effects regression model, referred to as 'adjusted Euclidean distance' (Nycz & Hall-Lew, 2014). The second approach allows the researcher to take into consideration the phonological environment in order to reduce skewing effects of allophony, as well as the random factor *word*, and provides a significance value for the estimated difference between the two vowels. Thus, the adjusted Euclidean distance is a more reliable way of using the Euclidean distance measure in determining the degree of merger, which is why, in this study, I made use of a mixed effect regression model in the calculation of the Euclidean distance in order to estimate the degree of merger of LOT and THOUGHT. The calculations are based on two models per speech style: one model for F1 and one model for F2. Each of these models contained fixed effects for a) the phonological environment (i.e. following manner, following place, following voice, preceding segment, following segment) and b) the vowel class (i.e. LOT and THOUGHT). The random factor *word* was not included in these models. For the wordlist and minimal pair tokens, it simply was not necessary to do so, as *word* was controlled for in the designing of these reading tasks (see Chapter 2.3.1). In spontaneous speech, the random factor *word* was not included because doing so likely would have resulted in an underestimation of the actual distance between the two phonemes (Dinkin, p.c.).

The second tool employed in this study to investigate the development of the COT-CAUGHT merger in production is the Bhattacharyya score. This score is a way to calculate the "overlap fraction that quantifies the extent of overlap between vowel[s] (for two-dimensional comparison)" (Di Paolo et al., 2011, p. 103), taking F1 and F2 into consideration simultaneously. For this study, the amount of overlap of LOT and THOUGHT tokens was calculated including all tokens in those vowel classes, with the exceptions of tokens in which the vowel is followed by /r/ or /lC/. Unlike the Euclidean distance, the Bhattacharyya score does not rely on definite or estimated means, but rather takes into consideration the distribution of all tokens of the two vowel classes. The more the two vowels overlap in phonetic space, the more merged a speaker is inferred to be in speech production. However, there is no consensus on how much overlap amounts to a partial or full merger. Beckford Wassink (2006) has provisionally defined the following cut-offs:

- No merger: 0-20%
- Partial merger: 20-40%
- Complete merger: >40%

Boberg (2001), on the other hand, took 50% overlap as an indication of a partial merger. Because there is no agreement on what counts as distinct, transitional or merged, neither for Euclidean distances nor overlap, I will refrain from labelling speakers with one or the other category and discuss their degree of merger in relative terms instead.

One way of combining both the Euclidean distance measure and the measure of overlap is the Pillai score. Similar to the two former techniques, the Pillai score reflects the degree to which speakers keep two vowel classes distinct. However, the Pillai score takes distance and overlap into consideration simultaneously and outputs a single number that can be used for statistical analysis. Therefore, Hay, Warren, et al. (2006, p. 467) argue that "as a summary of the degree to which two distributions are kept distinct, this is superior to taking Euclidean distances between means". However, the Pillai score measure was not employed in this study. While a single number does make statistical analysis easier, using two separate measures might be able to provide more insight into the mechanisms behind an advancing merger. For example, a merger by approximation would be reflected in decreasing Euclidean distances in apparent and real time. Increasing overlap, on the other hand, might be an indication of a merger by expansion (and/or approximation). A single score that combines both measures might obscure those processes, which is why I made use of the Euclidean distance measure and the Bhattacharyya score separately.

However, despite choosing two different methods of determining degree of merger, it was not possible to capture all potential differences in the production of LOT and THOUGHT. As discussed in Chapter 1.2.2, increasingly similar formant frequencies between the two vowel classes might be counteracted by other contrastive cues, e.g. a back upglide on THOUGHT, lip rounding, phonation, and duration. While duration is taken into consideration in the analysis, voice quality, formant trajectories and the degree of rounding are not included as potential factors in the analysis. However, for the purpose of this study, I regard a combination of adjusted Euclidean distance, Bhattacharyya scores and vowel length to present an adequate picture of the participants' degrees of merger in production. To complement these measures, I integrated two techniques to test the progress of the merger in perception.

The first measure I used to assess the progress of the merger in perception are the scores retrieved from the commutation test as explained in Chapter 2.3.2.2. The participants' responses to the auditory stimuli was coded as either right or wrong, and the number of correct responses were listed as their score for the test. I coded the responses to the participants' own production separately from their responses to the pre-recorded, clearly distinct line-up, so that it is possible for a participant to have e.g. a low score for their own production, and a high score for the distinct production of the pre-recorded speaker. The higher the scores for this test, the more successful a participant was in identifying the two sounds correctly, and the more distinct in perception the participant is inferred to be.

Another perception related technique, and the last one used in the study of the COT-CAUGHT merger, was self-judgment. Self-judgment here was the participants' response to the question of whether they thought the word pairs they read in the minimal pair reading sounded the same or different (see Chapter 2.4.4). Their responses were coded into three categories: "distinct", "transitional", and "merged". Participants who reported that none of the pairs sounded the same were coded as "distinct", whereas those who responded that all pairs sounded the same were coded as "merged". The "transitional" category contains those participants who were either unsure whether the pairs sounded the same or not, or thought that some of the pairs sounded more similar than others.

Chapter 3: The Rise and Fall of TRAP

3.1 TRAP – An Introduction

Variation in the realization of TRAP across varieties of English is not uncommon, and is likely the result of an unstable history of this vowel class. According to Dobson (1957, as referenced by Johnson, 2007), TRAP developed out of Middle English [a], which, in the 16th and 17th century, underwent a split into [æ] and [a:] through fronting and lengthening, respectively, as illustrated in Figure 14. Lengthening to [a:], which formed the START and BATH classes, occurred in limited phonological environments depending on the region: in positions preceding /r, f, s, θ/, e.g. *far, staff, pass, bath* in Southern British English and New England (though with exceptions such as in *gas* and *maths*) as well as in positions preceding /ns, ntʃ, nt, nd/ in a number of words, e.g. *dance* and *France* in RP. In US dialects outside of New England, lengthening to [a:] only occurred in positions preceding /r/. In most dialects, a number of words with different environments underwent this process as well, e.g. *father*. The majority of Middle English [a] words, however, came to be pronounced with [æ] (later [a] in most of England).

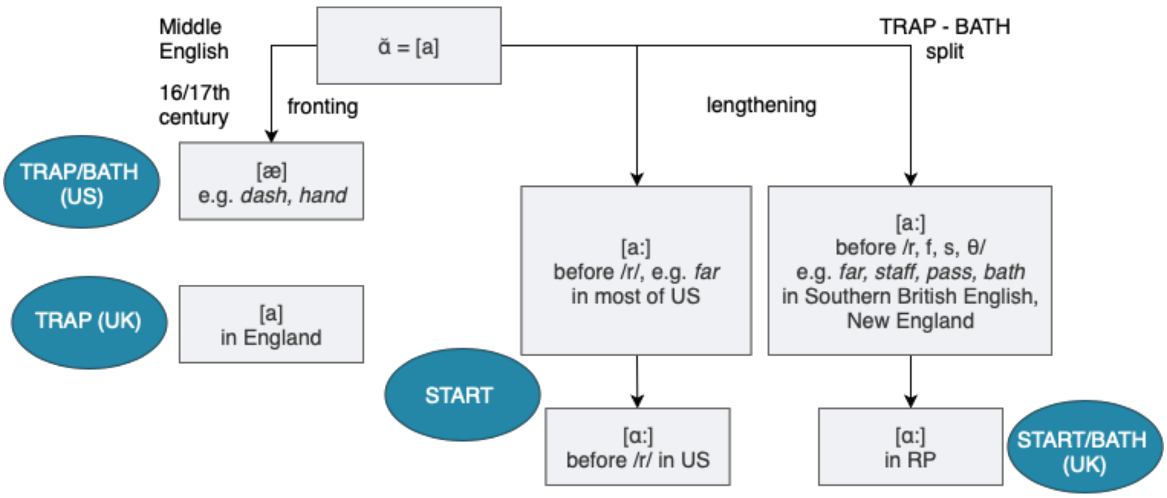


Figure 14: Historic development of TRAP.

Since this split of Middle English [a], [æ], i.e. TRAP, has undergone notable changes in US English, the direction of which has, until recently, predominantly been upward (*ANAE* p. 173; Callary, 1975), though frequently accompanied by fronting. Eastern New England, and cities like Pittsburgh, Columbus and Indianapolis in the Midland, for example, are characterized by a nasal TRAP system, in which pre-nasal TRAP is raised to a position distinct from low, pre-oral TRAP. This also applies to the split TRAP system found in the Mid-Atlantic States and New York City; however, in this system, raising of TRAP also occurs when preceding voiced stops and voiceless fricatives, and, in addition to being phonologically conditioned, TRAP raising in this system is lexically and grammatically conditioned as well (*ANAE*). The continuous TRAP system, found primarily in the Western and Midland States, also tends to raise pre-nasal TRAP, though, as the name implies, the transition from pre-oral to pre-nasal is continuous and not as sharp as in the nasal and split systems (*ANAE*). In addition to these three systems, Dinkin (2009, 2011) added the raised nasal system, the raised continuous system and the diffused system. The raised nasal and raised continuous systems resemble the traditional, (low) nasal and continuous systems, with the exception that both allophones, i.e. pre-oral and pre-nasal TRAP, are raised to a higher position. The diffused system resembles the split-system of New York City, but is less conditioned.

The NCS is the only system where TRAP is raised (and fronted) categorically, regardless of phonological environment, in US dialects. Although raising is generally preceded by fronting (Ito, 2001), the focus regarding NCS TRAP has generally been put on the degree of raising, which appears to depend on the community. NCS TRAP has been described as raised to a position in the range of the traditional mid vowels /e/ and /ɛ/ or higher (e.g. Gordon, 2001; Herndobler, 1993), and in some cases as developing an inglide in addition to raising, e.g. [eə] or [eæ] (e.g. *ANAE*; Gordon, 2001; E. R. Thomas, 2001), or even [iə] (Callary, 1975). In extreme cases, this is described as "Northern breaking", when the glide constitutes a second steady state that is of equal length and can be heard as clearly as the first nucleus (*ANAE*). The degree of raising can also vary within a community, which was the case in rural Michigan, where the acoustic range of TRAP covers the entire front half of the vowel space for some speakers (Ito, 2001). Labov (2007) defined two criteria to describe NCS-raised TRAP on both dimensions, height and front-backness: The first criterion by which NCS raised TRAP is defined is the AE1 criterion, which considers only the height of the vowel. According to this measure, TRAP with a mean F1 less than 700 Hz would be regarded as NCS shifted. The second criterion is the EQ criterion, which relates TRAP to DRESS and defines NCS TRAP as being both fronter and higher than DRESS.

Raised TRAP is one of the earliest and most distinctive features of the NCS, as well as one of the most studied elements of the shift. It was first evidenced by Fasold in 1969 in Detroit and has since been studied individually or as part of the shift, and was found to be dominant for example in Detroit (e.g. *ANAE*; Eckert, e.g. 1988, 1989; Labov, 1994; *LYS*), Chicago (e.g. *ANAE*; Durian, 2014; Herndobler, 1993; *LYS*; McCarthy, 2007, 2010), Lansing (e.g. Mason, 2018; Nesbitt, 2018; Wagner et al., 2016), and Buffalo and Rochester (e.g. *LYS*; King, 2017). In a larger survey of the Inland North, *ANAE* and Clopper et al. (2005) found a high degree of homogeneity in the raising of TRAP, although *ANAE* reported a particularly high degree of TRAP raising in Western New York. NCS-raised TRAP has also been reported in smaller communities. Examples are communities in Northern Illinois (Callary, 1975), Southern Illinois (Bigham, 2007), Michigan (Gordon, 2001; Ito, 2001; Ito & Preston, 1998) and New York (Dinkin, 2009, 2013). In Northwestern Wisconsin, Benson et al. (2011) and Fox (2014) reported that TRAP raising seemed to be a development that is independent of the NCS, as it was phonologically conditioned in a way that is atypical for the NCS, and that unconditioned NCS-raised TRAP seems to be absent from the community.

Most of these studies on TRAP have come to similar conclusions regarding the social distribution of TRAP in the Inland North. In none of them was TRAP raising reported as advancing in apparent time. Instead, *ANAE*, Gordon (2001) and Herndobler (1993) found older speakers to be leading over younger speakers in the frequency and degree of TRAP raising, though only *ANAE* interpreted this as TRAP raising being recessive. Despite the absence of apparent-time advancement, most studies found a stable female lead in the raising of TRAP (*ANAE*; Eckert, 1998; Fasold, 1969; Gordon, 2001; Herndobler, 1993; Ito, 2001; McCarthy, 2007). Regarding social class, findings are somewhat contradictory: Fasold (1969) found TRAP raising to be most advanced in the upper working and lower middle class in Detroit, while Herndobler (1993) and Gordon (2001) found that middle class speakers were leading over lower middle-class speakers in Chicago and small-town Michigan. McCarthy (2007), on the other hand, found no significant effect of education on TRAP raising in Chicago, and neither did Ito (2001) or Eckert (on social class) (e.g. 1998) in Michigan. *ANAE*, on a larger scale, found no effect of education on TRAP raising for the Inland North. Some of these earlier studies also reported a positive social perception of raised TRAP (e.g. Callary, 1975; Herndobler, 1993), though this was never tested in an experimental setting.

The most recent studies on NCS-raised TRAP, as well as some of the earlier ones, are indicating that communities are receding from the general raising of TRAP as part of the NCS, most likely because of negative evaluation (e.g. Bigham, 2010; Savage et al., 2016; Wagner et al., 2016). In most communities that are retracting TRAP, speakers appear to be transitioning toward a low nasal or continuous TRAP system, with younger female speakers generally leading this change (Morgan et al., 2017; Wagner et al., 2016). However, lowering and backing of TRAP has also been observed in North American dialects of English other than the Inland North, e.g. in Pennsylvania, Central Ohio, Texas, California and Canada, often to a position lower than LOT (E. R. Thomas, 2001).

In New York State, the earliest signs of categorical TRAP raising were observed in the late 1800s in Ithaca (Emerson, 1891). Later, in 1935 to 1937, C. K. Thomas reported striking raising and tensing of TRAP in New York. In his data from over 200 speakers[62] from all Upstate New York counties, TRAP is described as being heard as [ɛ] and slightly higher before /n/. However, C. K. Thomas did note that this is not a general occurrence, and that TRAP is also commonly heard as a low front vowel. More recent studies on raised TRAP in New York found that, much like in the rest of the Inland North, NCS TRAP is disappearing in this part of the Inland North as well (e.g. Driscoll & Lape, 2015), and is replaced with an unraised continuous or nasal system (Dinkin, 2009, 2013; King, 2017), seemingly led by young women (King, 2017).

In Ogdensburg, Dinkin (2009) found that four of the nine speakers had NCS-raised TRAP: Three of these speakers were found to fulfill the AE1 criterion, raising TRAP to a mean F1 less than 700 Hz, as well as the EQ criterion, i.e. TRAP was higher and fronter than DRESS. The fourth speaker met only EQ, without raising TRAP above 700 Hz. Although these four speakers were the youngest speakers in the sample, TRAP raising itself did not correlate to the speakers' *age* at a significant level. The remaining six speakers in the sample did not meet the 700 Hz mark of the AE1 criterion; however, they did produce TRAP with relatively small F1 means between 705 and 766 Hz, i.e. relatively high in the vowel space. For most speakers in the sample, pre-nasal TRAP was raised notably higher than pre-oral TRAP, suggesting a raised nasal system or a raised continuous system. In which direction TRAP has developed since then will be analyzed in the following subchapter.

[62] The number of informants continued to increase during the course of the 3 years in which C. K. Thomas' articles were written.

3.2 Results: TRAP and TRAMP in Production

Unlike in the 2008 data, speakers in the 2016 sample fail to meet the two criteria for NCS TRAP. Only for a few speakers does pre-oral TRAP reach the 700 Hz benchmark of the AE1 criterion, and, at least in spontaneous speech, TRAP does not seem to be raised above DRESS (EQ criterion) for most speakers. The following subchapters will present an in-depth analysis of apparent and real-time developments in F1 and F2 of TRAP, including an examination of a potential allophonic split between TRAP and TRAMP.

Figure 15 suggests a good amount of inter and intra-speaker variation in the realization of TRAP in the 2016 sample. In spontaneous speech, TRAP appears to be produced with relatively similar qualities in terms of height, while the frontness of TRAP varies notably across speakers in this speech style. TRAP and TRAMP appear to be clearly separated in this speech style, with TRAMP being produced notable higher and fronter in the vowel space. In wordlist style, inter-speaker variation in height, in addition to the frontness of TRAP[63], emerges from the data, suggesting also some degree of style shifting, especially among younger speakers.

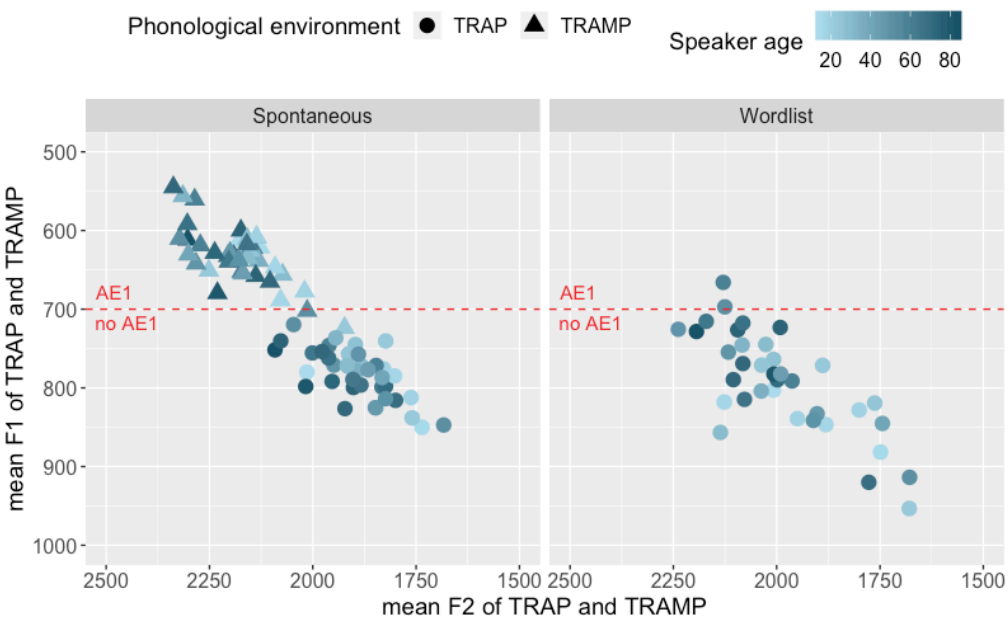


Figure 15: TRAP and TRAMP F1 and F2 means in 2016 across *speech styles* by *age*. Lighter shades represent younger speakers, darker shades older speakers. Note that there are no wordlist tokens for TRAMP in the data.

[63] Wordlist data is only available for TRAP, not for TRAMP.

3.2.1 TRAP and TRAMP in Spontaneous Speech

3.2.1.1 TRAP

Overall, spontaneous TRAP in the 2016 data seems to be marked by very little inter-speaker variation on the height dimension. The majority of speakers in this sample produce TRAP with an F1 between 750 and 800 Hz; consequently, none of the 2016 speakers meet the AE1 benchmark of 700 Hz in this speech style. However, it should be noted that TRAP is still produced relatively high in the vowel space by most speakers, as can be seen in Figure 16.

Figure 16 also suggests a difference in F1 between speakers of different educational backgrounds. Older speakers with a college education appear to have produced a somewhat higher TRAP than their peers without a college degree. For younger speakers, born after 1950, this pattern reverses owing to opposing apparent-time trends; college educated speakers seem to have consistently lowered TRAP, while those without a college degree continue to raise it. Consequently, college educated speakers now tend to produce a lower TRAP than those without a college degree.

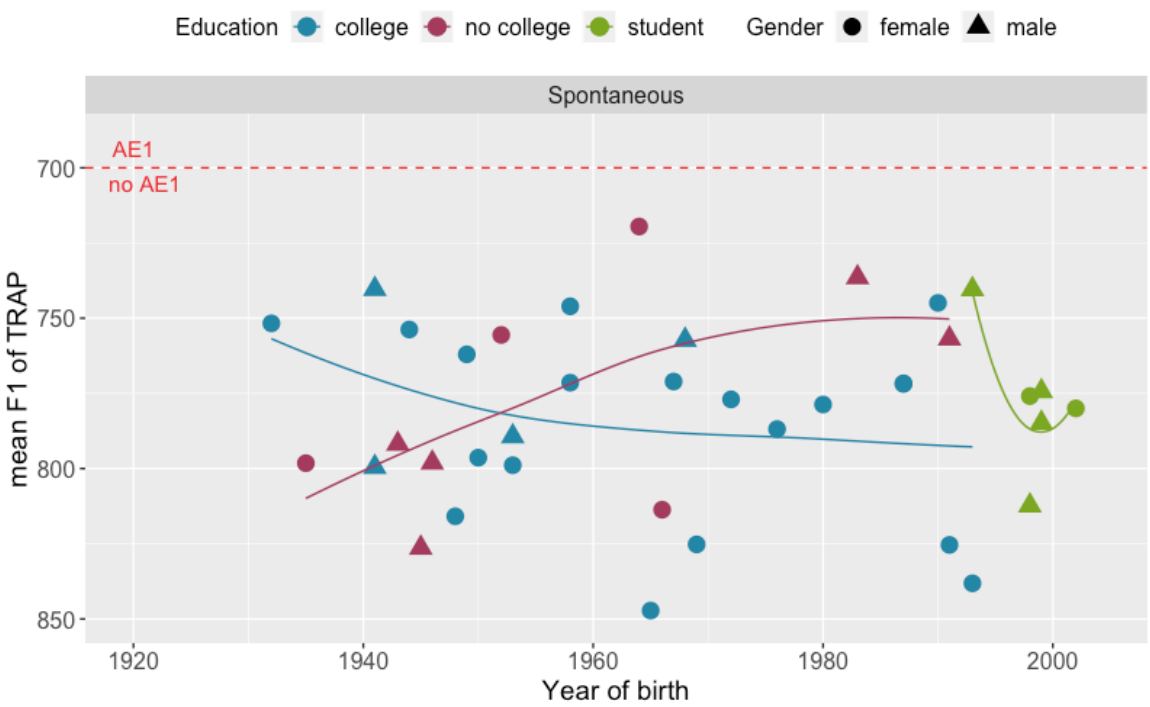


Figure 16: TRAP F1 means in spontaneous speech in 2016 by *education* and *gender*.

The regression model in Table 9 supports both of these observations. The main effect of *education* suggests that younger college educated speakers realize TRAP 89 Hz higher than their peers without a degree. The interaction between *age* and *education* in combination with the main effect of *age* indicates that, while TRAP is lowering among college educated speakers, it is raising at about double the pace for speakers without a college degree. However, this does not quite reach the level of statistical significance.

Predictor	Coefficient	*p*
(Intercept)	839.305 Hz	
Age	-0.442 Hz	
Gender (Male)	-1.337 Hz	0.929
Education (No college)	-89.263 Hz	
*Age*No college*	1.485 Hz	0.062
Environment		10^{-6}

Table 9: TRAP F1 in spontaneous speech in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word*. *n* = 3249

Variation in the frontness of TRAP appears to be less prominent than in F1 in the 2016 data. Figure 17 below shows that the majority of speakers produce TRAP relatively front in the vowel space, with F2 means between 1800 and 2000 Hz. For older speakers, TRAP appears to be somewhat fronter than for younger speakers, indicating an apparent-time retraction of TRAP. This trend appears to be more notable among speakers with a college education, however. For speakers without a college degree, on the other hand, no changes in F2 can be observed for TRAP. As a result, they produce TRAP somewhat fronter than speakers with a college education.

The regression model in Table 10 below supports these observations only partially. While it does predict TRAP to be backing in apparent time, this effect does not quite reach the level of statistical significance. Including students in the regression does not change that, though it does lower the *p* value to a nearly significant level (1.67 Hz, *p*= 0.056). The lack of a significant interaction between *age* and *education* suggests that apparent-time developments between speakers with and without a college degree do not differ greatly, and the main effect of *education* indicates that, overall, the two groups do not differ significantly in the frontness of TRAP.

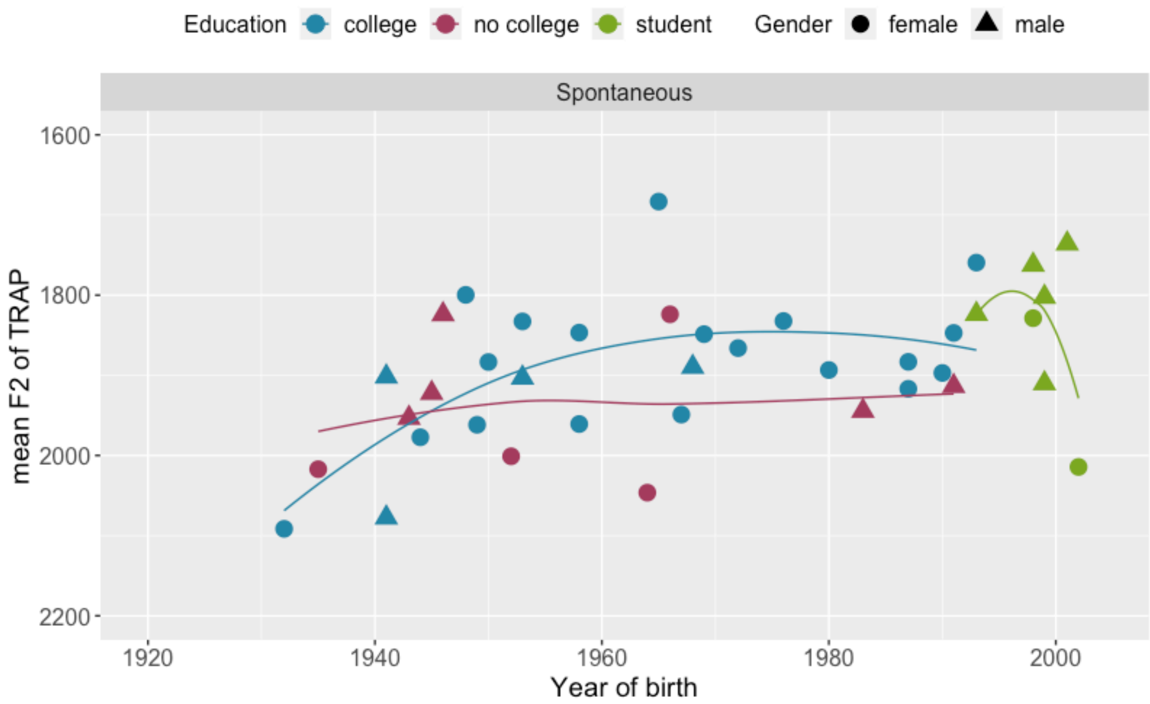


Figure 17: TRAP F2 means in spontaneous speech in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	1786.223 Hz	
Age	1.668 Hz	0.066
Gender (Male)	6.006 Hz	0.872
Education *(No college)*	38.519 Hz	0.301
Environment		0.016

Table 10: TRAP F2 in spontaneous speech in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word*. *n* = 3249

3.2.1.2 TRAMP

Similar to the observations for TRAP, the height of TRAMP in spontaneous speech seems to differ based on the speakers' educational backgrounds. As Figure 18 below indicates, the majority of speakers in the 2016 sample produce TRAMP with F1 means between 600 and 650 Hz. Older speakers with a college education fall in the lower end of this range, i.e. producing a higher TRAMP. Their peers without a college degree, on the other hand, have a slightly lower TRAMP. Again, the opposite pattern appears for younger speakers, i.e. those born after 1960, owing to contrary apparent-time trends. College educated speakers seem to have consistently lowered TRAMP, while for speakers without a college degree it has been raising. As a result, college educated speakers now tend to produce a lower TRAMP than speakers without a college degree.

Figure 18: TRAMP F1 means in spontaneous speech in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	672.616 Hz	
Age	-1.066 Hz	
Gender (Male)	6.869 Hz	0.652
Education (No college)	-146.755 Hz	
*Age*No college*	2.621 Hz	0.002
Environment		10^{-13}

Table 11: TRAMP F1 in spontaneous speech in 2016. Reference levels: females, college educated, /n/. Random effects: *speaker, word*. *n* = 1586

The regression model in Table 11 above supports these observations. Younger college educated speakers are estimated to realize TRAMP 147 Hz lower than their peers without a degree, as indicated by the main effect of *education*. The significant interaction between *age* and *education* supports the observation that the two educational groups are developing in opposing directions. Regression models that test for the effect of *age* in each of these groups separately confirm the significance of apparent-time trends in both groups; college educated speakers are lowering TRAMP by an estimated 12 Hz per 10 years

(p= 0.017), while speakers without a college education raise it by 14 Hz per 10 years (p= 0.018).

Variation in the frontness of TRAMP appears to be very similar to that in F1 in 2016. Figure 19 shows that the majority of speakers produce TRAMP relatively front in the vowel space, with F2 means between 2100 and 2300 Hz. However, while older speakers average in the higher end of this range, younger speakers seem to orient toward the lower end, i.e. producing a backer TRAMP than older speakers. Again, this seems to apply primarily to college educated speakers, for whom TRAMP seems to have been retracting continuously in apparent time. For speakers without a college education, on the other hand, no notable changes in the frontness of TRAMP can be observed. As a result, younger speakers from the two educational groups differ increasingly in their frontness of TRAMP.

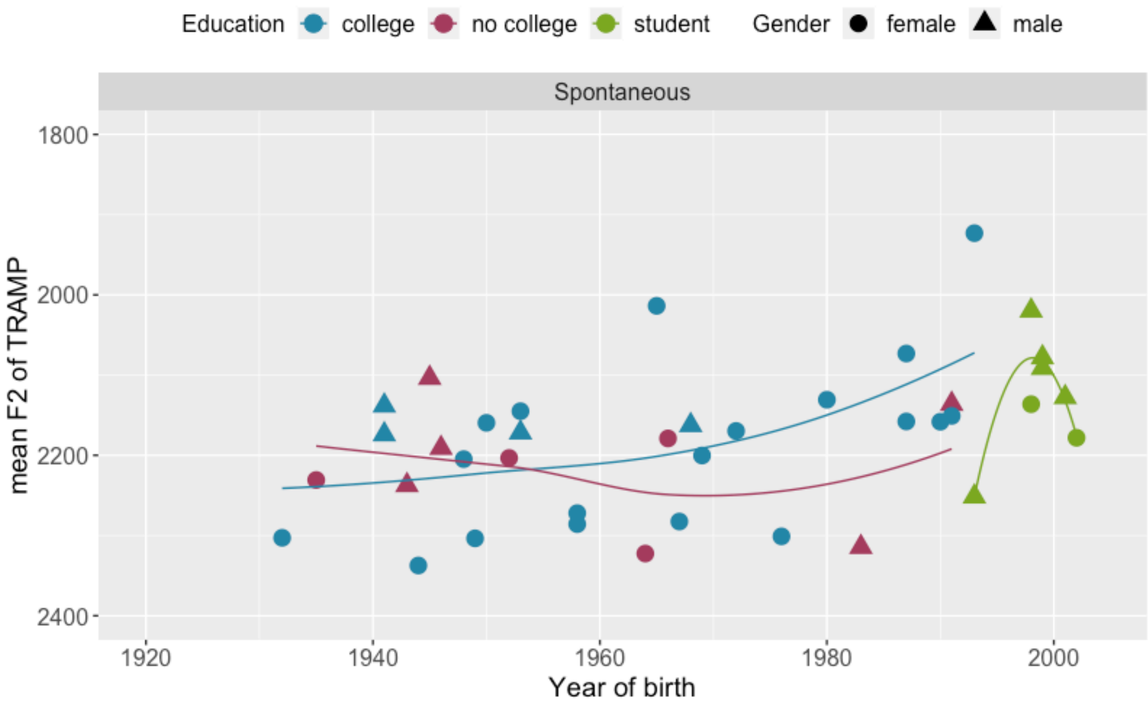


Figure 19: TRAMP F2 means in spontaneous speech in 2016 by *education* and *gender.*

The regression model in Table 12 below supports these observations. The significant interaction between *age* and *education* confirms that speakers with and without a college degree are undergoing different changes in apparent time, and the main effect of *education* predicts the F2 differences between the youngest speakers of these two groups to be quite high. However, a model that tests for the effect of *education* among speakers born after 1960 separately does not find the difference to be statistically significant

(62.274 Hz, *p*= 0.393). A model that tests for the effect of *age* among college educated speakers separately confirms the significance of the apparent-time backing of TRAMP for this group of speakers (3.72 Hz, *p*= 0.003).

Predictor	Coefficient	*p*
(Intercept)	2026.51 Hz	
Age	3.629 Hz	
Gender (Male)	-74.469 Hz	0.054
Education (No college)	281.015 Hz	
*Age*No college*	-4.318 Hz	0.031
Environment		$9x10^{-13}$

Table 12: TRAMP F2 in spontaneous speech in 2016. Reference levels: females, college educated, /n/. Random effects: *speaker, word*. *n* = 1586

Compared to pre-oral TRAP, pre-nasal TRAP appears to be produced notably higher and fronter in the vowel space by speakers in the 2016 sample. As described above, pre-oral TRAP is produced with F1 means between 750 and 800 Hz, and F2 means between 1800 and 2000 Hz in spontaneous speech by 2016 speakers. Pre-nasal TRAP, on the other hand, is realized at a position that is about 100 to 200 Hz higher and 200 to 400 Hz fronter by the majority of speakers, as shown in Figures 20 and 21 below.

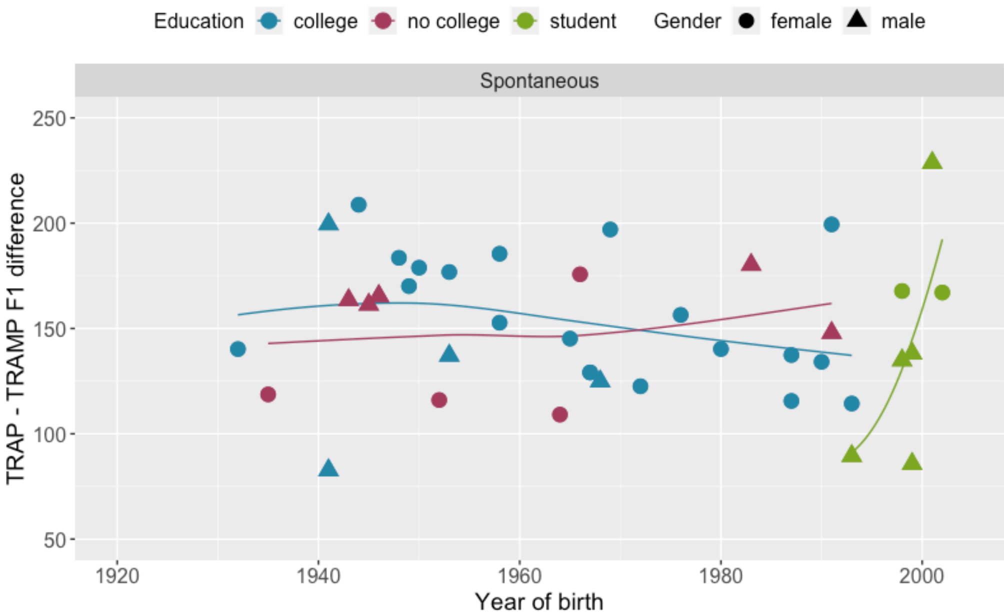


Figure 20: F1 difference between TRAP and TRAMP means in spontaneous speech in 2016 by *education* and *gender*. Positive values indicate that TRAP is lower than TRAMP.

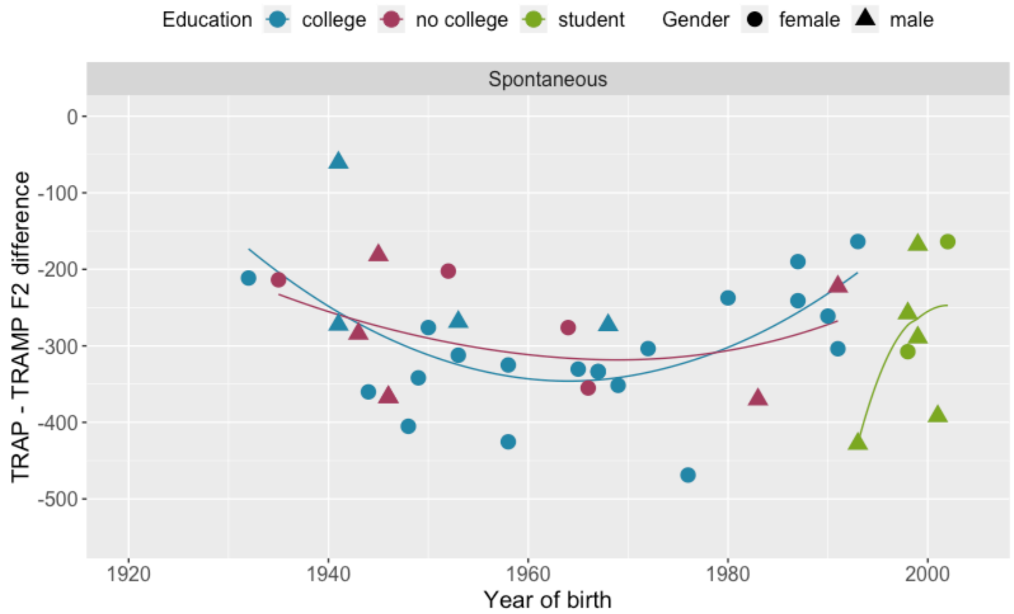


Figure 21: F2 difference between TRAP and TRAMP means in spontaneous speech in 2016 by *education* and *gender*. Negative values indicate that TRAP is backer than TRAMP.

The regression model in Table 13 confirms that the F2 differences between TRAP and TRAMP are statistically significant, and a separate model testing for the difference in F1 predicts this to be significant as well (-161.8 Hz, p= $2x10^{-16}$). Additionally, the significant interaction between *age* and *pre-nasal environment* in model in Table 14 suggests that the lowering of TRAMP proceeds at a slower rate than the lowering of TRAP, thereby increasing the difference between the two in the 2016 sample.

Predictor	Coefficient	*p*
(Intercept)	812.231 Hz	
Age	0.0179 Hz	
Gender *(Male)*	-1.888 Hz	0.877
Education *(No college)* *(Student)*	-1.636 Hz 7.457 Hz	0.907
Following manner *(pre-nasal)*	-97.585 Hz	
*Age*pre-nasal*	-0.369 Hz	0.002
Environment		$4x10^{-14}$

Table 13: TRAP and TRAMP F1 in spontaneous speech in 2016. Reference levels: females, college educated, pre-oral, /p/. Random effects: *speaker, word*. *n*= 5474

Predictor	Coefficient	*p*
(Intercept)	1790.489 Hz	
Age	1.837 Hz	0.017
Gender *(Male)*	-18.9 Hz	0.507
Education *(No college)* *(Student)*	38.206 Hz 0.713 Hz	0.452
Following manner *(pre-nasal)*	112.495 Hz	0.001
Environment		$3x10^{-10}$

Table 14: TRAP and TRAMP F2 in spontaneous speech in 2016. Reference levels: females, college educated, pre-oral, /p/. Random effects: *speaker, word*. *n*= 5474

3.2.1.3 TRAP and TRAMP in Real Time

In real time, TRAP and TRAMP in spontaneous speech appear to be undergoing changes on both the height and the front-back dimension. Not only do the data suggest that 2008 speakers have higher and fronter TRAP and TRAMP than 2016 speakers, it also seems that apparent-time trends are going in different directions in both samples.

Figure 22 suggests that especially younger 2008 speakers produce a higher TRAP and TRAMP than 2016 speakers. This appears to be due to a significant raising of TRAP in both environments in the 2008 sample, which seems to have been led by females, as the two male speakers have greater F1 means than their female peers. Statistically, the effects of *age* in the 2008 sample do not reach the level of statistical significance, neither for TRAP (1.18 Hz, *p*= 0.064) nor for TRAMP (1.14 Hz, *p*= 0.146), and the gender difference is only significant for TRAP. Nevertheless, not only is this raising from 2008 no longer present in the 2016 sample for the majority of speakers, but TRAP and TRAMP are *lower* than they had been in 2008. As a result, the only speakers who satisfy the AE1 criterion for TRAP in spontaneous speech are young females from the 2008 sample.

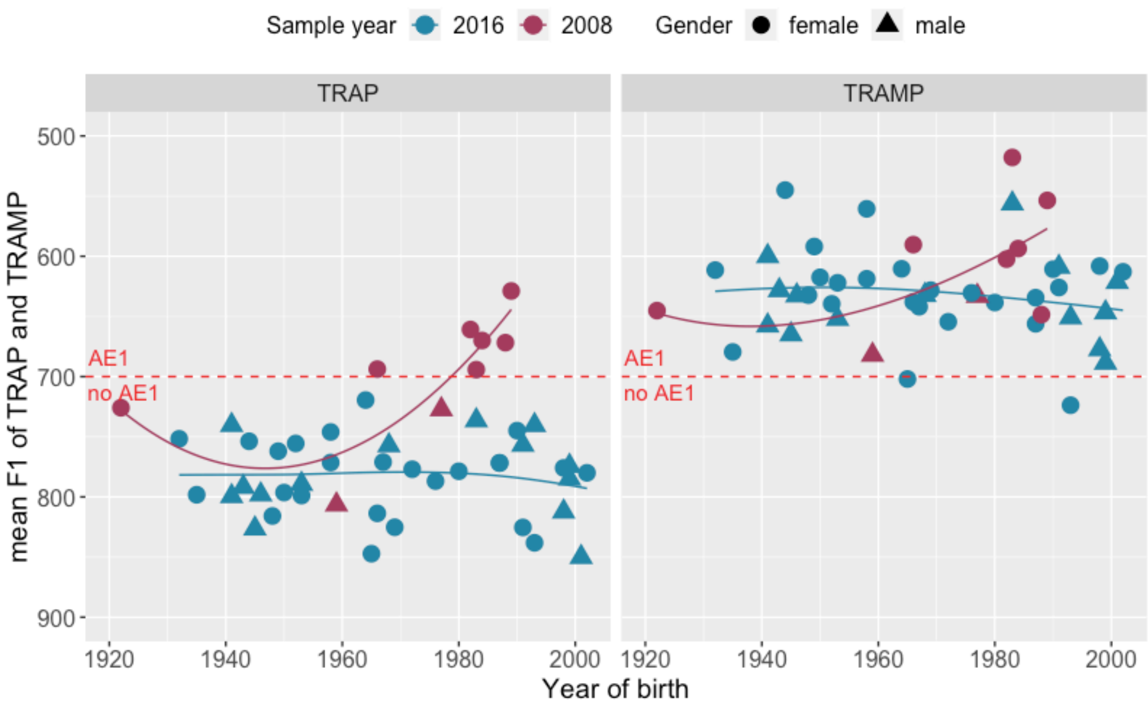


Figure 22: TRAP and TRAMP F1 means in spontaneous speech in 2008 and 2016 by *gender*. Note that the apparent-time profile for TRAP and TRAMP for 2016 speakers looks flat in these plots, as the trend lines combine speakers with and without college degrees, which appear to follow opposing trends in apparent time as detailed above.

The regression models in Tables 15 and 16 corroborate these observations. Younger 2008 speakers are predicted to have a notably higher TRAP and TRAMP than their 2016 peers. The significance of this difference is supported by a separate model testing for the effect of *sample year* for speakers born after 1960 (-100.75 Hz, p= 5x10^{-6}). Additionally, the significant interactions between *age* and *sample year* confirm that apparent-time trends differ notably across the two samples. Thus, the slight lowering of TRAP and TRAMP in apparent time among college educated speakers in 2016 is corroborated by real-time differences between the two samples. However, even speakers without a college degree in 2016, while seemingly still raising TRAP and TRAMP, realize both of them in a lower position than 2008 speakers did.

Predictor	Coefficient	*p*
(Intercept)	818.039 Hz	
Age	-0.139 Hz	
Gender (Male)	5.857 Hz	0.598
Sample year (2008)	-151.643 Hz	
*Age*2008*	1.719 Hz	0.019
Environment		2x10^{-6}

Table 15: TRAP F1 in spontaneous speech in 2008 and 2016. Reference levels: females, 2016, /p/. Random effects: *speaker, word*. *n*= 3840

Predictor	Coefficient	*p*
(Intercept)	633.858 Hz	
Age	-0.351 Hz	
Gender (Male)	11.841 Hz	0.331
Sample year (2008)	-111.286 Hz	
*Age*2008*	1.731 Hz	0.044
Environment		10^{-10}

Table 16: TRAMP F1 in spontaneous speech in 2008 and 2016. Reference levels: females, 2016, /n/. Random effects: *speaker, word*. *n*= 1842

Regarding the frontness of TRAP and TRAMP in real time, Figure 23 below presents a picture that is very similar to that for F1, with younger 2008 speakers producing a fronter TRAP and TRAMP than 2016 speakers. For TRAMP, this again appears to be due to significant fronting in the 2008 sample, which leads younger speakers in this sample to front TRAMP to an F2 greater than 2300 Hz, while in 2016, TRAMP is backing in apparent time for the majority of speakers. TRAP, on the other hand, seems to be backing slightly among 2008 speakers, similar to the trend observed in 2016.

The regression models in Tables 17 and below 18 corroborate these observations. Younger 2008 speakers are predicted to have a considerably fronter TRAMP than their 2016 peers, the significance of which is confirmed by a separate model (170.806 Hz, p= 0.001). Additionally, the significant interaction between *age* and *sample year* for TRAMP confirms that apparent-time trends differ notably across the two samples. However, a regression model that tests for the effects of *age* on F2 of TRAMP in the 2008 sample separately does not find this effect to be of statistical significance (-3.223 Hz, p= 0.121).

For TRAP, on the other hand, the regression model predicts no significant differences between the two samples, and the lack of a significant interaction between *age* and *sample year* confirms that apparent-time developments do not differ to a great extent across the two samples. Nevertheless, the real-time differences in the frontness of TRAP and TRAMP support the apparent-time backing observed in 2016.

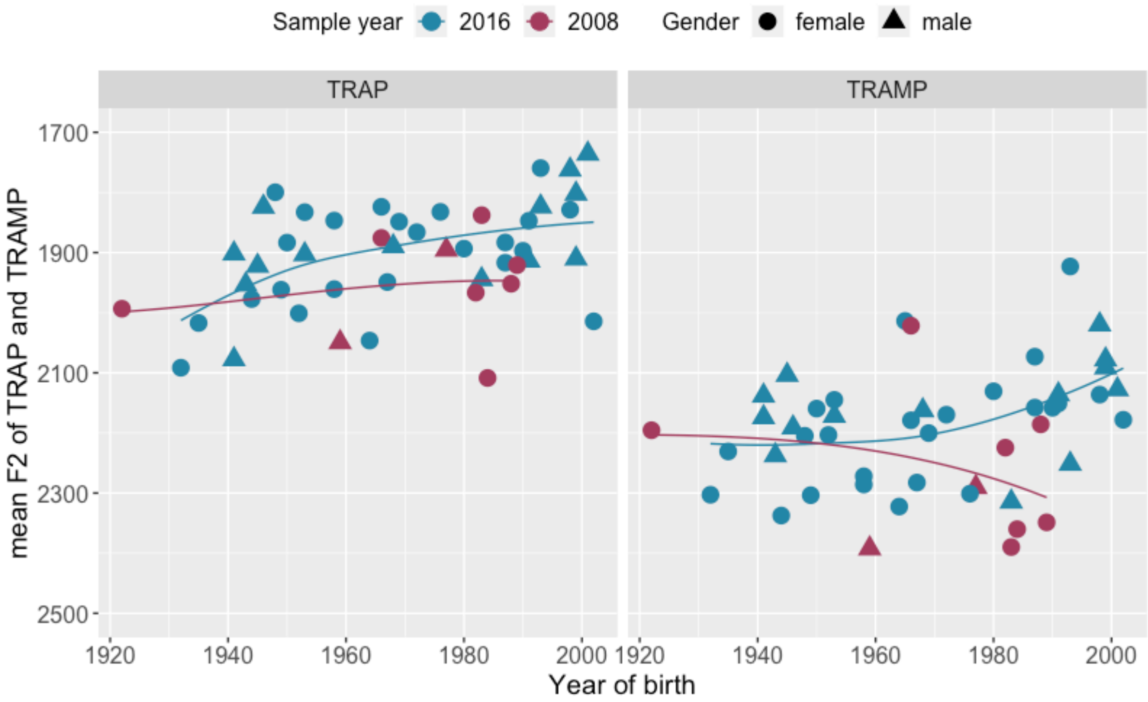


Figure 23: TRAP and TRAMP F2 means in spontaneous speech in 2008 and 2016 by *gender*.

Predictor	Coefficient	*p*
(Intercept)	1793.984 Hz	
Age	1.668 Hz	0.006
Gender (Male)	-3.147 Hz	0.904
Sample year (2008)	56.964 Hz	0.096
Environment		0.035

Table 17: TRAP F2 in spontaneous speech in 2008 and 2016. Reference levels: females, 2016, /p/. Random effects: *speaker, word*. *n*= 3840

Predictor	Coefficient	*p*
(Intercept)	2092.943 Hz	
Age	2.293 Hz	
Gender (Male)	-32.418 Hz	0.238
Sample year (2008)	341.522 Hz	
*Age*2008*	-5.035 Hz	0.01
Environment		$2x10^{-8}$

Table 18: TRAMP F2 in spontaneous speech in 2008 and 2016. Reference levels: females, 2016, /n/. Random effects: *speaker, word*. *n*= 1842

Because of the notably raised TRAP among younger 2008 speakers, most speakers in the 2008 sample appear to differentiate less between TRAP and TRAMP than 2016 speakers do. As was noted above and can be seen in Figure 24, 2016 speakers have, on average, a 100 to 200 Hz F1 difference between TRAP and TRAMP. For 2008 speakers, Figure 24 and the regression model in Table 19 below suggest a difference of less than 100 Hz. Thus, younger 2008 speakers have higher TRAP and TRAMP than 2016 speakers, and differentiate somewhat less between the two phonological environments than 2016 speakers do, indicating real-time lowering of TRAP and TRAMP, as well as an increasing difference between them.

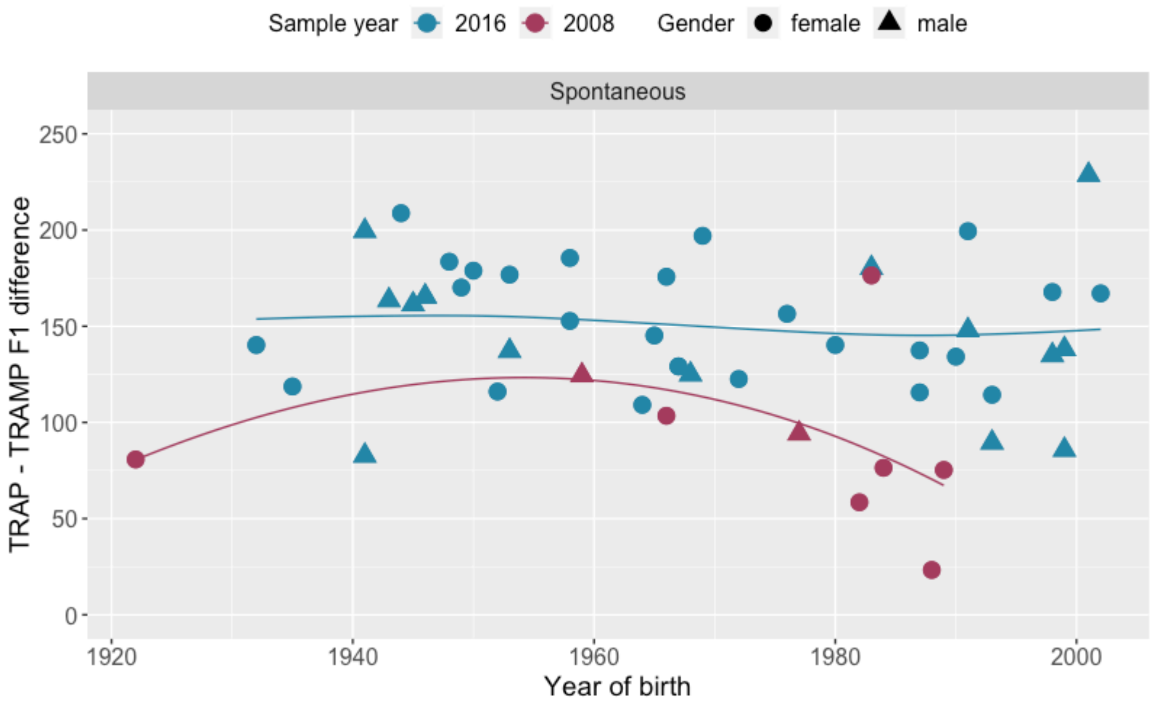


Figure 24: F1 difference between TRAP and TRAMP means in spontaneous speech in 2008 and 2016 by *gender*. Positive values indicate that TRAP is lower than TRAMP.

In terms of frontness, no real-time differences can be observed in the differences between TRAP and TRAMP. As shown in Figure 25 below, 2008 speakers realize TRAMP notably fronter than TRAP, to about the same extent as 2016 speakers do. The regression model in Table 20 confirms that the F2 differences between TRAP and TRAMP are statistically significant in 2008.

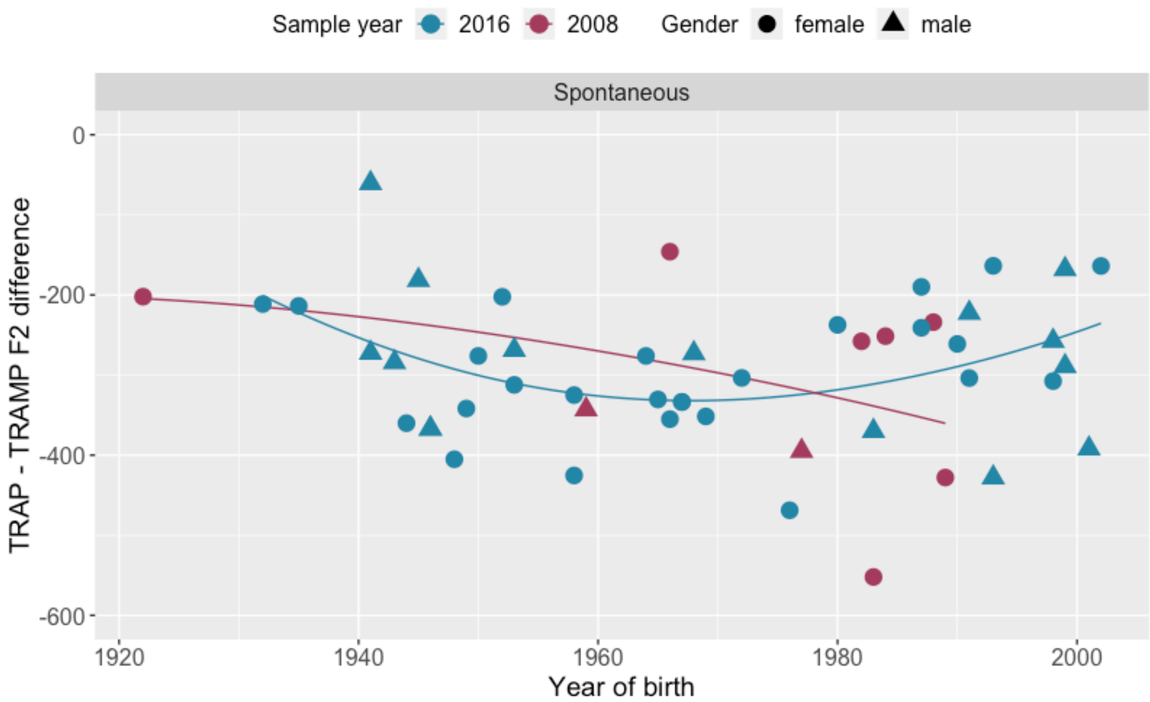


Figure 25: F2 difference between TRAP and TRAMP means in spontaneous speech in 2008 and 2016 by *gender*. Negative values indicate that TRAP is backer than TRAMP.

Predictor	Coefficient	*p*
(Intercept)	623.618 Hz	
Age	1.311 Hz	0.013
Gender *(Male)*	71.883 Hz	0.009
Following manner *(pre-nasal)*	-87.151 Hz	$2x10^{-9}$

Table 19: TRAP and TRAMP F1 in spontaneous speech in 2008. Reference levels: females, college educated, pre-oral. Random effects: *speaker, word*. *n*= 208

Predictor	Coefficient	*p*
(Intercept)	1972.213 Hz	
Age	-0.729 Hz	0.542
Gender *(Male)*	16.022 Hz	0.77
Following manner *(pre-nasal)*	286.499 Hz	$10x10^{-11}$

Table 20: TRAP and TRAMP F2 in spontaneous speech in 2008. Reference levels: females, college educated, pre-oral. Random effects: *speaker, word*. *n*= 208

3.2.1.4 TRAP and EQ

A second criterion that defines NCS-raised TRAP is the EQ criterion. This criterion relates TRAP to DRESS and defines NCS TRAP as being both fronter and higher than DRESS. Similar to AE1, there seems to be only moderate participation in this NCS criterion in spontaneous speech in Ogdensburg. Only 11 of the speakers in the combined 2008 and 2016 data fulfill this criterion in this speech style. However, as with the AE1 criterion, participation in EQ differs between the two sample years. While half of the 2008 speakers meet EQ, none of the 2016 speakers do so in spontaneous speech.

Height Distance

Figure 26 shows that only a few speakers seem to raise TRAP higher than DRESS in spontaneous speech. As the plot shows, the *age* and *gender* pattern for this are very similar to those observed for AE1: Only younger females (born after 1980) in the 2008 sample satisfy this criterion in spontaneous speech, which indicates that TRAP is increasing its distance from DRESS in an upward movement in the 2008 sample. This is supported by an apparent-time *decrease* in the distance between TRAP and DRESS among those 2008 speakers who do not raise TRAP above DRESS. For those speakers, the distance between both phonemes seems to be continuously decreasing, i.e. TRAP is approaching the height of DRESS in spontaneous speech, concomitant with the (not quite significant) apparent-time raising of TRAP in spontaneous speech described above. Thus, as 2008 speakers raise TRAP to an F1 of less than 700 Hz, they also seem to raise it above DRESS.

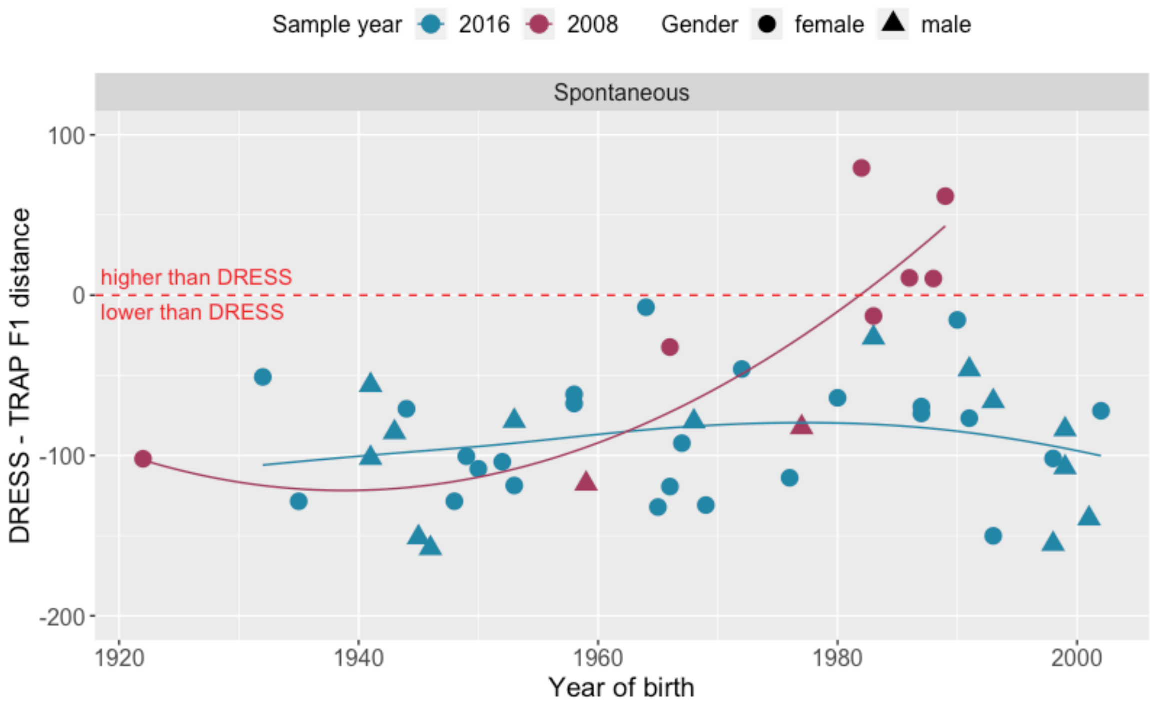


Figure 26: F1 distance between TRAP and DRESS means in spontaneous speech in 2008 and 2016 by *gender*. A positive value indicates that TRAP is raised above DRESS.

In the 2016 data, on the other hand, the F1 distance between TRAP and DRESS appears to lack any apparent-time trends in this speech style as shown in Figure 26 above, i.e. TRAP sits, apparently stably, about 80 Hz lower than DRESS. All 2016 speakers, regardless of *age*, produce TRAP with an F1 that is about 50 to 150 Hz higher than that of DRESS. Only three speakers approach the height of DRESS in their spontaneous speech production of TRAP, but there is no identifiable age pattern that would identify an apparent-time trend toward TRAP raising in this data set. Thus, the apparent-time differences between educational groups in the height of TRAP identified in 2016 data do not seem to translate to the TRAP-DRESS distance. As a result of these differences in apparent-time developments in the TRAP-DRESS distance in the two samples, notable real-time differences emerge for speakers born after 1980, as younger 2008 speakers raise TRAP above DRESS, while no other speakers do.

Front-back Distance

On the front-back dimension, the data presents a completely different picture in terms of EQ participation in spontaneous speech than on the height dimension. Overall, the majority of speakers produce spontaneous TRAP in a fronter position than DRESS, thus fulfilling the front-back aspect of the EQ criterion. Only a small number of speakers produces TRAP further back than DRESS in this speech style, all of which belong to the 2016 sample.

Speakers from the 2008 and 2016 data differ in terms of their absolute F2 TRAP-DRESS distance as well as in apparent-time trends in spontaneous speech. This can be seen in Figure 27 below. Four of the youngest 2008 speakers have a negative distance between -220 and -370 Hz, i.e. produce TRAP significantly fronter than DRESS, while only two of the 2016 speakers reach the -250 Hz mark; both of them older speakers. On the other hand, five mostly younger 2016 speakers produce TRAP backer than DRESS, and another 10 produce TRAP less than 50 Hz fronter than DRESS. Only two of the 2008 speakers fall within this range. Thus, it appears that 2016 speakers are reducing the F2 distance between TRAP and DRESS in apparent time. Older speakers from this sample tend to have a somewhat greater distance between both phonemes than younger speakers do, which is in agreement with the observed apparent-time backing of TRAP. In the 2008 sample, on the other hand, it is the youngest speakers who have the greatest F2 distance between both phonemes, while older speakers tend to produce them closer together. Thus, for the 2008 sample, the data indicates an apparent-time *increase* in the TRAP-DRESS F2 distance in spontaneous speech, i.e. TRAP being produced increasingly fronter than DRESS. This increase in distance in the 2008 data is unexpected, given the observation that spontaneous TRAP seems to be retracting, albeit slightly, in the 2008 data, so that a decrease in distance to DRESS on the front-back dimension would be expected. The reason for the observed increase in the TRAP-DRESS distance is the retraction of DRESS away from TRAP as shown in Figure 28 below. This will be explored further in Chapter 4.2.1.

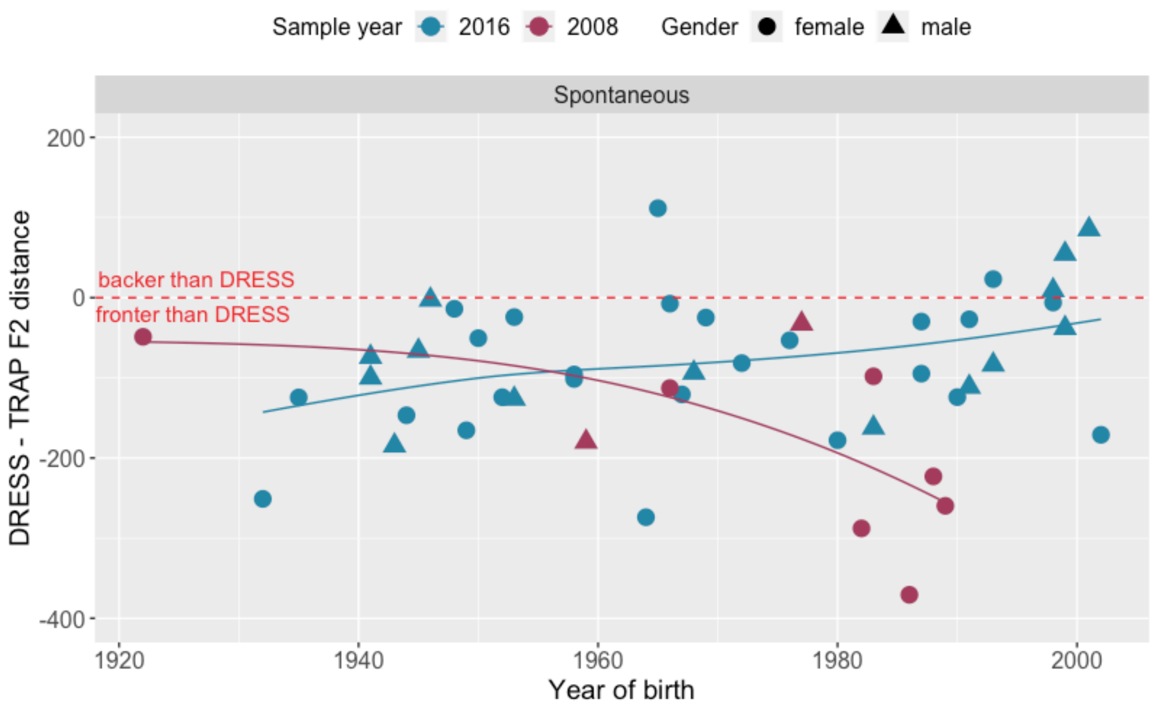


Figure 27: F2 distance between TRAP and DRESS means in spontaneous speech in 2008 and 2016 by *gender*. A negative value indicates that TRAP is fronter than DRESS.

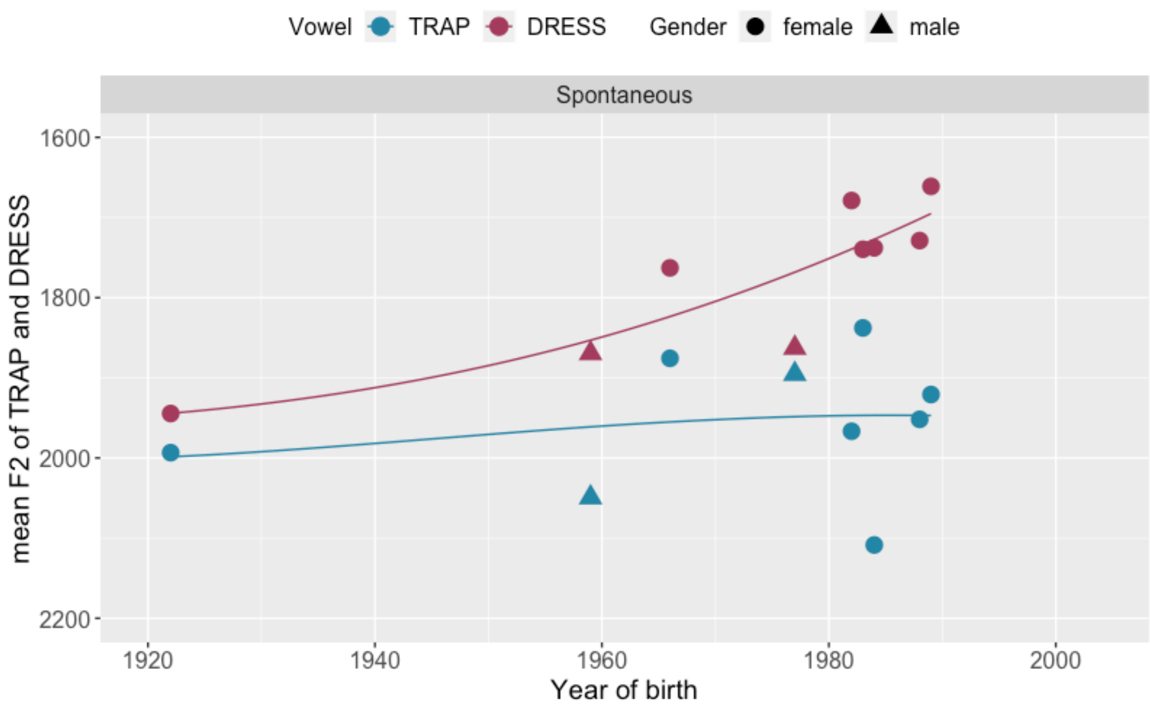


Figure 28: TRAP and DRESS F2 means in spontaneous speech in 2008 by *gender*.

3.2.2 TRAP in wordlist style

Among 2016 speakers, variation in the height of TRAP in wordlist style is relatively limited. As shown in Figure 29, most speakers produce wordlist TRAP relatively high in the vowel space, with an F1 between 700 and 850 Hz, but only two speakers in this sample cross the AE1 benchmark of 700 Hz in this speech style. Figure 29 further suggests that wordlist TRAP may be lowering slightly in apparent time; however, it appears that only speakers with a college degree participate in this trend, while TRAP has remained steady for speakers without a college degree. As a result, younger college educated speakers tend to produce a lower TRAP than their peers without a college degree.

The regression model in Table 21 below does not support these observations, however, and does not predict any of the tested social factors to have any significant impact on F1 of wordlist TRAP. Overall, however, wordlist TRAP *is* predicted to be lowering at a statistically significant level in the community by a model that tests for the effect of *age* in the combined 2016 sample and excludes education as a factor (-1.23 Hz, p= 0.013), likely owing to the students' low TRAP (Figure 29).

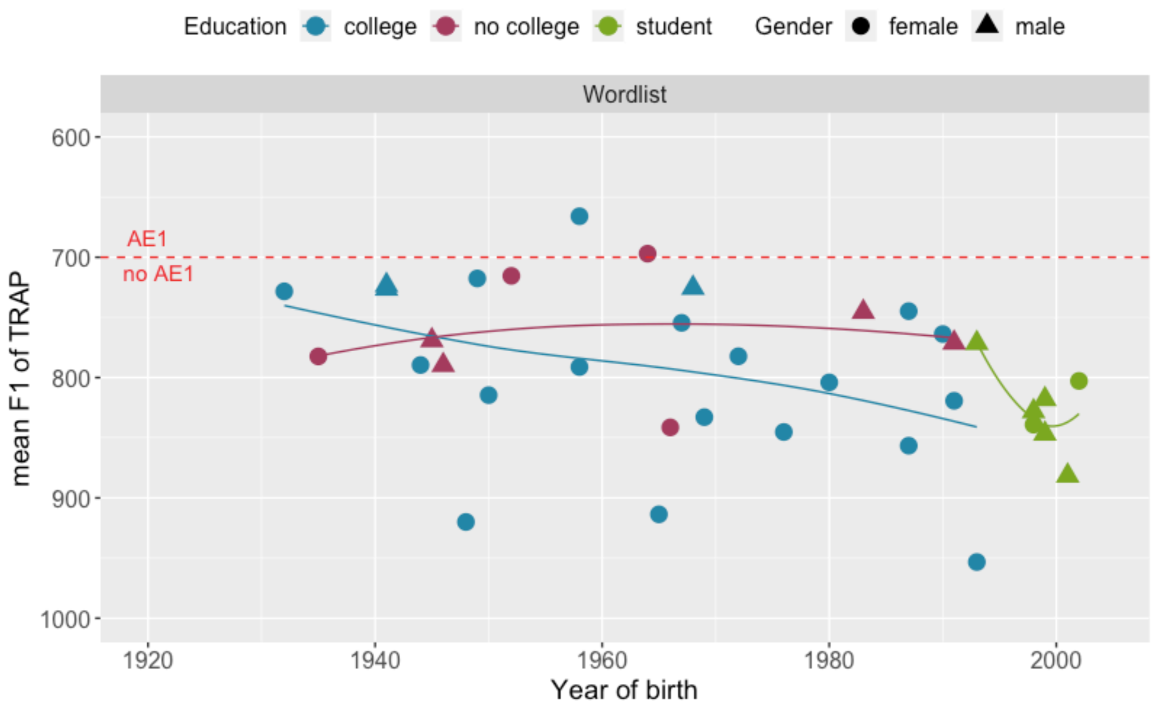


Figure 29: TRAP F1 means in wordlist style in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	876.759	
Age	-0.936	0.175
Gender (Male)	-37.081	0.242
Education (No college)	-11.661	0.697
Environment		0.039

Table 21: TRAP F1 in wordlist style in 2016. Reference levels: females, college educated, /s/. Random effects: *speaker, word*. *n* = 274

A very similar picture emerges for the frontness of TRAP in wordlist style. As shown in Figure 30, older speakers in the 2016 sample have F2 means of around 2100 Hz, while younger speakers realize wordlist TRAP as far back as 1650 Hz, indicating an apparent-time retraction of TRAP. However, as in spontaneous speech, only college educated speakers appear to participate in this trend in wordlist style. For speakers without a college education, F2 seems to remain steady, increasing the difference in the frontness of TRAP between these two groups. The regression model in Table 22 below, however, does not find these observations to be of statistical significance. A model that excludes education as a factor, on the other hand, does predict a significant *age* effect for the frontness of wordlist TRAP for the 2016 sample (2.67 Hz, *p*= 0.021).

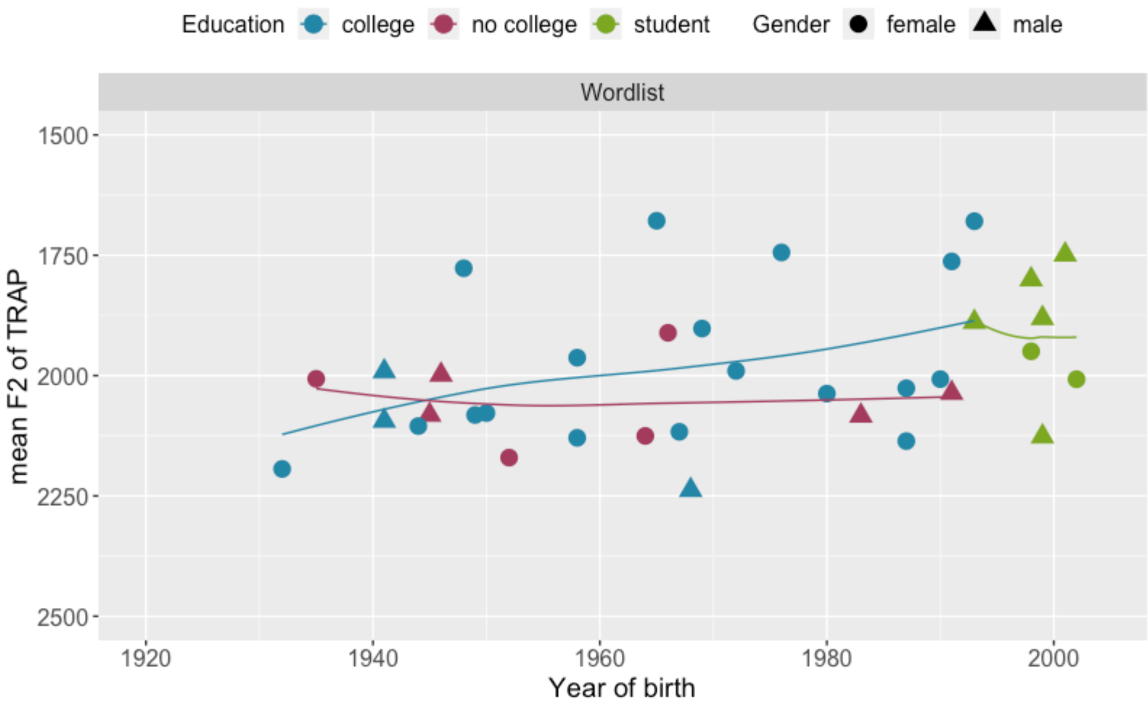


Figure 30: TRAP F2 means in wordlist style in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	1814.615 Hz	
Age	2.225 Hz	0.147
Gender (Male)	68.388 Hz	0.328
Education (No college)	28.709 Hz	0.665
Environment		$2x10^{-5}$

Table 22: TRAP F2 in wordlist style in 2016. Reference levels: females, college educated, /s/. Random effects: *speaker, word*. *n* = 274

In wordlist style, differences between the 2008 and 2016 samples are particularly notable on the height dimension. The data suggest that 2008 speakers have higher TRAP than 2016 speakers, while they do not differ to a great extent in terms of frontness.

Figure 31 suggests that 2008 speakers as a group produce a higher TRAP than the majority of speakers in the 2016 sample in wordlist style. Two of them realize TRAP with an F1 of less than 700 Hz, thus meeting the AE1 criterion. Although the same applies to two of the 2016 speakers, these numbers suggest that participation in AE1 is proportionately greater in 2008 than it is in 2016. All four of these speakers are born between 1958 and 1966.

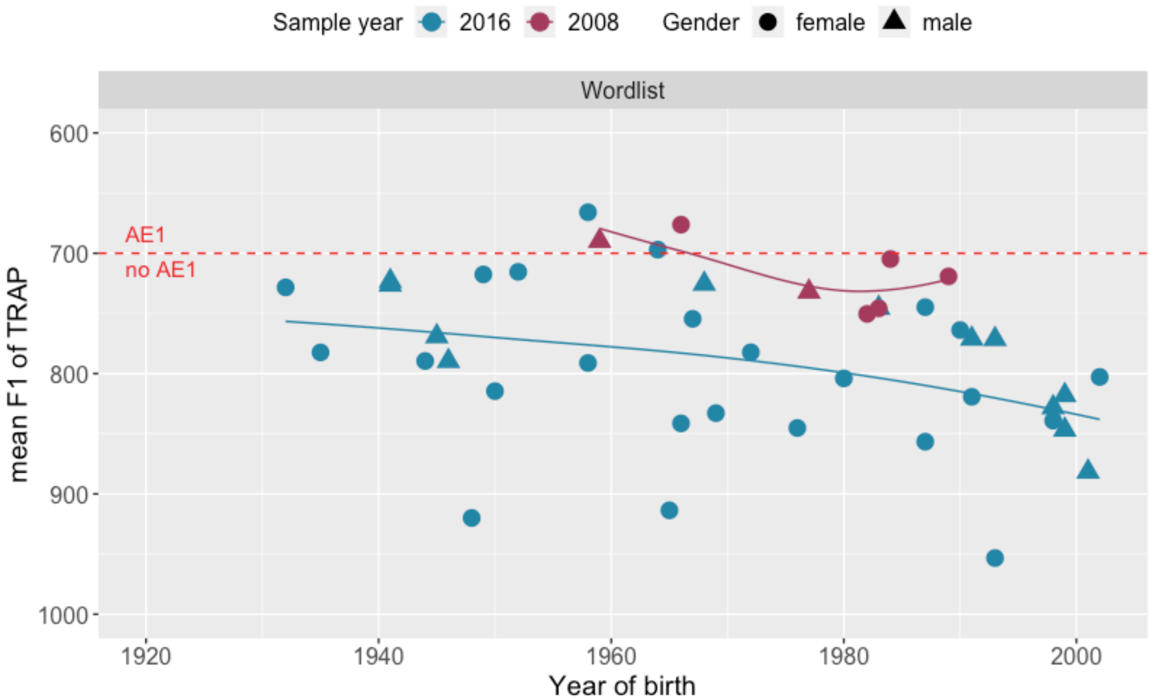


Figure 31: TRAP F1 means in wordlist style in 2008 and 2016 by *gender*.

Nevertheless, speakers in the 2008 sample appear to be undergoing the same apparent-time lowering of wordlist TRAP that was observed in the 2016 sample. Although this lowering does not reach the level of statistical significance in the 2008 data (-2.53 Hz, p= 0.11), it does lead to a reversal in AE1 participation in this sample: While the two relatively older speakers satisfy the AE1 criterion, younger 2008 speakers cease to participate in it. The regression model in Table 23 below predicts that the F1 differences between the 2008 and 2016 samples are statistically significant, so that the apparent-time lowering of TRAP is supported by real-time evidence.

In terms of frontness, 2008 and 2016 speakers do not differ greatly from each other in wordlist style. Although the trend line in Figure 32 suggest a potential apparent-time fronting of wordlist TRAP in 2008, which contrasts the apparent-time backing of TRAP in 2016, this appears to be due to one male speaker (born in 1959) with a relatively small mean F2. In a regression model that tests for the effect of *age* in 2008, this effect is not found to be statistically significant (-5.33 Hz, p= 0.443). The older outlier in the 2008 sample is also the only speaker who differs notably from the majority of 2016 speakers by producing a notably backer TRAP. The remaining speakers have F2 means that are very similar to those of their 2016 peers.

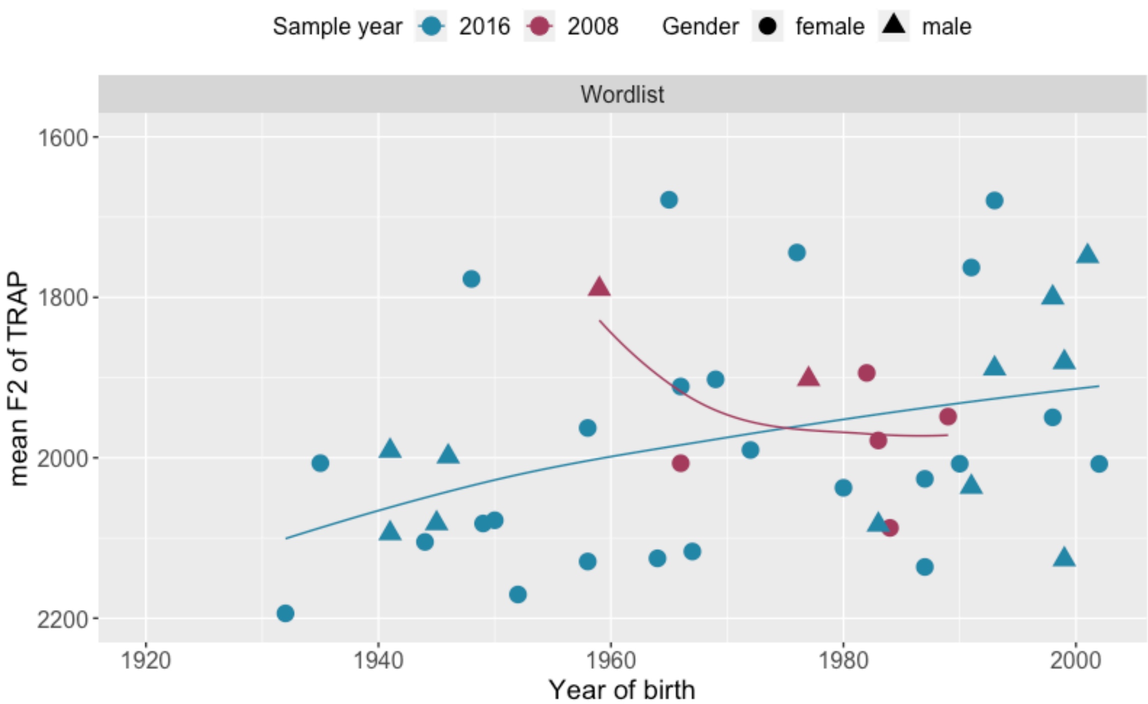


Figure 32: TRAP F2 means in wordlist style in 2008 and 2016 by *gender*.

However, in the regression model in Table 24, 2008 speakers are estimated to produce a nearly significantly *fronter* TRAP than 2016 speakers. If this effect is authentic, it lends real-time evidence to the apparent-time backing of TRAP, however, Figure 32 above does not suggest that this is the case.

Predictor	Coefficient	*p*
(Intercept)	890.84 Hz	
Age	-1.259 Hz	0.006
Gender *(Male)*	-21.987 Hz	0.246
Sample year *(2008)*	-85.017 Hz	0.002
Environment		0.452

Table 23: TRAP F1 in wordlist style in 2008 and 2016. Reference levels: females, 2016, /s/. Random effects: *speaker, word. n*= 400

Predictor	Coefficient	*p*
(Intercept)	1835.735 Hz	
Age	2.275 Hz	0.037
Gender *(Male)*	-0.203 Hz	0.996
Sample year *(2008)*	125.856 Hz	0.056
Environment		0.647

Table 24: TRAP F2 in wordlist style in 2008 and 2016. Reference levels: females, 2016, /s/. Random effects: *speaker, word. n*= 400

3.2.2.1 TRAP and EQ

Similar to AE1, participation in the EQ criterion in wordlist style seems to be rather limited. Only nine speakers in the combined 2008 and 2016 data fulfill the EQ criterion in this speech style; two of them from the 2008 sample, seven from the 2016 sample. Thus, while participation in AE1 is proportionately greater in 2008, these differences seem to neutralize for the EQ criterion, suggesting that the 2016 data offers conspicuously different results in wordlist style for the EQ criterion than for AE1.

Height distance

In wordlist style, speakers from the 2008 and 2016 sample do not differ notably in the TRAP-DRESS height distance. As shown in Figure 33 below, a total of nine speakers fulfill the F1 aspect of the EQ criterion in this speech style. Seven of them belong to the 2016 sample, while two of them are part of the 2008 data set. The two 2008 speakers, however, are not the same two speakers who satisfy AE1 in this speech style. Rather than the two oldest speakers, it is two of the youngest speakers who meet EQ in this speech style, suggesting an apparent-time increase in EQ participation in 2008. However, the other two 2008 speakers of similar ages do not meet the criterion, and neither do the majority of 2016 speakers. Thus, a trend toward higher TRAP in wordlist style does not seem likely. Instead, the distance between the two phonemes appears to remain consistent over time. This seems to contradict the wordlist findings for F1 TRAP to a certain extent. The slight tendency toward a lower TRAP in apparent time in both samples would have suggested a slight increase in the F1 distance between TRAP and DRESS, as TRAP moves further away from DRESS in a downward movement in apparent time. As the analysis of DRESS in Chapter 4.2.2 will show, the unexpected consistency in the F1 difference between TRAP and DRESS in wordlist style is likely caused by a simultaneous lowering of DRESS alongside TRAP.

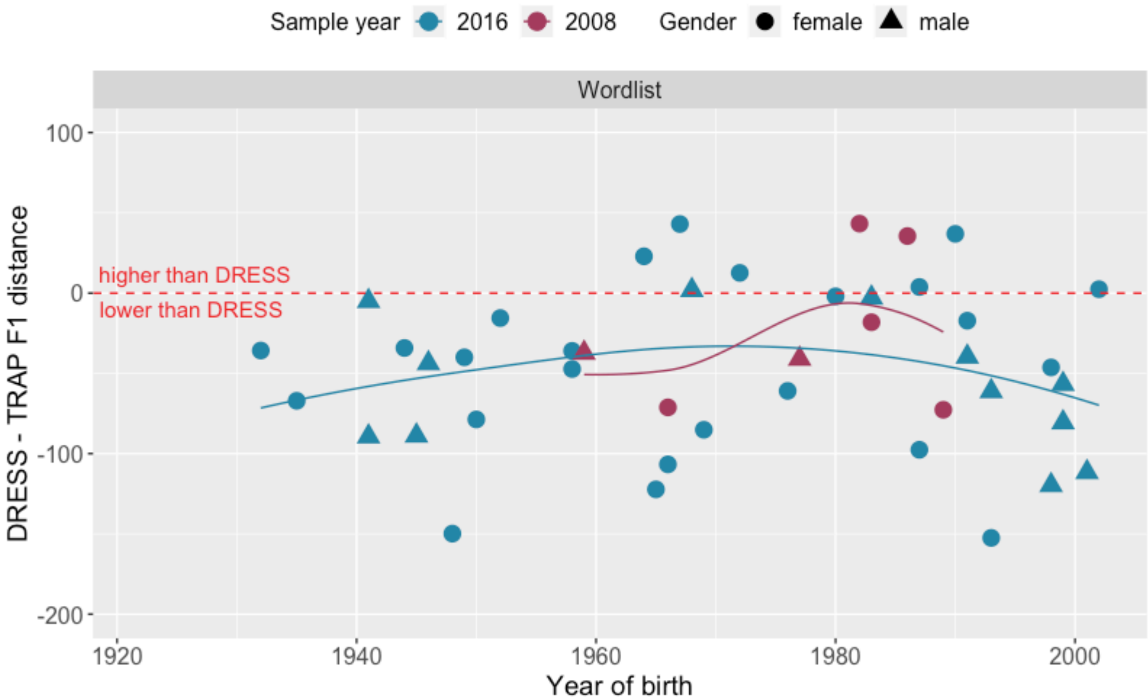


Figure 33: F1 distance between TRAP and DRESS means in wordlist style in 2008 and 2016 by *gender*. A positive value indicates that TRAP is raised above DRESS.

The majority of speakers in the 2008 and 2016 samples fall within the same distance range and realize TRAP between 0 to -100 Hz lower than DRESS. This lack of real-time difference contradicts the findings for wordlist TRAP F1, where 2008 speakers were found to produce TRAP about 85 Hz higher than 2016 speakers. However, developments in the height of DRESS can explain this seeming contradiction, as 2008 speakers produce not only wordlist TRAP, but also wordlist DRESS higher in the vowel space than 2016 speakers (see Chapter 4.2.2). Thus, the phonetic distance between the two phonemes does not differ notably between the two data sets in this speech style.

Front-back distance

On the front-back dimension, EQ participation in wordlist style presents a very different picture compared to the height dimension. As shown in Figure 34, the majority of speakers from both samples produce TRAP in a fronter position than DRESS, thus fulfilling the front-back aspect of the EQ criterion. Speakers in the 2008 sample are among those who produce TRAP the furthest front in relation to DRESS, with TRAP-DRESS F2 differences that are greater than -200 Hz in wordlist style. In the 2016 sample, only few speakers front TRAP to this extent. Furthermore, seven of the 2016 speakers produce wordlist TRAP *backer* than DRESS, and another two have TRAP no more than 50 Hz fronter than DRESS. Only one of the 2008 speakers stays within this range. Thus, 2016 speakers appear to produce wordlist TRAP with a lower F2 distance to DRESS than 2008 speakers, which is in agreement with the inter-set differences (i.e. those between 2008 and 2016) observed in F2 of wordlist TRAP.

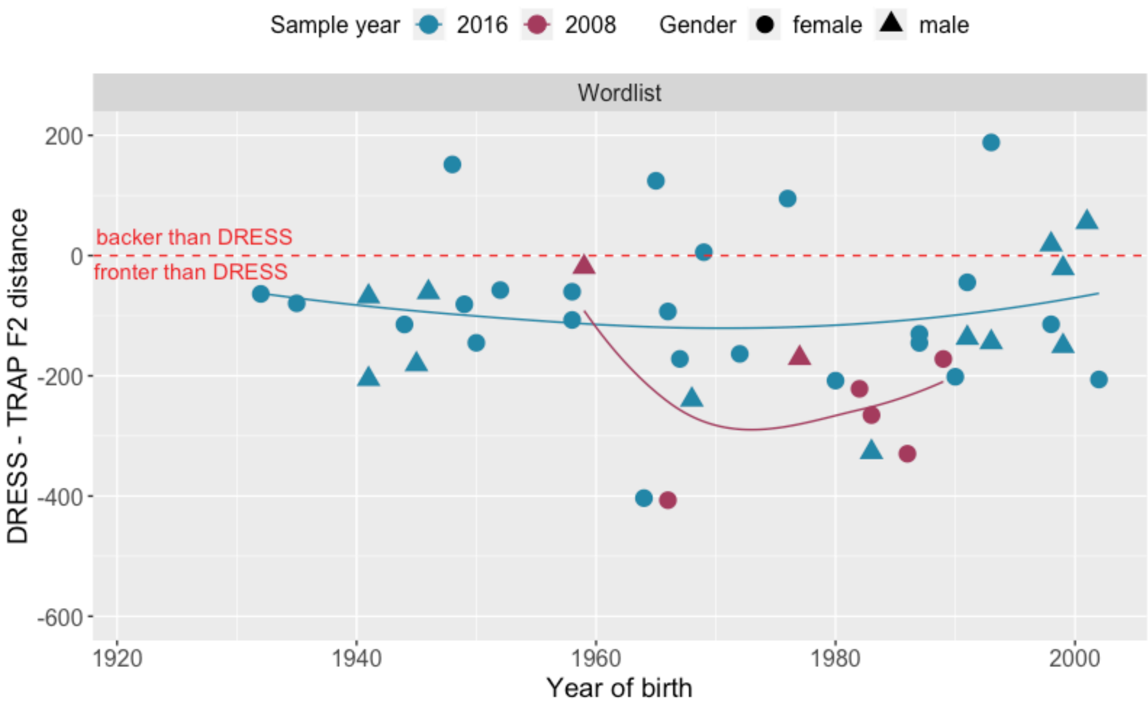


Figure 34: F2 distance between TRAP and DRESS means in wordlist style in 2008 and 2016 by *gender*. A negative value indicates that TRAP is fronter than DRESS.

Since the 2008 speakers with relatively fronter wordlist TRAP are among the youngest speakers in this sample, the data indicates an apparent-time increase in the F2 TRAP-DRESS difference in this style for speakers in the 2008 sample. However, this trend seems less sharp and not as linear as in spontaneous speech. Among 2016 speakers, on the other hand, no notable apparent-time trends in the relative frontness of TRAP and DRESS emerge from the data. This contradicts developments in the frontness of wordlist TRAP, which appears to be backing at least to some extent, which in turn would imply a decrease in its F2 distance to DRESS. However, as will be discussed in more detail in Chapter 4.2.2, wordlist DRESS is undergoing retraction in the 2016 sample as well, thereby offsetting the effect that TRAP retraction might have on its F2 distance to DRESS.

3.2.3 Style shifting TRAP

Style-shifting patterns for TRAP appear to differ greatly depending on the dimension in 2016: While the vast majority of speakers in this sample shift to a more fronted TRAP in wordlist style, style shifting in the height of TRAP seems to follow a more complex pattern.

Overall, style shifting in the height of TRAP seems to be minimal in the 2016 sample. As described above, the majority of speakers in this sample produce TRAP with an F1 between 750 and 800 Hz in spontaneous speech, and between 700 and 850 Hz in wordlist style, indicating very little intra-speaker variation. However, when considering the differences between each speaker's spontaneous and wordlist TRAP means, it appears that the range and direction of style shifting in the height of TRAP depends greatly on the speakers' *age*, as well as on *gender* in the 2016 data. As shown in Figure 35 below, the majority of older speakers in the 2016 sample shift toward more raised TRAP in wordlist style, while younger speakers in this sample shift away from it, i.e. realize a lower TRAP in more careful speech. This indicates an apparent-time change in progress affecting the direction of style shifting: The community is changing from treating raised TRAP as a target to avoiding it in more careful speech. The main effect of *style* in the regression model in Table 25 below confirms that young people realize TRAP lower in wordlist style than in spontaneous speech, while the interaction between *age* and *style* indicates that older speakers have TRAP higher in wordlist style. The effect of *style* switches from positive to negative around 1960 in apparent time, as shown in Figure 35 below.

Furthermore, Figure 35 below suggests a *gender* difference in the style shifting of the height of TRAP. While older men exclusively shift *toward* a higher TRAP in wordlist style, some of the older women do the opposite, similar to the style-shifting pattern of younger speakers. The interaction between *gender* and *style* in the regression model presented in

Table 25 below confirms the significance of the *gender* difference in style shifting, which indicates a female lead in the change in style shifting: Some older women seem to have anticipated the direction of the change in progress, while older men uniformly maintain the more conservative pattern of raised TRAP.

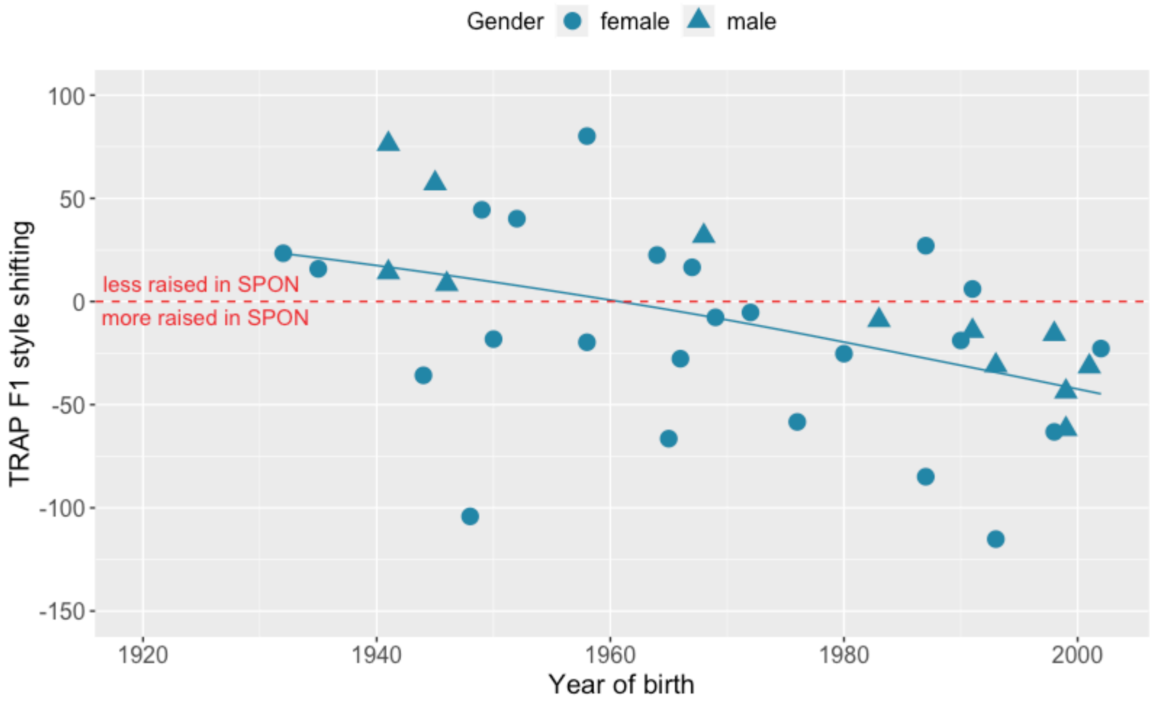


Figure 35: F1 difference between spontaneous and wordlist TRAP means in 2016 by *gender*. A positive value indicates that the vowel is more raised in wordlist style than in spontaneous speech.

In terms of frontness, the style-shifting pattern appears to be the same for the majority of speakers in the 2016 sample. Figure 36 below suggests that the majority of speakers shift to a slightly fronter TRAP in wordlist style. Only a few exceptional speakers do not seem to follow this pattern and shift to a slightly more retracted TRAP instead. This is somewhat surprising, as the shift to a *lower* TRAP in wordlist style among younger speakers should, theoretically, entail a shift toward *backer* TRAP instead of fronter. However, this shift in frontness is relatively small for these younger majority of speakers (and in fact do not reach the level of statistical significance as summarized in Table 26 below), and most of them do appear to minimize it even further, so that the directions of style shifting on the two dimensions do not necessarily contradict each other to a great extent.

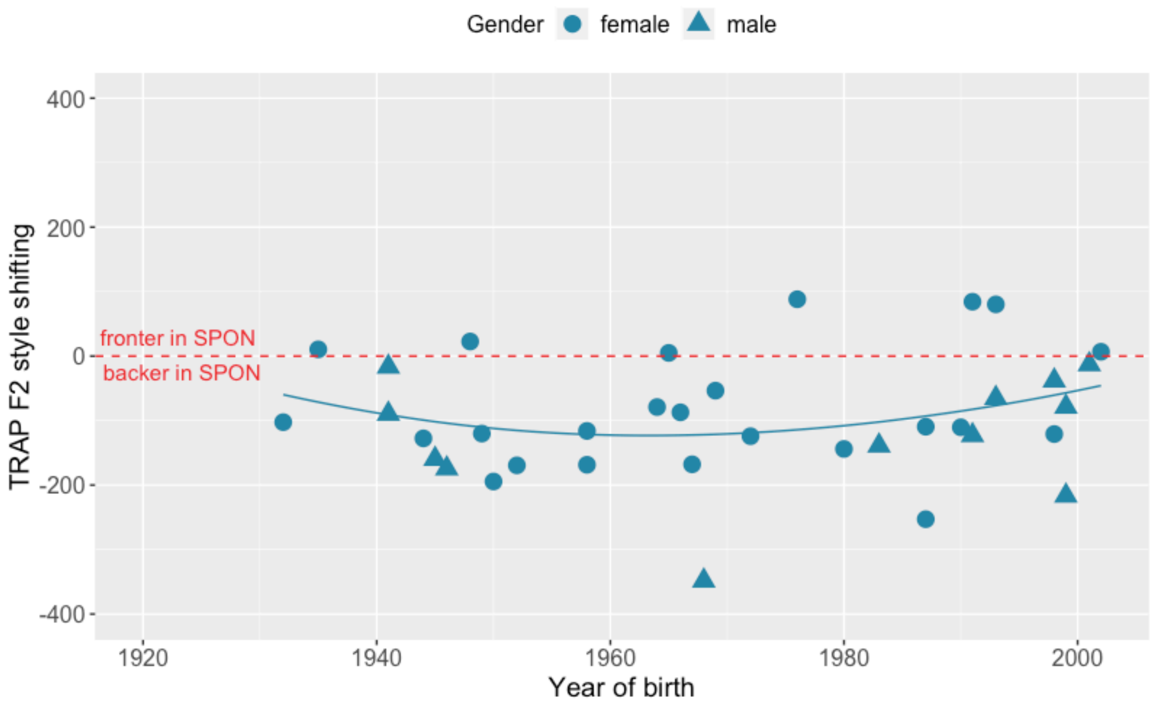


Figure 36: F2 difference between spontaneous and wordlist TRAP means in 2016 by *gender*. A positive value indicates that the vowel is backer in wordlist style than in spontaneous speech.

Predictor	Coefficient	*p*
(Intercept)	814.55 Hz	
Age	-0.021 Hz	
Gender (Male)	-1.628 Hz	
Education		
(No college)	-6.004 Hz	0.839
(Student)	7.489 Hz	
Style		
(Wordlist)	56.626 Hz	
*Age*wordlist*	-1.058 Hz	$2x10^{-7}$
*Wordlist*male*	-22.01 Hz	0.016
Environment		$9x10^{-6}$

Table 25: Effect of *style* on F1 of TRAP in 2016. Reference levels: females, college educated, spontaneous speech, /p/. Random effects: *speaker, word*. *n*= 4048

Predictor	Coefficient	*p*
(Intercept)	1788.141 Hz	
Age	1.689 Hz	0.061
Gender (Male)	-10.8 Hz	
Education		
(No college)	44.313 Hz	0.445
(Student)	-7.621 Hz	
Style		
(Wordlist)	1.275 Hz	
*Wordlist*male*	32.018 Hz	0.08
Environment		0.01

Table 26: Effect of *style* on F2 of TRAP in 2016. Reference levels: females, college educated, spontaneous speech, /p/. Random effects: *speaker, word*. *n*= 4048

The apparent-time raising of TRAP in the spontaneous speech among 2008 speakers, in contrast to the apparent-time lowering of wordlist TRAP in this data set, implies a significant amount of intra-speaker variation in this data set. Figure 37 below indicates that this is indeed the case on the height dimension. In terms of frontness, on the other hand, style shifting appears to be limited. On both dimensions, 2008 speakers appear to follow very similar style-shifting patterns as 2016 speakers.

In terms of height, the majority of older 2008 speakers shift toward more raised TRAP in wordlist style, while younger speakers in this sample shift away from it. Thus, like in 2016, younger speakers realize a lower TRAP in more careful speech, indicating an apparent-time change in progress toward avoiding raised TRAP in more careful speech. As was described above, this shift results in a negation of their AE1 participation in wordlist style.

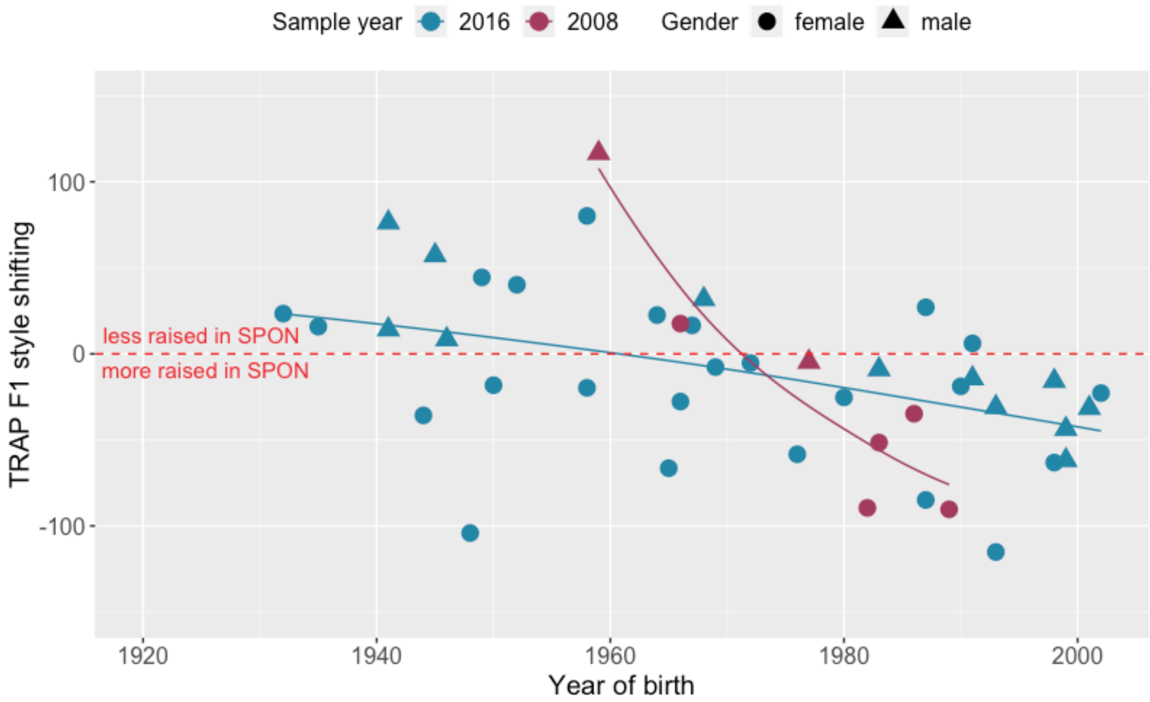


Figure 37: F1 difference between spontaneous and wordlist TRAP means in 2008 and 2016 by *gender*. A positive value indicates that the vowel is more raised in wordlist style than in spontaneous speech.

On the front-back dimension, style shifting in 2008 appears to be minimal. As shown in Figure 38 below, four of the speakers from this sample shift to a more fronted TRAP in wordlist style, much like 2016 speakers do; however, for all four of these speakers, the extent of this shift is relatively small, and for two of them it is negligible. The remaining three speakers shift to a more retracted TRAP in more careful speech, though again, the range of shifting is minimal, with the exception of the male speaker born in 1959.

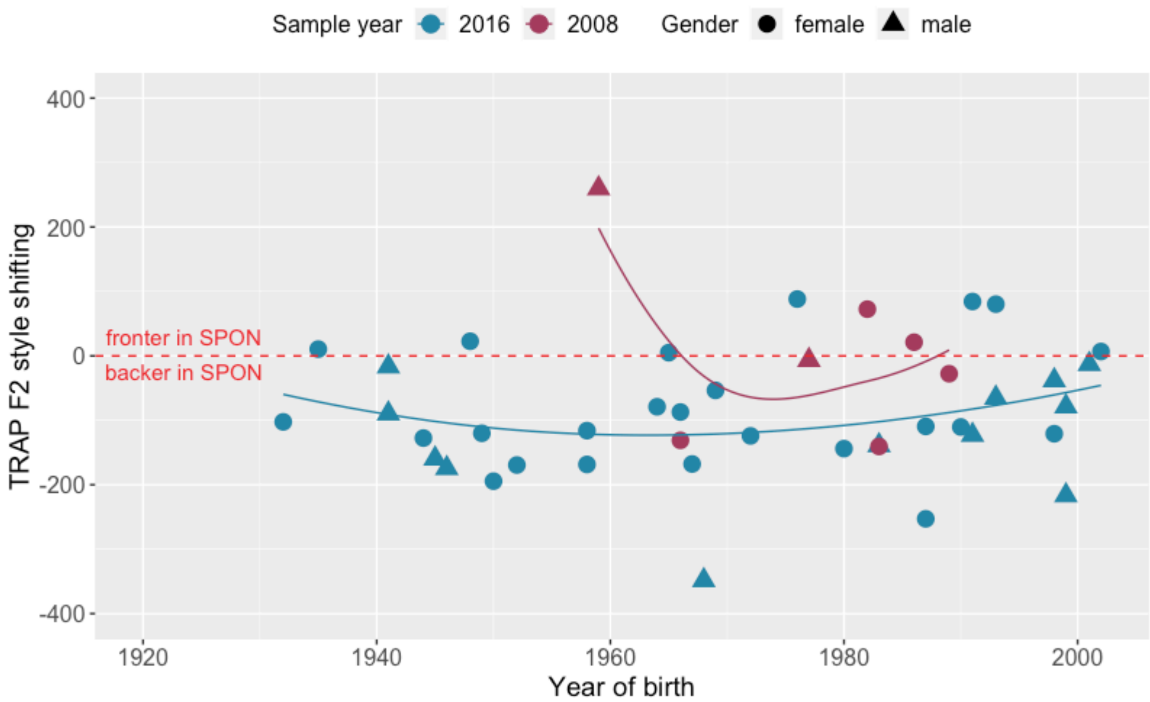


Figure 38: F2 difference between spontaneous and wordlist TRAP means in 2008 and 2016 by *gender*. A positive value indicates that the vowel is backer in wordlist style than in spontaneous speech.

These style-shifting patterns also affect the extent to which participants satisfy the EQ criterion in each speech style.

In the 2016 data, most speakers seem to slightly reduce the F1 distance between TRAP and DRESS in more careful speech, some to the extent of producing TRAP with a lower F1 than DRESS in wordlist style, as can be seen in Figure 39 below. Thus, the height aspect of the EQ criterion is fulfilled more often by 2016 speakers in wordlist style than it is in spontaneous speech. In terms of frontness, on the other hand, EQ participation does not seem to be affected by style shifting, as the majority of 2016 speakers continues to realize TRAP in a position that is fronter than DRESS in more careful speech, as can be seen in Figure 40 below.

In the 2008 data, style-shifting patterns found for F1 TRAP-DRESS distance oppose those observed in 2016, but are in line with those observed for TRAP F1 in this data set. Younger speakers in the 2008 sample seem to shift to a lower TRAP in more careful speech, to the extent of compromising their AE1 and EQ. Thus, while 2016 speakers shift toward more EQ participation in more careful speech, 2008 speakers retreat from it. In terms of F2, on the other hand, they continue to realize TRAP in a fronter position than DRESS, much like 2016 speakers do.

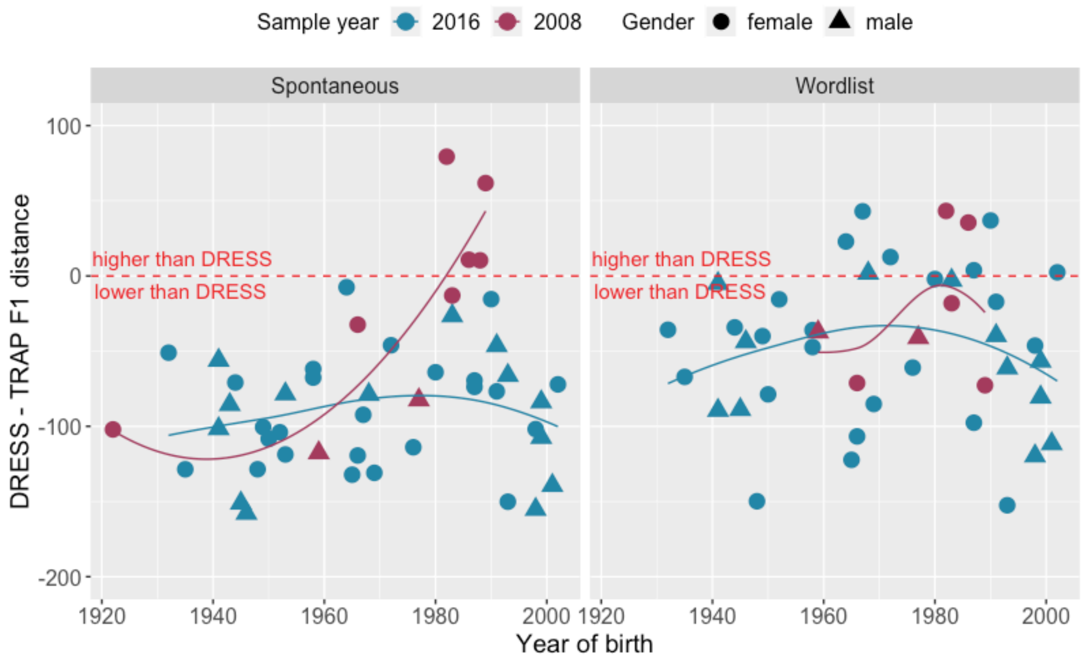


Figure 39: F1 distance between TRAP and DRESS means in 2008 and 2016 across *speech styles* by *gender*. A positive value indicates that TRAP is raised above DRESS.

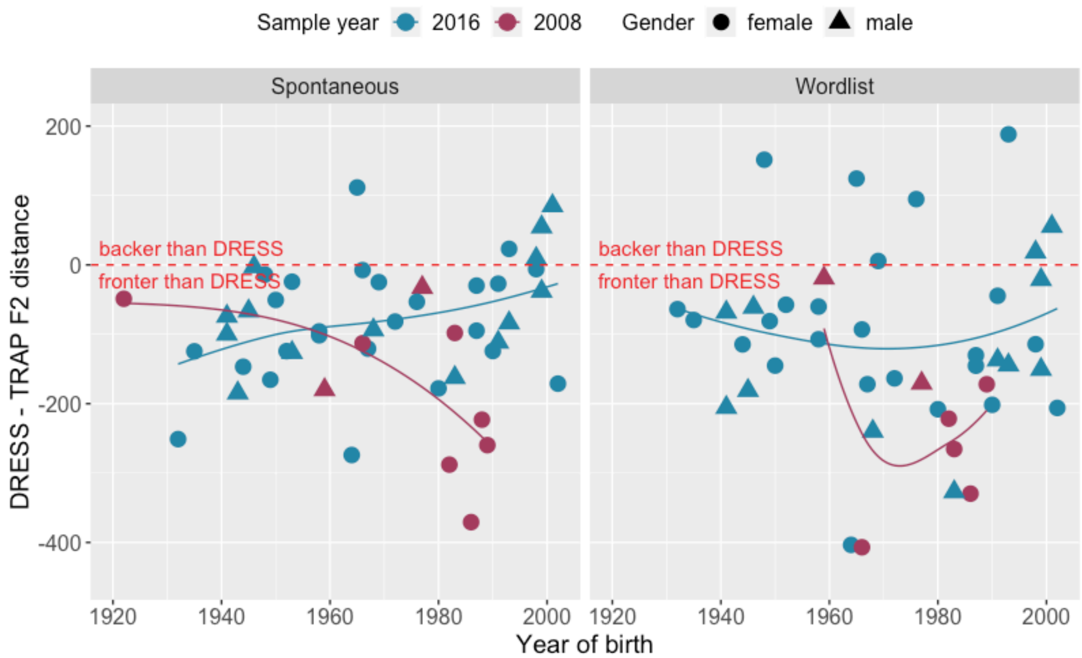


Figure 40: F2 distance between TRAP and DRESS F2 means in 2008 and 2016 across *speech styles* by *gender*. A negative value indicates that TRAP is fronter than DRESS.

The increased EQ participation in more careful speech among 2016 speakers in comparison to their spontaneous speech production contradicts the style-shifting patterns observed for F1 of TRAP. The analysis of TRAP revealed that wordlist TRAP is produced higher in the vowel space for older speakers, but lower for speakers born after 1960 in this sample. Additionally, wordlist TRAP was found to be lowering slightly in apparent time among 2016 speakers, thus moving away from DRESS on a downward trajectory, while no such lowering could be observed in spontaneous speech. Nevertheless, EQ participation, which requires TRAP to be produced above DRESS, is higher in wordlist style than it is in spontaneous speech. This contradiction resolves itself when taking into consideration changes in the height of DRESS. As will be described in more detail in Chapter 4.2.2, DRESS is lowering in apparent time in wordlist style alongside TRAP in 2016. However, DRESS lowering is progressing at a faster rate than that of TRAP: In the 70 years tracked in the data, the increase in F1 amounts to about 60 Hz for TRAP, and 100 Hz for DRESS. Thus, the phonetic distance between both phonemes is decreasing in apparent time, as shown in Figure 41. The same applies to spontaneous speech, however, DRESS does not seem to have surpassed TRAP on its downward trajectory in this speech style.

Figure 41: TRAP and DRESS F1 means in wordlist style in 2016 by *gender*.

3.2.4 Summary: TRAP and TRAMP in Production

The analysis presented above has shown that TRAP is undergoing significant changes in the community, especially in real time.

In spontaneous speech TRAP is lowering significantly in real time. Although, from the 2008 spontaneous speech data, more NCS raising of TRAP in 2016 would have been predicted, TRAP is less raised in 2016, with a trend in the opposite direction for the majority of speakers. As Figure 42 shows, speakers interviewed in 2016 produce TRAP notably lower than 2008 speakers, resulting in the lack of AE1 participation among 2016 speakers, while the majority of 2008 speakers do reach the 700 Hz threshold of this criterion. The 2016 data itself, however, indicates only insignificant apparent-time lowering and retraction of TRAP (seemingly led by college educated speakers), which suggests that the change away from NCS raised TRAP in this speech style was fairly sudden. Now, the relatively steady position of TRAP suggests a stable non-NCS TRAP system. While some of the speakers in 2016 still produce a relatively high TRAP in spontaneous speech, none of them produce it high enough to meet the 700 Hz mark, nor do any of them raise TRAP above DRESS in this speech style. All speakers, both in 2008 and 2016m realize TRAP in a fairly fronted position.

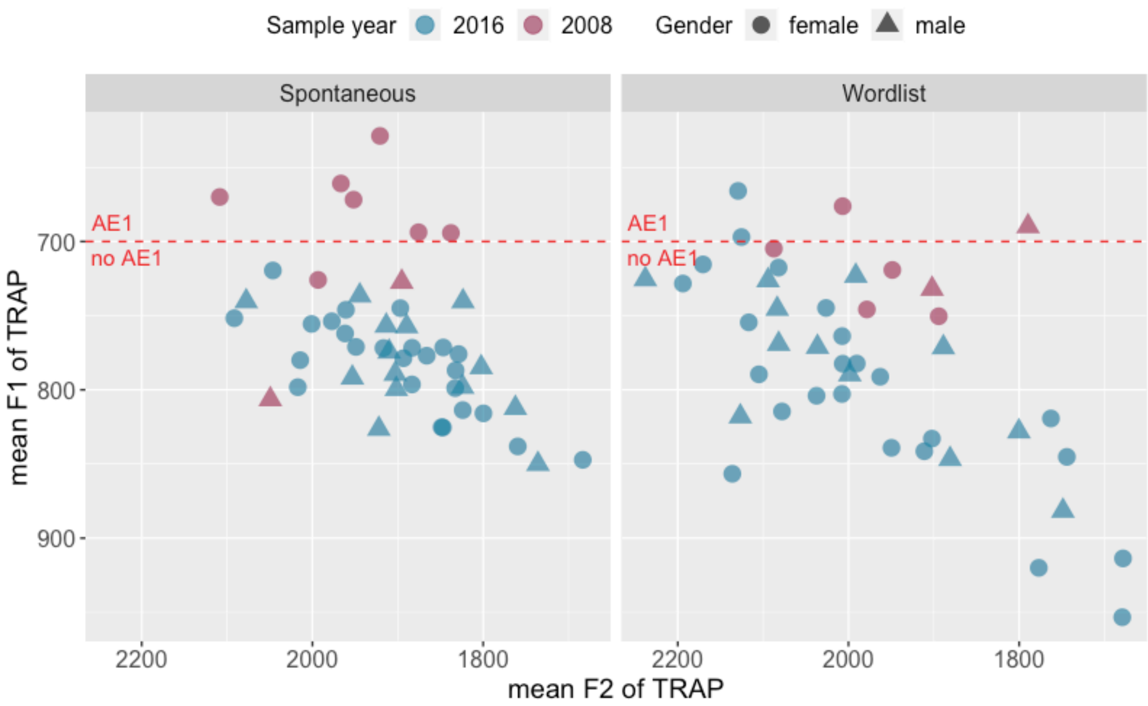


Figure 42: TRAP F1 and F2 means in 2008 and 2016 across *speech styles* by *gender*.

A similar picture emerges from the wordlist data. As Figure 42 above shows, 2016 speakers realize TRAP in a notably lower position than 2008 speakers in this speech style as well. Here, however, real-time lowering is accompanied by apparent-time lowering in both data sets. In other words, 2016 speakers appear to be continuing the trend that was suggested by the 2008 data. Additionally, 2016 speakers are retracting wordlist TRAP in apparent time, and the changes on both dimensions appear to be led by college educated speakers. As can be seen in Figure 42 above, wordlist TRAP tends to be produced lower and slightly fronter in the vowel space by the majority of speakers in both data sets than spontaneous TRAP. However, the analysis has shown that the shift to a lower TRAP applies only to younger speakers born after 1960. Those born before 1960 tend to shift in the opposite direction, especially males. Nevertheless, EQ participation is higher in this speech style than it is in spontaneous speech; however, this was found to be due to developments in the height of DRESS rather than TRAP.

The findings for the TRAP-DRESS distance in terms of height corroborate the findings for TRAP F1 in spontaneous speech, but not for wordlist TRAP. In spontaneous speech, the raising of TRAP in apparent time among 2008 speakers is reflected in an increasing positive distance between TRAP and DRESS, whereas the lack of such trend in the 2016 data is reflected in the unchanging distance between spontaneous TRAP and DRESS. The differences between 2008 and 2016 speakers regarding the height relation of TRAP and DRESS in spontaneous speech also match the differences found between the two samples for spontaneous TRAP F1.

In wordlist style, on the other hand, a slight increase in the F1 distance between TRAP and DRESS might have been expected, as wordlist TRAP seems to be lowering to a certain extent in apparent time in this style in both data sets. Furthermore, significant differences between speakers in the 2008 and 2016 data set were found in F1 of wordlist TRAP, with 2008 speakers producing a notably higher TRAP than most of the 2016 speakers. Neither of these observations, however, are not reflected in the TRAP-DRESS distance data, where 2008 and 2016 speakers appear to produce both phonemes at about the same height distance, and the TRAP-DRESS F1 distance does not appear to be undergoing apparent-time changes. These contradictions were found to be caused by changes in the height of DRESS.

Considering both the AE1 and EQ criterion in combination in Figure 43 below suggests even less NCS TRAP raising in the community than each criterion individually, as very few speakers fulfill both criteria. In spontaneous speech, out of the six speakers who meet the AE1 criterion, only four speakers also meet EQ. All of them belong to the 2008 sample. In wordlist style, only one speaker fulfills both criteria. She belongs to the 2016 sample. Thus, NCS TRAP raising seems to be more common among 2008 speakers than among 2016 speakers. However, none of the four 2008 speakers who meet both criteria do so in both speech styles, but in spontaneous speech only. In more careful speech, none of the four meet AE1, and only two of them meet EQ.

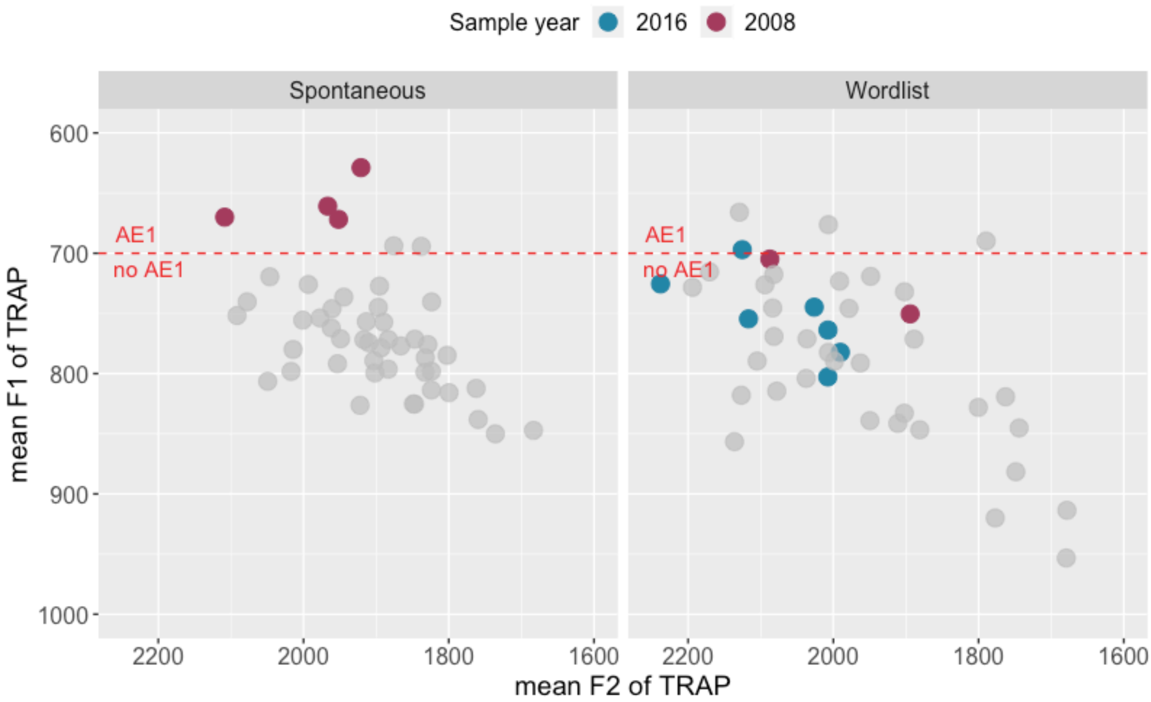


Figure 43: AE1 and EQ participation in 2008 and 2016 across *speech styles*. The colored dots represent speakers who meet the EQ criterion, the grey dots represent those who do not. The red dashed line marks the AE1 threshold.

TRAMP is undergoing changes in apparent and real time in the community as well. While, in 2008, TRAMP was raising significantly in apparent time, the majority of 2016 speakers no longer follow this trend. In fact, some of the speakers interviewed in 2016 appear to have reversed it and are in the process of lowering as well as retracting TRAMP. The majority of speakers participating in this trend, however, are those with a college education, while speakers without a college degree continue to raise TRAMP. In any case, TRAMP is realized significantly higher and fronter in the vowel space than TRAP and remains separate from its pre-oral counter-part in the community, as can be seen in Figure 44 below. In fact, the F1 distance between TRAMP and TRAP is increasing in real time. This indicates that TRAP is reconfiguring from a raised nasal or continuous system to a low nasal or continuous system.

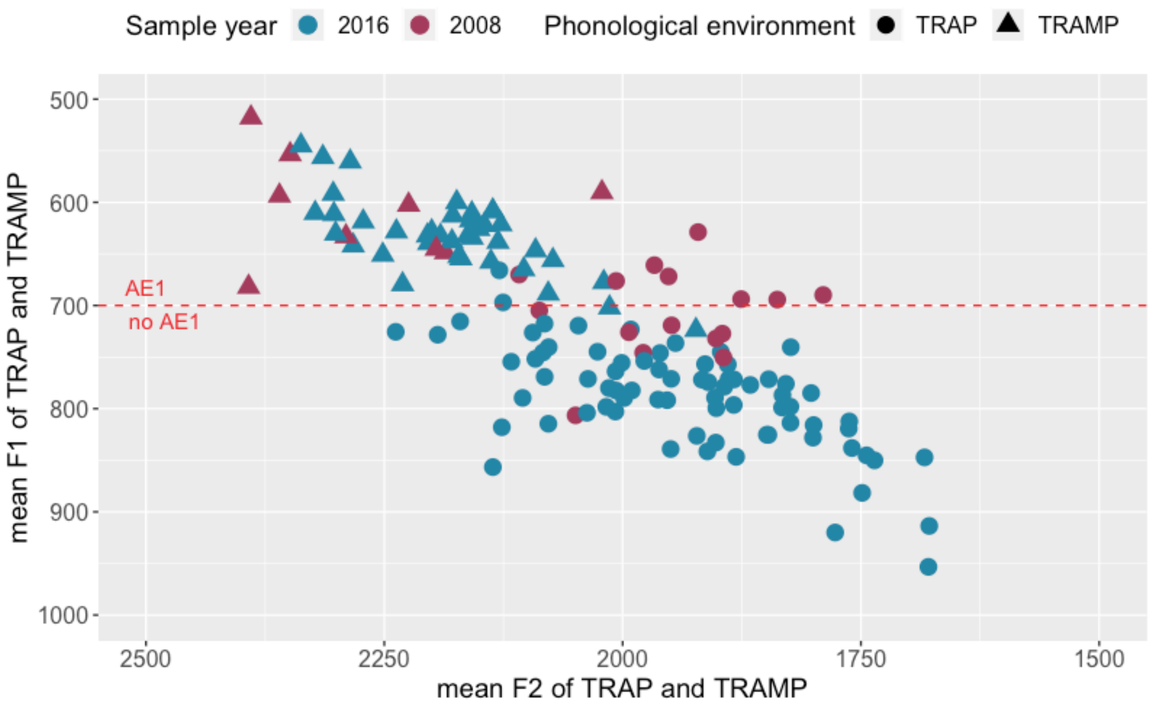


Figure 44: TRAP and TRAMP F1 and F2 means in 2008 and 2016.

3.3 The Social Evaluation of TRAP

Despite a tremendous amount of research into different production patterns for TRAP, its social perception has largely remained unexplored. It does, however, seem clear that TRAP has become a marker at least for some speakers in the Inland North. Savage et al. (2016) found evidence on social media that raised TRAP has attracted overt social commentary in the Inland North, and Driscoll and Lape (2015) found that, especially among younger speakers in Syracuse, raised TRAP, which has become known as "flat a", is a stigmatized variant. Their dislike of the raised variant appears to have increased to the extent that the local accent of Syracuse is described as "the worst" by Syracuse locals, and as one that at least one of Driscoll and Lape's participants likes to think they do not possess.

A similar awareness of "flat a" seems to have made its way to Ogdensburg in recent years. For example, Amber, born in 1967, went to school and has family in the Syracuse area, and describes their speech as sounding different:

> “*Their a’s sound different than ours.”*

Likewise, Richard, born in 1941, relies on raised realizations of TRAP to identify the local accents of Rochester and Syracuse:

> “*I can tell a Rochester accent and a, and a Syracuse accent, they have flat a’s … [beək] instead of [bæk], you know. My wife is from Syracuse, that's how I know.”*

Richard’s son, Patrick, born in 1968, also refers to the flat a’s in his mom’s speech, and extends the context of this pattern to Buffalo and Chicago. However, the example he provides does not refer to raised TRAP, but rather to fronted LOT (or potentially lowered and fronted THOUGHT, depending on whether *sausage* is more commonly realized with LOT or THOUGHT in the community, though impressionistically his imitation seemed to refer to fronted LOT rather than lowered and fronted THOUGHT):

> “*Well Syracuse had kind of a flat a and some slightly different pronounced words that, uh, I noticed growing up with my mom, since she was from Syracuse. Uh but ya, so I got the people in Buffalo kind of talked like a little bit … uh, like uh, it was kind of like the parodies you hear people talking in Chicago … You know [sa̟sɪdʒ].”*

These examples illustrate that there is at least some awareness of TRAP raising in Ogdensburg, and in combination with the change in progress in the effect of style shifting, they suggest that the retreat from raised TRAP in the community may be caused by people’s attitude toward raised TRAP: Raising is evaluated as less standard, leading speakers to retreat from it, especially in more careful speech. To test this hypothesis and examine whether variants of TRAP have attracted social meanings, the rating patterns of the matched guise data will be analyzed in the following subchapters. Out of the five categories tested, there are three for which the raised and unraised guises were rated differently: the perceived level of education, perceived localness and perceived age.

3.3.1 TRAP and its Perceived Level of Education

In the matched guise experiment, unraised TRAP received higher ratings, i.e. is perceived as sounding more educated, than raised TRAP. However, as Figure 45 suggests, older listeners do not seem to differentiate in their ratings between raised and unraised TRAP in terms of how educated they perceive it to sound. Only for participants born after 1960 does a differentiated rating pattern emerge.

Statistical analyses, summarized in Table 27 below, support these observations. The overall rating difference between the two guises is estimated at 0.3 units on a 6-point scale. This difference increases to 0.5 units in a model that considers only speakers born after 1960, while for those born before, the model finds no significant differences between the ratings of the raised and unraised guises. This suggests that younger listeners perceive lower TRAP to sounding more standard than its raised counterpart, while older listeners do not make this distinction.[64]

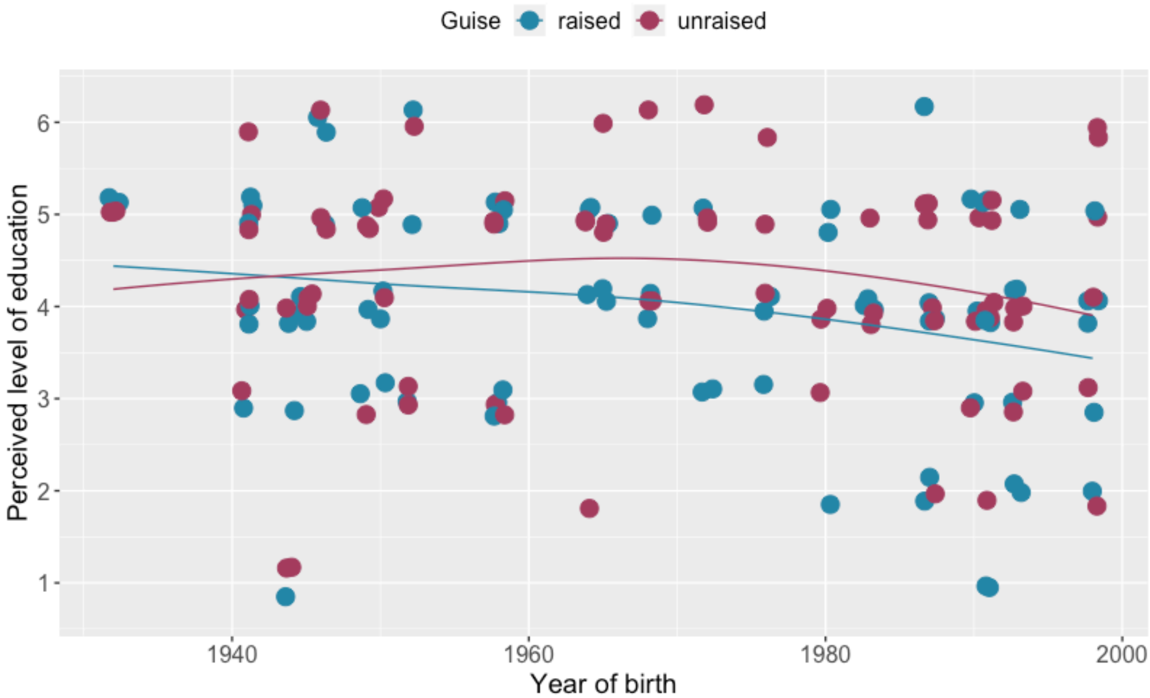


Figure 45: TRAP matched guise ratings for the perceived level of education.

[64] However, an interaction between *guise* and *age* does not reach the level of significance, and therefore is not included in the regression model shown here.

Predictor	Overall		born before 1960		born after 1960	
	Coefficient	*p*	Coefficient	*p*	Coefficient	*p*
(Intercept)	+7.079		−11.97		−4.882	
Age	+0.02	0.129	−0.051	0.752	+0.012	0.834
Gender (Male)	−0.523	0.221	−0.703	0.825	−1.626	0.455
Education *(No college)* *(Student)*	+0.066 +0.37	0.85	−1.185 ---	0.654	+0.345 +1.098	0.661
Spontaneous TRAP *F1*	+0.0003	0.98	+0.021	0.723	+0.014	0.612
Spontaneous TRAP *F2*	−0.002	0.509	+0.00003	0.998	−0.002	0.883
Style shifting F1	−0.001	0.788	−0.037	0.496	+0.005	0.749
Style shifting F2	−0.001	0.451	−0.009	0.729	−0.004	0.489
NCS score *(1)* *(2)* *(3)*	−0.93 +0.43 +0.252	0.115	+5.667 +4.606	0.534	+0.001 +0.763 +1.34	0.833
Voice *(R)* *(T)*	+0.22 +0.463	0.049	−0.108 +0.091	0.806	+0.438 +0.719	0.011
Guise *(unraised)*	+0.307	0.046	+0.019	0.939	+0.5	0.011

Table 27: TRAP matched guise ratings for the perceived level of education. Each pair of columns represents a different regression model: one on the whole data set (*n* = 161), and then separate models for speakers born before and after 1960. Reference levels: female, college educated, NCS score 0, voice J, raised guise. Random effect: *listener*.

The conclusion that younger listeners perceive raised TRAP to be less standard-sounding than unraised TRAP, while older listeners do not, needs to be preliminary, as the ratings for the perceived level of education among participants born after 1960 also appear to depend on *voice* (i.e. the three people who read the carrier phrases: J, R and T). As indicated in Table 27 above and visualized in Figure 46 below, the rating patterns for each individual voice deviate significantly from the combined results presented in Figure 45 above: The ratings for voice J roughly resemble the pattern from the overall results, in that older listeners rate his raised guise as more educated-sounding than his unraised guise, while for younger speakers, the ratings of the raised guise drop off sharply, leaving unraised TRAP as the guise with more favorable ratings. The ratings of voice R, on the other hand, show the opposite pattern, as the unraised guise is rated higher than the raised guise by listeners of most ages. For this voice, it is the unraised guise whose ratings decline for the youngest speakers, making the raised guise the higher rated one for this age group. As for voice T, the majority of listeners appear to agree that the unraised guise sounds more educated, though this difference in ratings for the raised and unraised guises increases for younger listeners. Thus, while voice R does not show the expected rating patterns, the ratings for voices T and J mirror the expected apparent-time change in the difference between the evaluation of raised and unraised TRAP, but from very different starting points.

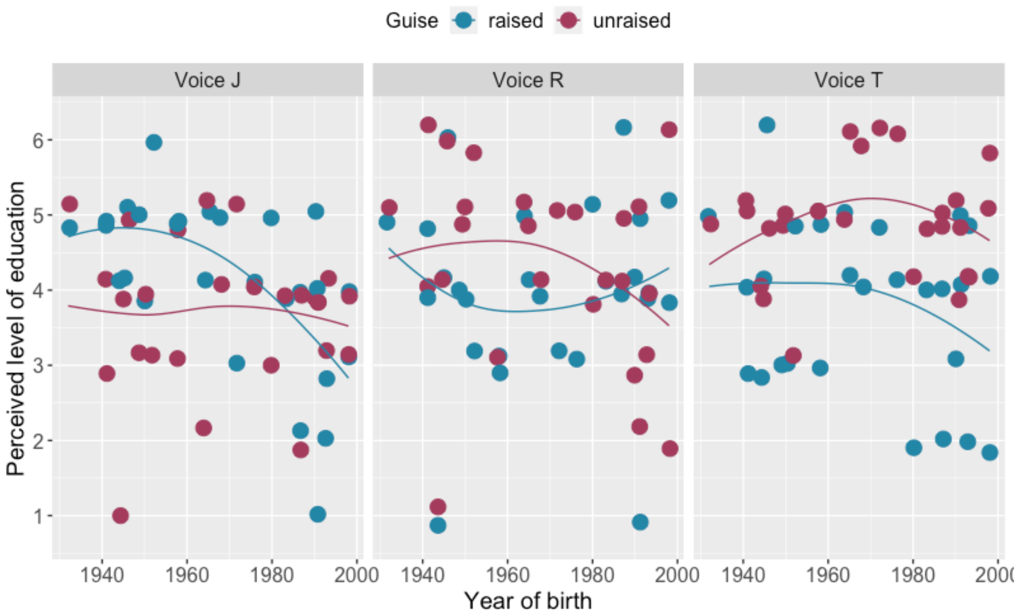


Figure 46: TRAP matched guise ratings for the perceived level of education across *voices.*

3.3.2 TRAP and Its Perceived Localness

The second category that yielded significant rating differences between the TRAP guises is how local to Ogdensburg the voices are perceived to be. The ratings for this category show the same overall pattern as education; however, here it is the raised realization of TRAP that generally received higher ratings (i.e. is perceived as more local-sounding). This can be seen in Figure 47 below. This effect appears to be mainly due to listeners born after 1960, the same cut-off year that was observed in the ratings for the perceived level of education. Participants born before 1960, on the other hand, do not differentiate between raised and unraised TRAP in their ratings for localness.

The regression models shown in Table 28 below support these observations.[65] The predicted overall rating difference between the raised and unraised guises is 0.5 units on a 6-point scale. For participants born before 1960, however, this differentiation in ratings nearly disappears, dropping to a non-significant 0.16 units, while for listeners born after

[65] Though, again, an interaction term between age of listener and guise does not reach the level of significance.

1960, it increases to an estimated 0.7 units. This suggests that raised TRAP came to be associated with the local dialect sometime in the mid–20th century, around the same time as it started to be perceived as the less educated way of producing TRAP.

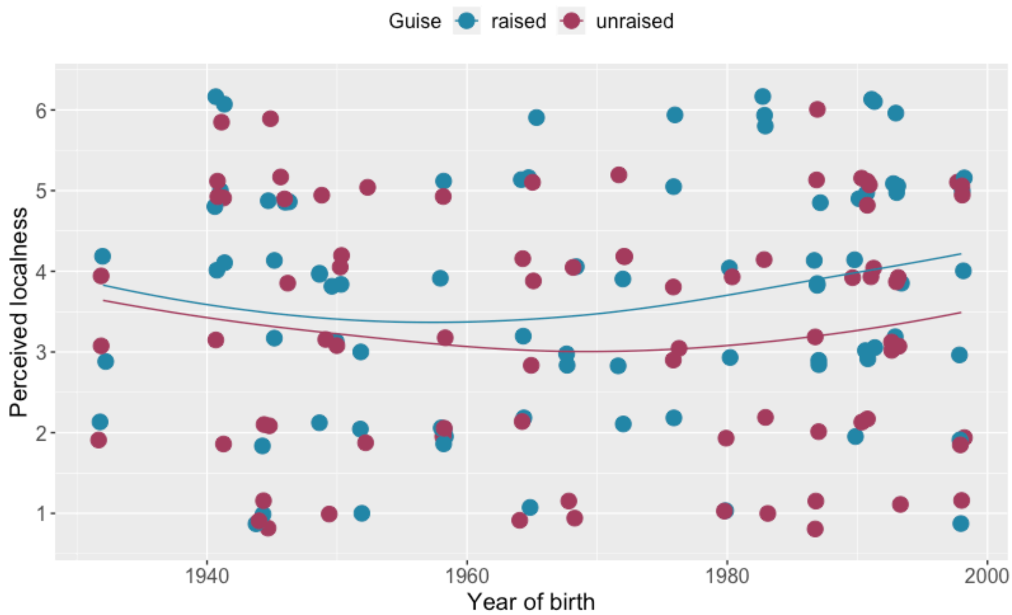


Figure 47: TRAP matched guise ratings for perceived localness.

Predictor	Overall		born before 1960		born after 1960	
	Coefficient	*p*	Coefficient	*p*	Coefficient	*p*
(Intercept)	+4.54		-20.28		+0.153	
Age	+0.002	0.863	−0.06	0.606	−0.05	0.399
Gender (Male)	+0.39	0.372	−0.89	0.675	+1.23	0.534
Education *(No college)* *(Student)*	+0.28 −0.52	0.544	−1.29 ---	0.487	+0.68 −0.66	0.682
Spontaneous TRAP F1	+0.004	0.7	+0.04	0.41	+0.003	0.903
Spontaneous TRAP F2	−0.002	0.577	−0.003	0.782	−0.007	0.985
ShiftingF1	+0.0003	0.948	−0.04	0.345	+0.006	0.693
ShiftingF2	+0.003	0.155	−0.005	0.756	−0.00009	0.985
NCS score *(1)* *(2)* *(3)*	−0.73 +0.53 −0.28	0.182	--- +6.96 +4.455	0.301	−0.81 +0.2 +0.78	0.679
Voice *(R)* *(T)*	+0.05 +0.53	0.079	+0.119 +0.45	0.421	$+6\times10^{-14}$ +0.56	0.192
Guise *(unraised)*	−0.49	0.023	−0.16	0.598	−0.71	0.016

Table 28: TRAP matched guise ratings for perceived localness. Each pair of columns represents a different regression model: one on the whole data set (n = 159), and then separate models for speakers born before and after 1960. Reference levels: female, college educated, NCS score 0, voice J, raised guise. Random effect: *listener*.

3.3.3 TRAP and Its Perceived Speaker Age

The third and last category that showed a significant amount of differentiation in the ratings of the TRAP guises is how old the voices are perceived to be. Again, the rating patterns for this category resemble those of the perceived level of education and localness. The raised guises of TRAP received higher ratings, suggesting that the voices are perceived as older when TRAP is realized in raised position. As can be seen in Figure 48, this effect again seems to be primarily due to listeners born after 1960, the same cut-off year that was observed for the other two categories. Participants born before 1960, on the other hand, do not differentiate between raised and unraised TRAP in their ratings for perceived age.

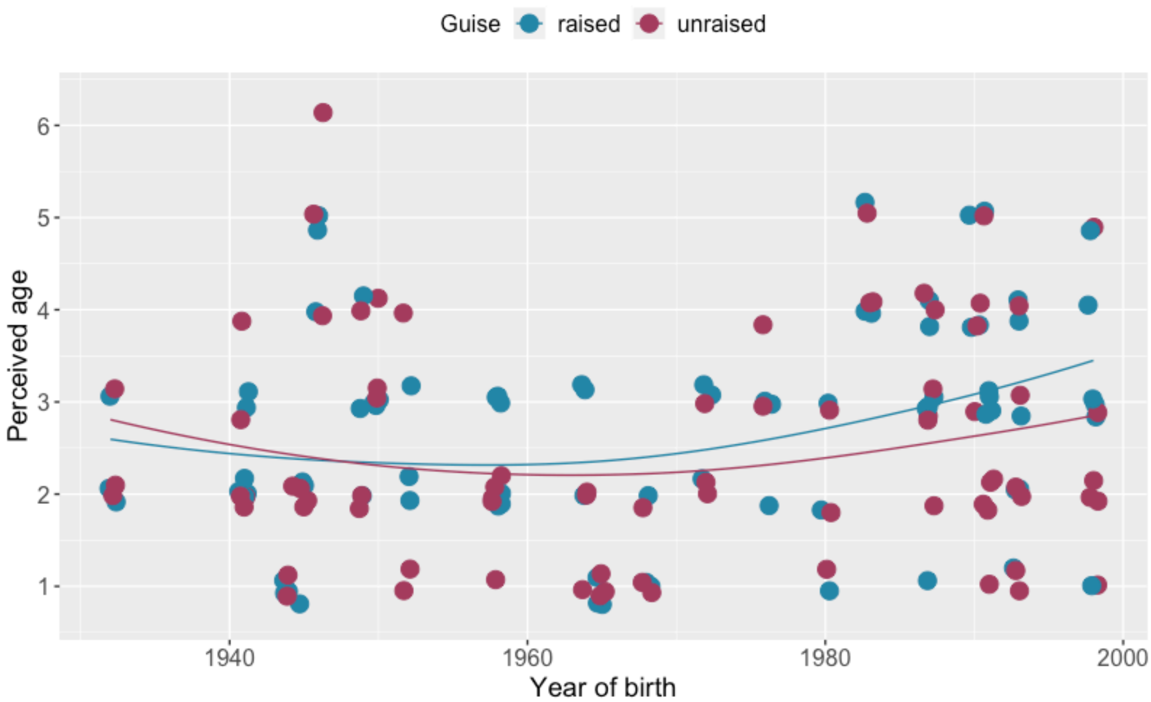


Figure 48: TRAP matched guise ratings for perceived age.

These observations are corroborated by the results of the regression models presented in Table 29 below.[66] The predicted rating difference between the raised and unraised guises is 0.2 units on a 6-point scale for the sample as a whole. While this differentiation minimizes to 0.07 units for the group of raters born before 1960, it increases to an

[66] Again, an interaction between *guise* and *age* does not reach the level of significance, and therefore is not included in the regression model.

estimated 0.4 units for raters born after 1960. This suggests that younger participants, i.e. those born after 1960, associate raised TRAP not only with less educated and local speakers, but also with older speakers.

Predictor	Overall		born before 1960		born after 1960	
	Coefficient	*p*	Coefficient	*p*	Coefficient	*p*
(Intercept)	2.218		1.049		-3.853	
Age	-0.014	0.433	-0.038	0.482	-0.027	0.599
Gender (Male)	0.108	0.853	-0.518	0.60	1.633	0.416
Education						
(No college)	0.791	0.376	-1.805	0.198	0.131	0.543
(Student)	-0.448		---		-1.187	
Spontaneous TRAP F1	0.001	0.92	0.015	0.471	-0.009	0.729
Spontaneous TRAP F2	-0.0001	0.985	-0.006	0.332	0.008	0.543
ShiftingF1	-0.003	0.662	-0.05	0.137	-0.014	0.386
ShiftingF2	-0.0001	0.973	-0.019	0.174	0.004	0.422
NCS score						
(1)	-0.033	0.698	---	0.063	-0.77	0.513
(2)	0.778		7.043		0.368	
(3)	0.225		5.48		-1.381	
Voice						
(R)	-0.283	0.097	0.189	0.056	-0.619	0.014
(T)	-0.292		-0.318		-0.282	
Guise (unraised)	-0.225	0.072	-0.025	0.881	-0.371	0.029

Table 29: TRAP matched guise ratings for perceived age. Each pair of columns represents a different regression model: one on the whole data set (*n* = 159), and then separate models for speakers born before and after 1960. Reference levels: female, college educated, NCS score 0, voice J, raised guise. Random effect: *listener*.

As indicated in Table 29 above, the ratings for perceived age among participants born after 1960 depend significantly on the voice in the guises, and to a nearly significant extent for participants born before 1960, so that the conclusion that younger listeners associate raised TRAP with older speakers, while older participants do not, needs to be preliminary. However, as illustrated in Figure 49 below, the rating patterns for all three voices closely resembles the pattern from the overall results presented above: Raised guises are rated higher, i.e. sounding older, than unraised guises by the majority of raters for all three voices. The main difference between the three voices is the cut-off year, i.e. the time when voices with a raised realization of TRAP came to be perceived as sounding older than those with unraised TRAP. For voice J, the timing appears to coincide roughly with the pattern presented above, with a cut-off year between 1960 and 1970. For voice R, raised guises came to be perceived as older sounding somewhat earlier, between 1930 and 1940. For voice T, on the other hand, the majority of raters, regardless of their own *age*, rate the raised guise as sounding older than the unraised guise. Thus, while the exact cut-off year might be difficult to determine owing to these differences across voices, the overall conclusion that voices with raised TRAP are perceived as sounding older than voices with unraised TRAP holds true for all three voices.

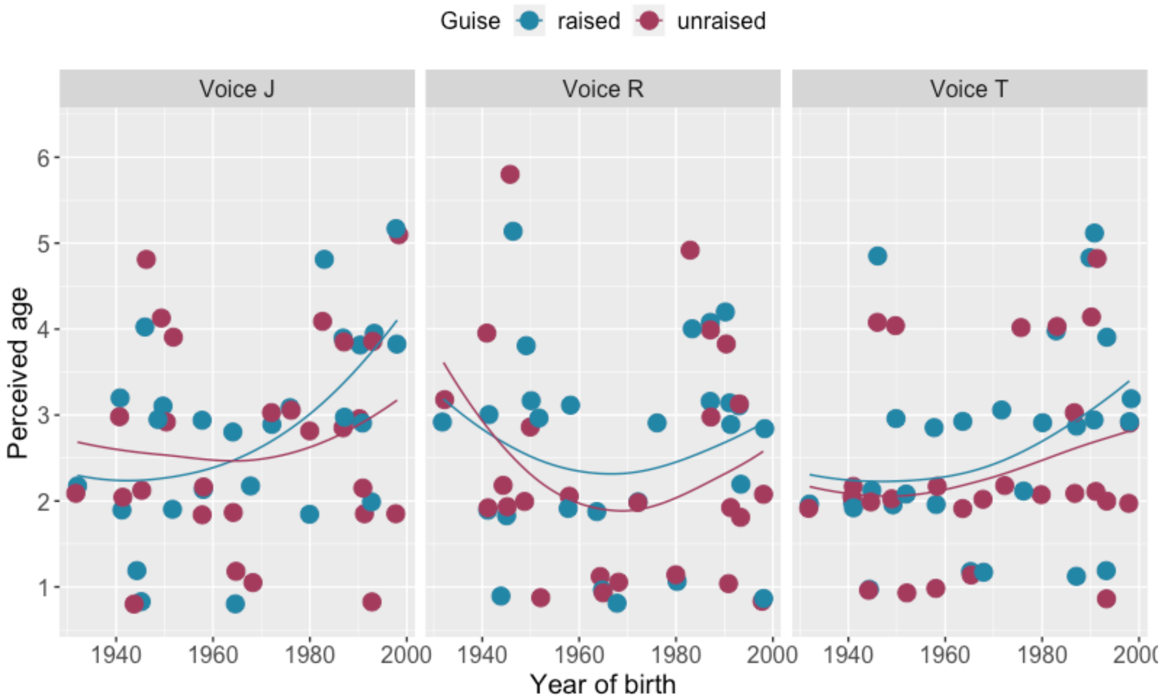


Figure 49: TRAP matched guise ratings for perceived age across *voices*.

3.3.4 Summary: Social Evaluation

The ratings for the perceived level of education, localness and age of raised and unraised TRAP indicate that social evaluation has started to form around raised realizations of TRAP. While older listeners do not seem to base their speaker judgments on the height of the speaker's TRAP vowel to a great extent, younger participants appear to associate the raised realization of TRAP with older, less educated locals. Interestingly, they do so despite the observation that raised TRAP by NCS standards seems to have been virtually abandoned in Ogdensburg in recent years, even among older speakers, as the speech data from the 2016 sample have suggested.

3.4 Discussion

The analyses of TRAP and TRAMP in production has shown that the community quit NCS TRAP raising quite abruptly, presumably sometime between 2008 and 2016, and that they are in the process of moving toward a low continuous or nasal system instead. Overall, these observations are in agreement with the developments in other Inland North communities, where general TRAP raising is disappearing, and a low nasal or continuous system is developing instead, and provide further evidence for an increasing social stigma that has formed around the raised realization of TRAP. The data do, however, raise questions about the time frame of the abandonment of raised TRAP.

3.4.1 Communal Change

The absence of any significant apparent-time trend away from raised TRAP in spontaneous speech is a puzzle, given the trend in wordlist style and the seemingly rapid real-time change from 2008 to 2016. Wordlist style shows what appears to be generational change; spontaneous speech, on the other hand, appears to display communal change due to virtually the majority of members of the community simultaneously adopting a linguistic innovation.

It seems likely that the apparent lack of an *age* correlation in the spontaneous speech of 2016 speakers is merely a statistical underestimation of the effect of *age* in the respective educational groups. As explained above, NCS-raised TRAP seems to be following two contrary trends: TRAP in spontaneous speech is lowering in apparent time among speakers with a college education, but is still raising among speakers without one. Although neither the difference between these two trends, nor the apparent lowering among college educated speakers reach the level of statistical significance, this pattern fits the expected profile of speakers with more access to social prestige being the first to retreat from a newly stigmatized variant. Recent research on the loss of NCS in communities such as Buffalo, Chicago, and Lansing has shown indices of TRAP-raising persisting longer in speakers with less education (Milholland, 2018) or blue-collar occupations (Durian & Cameron, 2018, Nesbitt, 2018). It seems likely that the same would apply to Ogdensburg, and that the significance of *age* and *education* is underestimated by the regression models. However, even taking seriously the different apparent-time developments among speakers with and without a college background in 2016 does not account for the sharp real-time differences between the 2008 and 2016 data.

A first potential explanation for differences observed between the 2008 and 2016 data are differences in methodological choices between the two data sets. As was explained in Chapter 2.4.1, the 2008 data were collected under different circumstances and analyzed using different techniques than the 2016 data. It is therefore possible that the differences observed between the two data sets are not real-time changes, but simply results of different methodologies. However, the observed patterns suggest that methodology hardly seems likely to account alone for the magnitude of the difference between the TRAP F1 measurements, which will be explored further in Chapter 8.2.1.

Thiel and Dinkin (under review) provide another potential explanation: The observed patterns might be an indication that the NCS was relatively new to Ogdensburg, even in 2008. The 2008 data indicates an apparent-time trend toward TRAP-raising in spontaneous speech, suggesting that younger speakers were in the process of acquiring the NCS at this time. However, soon after, they abandoned it via communal change due to its emerging stigma, first in more careful speech, then in spontaneous speech, while older speakers never had advanced TRAP-raising to begin with. Thus, the flat-seeming apparent-time profile is a result of the retreat from TRAP-raising among younger speakers' in combination with the unchanging, less-raised TRAP of older speakers.

3.4.2 Unraised TRAP as the Incoming Norm

As in other NCS communities, the reconfiguration of the TRAP system appears to be socially motivated. The results from the matched guise experiment suggest that unraised TRAP is perceived as the more standard variant in Ogdensburg, associated with younger, less local and more educated speakers. These findings are corroborated by production patterns, which becomes apparent by the slight lead of college educated speakers in the lowering and retracting of TRAP, as well as the style shifting data, which also include a female lead. Additionally, the apparent-time changes in attitudes toward raised and unraised TRAP as indicated by the matched guise ratings is in agreement with the apparent-time change observed in the direction of style shifting: Younger speakers favor unraised TRAP in their ratings, and are more likely to avoid raised TRAP in the more careful style.

The confidence in the validity of the rating patterns for the perceived level of education of raised and unraised TRAP is weakened because of the different rating patterns associated with the three different voices. The source of the differences in rating patterns between the three voices cannot be identified in the data. It is possible that they result from problems with authenticity or differences in the quality of the auditory stimuli

(Kircher, 2015). However, all TRAP stimuli received the same quality ratings in the design phase of the experiment, so that it seems unlikely that sound quality had a significant effect on the ratings. The validity of the rating patterns for the perceived level of education is *strengthened* by the observation that, in all three categories, i.e. perceived level of education, localness, and age, the average point in apparent time at which the two guises begin to be rated differently is about 1960, around the same year at which the direction of style shifting reverses in production, which suggests that the three perceptual categories are affected by the same underlying change. Thus, the rating patterns can be taken as evidence that raised TRAP came to be perceived as less standard than unraised TRAP in Ogdensburg starting with speakers born around 1960, while unraised TRAP has become the new perceived standard in the community. The cut-off year of 1960 in the evaluation data matches the apparent-time year when speakers in the 2016 data become more likely to produce a lower TRAP in more careful speech than they do in spontaneous speech. Thus, a growing negative evaluation of raised TRAP relative to unraised TRAP does seem to have been a likely motivation for speakers to have abandoned raised TRAP in favor of the non-local standard, unraised TRAP. This will be discussed further in Chapter 8.3.

If access to social prestige is in fact a relevant factor in the abandonment of raised TRAP in Ogdensburg, this raises the question of where this prestige is located, i.e. which dialect speakers in Ogdensburg are orienting toward. The most likely source appears to be the neighboring North Country, where an unraised nasal system has been found to be the predominant TRAP configuration (Dinkin, 2009). In other words, it seems that the unraised nasal system of the North Country is perceived as the new norm, and that it is adopted by Ogdensburgers through contagious diffusion.

However, if the unraised nasal configuration is in fact spreading to Ogdensburg by means of diffusion, a different *age* pattern would be expected. As the analysis has shown, pre-oral TRAP is lowering gradually in apparent time in wordlist style, with the youngest speakers leading this change. In spontaneous speech, evidence of gradual lowering is weaker, and it seems more likely that raised TRAP was abandoned by all speakers simultaneously regardless of *age*. Neither of these *age* patterns is what would be expected for a change that is adopted through diffusion, which is typically led by adults. The implications of this will be explored further in Chapter 8.3.2.

The hypothesis that the dialectological North Country is the source of the nasal TRAP configuration, and that this configuration is perceived as the new standard, presumes not only regular contact with this area, but also a positive orientation toward the North Country dialect area among Ogdensburgers. That residents of Ogdensburg have indeed been in continuous contact with the dialectological North Country has been established in Chapter 1.8, as many of them commute to the neighboring towns of Canton and Potsdam for work, school or leisure. Chapter 8.3.1 will expand on this argument and explore the possibility of a regional reorientation toward the North Country dialect area as a potential cause for the changes observed in the realization of TRAP.

Chapter 4: The DRESS Code

4.1 DRESS – An Introduction

Realization of DRESS vary across dialects throughout North America. While raising is one of the potential directions for DRESS shifting, found primarily in the South (here often merging with KIT in pre-nasal environments as part of the PIN-PEN merger), among African Americans and Mexican Americans, lowering and backing are the more commonly observed shifts for this vowel (E. R. Thomas, 2001). These trajectories have been observed in, e.g. Western New England, in areas affected by the Elsewhere/California/Canada Shift (*ANAE*) and, of course, in the Inland North. Nevertheless, lowering and backing of DRESS has been studied notably less than the raising of TRAP or the fronting of LOT.

As part of the NCS in the Inland North, the shifting of DRESS is believed to date back to the 1960s (*ANAE*), but was first reported in *LYS*, who found that it was not a particularly regular process. This conclusion is supported by e.g. Gordon (2001), who found very little shifting of DRESS in his Michigan data. NCS DRESS shifting is a multidirectional process, including backing, lowering, or a combination of both. The shift of DRESS was originally described as lowering toward TRAP (*LYS*), which is the pattern that was also found by e.g. McCarthy (2007), though she reports that some retraction is associated with lowering as well. Combinations of both directions were also found by *ANAE*, Dinkin (2009), and Gordon (2001). However, Gordon observed lowering and backing as separate processes in his data as well, and in fact more frequently than in combination, with backing toward schwa being the most dominant direction, which was also reported by *ANAE*. In some cases, *ANAE* reported, DRESS was still front of center and fronter than STRUT, but in the most extreme cases, DRESS was found to overlap with STRUT. Pure backing was observed only by Eckert (e.g. 1988). Labov (1994) argues that these differences in the direction of DRESS shifting reflect the development of this shift over time, with DRESS initially being lowered, but lowering being replaced by backing at a later stage. Eckert, on the other hand, suggests that different directions represent communal differences. Gordon (2001) argues that it appears more likely that various variants are available in NCS communities, and that the activation of one or the other depends on attitudinal factors.

NCS-like variants of DRESS in New York State were first observed by Emerson in 1891 in Ithaca. In a sample of 10 speakers, most likely born between 1811 and 1841[67], he identified some degree of DRESS lowering. A few decades later, C. K. Thomas (1935-1937) also reported centralizing tendencies for DRESS, in some cases interchanging with TRAP. Both directions of change were also identified by *ANAE* in New York State. In fact, all speakers in the New York part of the Inland North showed the highest degree of lowering and backing, while the rest of the Inland North was more varied in the *ANAE* data. Likewise, Dinkin (2009) found that the backing of DRESS was considerably more advanced in New York compared to other Inland North communities. In Ogdensburg, shifting of DRESS was found to still be ongoing in 2008, moving both back and downward in the vowel space (Dinkin, 2009). Only for Syracuse was DRESS found to be *fronting* among younger speakers (Driscoll & Lape, 2015).

The majority of studies that have examined the social stratification of DRESS in the Inland North have found similar patterns. In all studies, the lowering/backing of DRESS appeared to be an ongoing change, advancing in apparent time. One slight exception here is Gordon (2001), who found an overall lead of adults over younger speakers. However, this pattern referred only to backing; in the lowering or combined movement, younger speakers were more advanced than adults. An overall female lead was reported for the shifting of DRESS (*ANAE*; Clopper et al., 2005; Gordon, 2001; McCarthy, 2007). Again, Gordon's results are somewhat of an exception. Although he did find an overall female lead, he also reported a differentiation between males and females in the direction of shifting, with males preferring backed variants, while females employed all three options of DRESS shifting, i.e. backing, lowering or a combination of the two. Regarding DRESS shifting in relation to social class, studies have reported contradictory results. Eckert (e.g. 1988) found that working class-oriented students led over middle class-oriented students in the shifting of DRESS, while Gordon (2001) reported the opposite patterns for the girls in his study, and *ANAE* found higher levels of education to correlate with increased shifting of DRESS. In terms of social evaluation, a recent study by Savage and Mason (2018) found that lowering of DRESS is associated with positive characteristics such as intelligence, confidence, articulateness, and friendliness. They also reported positive effects of style, with increased lowering of DRESS in more formal speech styles. No one,

[67] Emerson (1891) provides, if any, only estimated ages or birth years for his participants.

however, has reported any overt social commentary on different realizations of DRESS. Hickey (2008) claims that lowered DRESS has reached a certain degree of social awareness based on discussion of its realization as TRAP discussed on the internet. However, he provides no evidence for this claim, and I have not been able to verify it in a google search.

4.2 Results: DRESS

Labov (2007) employs two criteria to define NCS-shifted DRESS on the height and the front-back dimension: the EQ and the ED criterion, which relate DRESS to the height and frontness of TRAP, and the frontness of LOT, respectively. As was described in Chapter 3.2, only few speakers fulfill the EQ criterion, indicating not only limited participation in NCS-TRAP, but also in NCS-DRESS. As the analysis showed, this lack of participation in the EQ criterion is mostly due to insufficiencies on the height dimension, since only few speakers raise TRAP above DRESS. In terms of F2, on the other hand, most participants fulfill this criterion in both spontaneous speech and wordlist style. In this chapter, apparent and real-time developments in F1 and F2 of DRESS will be analyzed in order to determine their role in the limited participation in EQ. Additionally, DRESS will be analyzed with respect to the ED criterion, i.e. its frontness in relation to LOT. Since the evaluation ratings from the matched guise experiment did not produce any noteworthy results for DRESS, the social evaluation of shifted and unshifted DRESS will not be discussed in this chapter.

Figure 50 below suggests a good amount of inter and intra-speaker variation in the realization of DRESS in the 2016 sample. While, in spontaneous speech, SUPPORTS THESE FINDS.ly similar spectral qualities by the majority of speakers regardless of *age*, it shows considerable variation in wordlist style. Older speakers tend to produce wordlist DRESS fronter and lower than spontaneous DRESS, while younger participants shift to a lower and retracted DRESS in more careful speech, creating a noticeable divide between speakers of different generations in wordlist style.

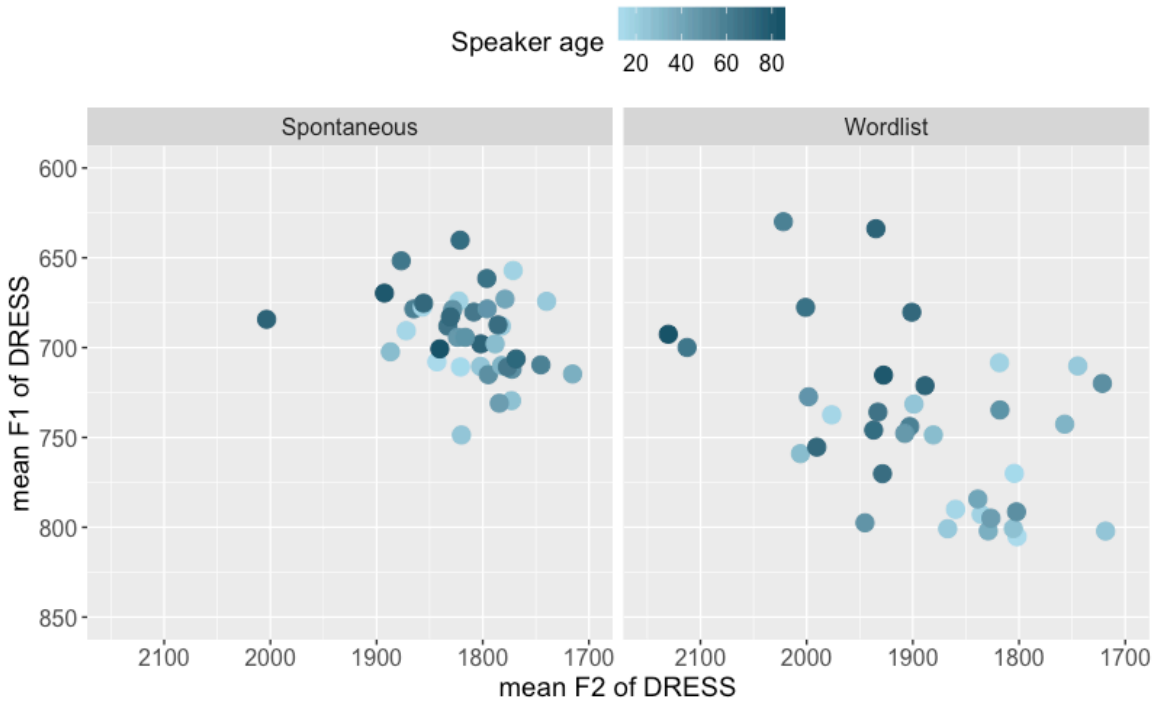


Figure 50: DRESS F1 and F2 means in 2016 across *speech styles* by *age*. Lighter shades represent younger speakers, darker shades older speakers.

4.2.1 DRESS in Spontaneous Speech

Figure 51 below illustrates that the majority of speakers in 2016 produce DRESS with an F1 between 650 and 750 Hz in spontaneous speech, with very few exceptions. However, Figure 51 suggests a slight lowering of spontaneous DRESS in apparent time, as younger speakers have slightly greater mean F1 than older speakers. What is striking, however, is that students, i.e. the youngest speakers in the sample, do not seem to participate in this trend at all; in fact, they appear to have reversed it quite abruptly. Although their F1 means still fall within the range of older speakers, there is a notable difference between students on the one hand, and non-students born after 1980 on the other.

The regression model presented in Table 30 below supports these observations: It predicts that spontaneous DRESS is lowering significantly in apparent time despite the slightly raised realizations among students, and that students produce a somewhat higher DRESS than other speakers. However, the estimated difference is relatively small and does not quite reach the level of statistical significance. The lack of a significant interaction between *age* and *education* suggests that the developments over time do not differ significantly across educational groups. This may, however, simply be a result of the small number of students in the 2016 sample.

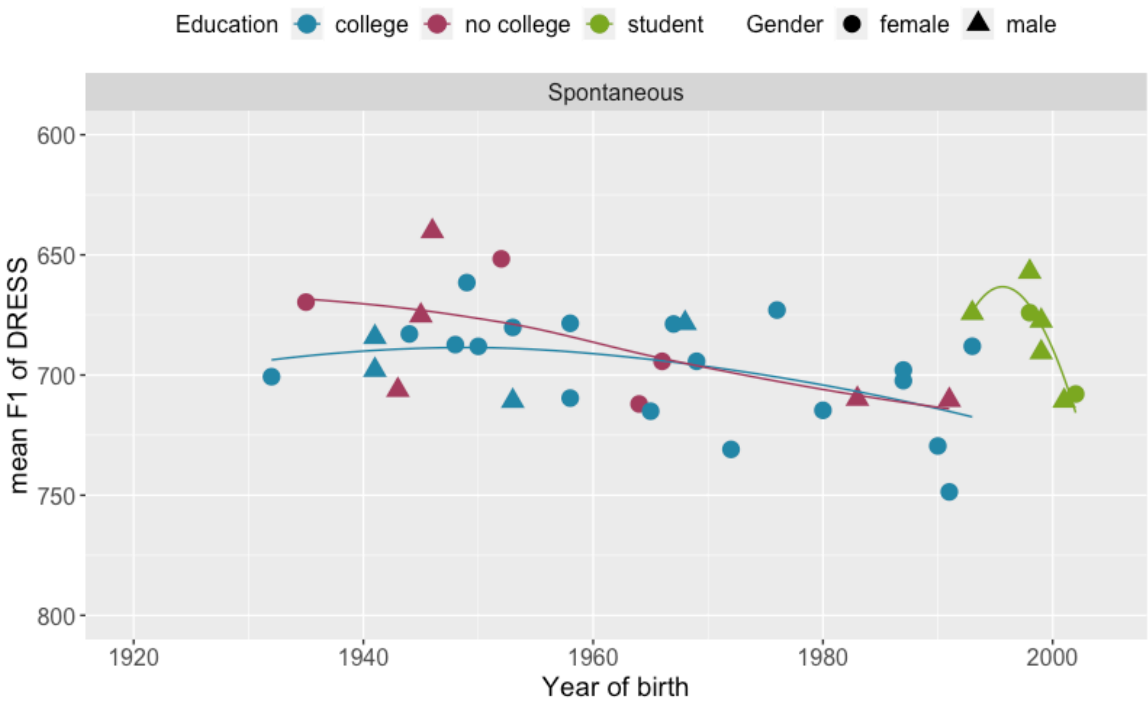


Figure 51: DRESS F1 means in spontaneous speech in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	782.465 Hz	
Age	-0.679 Hz	0.0001
Gender (Male)	-1.18 Hz	0.848
Education *(No college)* *(Student)*	-3.48 Hz -23.69 Hz	0.074
Environment		$2x10^{-16}$

Table 30: DRESS F1 in spontaneous speech in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word. n*= 8324

Variation in F2 of spontaneous DRESS mirrors the developments observed in F1 in the 2016 data. Figure 52 below shows that most speakers' means range from 1750 to 1850 Hz, though there appears to be a slight decrease in F2, i.e. backing of spontaneous DRESS in apparent time. Again, the students in the sample seem to behave slightly differently, most of them not following the apparent-time backing and instead producing a somewhat fronter DRESS than other speakers. With the exception of the abnormal behavior of students, the regression model in Table 31 below supports these observations, predicting significant backing of spontaneous DRESS in apparent time among 2016 speakers. The estimated differences between students and the two other educational groups are minor, and far from being statistically significant. Again, the lack of a significant interaction between *age* and *education* suggests that the differences in apparent-time trends across the three groups are not particularly substantial.

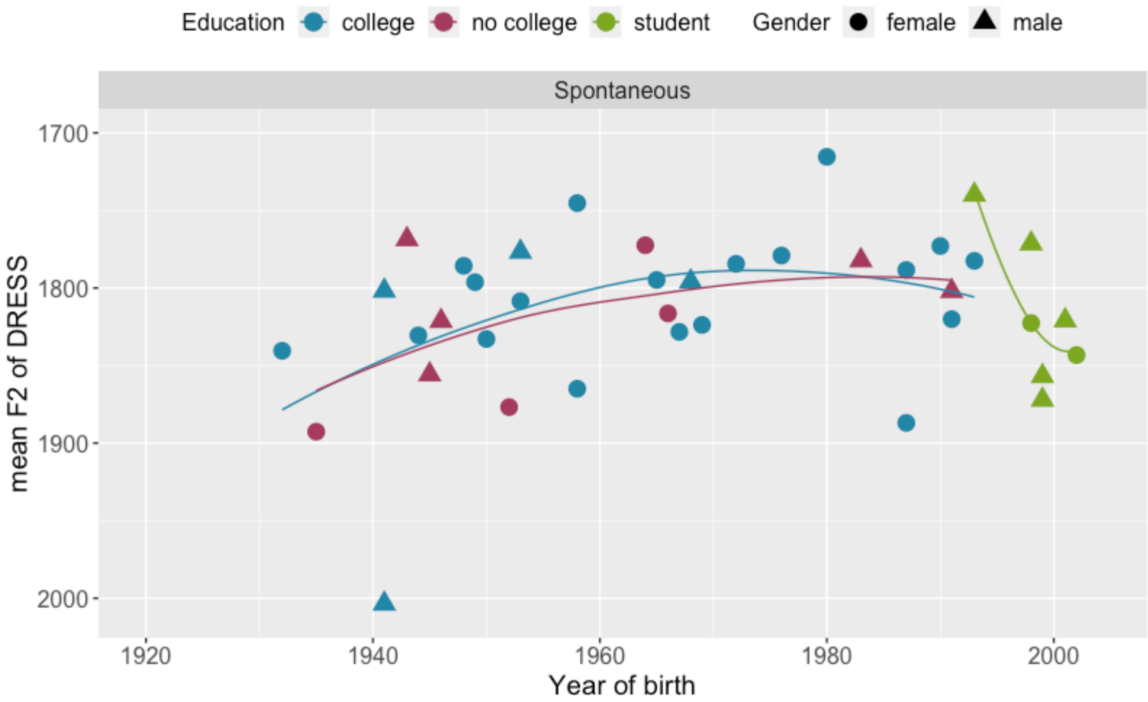


Figure 52: DRESS F2 means in spontaneous speech in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	1622.773 Hz	
Age	1.336 Hz	0.0002
Gender (Male)	15.493 Hz	0.216
Education *(No college)* *(Student)*	-11.367 Hz 9.476 Hz	0.55
Environment		2x10^{-16}

Table 31: DRESS F2 in spontaneous speech in 2016.
Reference levels: females, college educated, /p/.
Random effects: *speaker, word. n*= 8324

In real time, DRESS in spontaneous speech appears to be undergoing changes primarily on the front-back dimension, while in terms of height, differences between speakers from the 2008 and 2016 samples are minor, both in terms of mean formant values and apparent-time trends.

Although Figure 53 below suggests that 2008 speakers may have produced DRESS higher in the vowel space than 2016 speakers, this appears to be due to only one speaker, the oldest speaker in the entire sample, who produced DRESS with a notably lower F1 than any other speaker in this speech style. Speakers in the 2008 sample appear to be undergoing a similar slight lowering of DRESS as the speakers in 2016, though the extent of this shift appears to be minor. The regression model in Table 32 below supports these findings. Spontaneous DRESS is found to be lowering only slightly, but at a statistically significant level, and 2008 speakers produce a slightly, not quite significantly higher DRESS than 2016 speakers.

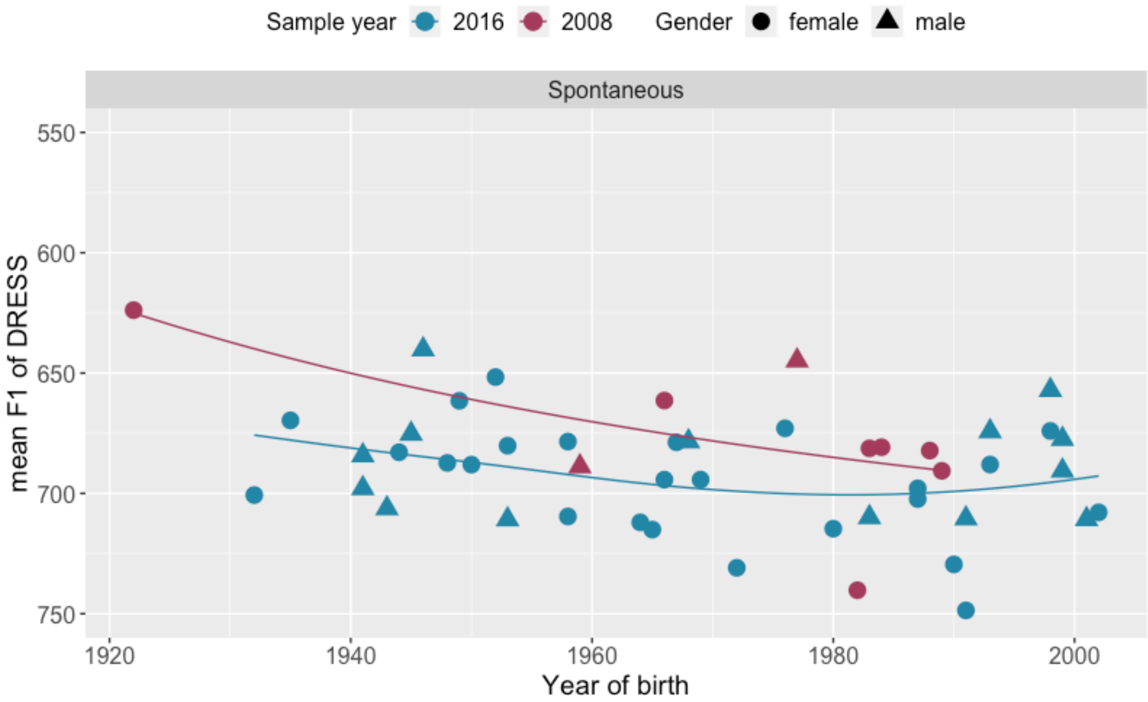


Figure 53: DRESS F1 means in spontaneous speech in 2008 and 2016 by *gender*.

In terms of frontness, there do appear to be notable differences in the realization of DRESS between speakers from the 2008 and 2016 sample, especially among younger speakers, as can be seen in Figure 54 below. This appears to be due to a faster paced apparent-time decrease in F2 in the 2008 sample compared to the 2016 sample. While spontaneous DRESS in the 2016 data is backing by about 13 Hz per 10 years in spontaneous speech, in the 2008 data, DRESS backing is proceeding notably faster at 31 Hz per 10 years (p= $9x10^{-7}$). Thus, younger speakers in 2008 produce a notably backer DRESS than 2016 speakers. The regression model in Table 33 below supports this finding. It estimates that 2008 speakers are backing DRESS about three times as fast as 2016 speakers, a difference that is found to be statistically significant. As a result, the entire 2008 sample is predicted to produce a significantly backer DRESS than 2016 speakers.

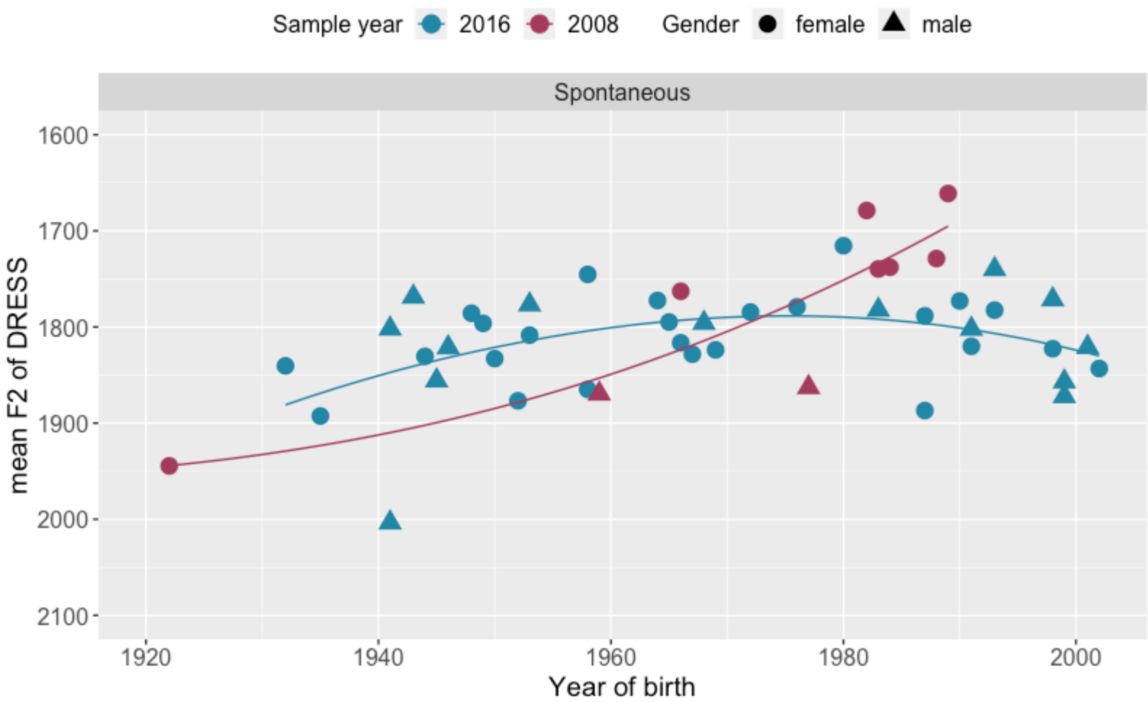


Figure 54: DRESS F2 means in spontaneous speech in 2008 and 2016 by *gender*.

Predictor	Coefficient	*p*
(Intercept)	770.94 Hz	
Age	-0.477 Hz	0.0002
Gender (Male)	-6.56 Hz	0.21
Sample year (2008)	-14.86 Hz	0.059
Environment		$2x10^{-16}$

Table 32: DRESS F1 in spontaneous speech in 2008 and 2016. Reference levels: females, 2016, /p/. Random effects: *speaker, word*. *n*= 8578

Predictor	Coefficient	*p*
(Intercept)	1625.26 Hz	
Age	1.184 Hz	
Gender (Male)	16.012 Hz	0.118
Sample year (2008)	-127.708 Hz	
*Age*2008*	2.088 Hz	0.007
Environment		$2x10^{-16}$

Table 33: DRESS F2 in spontaneous speech in 2008 and 2016. Reference levels: females, 2016, /p/. Random effects: *speaker, word*. *n*= 8578

These F2 differences between speakers in 2008 and 2016 in spontaneous speech impact participation in the NCS criteria that involve DRESS: EQ in relation to TRAP, and ED in relation to LOT. Participation in EQ has been discussed in Chapter 3.2 and is summarized again alongside ED in Figure 55 below. As the analysis of EQ in spontaneous speech has shown, TRAP tends to be higher and fronter than DRESS for many 2008 speakers. On the other hand, for 2016 speakers, spontaneous TRAP is fronter, but not higher than DRESS.

In relation to LOT, Figure 55 below suggests a relatively small F2 distance for some of the speakers in 2008, indicating at least some ED participation, while for 2016 speakers, DRESS and LOT appear to be more or less securely distinct, though still close. According to the ED criterion, speakers with a DRESS-LOT F2 distance of less than 375 Hz are regarded as NCS shifted. In the present data sets, a total of 32 speakers fulfill this criterion in spontaneous speech, 25 of them from the 2016 sample, seven from the 2008 sample.

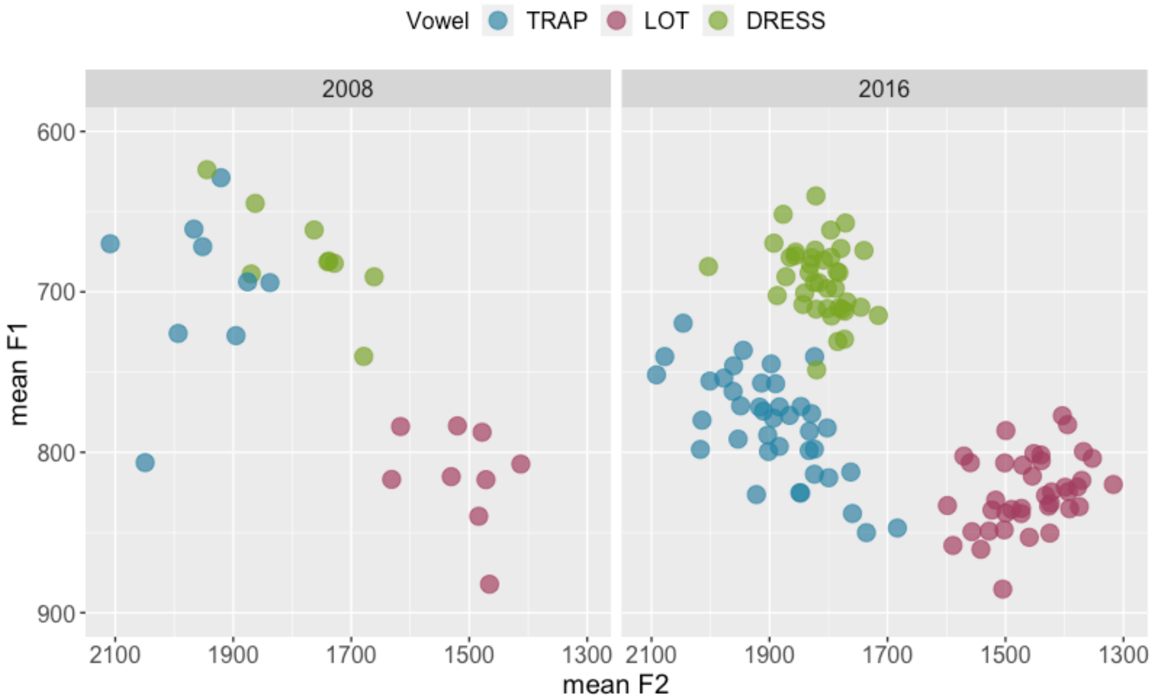


Figure 55: DRESS in relation to TRAP and LOT in 2008 and 2016 in spontaneous speech.

In the 2016 data, ED participation in spontaneous speech appears to depend primarily on the speakers' *level of education*. As can be seen in Figure 56 below, college educated speakers tend to have greater F2 distances between DRESS and LOT than speakers without a college degree in both spontaneous speech and wordlist style. On average, the difference between the two educational groups is about 80 Hz. While this is a relatively small number on the F2 dimension, it results in proportionally less participation in the ED criterion among college educated speakers than speakers without a college degree. In fact, with the exception of the oldest male, all speakers in the latter group meet the ED criterion in this speech style, and it seems that this group of speakers might be in the process of reducing this distance further in apparent time, while no such trend can be observed for college educated speakers.

These differences between the two educational groups contradict the findings for F2 of DRESS for 2016 speakers, where no significant differences in DRESS F2 could be found between the two groups; however, inter-speaker variation in the frontness of LOT explains these differences. As will be presented in detail in Chapter 6.2.3.1.1, college educated speakers produce a backer LOT and are retracting LOT in apparent time, while speakers without a college degree produce a fronter LOT without any indications of retraction. Consequently, by retracting both vowels, college educated speakers are maintaining the distance between DRESS and LOT in spontaneous speech and thus do not meet the ED criterion as frequently. Speakers without a college degree, on the other hand, are backing DRESS without retracting their somewhat fronted LOT, thus decreasing the distance between both vowels, resulting in higher ED participation. The students in the 2016 sample are relatively evenly distributed with regard to their ED participation in spontaneous speech. Four of them meet this criterion, while three of them do not. This pattern is similar to the patterns that can be observed in the production of both DRESS and LOT (see Chapter 6.2.3.1.1) for students.

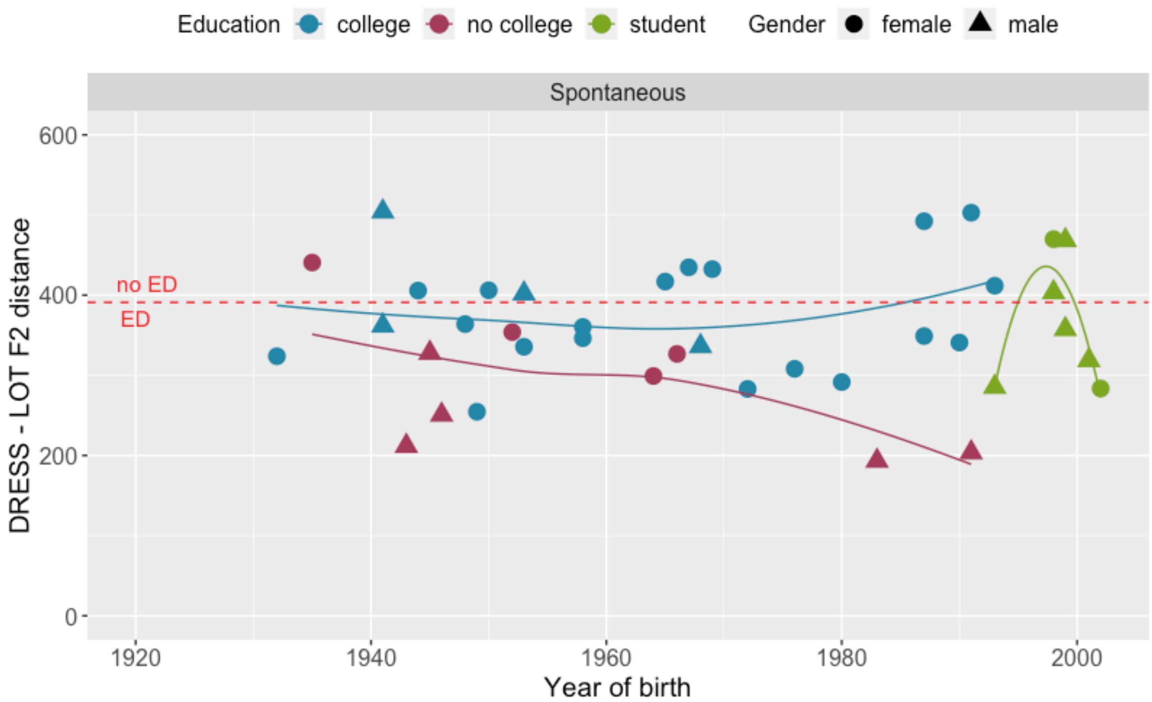


Figure 56: F2 distance between DRESS and LOT means in spontaneous speech in 2016 by *education* and *gender*.

In comparison to the 2016 data, 2008 speakers appear to produce spontaneous DRESS and LOT at a smaller and continuously decreasing distance on the front-back dimension. Thus, as shown in Figure 57 below, all but two of the 2008 participants fulfill ED in spontaneous speech. The older speakers among them seem to produce spontaneous DRESS and LOT with an F2 distance of 200 to 500 Hz. The youngest speakers in the 2008 sample, on the other hand, produce DRESS and LOT at a distance as little as 120 Hz in this speech style. This decrease in the F2 distance between DRESS and LOT is similar to that observed for speakers without a college degree in the 2016 data. These findings match the differences in the frontness of DRESS between speakers from the two samples in spontaneous DRESS presented above. While F2 DRESS means do not differ notably between both samples for older speakers, they diverge to a certain extent as speakers get younger due to the faster paced apparent-time backing of DRESS in the 2008 data. LOT does not appear to influence this pattern to a great extent, as there are only slight differences between the two samples in the frontness of LOT in spontaneous speech (see Chapter 6.2.3.1.1).

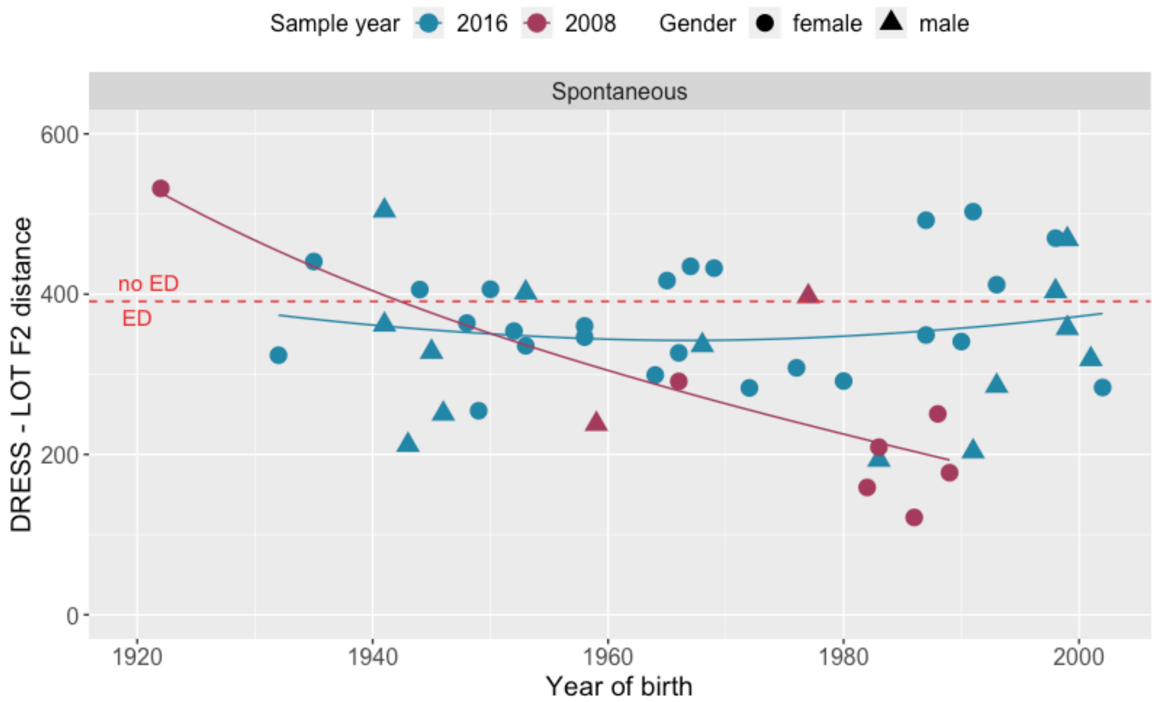


Figure 57: F2 distance between DRESS and LOT means in spontaneous speech in 2008 and 2016 by *gender*.

4.2.2 DRESS in Wordlist Style

In the 2016 wordlist data, substantial apparent-time changes in the height of DRESS can be observed. Figure 58 below illustrates that, while the oldest 2016 produce wordlist DRESS with an F1 of about 700 Hz, many of the younger speakers in this data set produce it as low as 800 Hz. The regression model presented in Table 34 below confirms this observation, predicting a significant lowering of DRESS in apparent time in wordlist style for the 2016 sample. As was noted in Chapter 3.2, this significant lowering entailed the reversal of the relative height of TRAP and DRESS for some of the speakers in the sample, i.e. they lowered their mean DRESS to a position that is lower than their mean TRAP.

Figure 58 suggests that this lowering of DRESS is somewhat more substantial among college educated speakers and/or females than it is for (male) speakers without a college degree. Whether *education* or *gender* is a more influential factor is difficult to discern from the representation in Figure 58. While it does seem that speakers without a college degree are not participating notably in DRESS lowering, the two youngest speakers in this group are males, which may indicate an effect of *gender* rather than education. The *gender* distribution in the student group supports this to a certain extent. While male students range in their F1 means between those of younger speakers with and without a college degree, the two female students clearly align with the youngest college educated females.

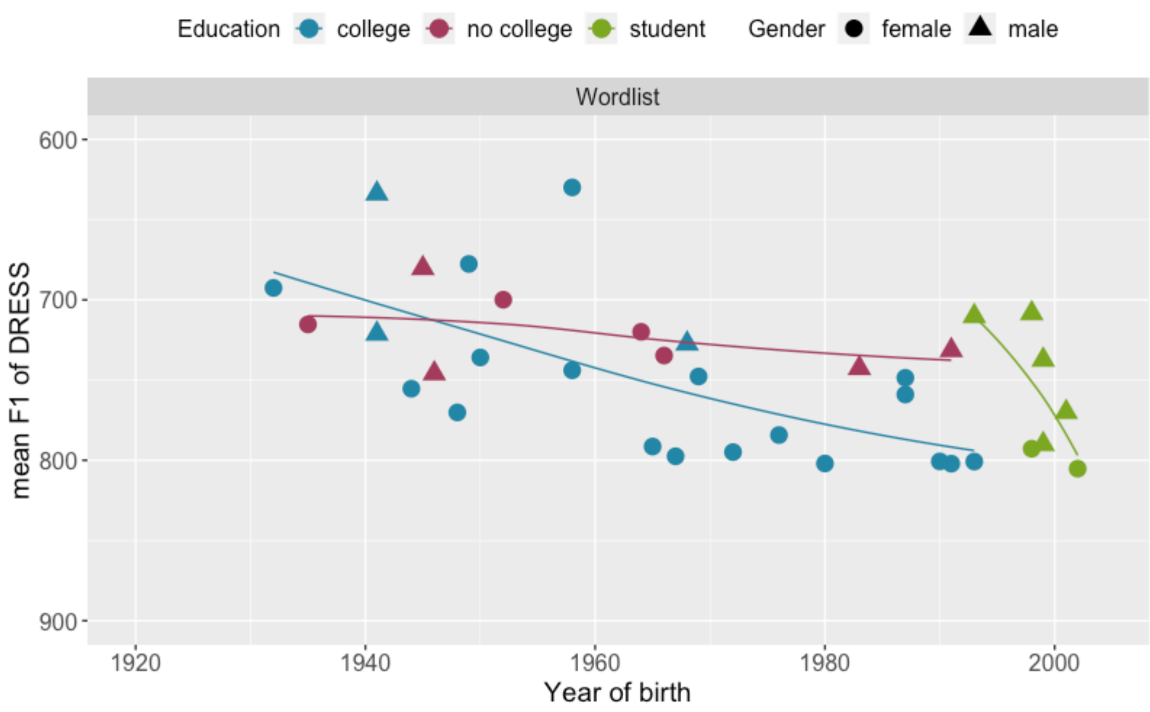


Figure 58: DRESS F1 means in wordlist style in 2016 by *education* and *gender*.

The regression model in Table 34 below assigns the bulk of the difference to *gender*, and only a small amount to *education*, suggesting that this is a female-led change, rather than one led by more educated speakers. However, the estimated coefficients for both factors are relatively small, and not statistically significant. A model that excludes *education*, however, does find *gender* to have a significant effect (-37.66 Hz, p= 0.006), while a model that excludes *gender* does not find a significant effect of *education*. Thus, a slight female-led change appears to be more likely than a change led by college educated speakers, though the significance of this questionable. Either way, a similar and significant apparent-time lowering of wordlist DRESS is predicted by all of these models, regardless of who is or is not leading this trend.

Predictor	Coefficient	*p*
(Intercept)	876.928 Hz	
Age	-1.43 Hz	0.001
Gender (Male)	-23.771 Hz	0.191
Education (No college)	-9.775 Hz	0.568
Environment		0.064

Table 34: DRESS F1 in wordlist style in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker*, *word*. n= 294

On the front-back dimension, inter-speaker variation in wordlist DRESS appears to relate mostly to *age*, regardless of *gender* or *education* in 2016. Figure 59 shows that younger 2016 speakers produce a notably backer wordlist DRESS than older speakers, suggesting apparent-time DRESS backing. While older speakers averaged around 1950 Hz, the majority of younger speakers have F2 means that are around 100 Hz lower. The regression model in Table 35 below confirms the significant of this apparent-time trend.

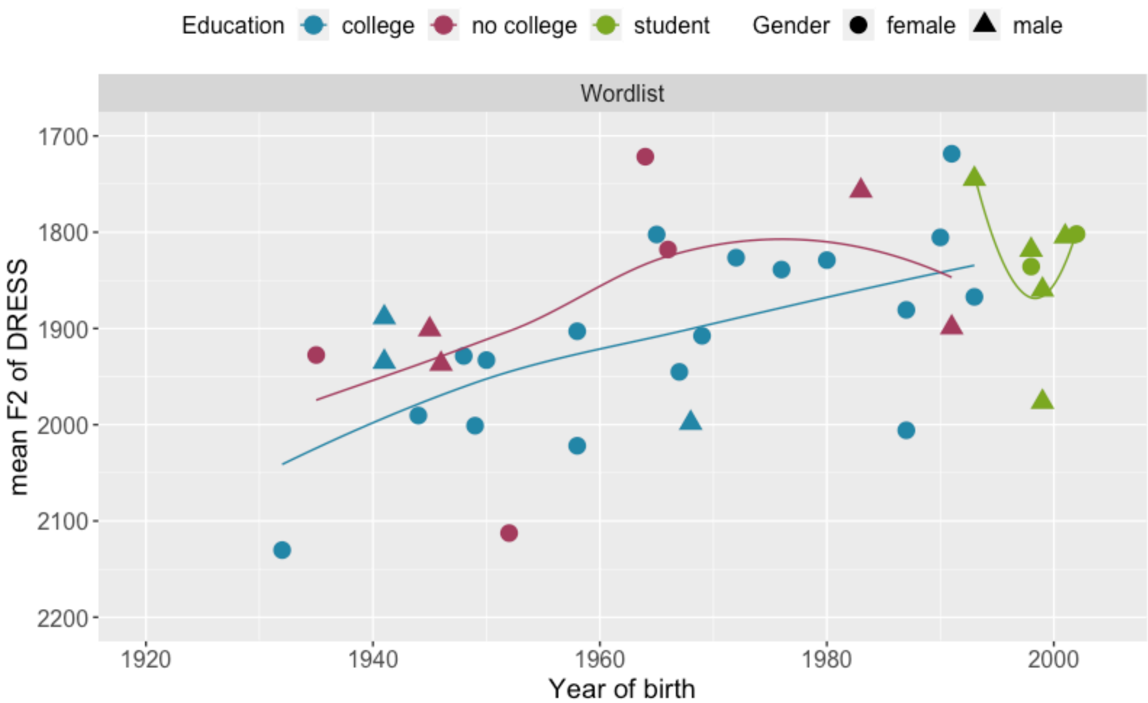


Figure 59: DRESS F2 means in wordlist style in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	1636.573 Hz	
Age	3.089 Hz	0.002
Gender (Male)	-13.299 Hz	0.742
Education (No college)	-48.096 Hz	0.22
Environment		0.041

Table 35: DRESS F2 in wordlist style in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word. n*= 294

In real time, the realization of DRESS does not appear to be changing significantly in wordlist style, both in terms of height and front-backness.

In terms of F1, 2008 speakers produce DRESS with a slightly smaller F1, i.e. higher in the vowel space, than most 2016 speakers. Additionally, Figure 60 suggests that 2008 speakers participate in the same lowering of DRESS in wordlist style as 2016 speakers, lending real-time evidence to the apparent-time trends in the data.

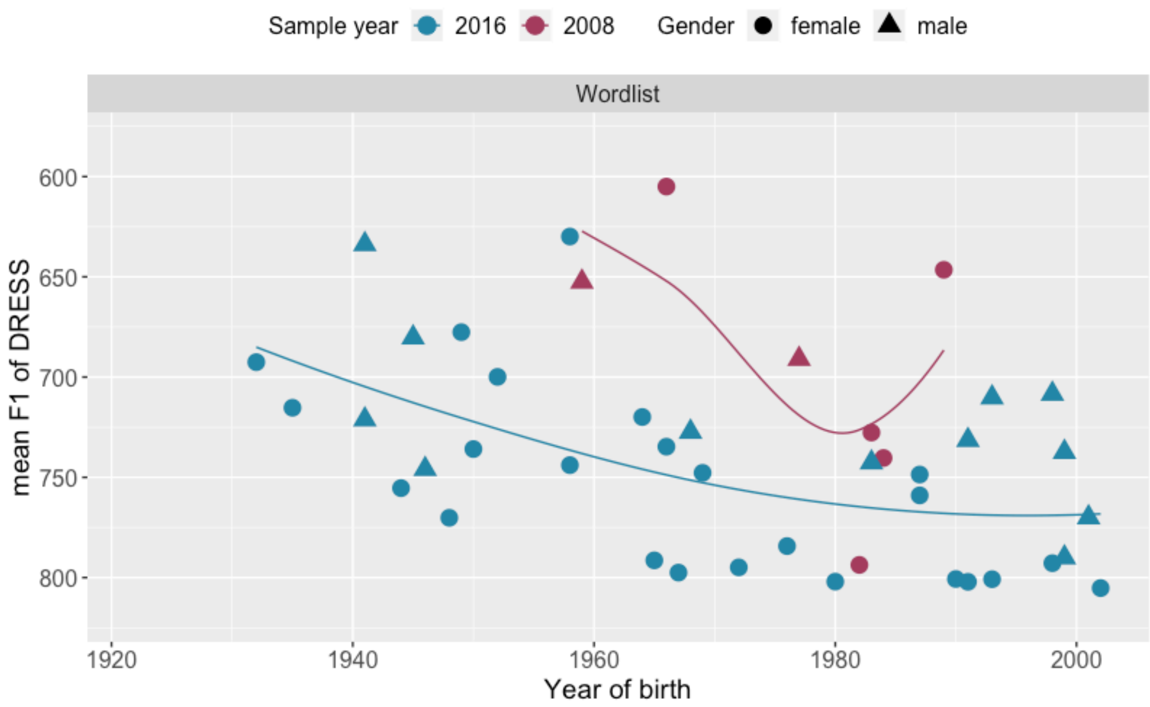


Figure 60: DRESS F1 means in wordlist style in 2008 and 2016 by *gender*.

The regression model in Table 36 below supports this conclusion of DRESS lowering. Adding the 2008 sample to the regression appears to strengthen the effect of *gender* on the height of wordlist DRESS, with females producing a significantly lower DRESS than males.

Though the estimated difference in Table 36 is still relatively small, it does suggest that DRESS lowering is a female-led change.[68]

Real-time differences on the front-back dimension for wordlist DRESS are similar to those observed for its height. As can be seen in Figure 61, two of the 2008 speakers clearly stand out from the rest of the sample, producing a notably backer DRESS than any other speakers. These are the same 2008 speakers who also produce a notably higher DRESS than most other speakers in this speech style. Otherwise, however, the two samples do not appear to differ significantly from each other. Although the regression model in Table 37 below estimates a substantial F2 difference between 2008 and 2016 speakers, this appears to be due to the two outliers in 2008 and is not statistically significant. In a model that excludes these two speakers, the estimated difference reduces to -130.81 Hz (*p*= 0.346), which is still higher than what Figure 61 suggests, but notably lower than the 210 Hz estimated by the model in Table 37 below. The apparent-time trend of DRESS retraction in this speech style holds for the combined samples at a statistically significant level.

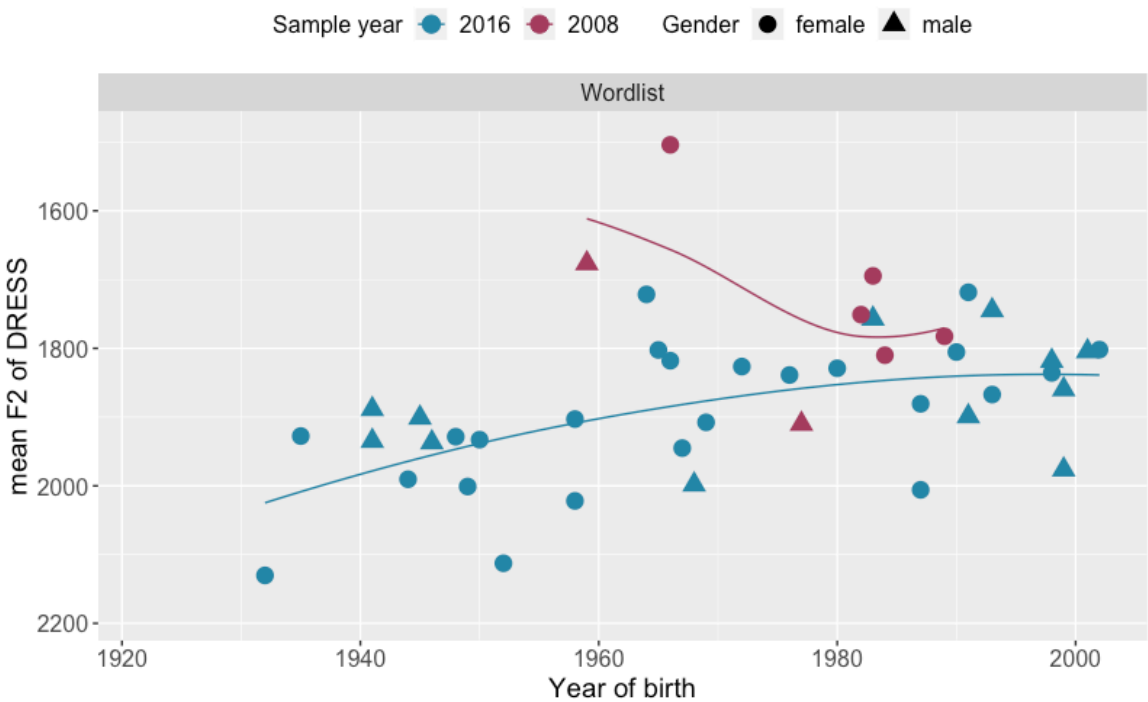


Figure 61: DRESS F2 means in wordlist style in 2008 and 2016 by *gender*.

[68] The same effect of *gender* holds if *education* is added to the model (-32.923 Hz, *p*= 0.015), while for education, the model does not predict a significant effect (-3.464/-15.886 Hz, *p*= 0.696).

Predictor	Coefficient	*p*
(Intercept)	864.634 Hz	
Age	-1.25 Hz	$4x10^{-5}$
Gender (Male)	-36.519 Hz	0.003
Sample year (2008)	-44.109 Hz	0.382
Environment		0.229

Table 36: DRESS F1 in wordlist style in 2008 and 2016. Reference levels: females, 2016, /p/. Random effects: *speaker, word*. *n*= 415

Predictor	Coefficient	*p*
(Intercept)	1684.838 Hz	
Age	2.316 Hz	0.001
Gender (Male)	-5.273 Hz	0.848
Sample year (2008)	-210.007 Hz	0.162
Environment		0.073

Table 37: DRESS F2 in wordlist style in 2008 and 2016. Reference levels: females, 2016, /p/. Random effects: *speaker, word*. *n*= 415

Although F2 differences in wordlist DRESS are minor, participation in the NCS criteria involving DRESS differs notably among speakers of the two different samples. Participation in EQ in wordlist style has been discussed in Chapter 3.2.2 and is summarized again alongside ED in Figure 62. As the analysis of EQ has shown, TRAP is frequently realized in a position higher and fronter than DRESS among 2008 and 2016 speakers.

In relation to LOT, DRESS appears to be produced only slightly fronter by some of the 2008 speakers, indicating at least some ED participation, while in 2016, DRESS and LOT appear to be more or less securely distinct. In the combined data, a total of 12 speakers have an F2 distance of less than 375 Hz between the two vowels in wordlist style to meet the ED criterion. Five of them belong to the 2016 sample, seven to the 2008 sample.

Figure 62: DRESS in relation to TRAP and LOT in wordlist style in 2008 and 2016.

In 2016, participation in the ED criterion is rare in wordlist style, as can be seen in Figure 63. Although males tend to have slightly smaller F2 distances between DRESS and LOT, this difference does not notably effect ED participation. In fact, three of the five speakers who do meet ED are female. However, a difference between males and females in 2016 does emerge in apparent-time trends regarding the DRESS-LOT distance in wordlist style. Figure 63 suggests that female speakers are reducing their distances between DRESS and LOT, while this does not appear to be the case for male speakers, with the exception of students. This contradicts the observation that DRESS backing in wordlist style does not appear to be conditioned by any social factor other than *age*. These differences in apparent-time trends between male and female speakers regarding the DRESS-LOT F2 distance in this speech style appear to be the result of notably fronted LOT among the two youngest male speakers (see Chapter 6.2.3.1.1). Because their wordlist LOT is much fronter than that of any other speakers, they also have considerably lower F2 distances between DRESS and LOT.

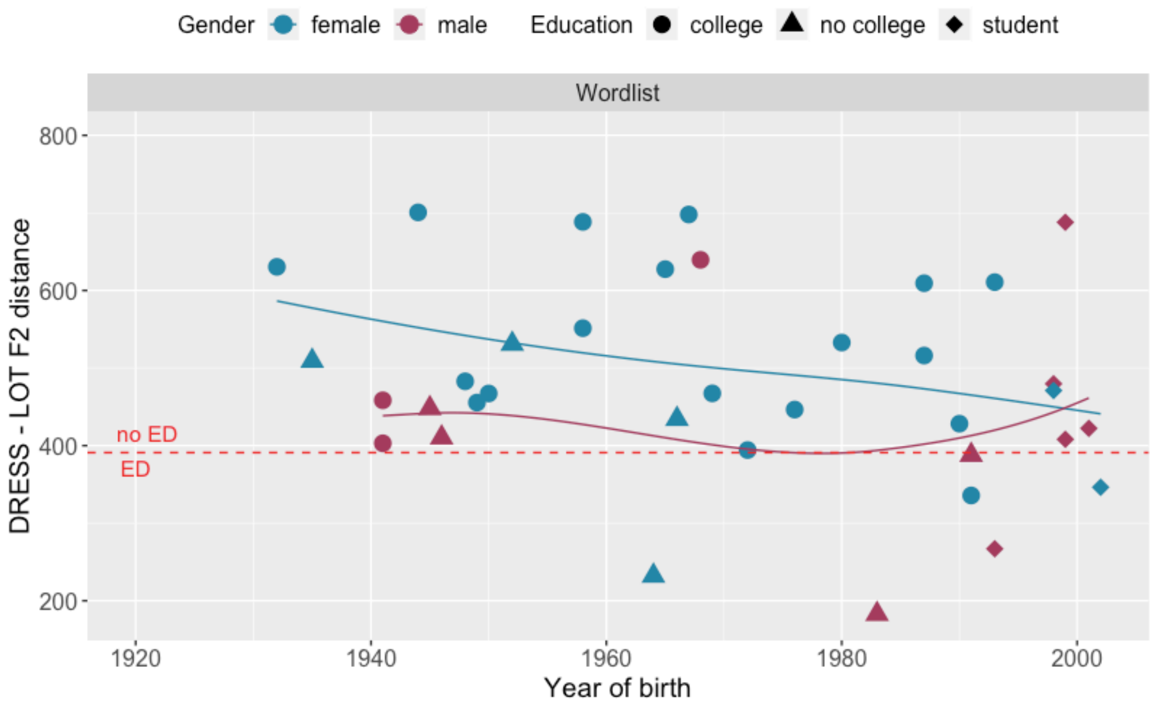


Figure 63: F2 distance between DRESS and LOT means in wordlist style in 2016 by *gender* and *education*.

In 2008, speakers participate significantly more in the ED criterion than 2016 speakers do in wordlist style. As shown in Figure 64 below, all 2008 speakers who participated in the reading of the wordlist meet this criterion, with much smaller F2 distances between

DRESS and LOT than most of the 2016 speakers. In fact, it appears that four of the seven 2008 speakers reverse the relative positions of DRESS and LOT on the front-back dimension, thus producing DRESS further back in the vowel space than LOT. Because all 2008 speakers meet this criterion, *gender* does not appear to play a significant role, as the developments of F2 of DRESS in 2008 wordlist style might suggest.

This remarkable difference in ED participation in wordlist style between 2008 and 2016 speakers contradicts the absence of real-time differences in F2 of DRESS in this speech style. This can be explained by the observation that 2008 speakers produce a significantly fronter wordlist LOT than 2016 speakers (see Chapter 6.2.3.1.1). Thus, despite similar F2 means for DRESS, the significantly higher F2 of LOT leads to much smaller or negative distances between DRESS and LOT for 2008 speakers compared to speakers recorded in 2016.

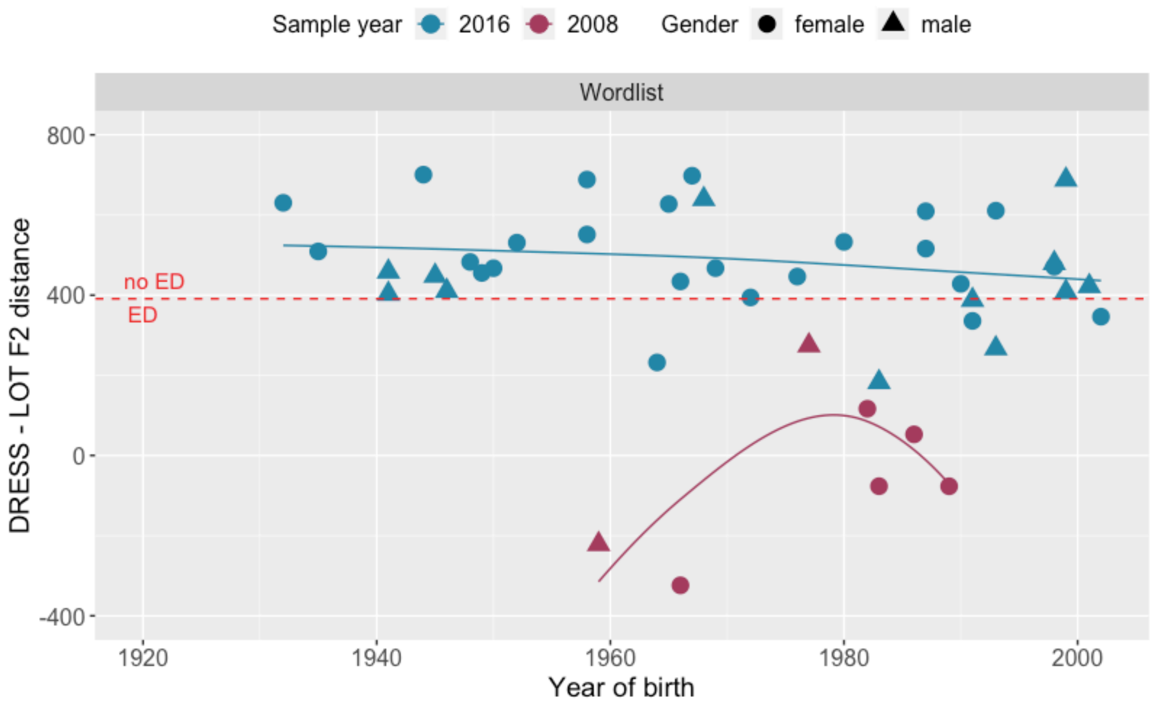


Figure 64: F2 distance between DRESS and LOT means in wordlist style in 2008 and 2016 by *gender*.

4.2.3 Style Shifting DRESS

The vast majority of speakers in the 2016 sample appear to produce a slightly higher and considerably backer DRESS in spontaneous speech than they do in wordlist style, as can be seen in Figures 65 and 66 below. However, the effect of *age* in the extent of style shifting appears to be increasing on the height dimension, while it is minimizing on the front-back dimension. In other words, younger speakers are realizing wordlist DRESS increasingly lower than spontaneous DRESS, while they seem to make no, or only minimal style differences in terms of frontness. Some of the speakers born after 1980 appear to reverse the direction of style shifting and shift to a slightly backer DRESS in the more careful speech style. These differences in the extent of intra-speaker variation in the realization of DRESS appear to stem from differently paced apparent-time developments in the two speech styles, as DRESS is lowering and backing more than twice as fast in wordlist style than it is in spontaneous speech.

The regression models presented in Tables 38 and 39 below support these observations. They predict that, for younger speakers, F1 style differences are in fact significant, while F2 style differences are not. They also predict the apparent-time developments in the two styles to be significantly different on both dimensions.

Figure 65: F1 difference between spontaneous and wordlist DRESS means in 2016 by *gender*. A negative value indicates that the vowel is lower in wordlist style than in spontaneous speech.

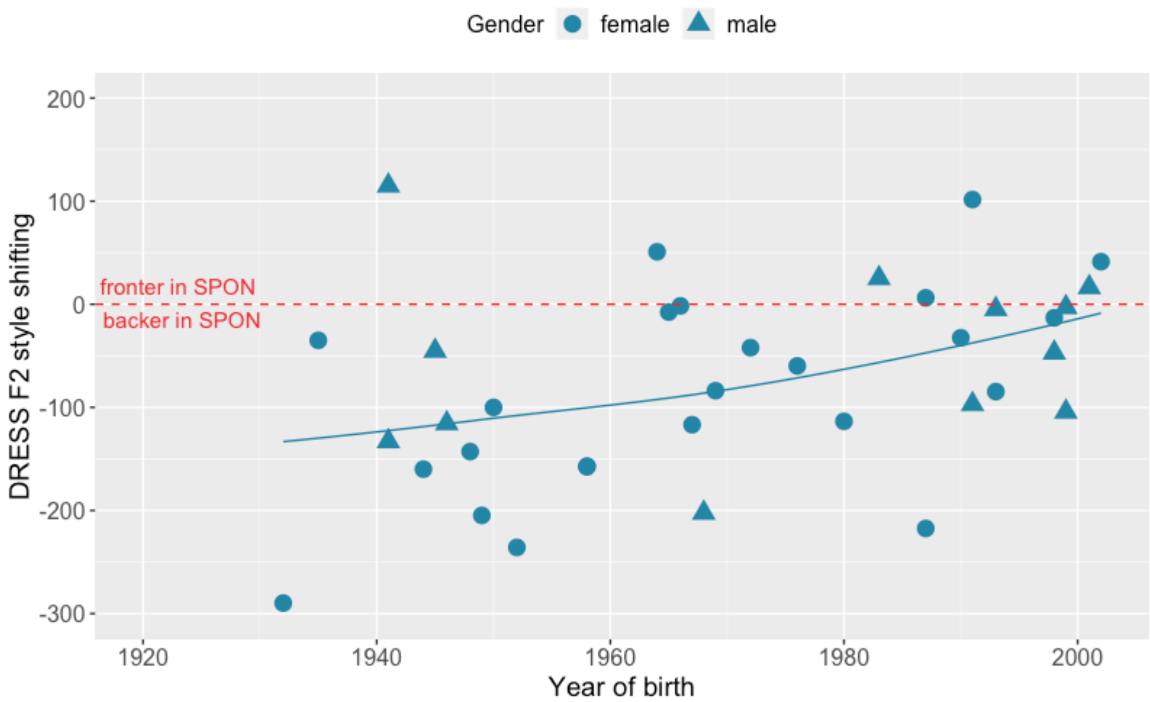


Figure 66: F2 difference between spontaneous and wordlist DRESS means in 2016 by *gender*. A positive value indicates that the vowel is backer in wordlist style than in spontaneous speech.

Predictor	Coefficient	*p*
(Intercept)	784.847 Hz	
Age	-0.68 Hz	
Gender (Male)	-2.66 Hz	0.664
Education *(No college)* *(Student)*	 -3.328 Hz -23.584 Hz	0.07
Style *(Wordlist)*	 48.315 Hz	
*Age*wordlist*	-0.69 Hz	0.0003
Environment		$2x10^{-16}$

Table 38: Effect of *style* on F1 of DRESS in 2016. Reference levels: females, spontaneous speech, /p/. Random effects: *speaker, word*. *n*= 8687

Predictor	Coefficient	*p*
(Intercept)	1619.685 Hz	
Age	1.35 Hz	
Gender (Male)	15.005 Hz	0.229
Education *(No college)* *(Student)*	 -13.554 Hz 10.782 Hz	0.438
Style *(Wordlist)*	 -29.199 Hz	
*Age*wordlist*	1.342 Hz	0.001
Environment		$2x10^{-16}$

Table 39: Effect of *style* on F2 of DRESS in 2016. Reference levels: females, spontaneous speech, /p/. Random effects: *speaker, word*. *n*= 8687

Speakers from the 2008 and 2016 samples appear to follow similar style-shifting patterns in terms of height. As shown in Figure 67 below, in 2008, four of the seven speakers who produced wordlist tokens shift in the same direction as most 2016 speakers in F1, i.e. spontaneous DRESS is higher than wordlist DRESS. The other three, however, seem to shift to a higher DRESS in the more careful speech style. While this can also be observed for some 2016 speakers, it seems to be the exception rather than the rule in this sample.

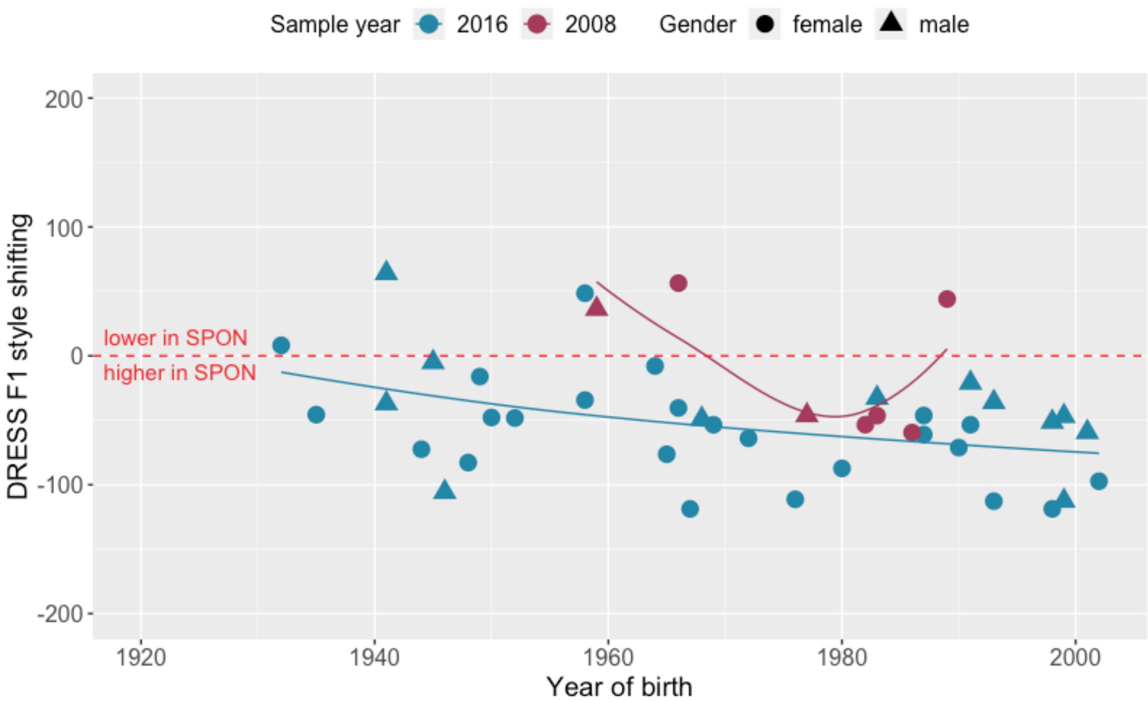


Figure 67: F1 difference between spontaneous and wordlist DRESS means in 2008 and 2016 by *gender*. A negative value indicates that the vowel is lower in wordlist style than in spontaneous speech.

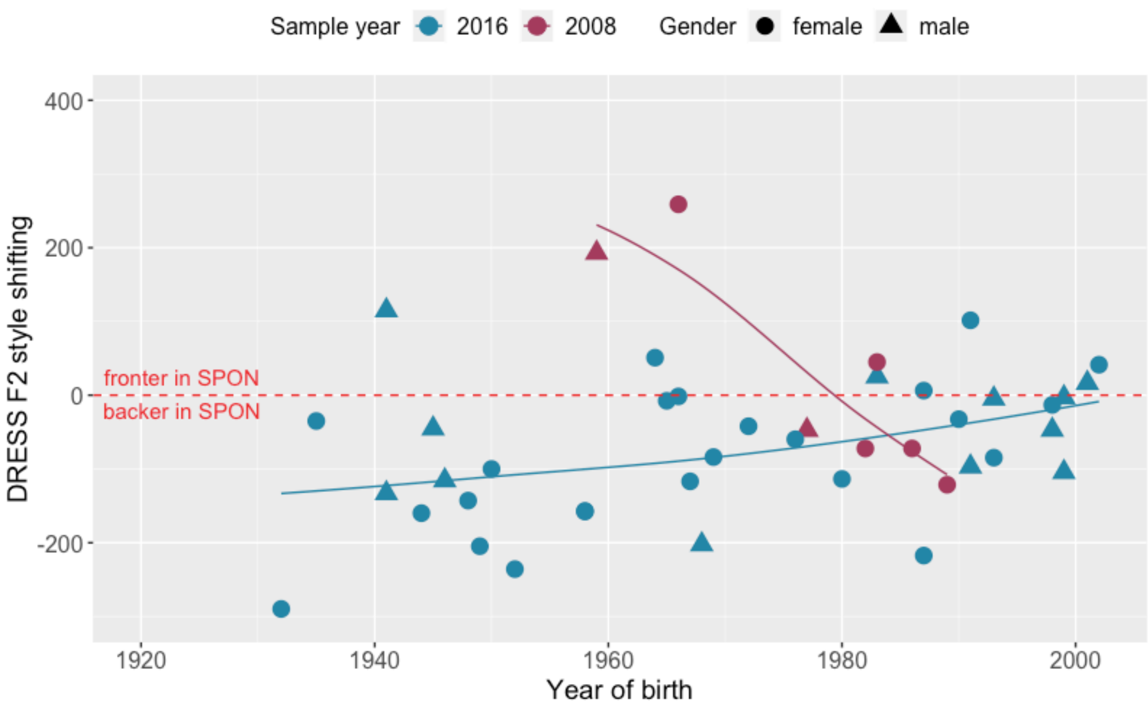


Figure 68: F2 difference between spontaneous and wordlist DRESS means in 2008 and 2016 by *gender*. A positive value indicates that the vowel is backer in wordlist style than in spontaneous speech.

On the front-back dimension, speakers in the 2008 sample seem to follow opposing style-shifting patterns: While four of them shift to a fronter DRESS in wordlist style, much like 2016 speakers, three shift to a backer DRESS in more careful speech, as Figure 68 above shows. Two of the speakers who shift to a backer DRESS, born in 1959 and 1966, are the same two speakers who also shift to a higher DRESS in wordlist style. With the exception of these two speakers, however, the extent of style shifting on this dimension is quite small in 2008. Furthermore, *age* appears to have the opposite effect on speakers in the 2008 sample as it does on the 2016 sample. Rather than minimizing style differences in apparent time, 2008 speakers appear to increase them, i.e. wordlist DRESS is realized increasingly fronter than spontaneous DRESS.

Because of these different style-shifting patterns, participation in the ED criterion differs notably between the two speech styles and the two samples.

As can be seen in Figure 69, more than half of the 2016 speakers meet the ED criterion in spontaneous speech, but most of them cease to do so in the more careful speech style. This appears to be a result of shifting to fronter DRESS in careful speech, especially among older speakers. Because this shift is minimized among younger speakers, all but one of the 2016 speakers who continue to meet the ED criterion in more careful speech are those born after 1980.

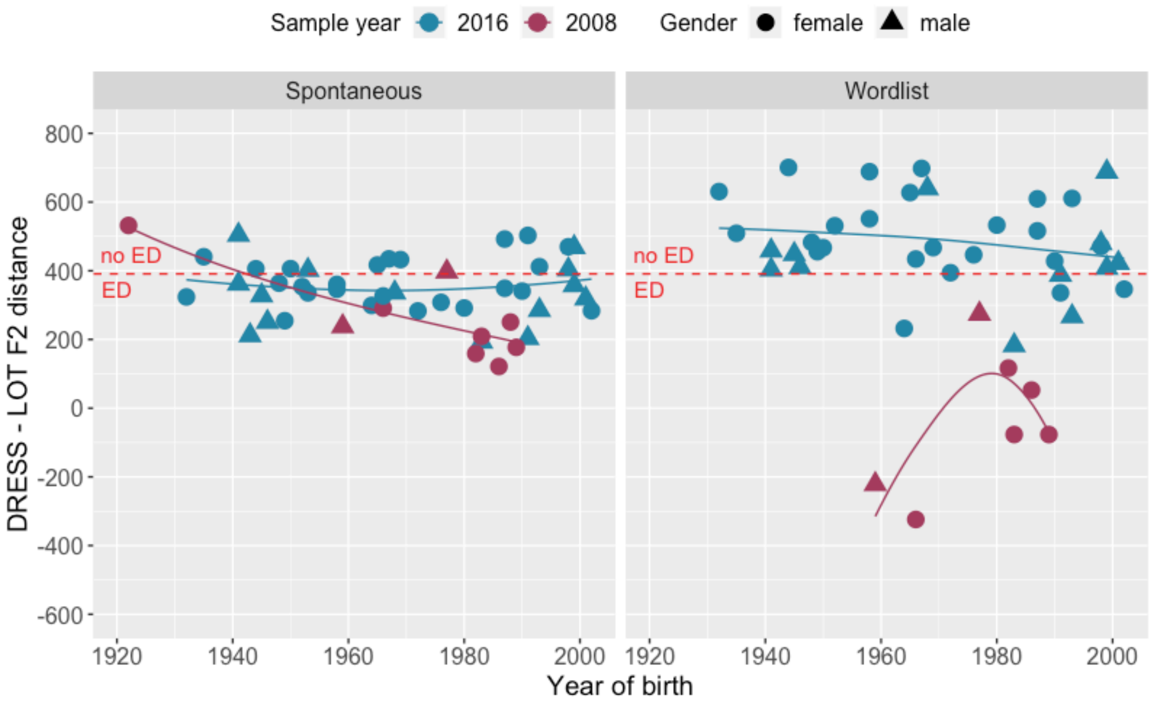


Figure 69: F2 distance between DRESS and LOT means in 2008 and 2016 across *speech styles* by *gender*.

In the 2008 data, on the other hand, participation in the ED criterion *increases* in the more careful style. While speakers in this sample predominantly meet the ED criterion in both styles, they have slightly lower F2 differences between LOT and DRESS in spontaneous speech than they do in wordlist style, where more than half of them appear to have reversed the relative positions of the two vowels, with DRESS being produced further back in the vowel space than LOT. Because DRESS itself seems to be style-shifted on the front-back dimension only to a small extent in the 2008 sample, the ED style differences appear to be a result of style shifting in the realization of LOT (see Chapter 6.2.3.1.1).

Indeed, as Figure 70 shows, there is virtually no overlap in F2 of DRESS and LOT in spontaneous speech for any of the 2008 speakers, though for some of them they are quite close. In wordlist style, on the other hand, LOT is produced much fronter by most speakers than it is in spontaneous speech, while the position of DRESS barely changes across the two speech styles, which explains the reduced (or negatively increased) F2 distance between the two vowels.

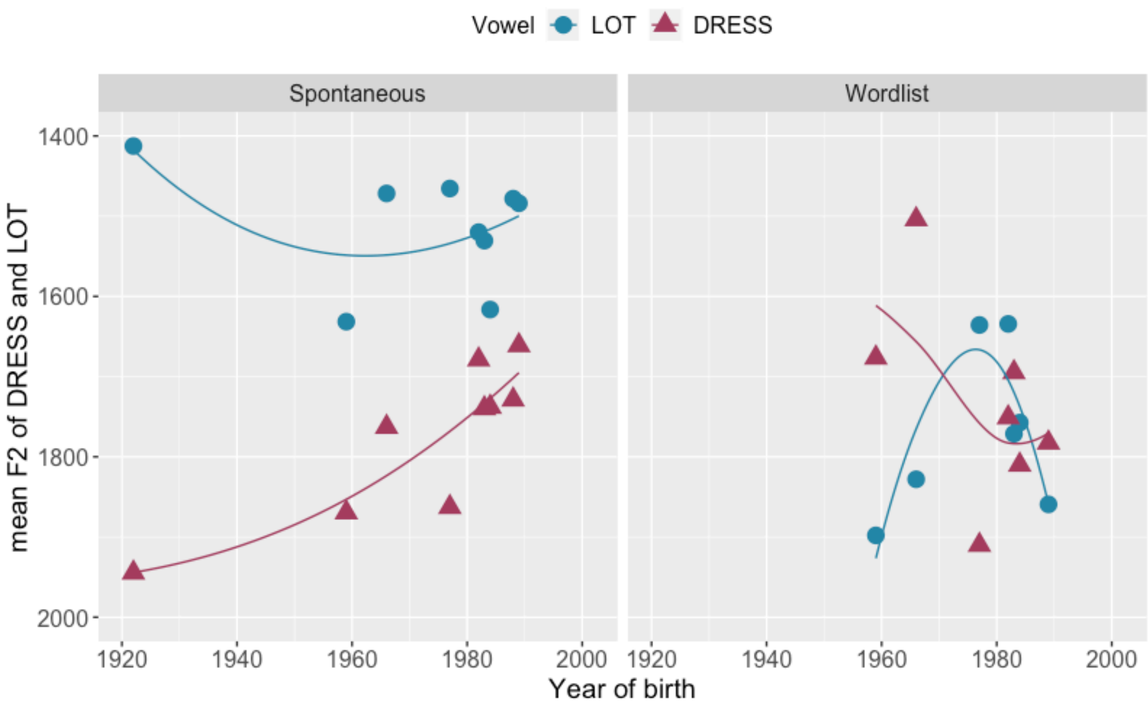


Figure 70: Style shifting the frontness of DRESS and LOT in 2008.

4.2.4 Summary: DRESS in Production

The analysis of DRESS presented above has shown that DRESS is undergoing significant changes in apparent and real time in the community. In both spontaneous speech and wordlist style, DRESS is lowering and backing in the 2016 data, though the pace of these processes differs in the two different speech styles.

In spontaneous speech, DRESS lowering and backing is progressing relatively slowly in the 2016 data. These apparent-time trends are only supported by real-time evidence on the height dimension, as 2008 speakers produce a slightly higher DRESS than 2016 speakers in this speech style (Figure 71 below). In terms of frontness, however, especially younger 2008 speakers far exceed 2016 speakers, producing a much backer DRESS than most other speakers in the two samples. While the majority of speakers seem to participate in apparent-time trends regardless of *gender* or *educational background*, a noteworthy exception are the students in the 2016 sample, who appear to have discontinued, or in fact reversed, both the lowering and the backing of DRESS.

DRESS lowering and backing in wordlist style is proceeding at a significantly faster rate than in spontaneous speech in the 2016 data. However, there is no significant real-time support for these trends in this speech style; in fact, 2008 speakers realize wordlist DRESS somewhat *backer* than 2016 speakers do, as shown in Figure 71 below. While all speakers seem to participate in both lowering and fronting, including students, lowering appears to be a female led change. Compared to spontaneous speech, wordlist DRESS is produced lower, but fronter by the majority of 2016 speakers. The extent of this shift from casual to more careful speech appears to depend on the speaker's *age* in this sample. While older style speakers shift only slightly on the height dimension but drastically on the front-back dimension, the opposite appears to be the case for younger participants. In the 2008 data, on the other hand, style-shifting patterns are reversed for some speakers.

Because DRESS is backing alongside LOT (see Chapter 6.2.3.1.1) for most speakers, participation in the ED criterion does not appear to be influenced by changes in the realization of DRESS. While there are differences in ED participation depending on the speaker's educational background in the 2016 data, and between 2008 and 2016 speakers, they result primarily from inter-speaker and inter-set variation in the frontness of LOT, not DRESS.

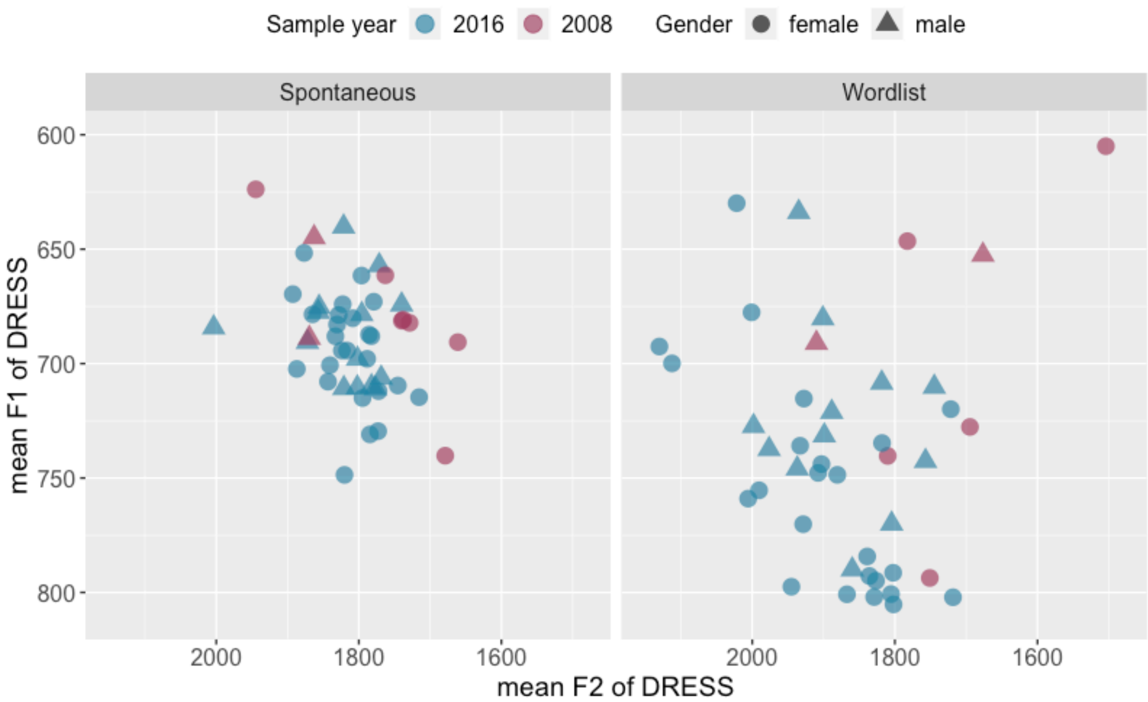


Figure 71: DRESS F1 and F2 means in 2008 and 2016 across *speech styles* by *gender*.

4.3 Discussion

The results presented in the analysis of DRESS, overall, match the developments that have been reported for this vowel in and outside of the Inland North. In both speech styles, DRESS is backing and lowering in apparent, and in some cases in real time, with a potential female advantage in more careful speech. Whether these developments are part of the NCS or are continuing as part of the reconfiguration toward the Elsewhere system is a question that cannot be answered based on the data presented in this study; however, the results presented leave room for the discussion of two topics: the motivation behind the continued lowering and backing of DRESS, and the reason for the unexpected behavior of students in spontaneous speech.

4.3.1 The Motivation for the Shifting of DRESS

What could not be confirmed in the present study is the positive social evaluation of retracted and/or lowered DRESS that was reported by Savage and Mason (2018). None of the categories tested in the matched guise experiment in this study suggested any significant differences in the ratings for shifted and unshifted DRESS.

One reason for the conflicting results in this study compared to those of earlier research might be of methodological nature. In the matched guise experiment in Savage and Mason's study in Lansing, shifted DRESS guises contained a *lowered* stimulus for DRESS. In the present study, on the other hand, shifted DRESS guises had a *backed* rather than lowered DRESS stimulus. Given that height differences tend to be more easily perceptible than front- and backness, it is possible that the participants in the present study simply did not react to variation in the realization of DRESS because this variation related to frontness rather than height. Whether or not they would have reacted differently to DRESS stimuli that differed in height cannot be answered here. It is possible, of course, that, as in Lansing, they would have rated the lower guises as more standard sounding. However, the production data provides little evidence for any social meaning attached to shifted DRESS. The slight female lead in combination with increasing style shifting toward lower DRESS in more careful speech may be indications that lowered (and/or backed) DRESS is the perceived norm that speakers are striving toward. The direction of style shifting on the front-back dimension, however, may suggest the opposite, as the majority of speakers shift to a fronter, rather than backer DRESS in more careful speech. Only the youngest speakers seem to minimize these style differences, and some of them do shift to a backer DRESS in wordlist style.

While these style-shifting patterns, i.e. fronter and lower DRESS in wordlist style, are somewhat arbitrary, they are very similar to those identified in Lansing, where speakers did favor lowered DRESS in their matched guise ratings. Thus, there is some slight evidence that shifted DRESS, in both directions, is perceived as a more standard variant at least by some speakers in Ogdensburg, and that this change therefore continues to be advanced in the community. If this was the case, variation in the realization of DRESS might be expected to have attracted some overt commentary as well. While there is little evidence that different realizations of DRESS have reached the level of conscious awareness in Ogdensburg, one of the participants does comment on it in her interview: Ruth, born in 1948, seems to rely, among others, on the realization of DRESS in describing the speech of Canadian co-workers:

> “*The e's and a's sounds were different, I can't remember any particular words, but they just seemed to have a – just a little bit different slant on things.”*

Although she does not specify how the Canadians' DRESS is different from what she is used to, it is possible that she is referring to DRESS being lowered and/or backed as part of the Canadian Shift in the speakers she has come across. It is, however, questionable *how* different DRESS in Canadian speech really is from that in Ogdensburg. *ANAE* describes lowered DRESS as part of the Canadian Shift with an F1 greater than 660 Hz[69], and the vast majority of speakers in Ogdensburg were found to realize DRESS with similar, or even smaller F1 means. Thus, it is possible that this account of differences in the articulation of DRESS is simply false.

Another explanation for the lack of significant results in the matched guise experiment might be that DRESS has remained an indicator, as defined by Labov, in Ogdensburg and that participants therefore simply do not perceive shifted DRESS, whether lowered or backed, in the same way that speakers in Lansing do. In this case, DRESS backing would be proceeding as a change from below, and likely be purely internally motivated, potentially by a pull chain triggered by a relatively high TRAP (see Chapter 3.2) or a drag chain that was triggered by a retracted realization of STRUT (see Chapter 5.2). The observed style differences, under this scenario, might simply be a result of hyper-articulation in more careful speech rather than an indication of perceived standard.

4.3.2 Students Have Their Own DRESS Code

A second point worth discussing is the unexpected behavior of the students in their realizations of DRESS in spontaneous speech. Neither on the height, nor on the front-back dimension do they seem to continue the apparent-time trends that have been set by the community in this speech style, but instead raise and front DRESS back to its original position.

There are, of course, speakers in the community who realize spontaneous DRESS with similar spectral qualities as students do, and it is possible that students are simply following the patterns of older generations, potentially their parents. This is in agreement with Labov's (2001) argument that sound change in progress tends to peak in apparent time around late adolescence, as younger speakers are still in the process of acquiring the

[69] Although the 660 Hz threshold is based on a different normalization method, it can be assumed that it would not be significantly higher in the present study (see Chapter 2.4.2).

innovative form. However, if the parents' production of DRESS did in fact have a direct impact on younger speakers, it would be expected that this affects the youngest speakers outside of the student group as well. This is, however, not the case, and in fact, non-students born after 1980 are notably different from their parents' generation on both dimensions. Thus, unless we believe that the parents' production affects younger speakers all throughout high school and/or college, and that this effect suddenly reverses once they graduate, this seems to be an unlikely cause of the students' behavior.

A more likely explanation for the production patterns observed among students is age-grading, i.e. students are simply going through a phase of producing a higher and fronter DRESS. The sharp divide between students and younger speakers who are not students would support this hypothesis. The question that remains is why these kinds of age-grading patterns do not emerge in all vowels examined in this study. Students are, for example, among the most advanced in the lowering of TRAP (Chapter 3.2), the backing of LOT and the fronting of THOUGHT, and thus in the adoption of the COT-CAUGHT merger (Chapter 6.2), as well as in the fronting of GOOSE (Chapter 7.2) and GOAT (Chapter 7.4), all the while following community-wide trends. The difference between DRESS and these other vowels might be the lack of social evaluation attached to fronter and higher DRESS. While raised TRAP and fronted LOT appear to be perceived less favorably in the community than their unraised and unfronted counterparts, this kind of judgment seems to be absent from variation in DRESS. While the style-shifting patterns and previous research do suggest that especially lowered DRESS is the favored variant, this may not be as conscious a judgment as it appears to be for TRAP and LOT. Thus, it is possible that the lack of conscious awareness in variations of DRESS allows students to stray from the norm without having to use a nonstandard variant, which may not be the case with raised TRAP or fronted LOT. The observation that students do not show this diverging behavior in wordlist DRESS might signal awareness of the communal norm, in which they do participate in more formal speech. However, Boberg (2004) suggests that, in patterns affected by age-grading, younger speakers typically use the *innovative* form more frequently, which is the opposite of the pattern observed in the realization of DRESS among students, who seem to use the more *conservative* form instead.

It is, of course, also possible that differences between students and the rest of the sample are not meaningful at all. Although this is admittedly somewhat less exciting than potential age-grading, it should be kept in mind that the differences between students and younger non-student speakers do not exceed 100 Hz on either dimension. For F1, the estimated difference is a mere 24 Hz, and for F2 an even lower 9 Hz, neither of which are found to be statistically significant. Also, very similar patterns can be found for STRUT (Chapter 5.2.1) and KIT (Chapter 5.3), which are both realized slightly higher in the vowel space by students than most other speakers in spontaneous speech, so that it seems unlikely that the students' slightly deviating realizations have any kind of social meanings.

Chapter 5: Moving Forward: STRUT & KIT

5.1 STRUT and KIT – An Introduction

Although variation in the realizations of STRUT and KIT are not uncommon in American English, both have received relatively little attention in the literature, which is why they are combined here into one chapter. E. R. Thomas (2001) describes the most widespread variant of STRUT in US dialects as a mid-central [ɜ], with occasional fronting toward /æ/ and /e/ and, in the South, potential raising to [ə]. Backed variants of STRUT, identified by an F2 that is back of center, are less common, and have only been observed in Newfoundland, the Caribbean, in parts of the South, Southern New England, and the Inland North (*ANAE*; E. R. Thomas, 2001). KIT is generally realized as high front [ɪ], shorter and somewhat centralized from FLEECE in US English, and E. R. Thomas (2001, p. 16) explains that most diaphones "involve varying degrees of fronting and backing, though lowering is also reported". However, backing appears to be the more frequently observed movement, and has been reported in e.g. Florida, Western New England, and the Inland North (*ANAE*). Lowering has been reported in the Inland North, as well as in San Francisco, Canada (E. R. Thomas, 2001) and the Mid-Atlantic States (*ANAE*). As pointed out before, KIT often merges with DRESS in pre-nasal position in some varieties of English as part of the PIN-PEN merger, though this generally affects the phonetic realization of DRESS rather than KIT.

In the Inland North, the directions of STRUT and KIT shifting have been found to vary. For both, retraction appears to be the more dominant movement; however, in some parts of the Inland North, backing has been accompanied by lowering, or lowering occurred individually. The outcome of backed STRUT has most commonly been described as a shorter and potentially unrounded version of /ɔ/ (Gordon, 2001; Labov, 1994; McCarthy, 2007; E. R. Thomas, 2001), or, in combination with lowering as [ɑ, ɒ] (Eckert, 1988; Gordon, 2001). Additionally, Eckert (1988) described STRUT raising to [ʊ]. Regarding shifted KIT, *LYS* described the lowered variant as approaching /e/, and Gordon (2001) described the outcome of backing as [ɨ]. Gordon (2001) also reported accounts of KIT centralization toward [ʌ].

In general, KIT shifting appears to be rather rare and does not seem to be an integral part of the NCS (*ANAE*; Durian & Cameron, 2018; Gordon, 2001), and in New York it has rarely been examined outside of *LYS* and *ANAE*. Emerson (1891) reported lowering of KIT to [ɛ] only in a few isolated words, and C. K. Thomas (1935-1937) observed frequent centralization, often accompanied by lowering. More recently, Driscoll and Lape (2015) reported *raising* and *fronting* of KIT among younger speakers in Syracuse. An analysis of KIT is included here to fill this gap in the research on potential variation in the realization of KIT in New York. Although the results of previous research suggest that KIT shifting as part of the NCS is limited, KIT is expected to shift along the same trajectories in the Elsewhere Shift, which may be affecting the community of interest in this study, potentially increasing the amount of variation in the realization of this vowel.

STRUT shifting, on the other hand, appears to be a more common change in New York State. Frequent backing and lowering as part of the NCS have been reported in the literature (*ANAE*; Clopper et al., 2006; Dinkin, 2009). More recently, however, a minimal amount of STRUT *fronting*, which is commonly associated with the Elsewhere/California/Canadian Shift, and which runs counter to the NCS trajectory of STRUT, has been observed in the Inland North (Driscoll & Lape, 2015; Wagner et al., 2016).[70] The same was found in an early study in New York (Emerson, 1891). The analysis of STRUT in this chapter will provide insight into which, if any, of these two opposing developments STRUT may be following in Ogdensburg.

Few studies have examined different realizations of STRUT and KIT, and even fewer have taken into account the social distribution of their variants. Some studies reported ongoing shifting in apparent time for STRUT (*ANAE*; Gordon, 2001; McCarthy, 2007, 2010) and KIT (Morgan et al., 2017). STRUT backing appears to have been led by working-class oriented speakers (Eckert, 1989) and females (Clopper et al., 2006; Gordon, 2001;), although *ANAE* found no effect of *gender* on the realization of STRUT. For KIT, no clear lead of a particular social group has been reported. Gordon (2001) found that, when shifting was observed at all, male speakers seemed to prefer backing, while females show more variation in the direction of shifting. To the best of my knowledge, social evaluations of variation in the realizations of KIT and STRUT have not yet been reported.

[70] An exception to this appears to be Chicago, where STRUT backing was found to be an ongoing process (Durian & Cameron, 2018).

5.2 Results: STRUT

Labov (2007) formally describes the shifting of STRUT in the NCS with the UD criterion, which relates the position of STRUT with respect to LOT on the front-back dimension. According to this criterion, STRUT with an F2 lower than that of LOT, i.e. STRUT that is produced further back in the vowel space than LOT, is considered shifted by NCS standards. If and in which direction STRUT is shifting in Ogdensburg, and how it relates to LOT, will be the focus of this chapter. Since the evaluation ratings from the matched guise experiment did not produce any noteworthy results for STRUT, they will not be discussed in this chapter.

Figure 72 suggests a good amount of intra-speaker variation in the realization of STRUT in the 2016 sample, as well as variation between speakers in the more careful speech style. In spontaneous speech, the majority of speakers seem to produce STRUT with relatively similar qualities, just back of center. In wordlist style, on the other hand, speakers appear to deviate considerably from each other, without an immediately noticeable *age* pattern. Figure 72 also suggests substantial style differences, with wordlist STRUT being produced lower and backer in the vowel space. These patterns will be explored further in the following subchapters.

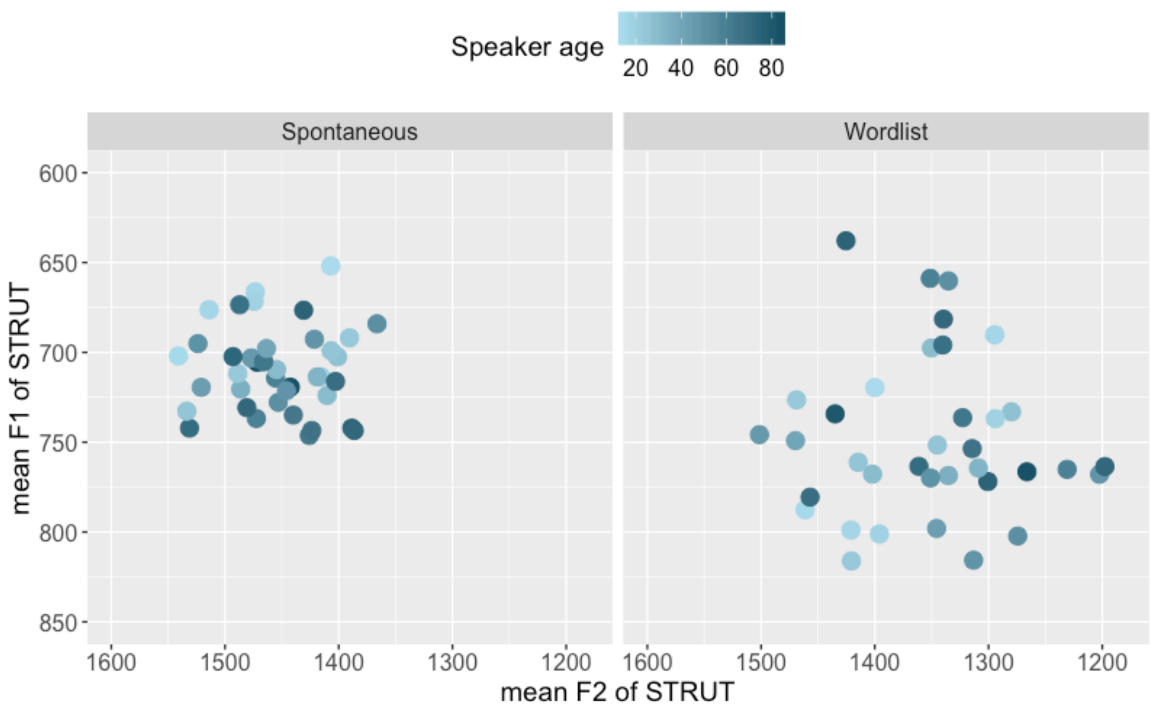


Figure 72: STRUT F1 and F2 means in 2016 across *speech styles* by *age*. Lighter shades represent younger speakers, darker shades older speakers.

5.2.1 STRUT in Spontaneous Speech

On the height dimension, variation in the realization of STRUT appears to be rather limited in spontaneous speech in the 2016 sample. As shown in Figure 73, the majority of speakers produce spontaneous STRUT within a very narrow F1 range of less than 100 Hz, most of them between 700 and 750 Hz. Only students appear to have a slightly higher STRUT, with F1 means ranging from 650 to 700 Hz. Otherwise, there appears to be complete homogeneity in the height of STRUT in the community, with no apparent-time trends in either direction. The lack of significant effects in the regression model presented in Table 40 supports this observation, though the small differences between speakers of different educational levels is found to be statistically significant.

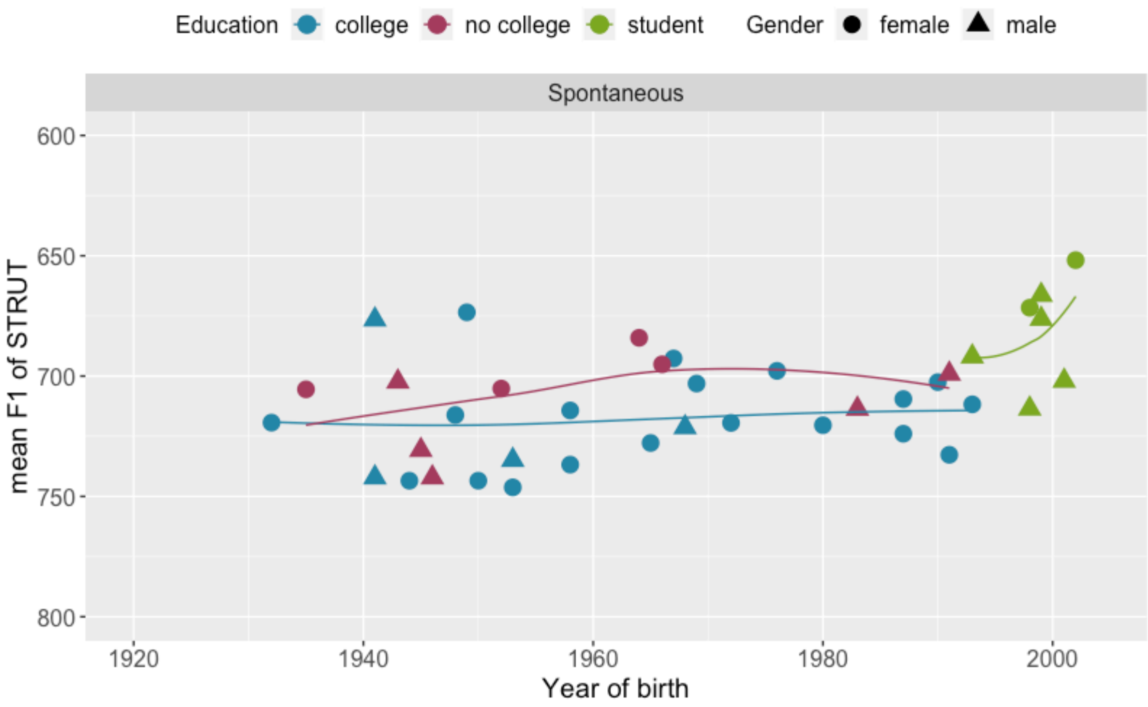


Figure 73: STRUT F1 means in spontaneous speech in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	750.718 Hz	
Age	0.118 Hz	0.563
Gender (Male)	7.532 Hz	0.339
Education *(No college)* *(Student)*	 -14.344 Hz -28.33 Hz	0.049
Environment		0.002

Table 40: STRUT F1 in spontaneous speech in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word*. *n*= 5341

On the front-back dimension, inter-speaker variation in the realization of spontaneous STRUT appears to be relatively limited as well. Figure 74 shows that STRUT is produced with an F2 between 1400 and 1500 Hz by the majority of speakers in the 2016 sample. However, Figure 74 suggests that STRUT might be undergoing two opposing trends in apparent time depending on the speakers' *educational background*: While speakers without a college degree are backing STRUT in apparent time, those with a college degree seem to be fronting it slightly. The significant interaction between *age* and *education* in Table 41 confirms the two opposing developments between these two educational groups.

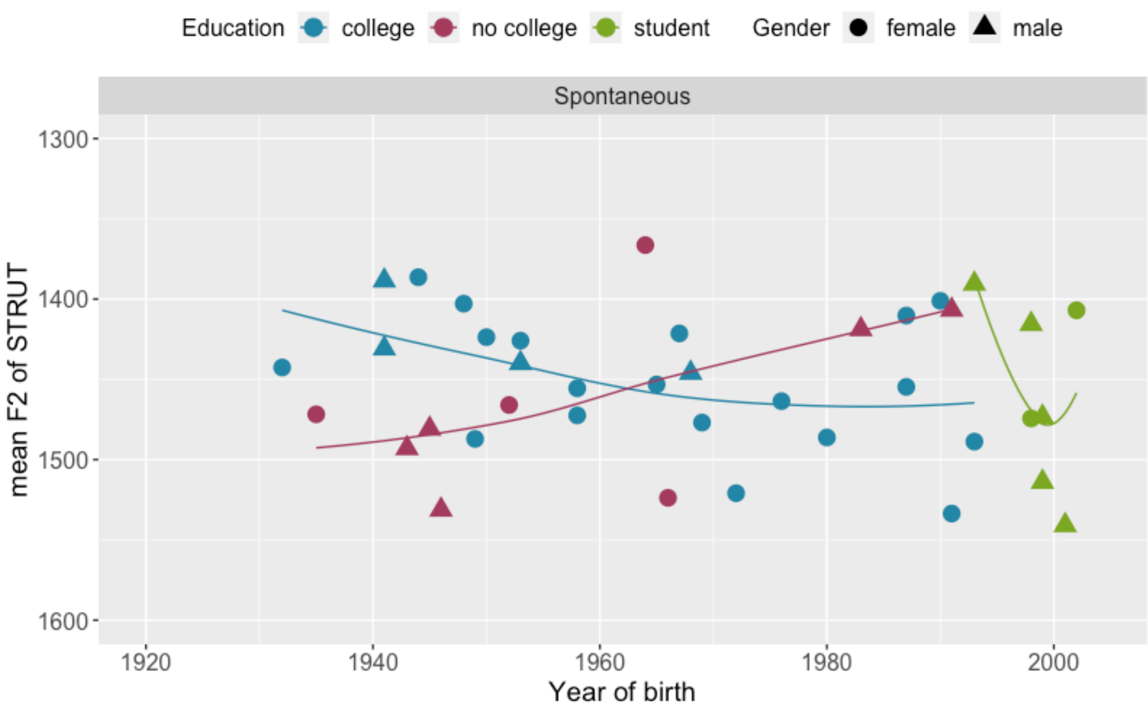


Figure 74: STRUT F2 means in spontaneous speech in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	1439.614 Hz	
Age	-0.55 Hz	
Gender (Male)	-2.477 Hz	0.851
Education (No college)	-112.707 Hz	
Age No college*	2.04 Hz	0.005
Environment		$2x10^{-6}$

Table 41: STRUT F2 in spontaneous speech in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word*. *n*= 4741

In a comparison of the 2016 and 2008 data, only minimal differences can be observed in the height and frontness of STRUT in spontaneous speech.

Compared to the 2016 sample, speakers in the 2008 data set produce STRUT with slightly smaller F1 means, i.e. higher in the vowel space. As can be seen in Figure 75, this is particularly noticeable among younger speakers, though the actual difference is extremely small. Overall, the F1 differences between the two samples are not found to be statistically significant (Table 42 below), however, a model that considers only speakers born after 1980, excluding students, does predict a significant difference of 45 Hz between the two samples (*p*= 0.036).

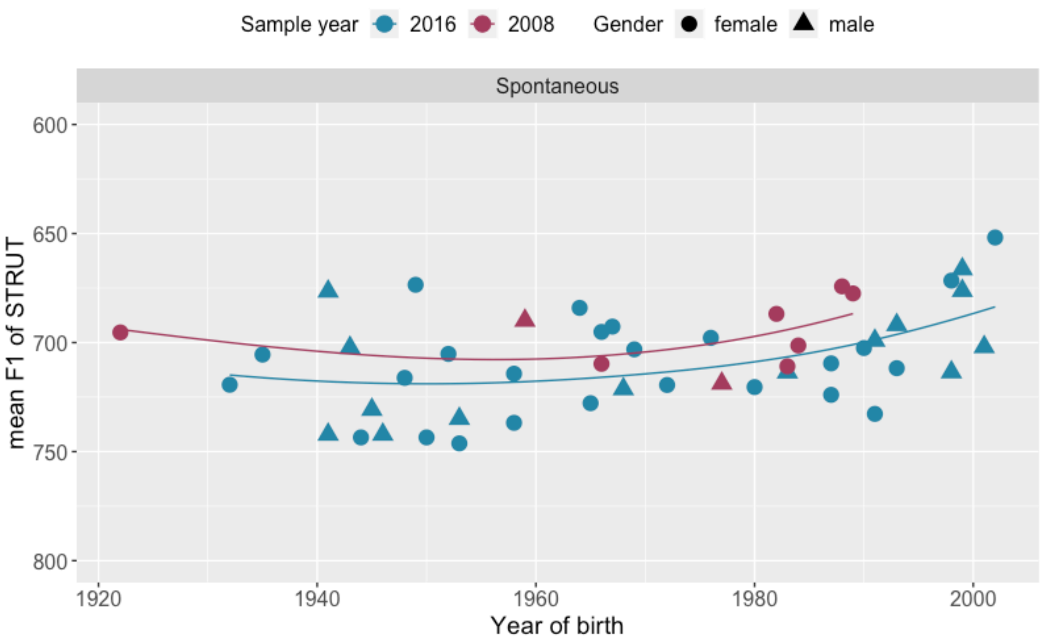


Figure 75: STRUT F1 means in spontaneous speech in 2008 and 2016 by *gender*.

A similar, but more pronounced pattern of real-time differences can be observed for the frontness of STRUT in spontaneous speech. As can be seen in Figure 76 below, the majority of speakers in the combined 2008 and 2016 data seem to produce STRUT with relatively comparable F2 means between 1400 and 1500 Hz. However, younger speakers in the 2008 sample have somewhat smaller F2 means than most others. Not only does this suggest real-time fronting, it also indicates an apparent-time backing of STRUT in the 2008 data that cannot be observed in 2016.

Younger speakers in the 2008 sample produce STRUT in spontaneous speech about 50 to 100 Hz further back than older 2008 speakers. While older speakers in this sample average around 1450 Hz, younger participants have F2 means between 1350 and 1400 Hz. However, this difference between younger and older 2008 speakers does not reach the level of statistical significance (0.67 Hz, *p*= 0.44). Despite the lack of statistical significance, the retracted realization of STRUT among younger 2008 speakers leads to notable real-time differences among speakers born between 1970 and 1990, as shown in Figure 76. However, neither the regression model presented in Table 43, nor a model that considers only speakers born after 1970 find significant difference between the two samples (-34.777 Hz, *p*= 0.252).

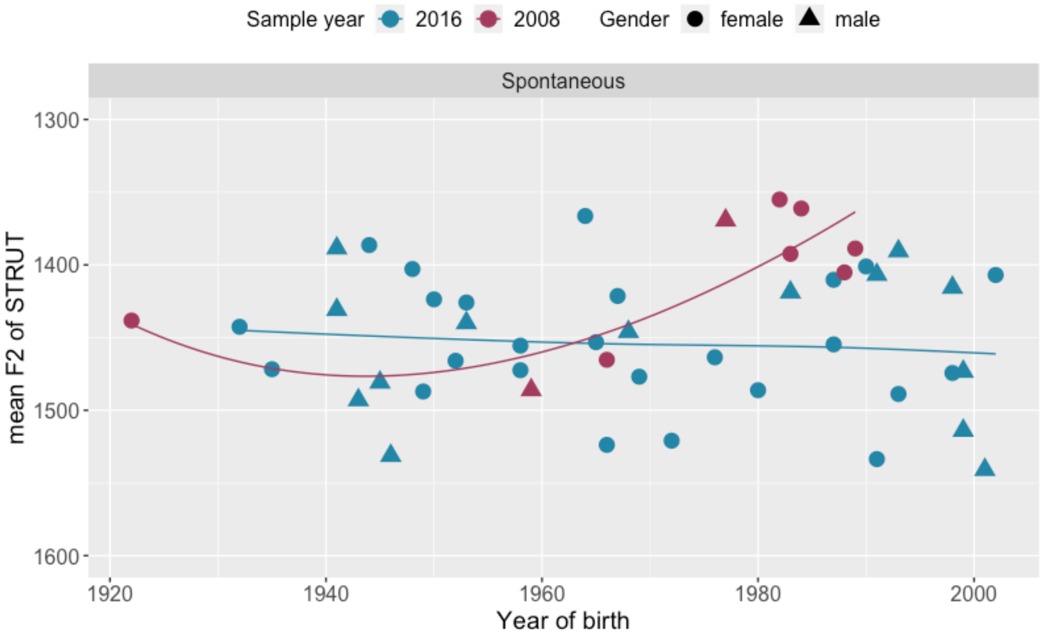


Figure 76: STRUT F2 means in spontaneous speech in 2008 and 2016 by *gender*.

Predictor	Coefficient	*p*
(Intercept)	754.26 Hz	
Age	0.063 Hz	0.731
Gender (Male)	2.392 Hz	0.76
Education (No college)	9.138 Hz	0.243
Sample year (2008)	-20.653 Hz	0.064
Environment		0.025

Table 42: STRUT F1 in spontaneous speech in 2008 and 2016. Reference levels: females, college educated, 2016, /p/. Random effects: *speaker, word*. *n*= 4910

Predictor	Coefficient	*p*
(Intercept)	1399.725 Hz	
Age	0.126 Hz	0.695
Gender (Male)	-1.021 Hz	0.934
Education (No college) (Student)	-2.771 Hz -5.286 Hz	0.95
Sample year (2008)	-19.125 Hz	0.252
Environment		10^{-7}

Table 43: STRUT F2 in spontaneous speech in 2008 and 2016. Reference levels: females, college educated, 2016, /p/. Random effects: *speaker, word*. *n*= 4910

Despite few F2 differences between 2008 and 2016 speakers, participation in the UD criterion, which defines NCS-shifted STRUT as being backer than LOT, appears to differ between the two samples in spontaneous speech. As can be seen in Figure 77, all but one speaker fulfill UD in 2008, i.e. they produce STRUT in a backer position than LOT. In the 2016 sample, UD participation is more balanced. Of the 39 speakers in the 2016 data, 22 fulfill the criterion in spontaneous speech, while 17 fail to do so. Thus, proportionally, UD participation is much higher in the 2008 sample than it is in the 2016 sample. This difference is somewhat surprising, given the lack of any significant differences between 2008 and 2016 speakers in F2 of STRUT in this speech style. Taking into account different realizations of LOT, however, explain the differences in UD participation, as 2008 speakers produce a slightly fronter LOT in spontaneous speech than 2016 speakers do (see Chapter 6.2.3.1.1), thereby reversing the relative frontness of LOT and STRUT.

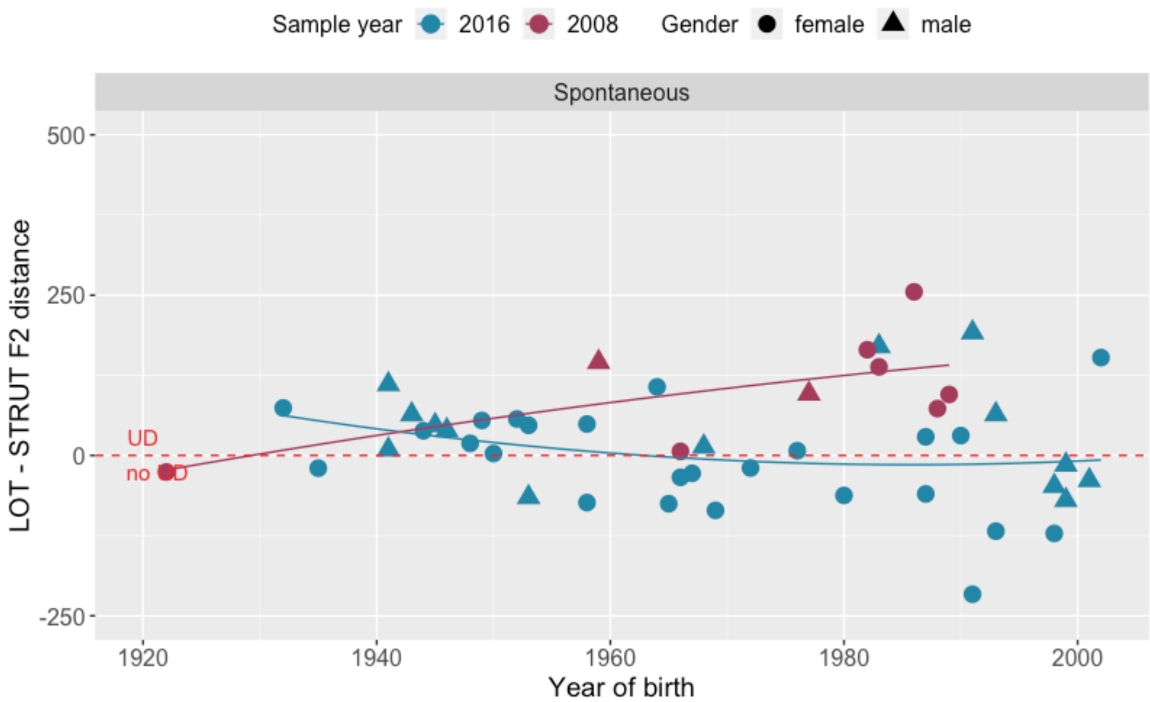


Figure 77: F2 distance between LOT and STRUT means in spontaneous speech in 2008 and 2016 by *gender*. A positive value indicates that STRUT is backer than LOT.

Furthermore, Figure 77 above suggests opposing apparent-time changes in the STRUT-LOT distance in spontaneous speech between the two data sets. It seems that, for 2008 speakers, STRUT in spontaneous speech is produced increasingly further back than LOT, while 2016 speakers are returning to a configuration where STRUT is fronter than LOT.

Thus, inter-set differences in the LOT-STRUT distance are increasing in apparent time, and younger speakers in the two samples differ notably from each other in their F2 distances between both vowels. For STRUT itself, however, no significant real-time differences in F2 emerge from the data. For 2008 speakers, this trend can be explained with a notably backer STRUT among younger speakers as discussed above, while F2 of LOT remains steady across all ages in this sample (see Chapter 6.2.3.1.1). Decreasing UD participation among 2016 speakers, on the other hand, appears to be primarily a result of retraction of LOT in apparent time (see Chapter 6.2.3.1.1). For college educated speakers and students, STRUT fronting is likely affecting receding UD participation as well.

5.2.2 STRUT in Wordlist Style

Figure 78 indicates that, in 2016, there is little variation in the height of STRUT in wordlist style, most speakers realizing the vowel with F1 means between 700 and 800 Hz. Only very few speakers deviate slightly up or downwards, however, the amount of this deviation is minimal, and no patterns regarding *age*, *gender* or *education* can be identified in Figure 78 or the regression model in Table 44 below. Thus, it appears that wordlist STRUT has been produced at a consistent height over the 70 years tracked in the 2016 data.

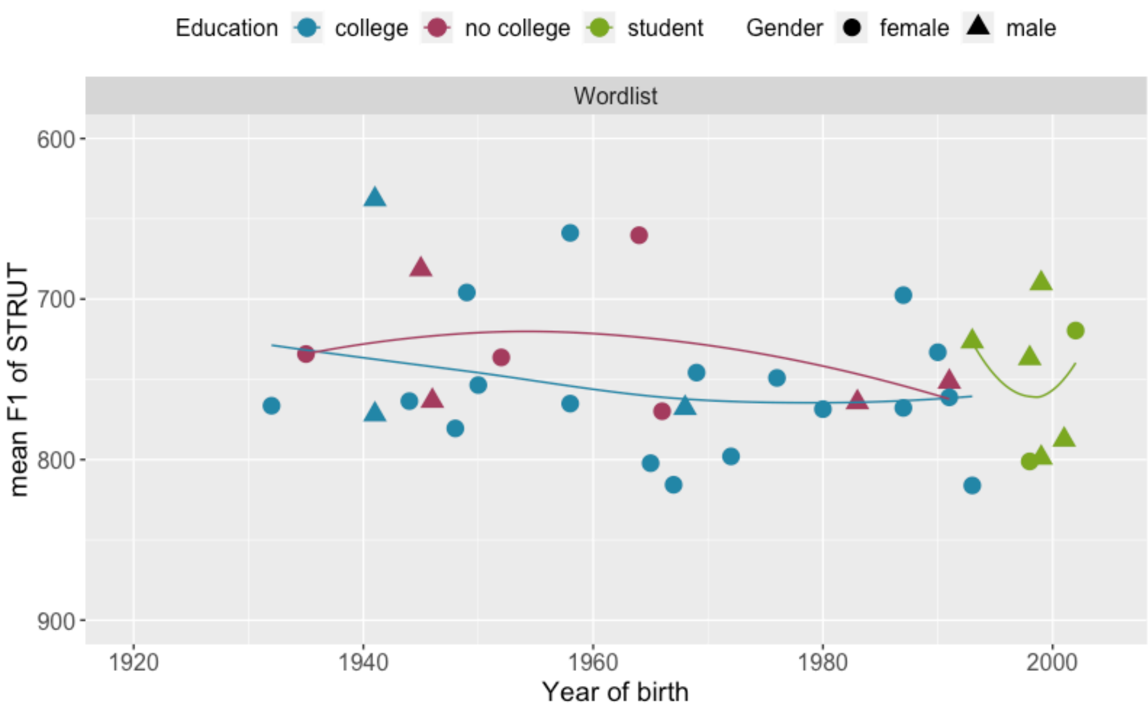


Figure 78: STRUT F1 means in wordlist style in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	818.857 Hz	
Age	-0.539 Hz	0.252
Gender (Male)	-8.081 Hz	0.667
Education *(No college)* *(Student)*	-13.75 Hz -16.313 Hz	0.724
Environment		0.494

Table 44: STRUT F1 in wordlist style in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word*. *n*= 313

A very similar pattern emerges for the frontness of STRUT in wordlist style for 2016 speakers. Most of them seem to produce STRUT with an F2 between 1300 and 1400 Hz. Although this is a relatively narrow range, the visualized data in Figure 79 indicates that the youngest college educated speakers and students in the sample seem to have a tendency to produce a slightly fronter STRUT than older speakers. However, the regression model in Table 45 below does not find *age*, or any other social factor, to be of any significance to the frontness of wordlist STRUT.

Figure 79: STRUT F2 means in wordlist style in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	1443.036 Hz	
Age	-0.593 Hz	0.46
Gender (Male)	-20.887 Hz	0.518
Education		0.647
(No college)	21.268 Hz	
(Student)	40.239 Hz	
Environment		0.528

Table 45: STRUT F2 in wordlist style in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word*. *n*= 313

In wordlist style, notable real-time differences can be observed in the data, both on the height and the front-back dimension.

In terms of height, real-time differences appear to depend on the speaker's *age*. Figure 80 suggests that the two older 2008 speakers have a notably higher wordlist STRUT than most 2016 speakers. Additionally, while the *age* pattern in the 2008 sample suggests a strong apparent-time trend toward lower STRUT, no such developments can be observed in the 2016 data.

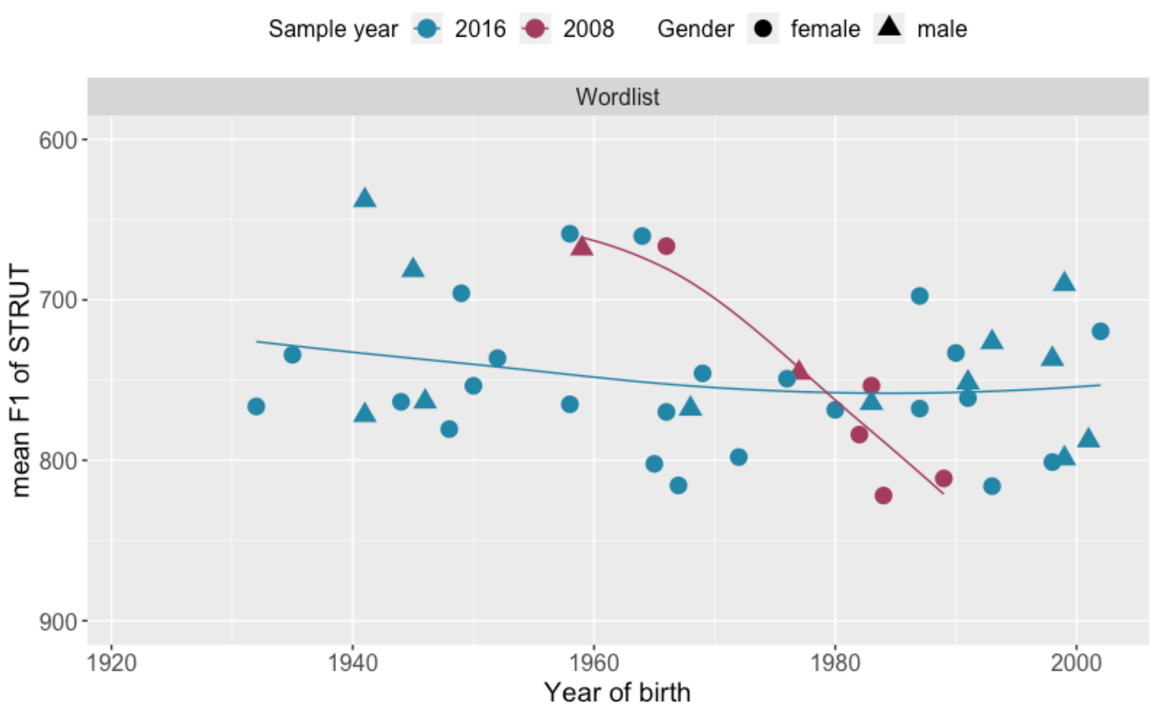


Figure 80: STRUT F1 means in wordlist style in 2008 and 2016 by *gender*.

The regression model in Table 46 below corroborates these observations, and a model that tests for the effect of *age* for the 2008 sample separately lends further evidence to the observed apparent-time lowering of STRUT (-7.12 Hz, *p*= 0.006). However, there are only two older speakers with relatively small F1 means in the 2008 sample, so the reliability of this trend is debatable. While these two 2008 outliers are joined in their small F1 by two 2016 speakers, they seem to be outliers in the 2016 sample. Thus, it seems unlikely that *raised* STRUT was ever a stable feature in Ogdensburg. Even if STRUT lowering in 2008 was a significant development, because it started out at a relatively small F1 with those two older speakers, the result is simply that younger speakers in the 2008 sample produce STRUT at the same height as the majority of speakers in the 2016 sample.[71]

In terms of frontness, Figure 81 suggests that 2008 speakers produce a notably fronter STRUT in wordlist style than 2016 speakers do. In this speech style, the majority of 2008 speakers seem to produce STRUT with an F2 of about 1500 Hz, while the majority of 2016 speakers have F2 means between 1300 and 1400 Hz, creating a 100 to 200 Hz difference between both samples.

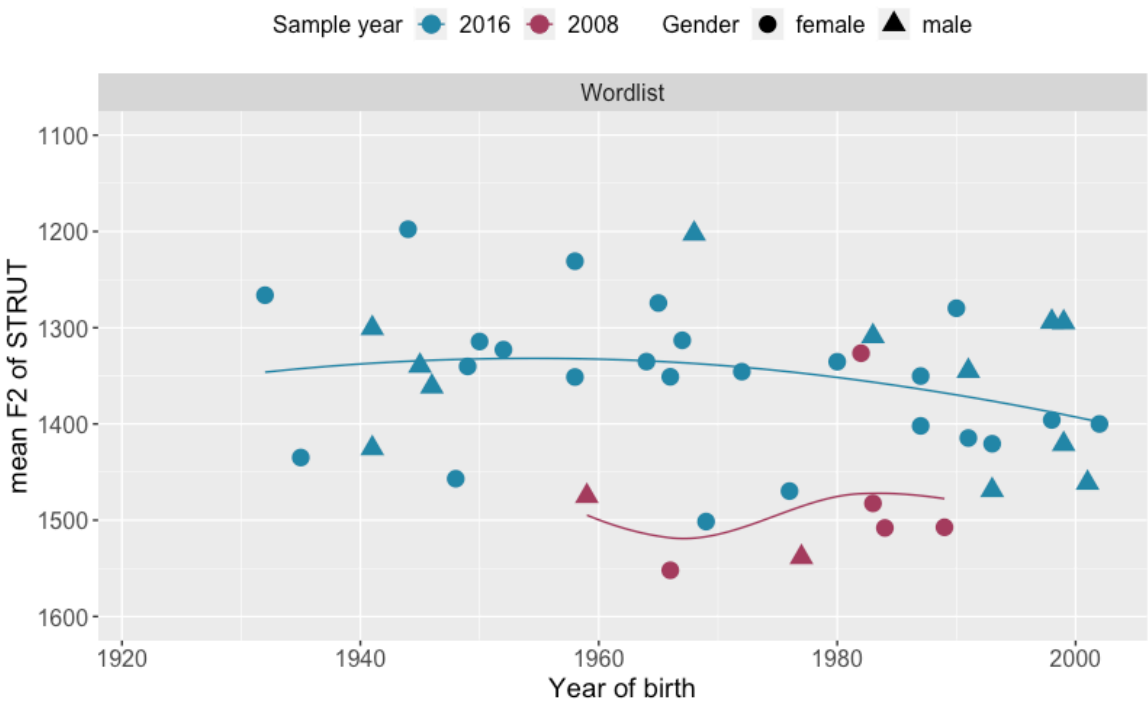


Figure 81: STRUT F2 means in wordlist style in 2008 and 2016 by *gender*.

[71] The main effect of *sample year* in Table 46 predicts a significant difference between the younger speakers in the two samples; however, the estimated amount of 213 Hz seems far-fetched based on the representation of the data in Figure 80 above, so that this does not appear to be a reliable prediction.

The regression model in Table 47 supports the observation that 2008 speakers produce a notably fronter wordlist STRUT than 2016 speakers; however, this difference is not found to be statistically significant. Interestingly, no apparent-time trend that would have led to this real-time difference can be observed in either of the two data sets.

Predictor	Coefficient	*p*
(Intercept)	817.217 Hz	
Age	-0.512 Hz	
Gender (Male)	-5.374 Hz	0.752
Education *(No college)* *(Student)*	 -16.12 Hz -15.919 Hz	0.626
Sample year *(2008)*	 212.979 Hz	
*Age*2008*	-5.162 Hz	0.016
Environment		0.541

Table 46: STRUT F1 in wordlist style in 2008 and 2016. Reference levels: females, college educated, 2016, /p/. Random effects: *speaker, word*. *n*= 329

Predictor	Coefficient	*p*
(Intercept)	1456.434 Hz	
Age	-0.848 Hz	0.262
Gender (Male)	-4.639 Hz	0.87
Education *(No college)* *(Student)*	 15.45 Hz 5.228 Hz	0.886
Sample year *(2008)*	 179.656 Hz	0.226
Environment		0.576

Table 47: STRUT F2 in wordlist style in 2008 and 2016. Reference levels: females, college educated, 2016, /p/. Random effects: *speaker, word*. *n*= 329

While the majority of speakers from the entire sample meet the UD criterion in wordlist style, the F2 relation between STRUT and LOT does seem to differ notably between the two data sets. This is illustrated in Figure 82 below.

In the 2016 data, most of the participants meet the UD criterion in wordlist style. However, there appears to be a slight decrease in UD participation, as many of the younger speakers cease to meet this criterion, indicating that, as a whole, the community is moving away from the reversal of STRUT and LOT on the front-back dimension in apparent time. This reconfiguration of STRUT and LOT to their traditional relative positioning seems to be primarily a result of LOT retraction in wordlist style among 2016 speakers (see Chapter 6.2.3.1.1), though for college educated speakers and students, STRUT fronting may be contributing to this reversal.

In 2008, all speakers meet the UD criterion in wordlist style, and their distances between LOT and THOUGHT are much greater than they are for most 2016 speakers, i.e. 2008 speakers produce wordlist STRUT notably further back than wordlist LOT, while this difference is much less severe for most 2016 speakers. Given that it was observed above that 2008 speakers actually produce a *fronter* STRUT than 2016 speakers in this speech style, this is quite a surprising finding. Counteracting real-time differences in the frontness of LOT, however, explain these surprising patterns, as 2008 speakers produce wordlist LOT significantly fronter than 2016 speakers (see Chapter 6.2.3.1.1). Thus, despite wordlist STRUT being fronter in 2008 than in 2016, UD participation in this speech style is higher in 2008, owing to an extremely fronted wordlist LOT among 2008 speakers.

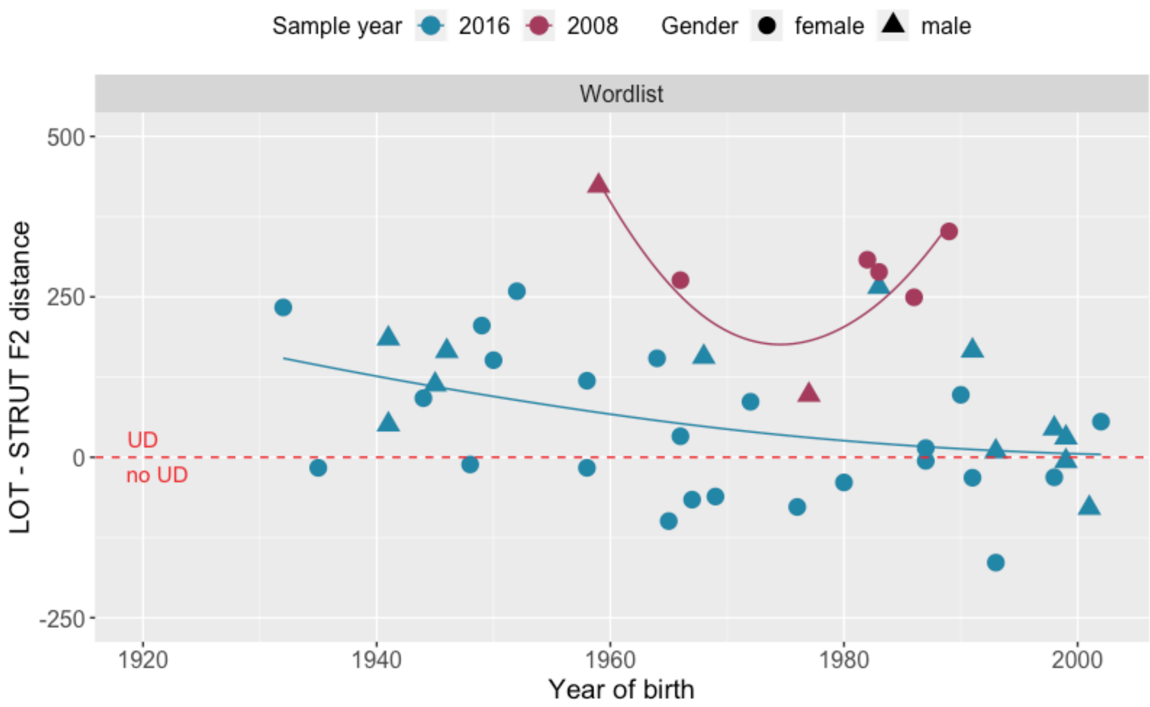


Figure 82: F2 distance between LOT and STRUT means in wordlist style in 2008 and 2016. A positive value indicates that STRUT is backer than LOT.

5.2.3 Style Shifting STRUT

The vast majority of speakers in the 2016 sample appear to produce a slightly higher and somewhat fronter STRUT in spontaneous speech than they do in wordlist style, as can be seen in Figures 83 and 84 below. For F2, this appears to have been a stable variation between the two speech styles; for F1, on the other hand, the range of style shifting appears to depend on the speaker's *age* in the 2016 sample.

Figure 83 below indicates that intra-speaker height variation is somewhat greater for younger speakers than it is for older speakers. This is to be expected, as the analysis above has found younger speakers, i.e. students, to produce a slightly higher spontaneous STRUT than the rest of the speakers in the sample, while no such differences emerged from the wordlist data. The regression model in Table 48 below supports these observations and suggests that F1 of spontaneous STRUT is lowering significantly in comparison to wordlist STRUT, leading to significant intra-speaker variation among younger participants.

Regression models that test for the effect of *style* among these younger speakers confirm the statistical significance of this variation, both for students (62 Hz, *p*= 0.001) and for speakers born after 1980 regardless of *educational status* (41 Hz, *p*= 0.002).

Figure 83: F1 difference between spontaneous and wordlist STRUT means in 2016 by *gender*. A positive value indicates that the vowel is higher in wordlist style than in spontaneous speech.

On the front-back dimension, Figure 84 below suggests that, compared to spontaneous speech, wordlist STRUT tokens are produced with a smaller F2, i.e. wordlist STRUT is realized about 100 Hz backer than spontaneous STRUT by most 2016 speakers. As can be seen in Figure 84, these style differences are minimized for the youngest 2016 speakers, owing to a slightly fronter wordlist STRUT for these speakers. The regression model in Table 49 below corroborates this impression, as it does not find any significant differences in F2 of STRUT between the two speech styles for the youngest speakers. A model that tests for this effect separately for speakers born after 1980 confirms this prediction (-40 Hz, *p*= 0.159). However, a model that considers only speakers born *before* 1980 estimates similarly insignificant style differences (-45 Hz, *p*= 0.056). Although for this group of speakers the difference just barely misses the level of significance, these results suggest that, on the F2 dimension, style differences appear to be a more or less stable variation.

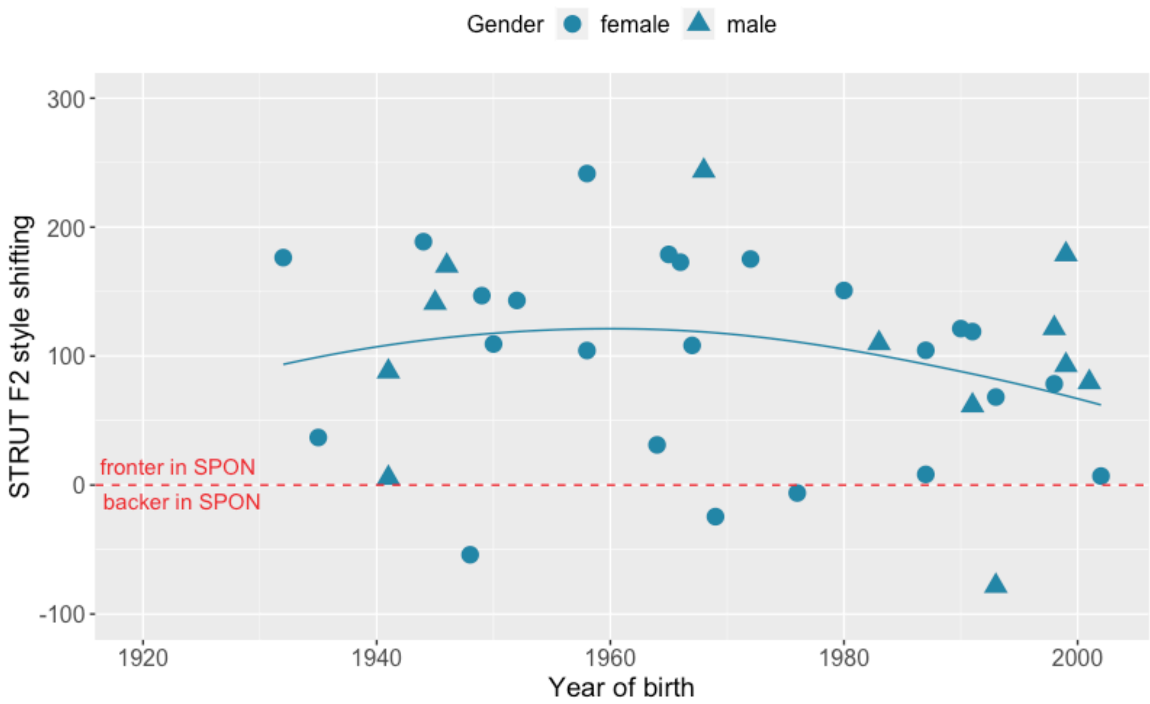


Figure 84: F2 difference between spontaneous and wordlist STRUT means in 2016 by *gender*. A positive value indicates that the vowel is backer in wordlist style than in spontaneous speech.

Predictor	Coefficient	*p*
(Intercept)	750.032 Hz	
Age	0.134 Hz	
Gender (Male)	7.16 Hz	0.37
Education *(No college)* *(Student)*	 -14.581 Hz -28.542 Hz	0.05
Style (Wordlist)	52.341 Hz	
*Age*wordlist*	-0.707 Hz	0.001
Environment		0.002

Table 48: Effect of *style* on F1 of STRUT in 2016. Reference levels: females, college educated, spontaneous speech, /p/. Random effects: *speaker, word*. *n*= 5654

Predictor	Coefficient	*p*
(Intercept)	1404.419 Hz	
Age	0.036 Hz	
Gender (Male)	-8.002 Hz	0.557
Education *(No college)* *(Student)*	 5.696 Hz 0.371 Hz	0.925
Style (Wordlist)	9.556 Hz	
*Age*wordlist*	-0.943 Hz	0.008
Environment		$9x10^{-7}$

Table 49: Effect of *style* on F2 of STRUT in 2016. Reference levels: females, college educated, spontaneous speech, /p/. Random effects: *speaker, word*. *n*= 5654

In terms of style shifting the height of STRUT, 2008 speakers appear to behave in similar ways as speakers from the 2016 sample. This can be seen in Figure 85 below. While style differences are minimal for older 2008 speakers, younger speakers in this sample appear to shift to a notably lower STRUT in the more careful style, the same pattern that was observed in 2016.

For style shifting in F2 of STRUT, on the other hand, opposing patterns can be observed in the 2008 and 2016 samples. As Figure 86 below shows, especially younger

2008 speakers appear to shift to a notably fronter STRUT in wordlist style compared to their spontaneous speech, while 2016 speakers shift to more retracted STRUT.

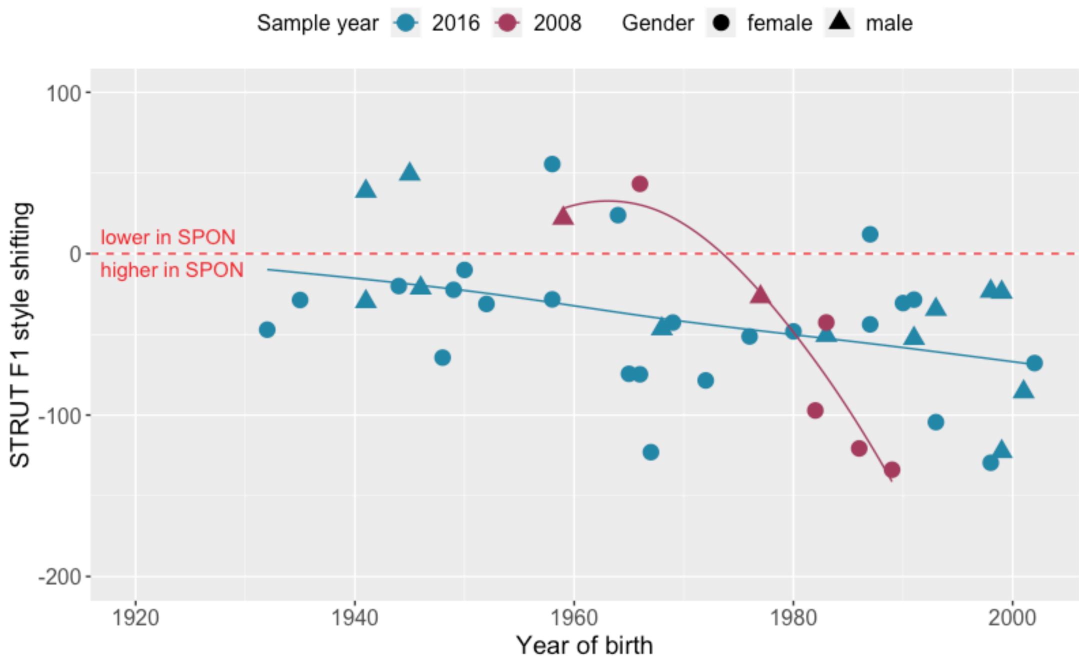


Figure 85: F1 difference between spontaneous and wordlist STRUT means in 2008 and 2016 by *gender*. A positive value indicates that the vowel is higher in wordlist style than in spontaneous speech.

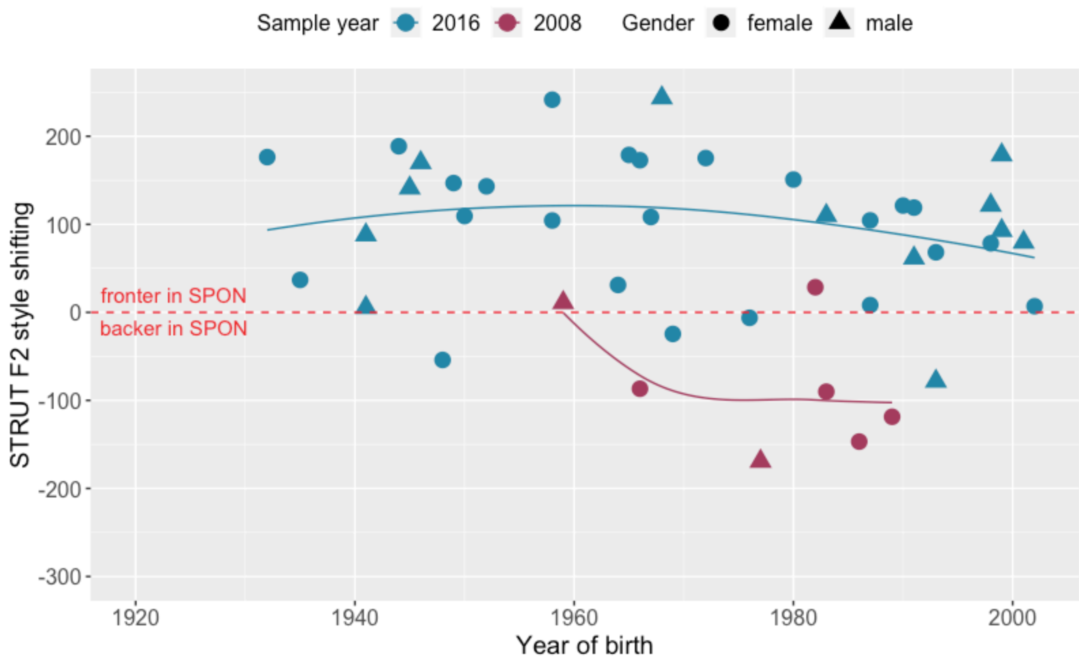


Figure 86: F2 difference between spontaneous and wordlist STRUT means in 2008 and 2016 by *gender*. A positive value indicates that the vowel is backer in wordlist style than in spontaneous speech.

Because of these different style-shifting patterns regarding the frontness of STRUT, participation in the UD criterion differs between the two speech styles and the two samples. As can be seen in Figure 87, participation in the UD criterion presents a very similar picture across the two speech styles for 2016 speakers. Overall, slightly more speakers from this sample participate in UD in wordlist style than spontaneous speech, and in both styles, older speakers are found to meet this criterion more commonly than younger speakers. Younger speakers have comparable F2 distances between STRUT and LOT across the two styles, and many of them cease to meet the UD criterion both styles, indicating that, as a whole, the community is moving away from the reversal of STRUT and LOT on the front-back dimension in apparent time. On the other hand, for most older speakers, the reversal of STRUT and LOT seems to be more secure in the more careful speech style than it is in spontaneous speech. This difference between younger and older generations is somewhat surprising, as no significant age differences were found in the F2 style-shifting patterns for STRUT in the 2016 sample. While older speakers were found to shift STRUT to a backer position in wordlist style slightly more than younger speakers from this sample, this difference alone seems unlikely to account for the style differences in UD participation. Style-shifting patterns in LOT resolve this contradiction: While younger speakers shift to a backer LOT in wordlist style (see Chapter 6.2.3.1.1), older speakers do not, so that their style shifting in STRUT leads to an increased distance to LOT.

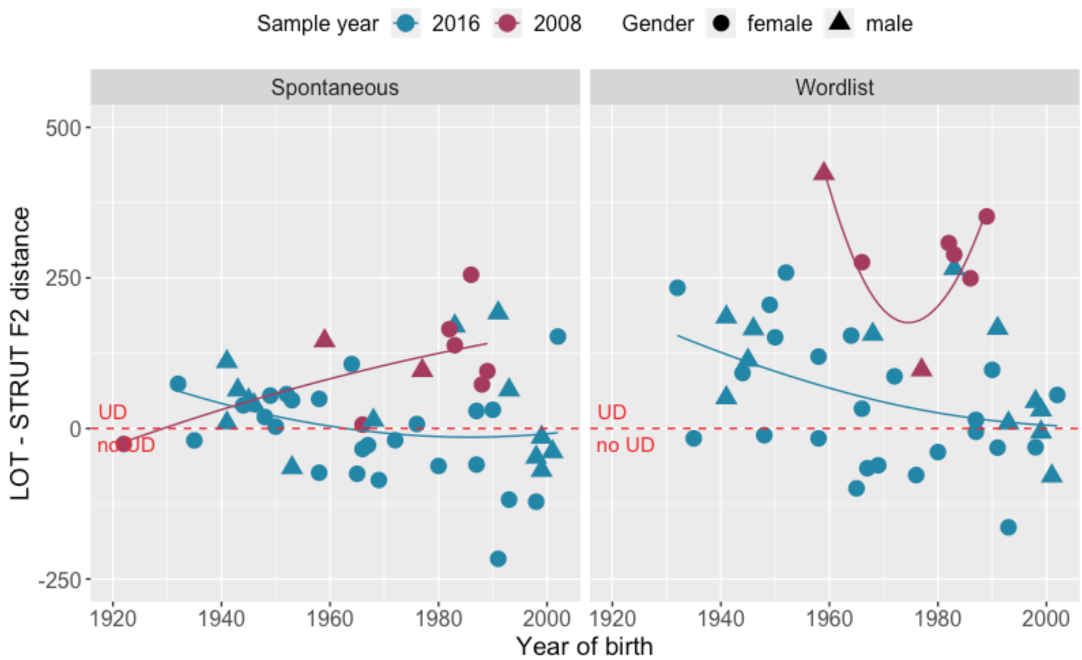


Figure 87: F2 distance between LOT and STRUT means in 2008 and 2016 across *speech styles* by *gender*.

Figure 87 above shows that, for most 2008 speakers, the F2 distance between STRUT and LOT is notably greater in wordlist style than it is in spontaneous speech. This is similar to the pattern observed for older speakers in 2016 but differs from the majority of speakers in the sample. This observation is unexpected, considering that 2008 speakers shift to a fronter STRUT in wordlist style, which would suggest that its distance to LOT would be reduced rather than amplified. Style-shifting patterns in the realization of LOT provide an explanation for these findings: 2008 speakers produce wordlist LOT significantly fronter than spontaneous LOT. Thus, despite wordlist STRUT being fronter than their spontaneous STRUT their LOT-STRUT distances increase in the more careful speech style owing to their extreme fronting of wordlist LOT (see Chapter 6.2.3.1.1).

5.3 Results: KIT

Figure 88 suggests little inter-speaker variation in the realization of KIT in the 2016 sample in spontaneous speech (wordlist data for KIT is only available for the 2008 speakers). Variation is particularly limited on the height dimension, while the frontness of KIT varies slightly more in the sample, though all speakers realize KIT well front of center, suggesting that it has never undergone centralization in the community.

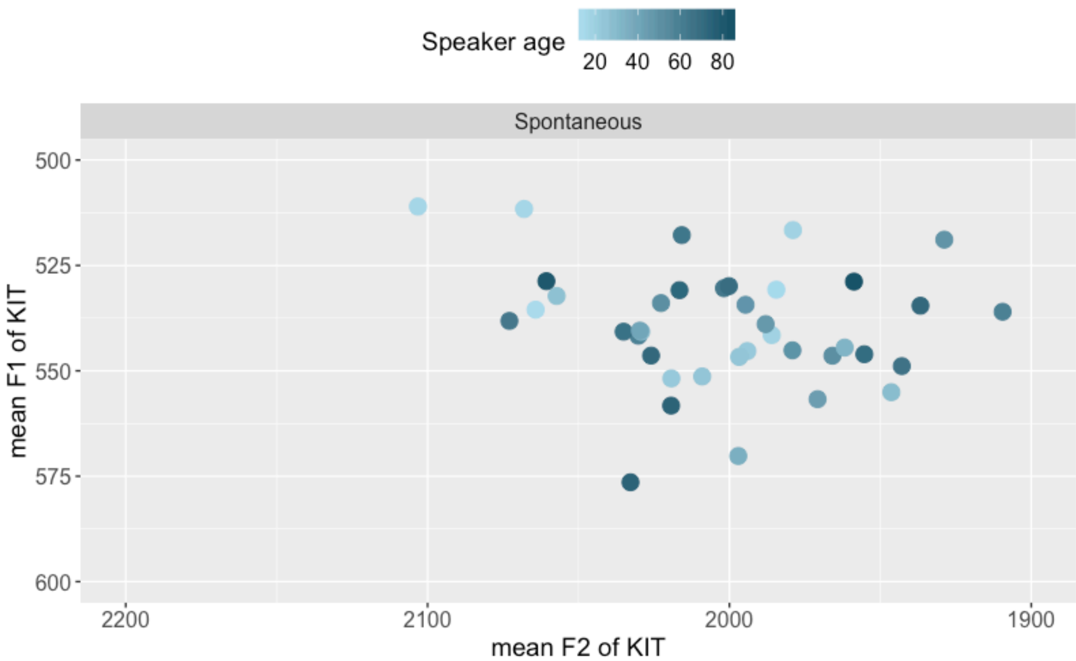


Figure 88: KIT F1 and F2 means in 2016 by *age*. Lighter shades represent younger speakers, darker shades older speakers.

As shown in Figure 89, the majority of speakers produce KIT with F1 means between 500 and 600 Hz, i.e. within a very narrow F1 range of less than 100 Hz. Figure 89 indicates a very slight trend toward lower KIT in apparent time, however, the students in the sample do not seem to participate in this trend. Instead, they realize KIT with F1 means similar to those of older speakers. Neither the potential apparent-time lowering nor the difference between students and non-student speakers is found to be statistically significant by the regression model in Table 50 below. Overall, the model corroborates the impression of relative homogeneity in the height of KIT in 2016.

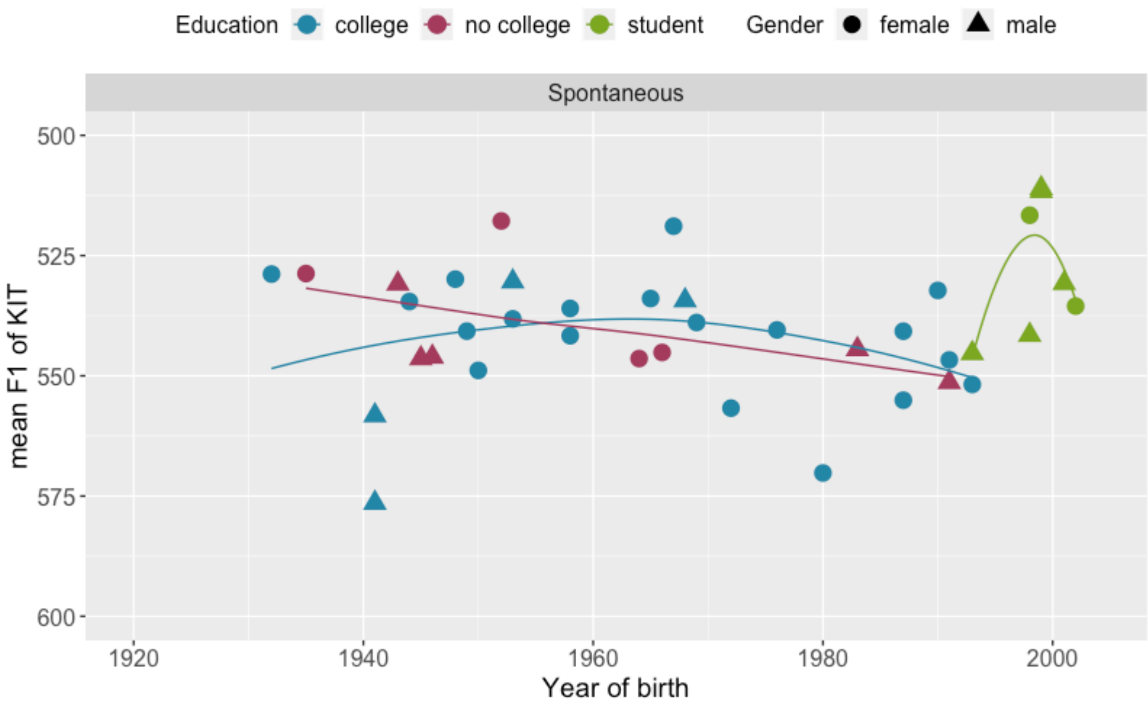


Figure 89: KIT F1 means in spontaneous speech in 2016 by *education* and *gender*.

On the front-back dimension, KIT shows a similar lack of inter-speaker variation as for its height. Figure 90 below shows that KIT is produced with an F2 between 1900 and 2100 Hz by the majority of speakers in the 2016 sample. Only a few speakers diverge from this F2 range, however, no noticeable patter in this deviation emerges from Figure 90 or the regression model in Table 51 below. Thus, the frontness of KIT does not appear to be subject to social variation in the community, or to be undergoing any changes in apparent time.

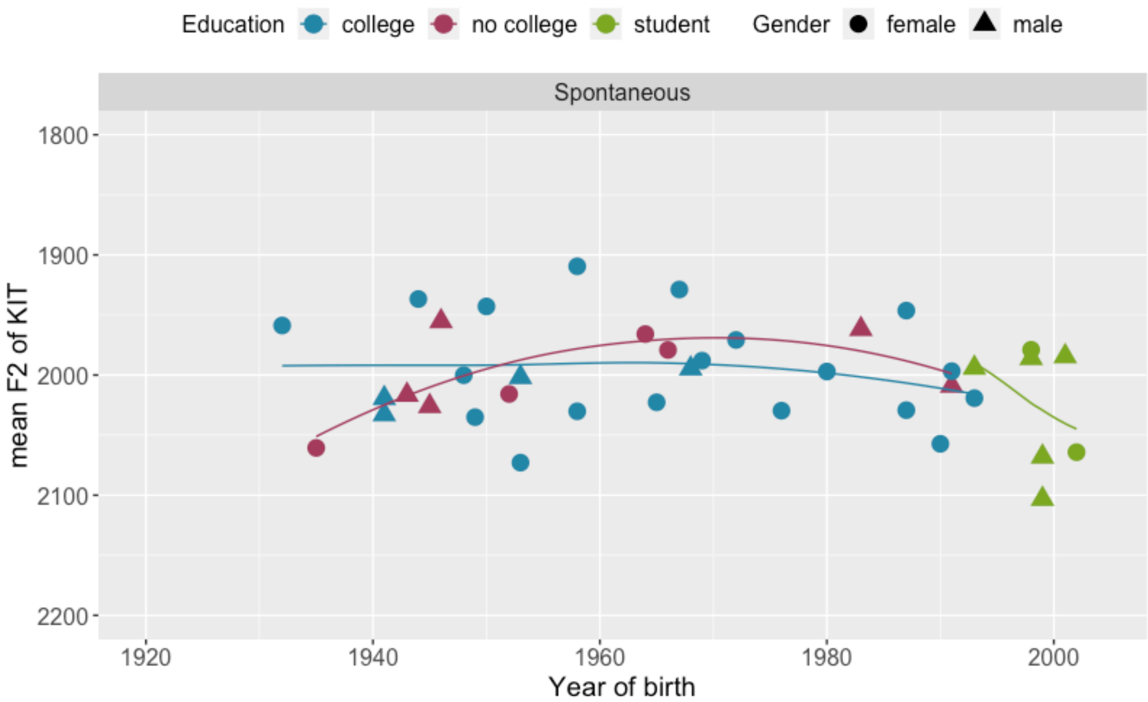


Figure 90: KIT F2 means in spontaneous speech in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	580.3 Hz	
Age	-0.138 Hz	0.295
Gender (Male)	8.263 Hz	0.105
Education *(No college)* *(Student)*	-0.01 Hz -13.9 Hz	0.224
Environment		2x10-16

Table 50: KIT F1 in spontaneous speech in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word. n*= 6467

Predictor	Coefficient	*p*
(Intercept)	1890.228 Hz	
Age	0.031 Hz	0.928
Gender (Male)	-4.887 Hz	0.708
Education *(No college)* *(Student)*	-14.037 Hz 21.342 Hz	0.282
Environment		2x10-16

Table 51: KIT F2 in spontaneous speech in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word. n*= 6467

In a comparison of the 2016 and 2008 data, only minimal differences can be observed in the height and frontness of KIT in spontaneous speech.

Compared to the 2016 sample, speakers in the 2008 data set produce KIT with very similar F1 means, i.e. at the same height, as can be seen in Figure 91 below. This lack of real-time differences, confirmed by the regression model in Table 52 below, corroborates the relative stability of the height of KIT in the community.

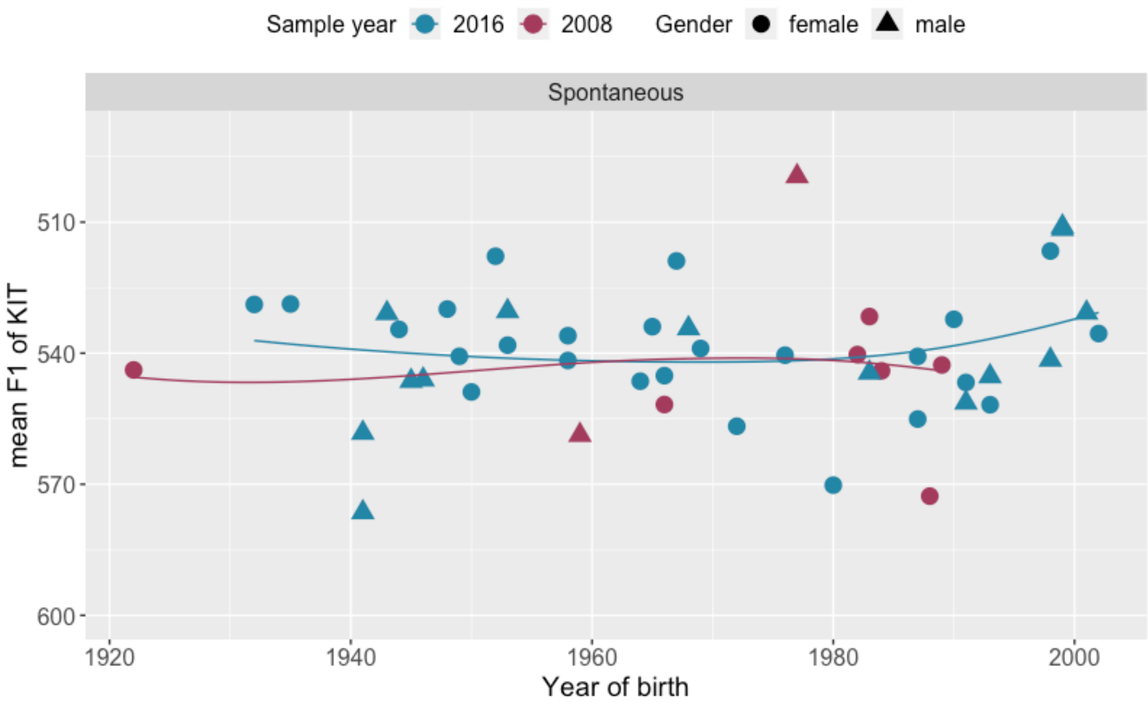


Figure 91: KIT F1 means in spontaneous speech in 2008 and 2016 by *gender*.

Predictor	Coefficient	*p*
(Intercept)	576.621 Hz	
Age	-0.048 Hz	0.686
Gender (Male)	3.735 Hz	0.417
Education *(No college)* *(Student)*	 -1.086 Hz -5.949 Hz	0.694
Sample year *(2008)*	1.281 Hz	0.836
Environment		$2x10^{-16}$

Table 52: KIT F1 in spontaneous speech in 2008 and 2016. Reference levels: females, college educated, 2016, /p/. Random effects: *speaker, word*. *n*= 6751

In terms of frontness, some differences do emerge between the 2008 and 2016 samples. While no changes over time could be observed in the 2016 data, Figure 92 below suggests an apparent-time backing of KIT in 2008. While the oldest speaker in the 2008 sample has a mean F2 of more than 2000 Hz, the youngest speakers average below 1900 Hz. In a real-time comparison, however, this backing cannot be confirmed, as 2016 speakers produce a notably fronter KIT than most 2008 speakers. The regression model in Table 53 below corroborates these impressions. Speakers from the two samples are found to differ significantly in the frontness of KIT, as well as in apparent-time developments of this frontness. A model that tests for the effect of *age* in the 2008 data separately does not find it to be statistically significant (0.611 Hz, p=0.501).

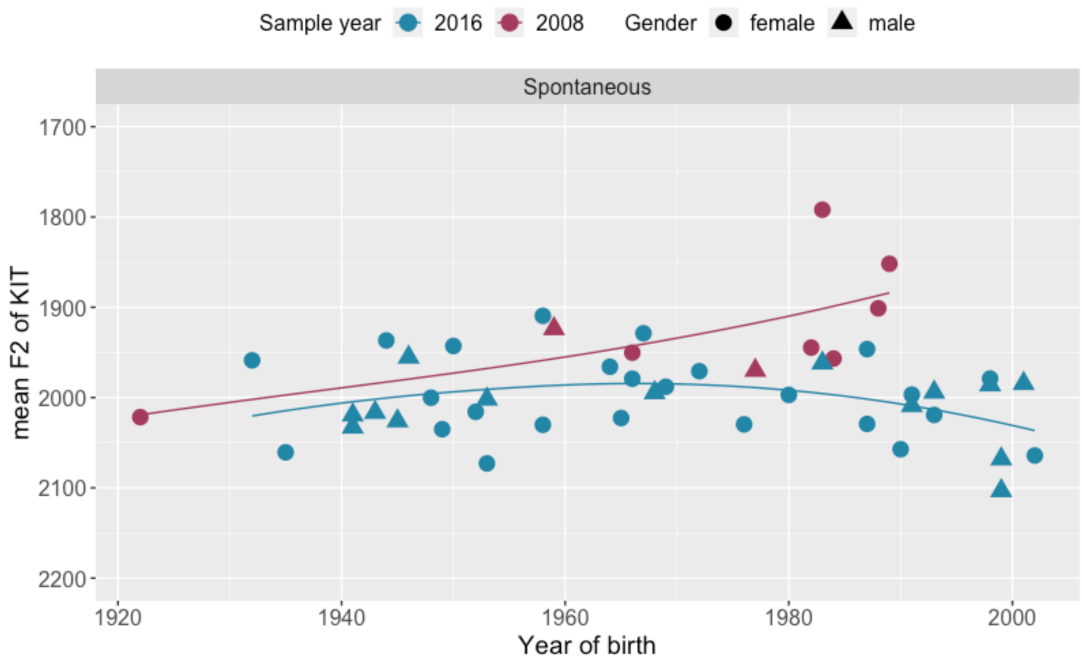


Figure 92: KIT F2 means in spontaneous speech in 2008 and 2016 by *gender*.

Predictor	Coefficient	p
(Intercept)	1895.397 Hz	
Age	-0.1 Hz	
Gender (Male)	2.108 Hz	0.86
Education *(No college)* *(Student)*	 -21.517 Hz 4.144 Hz	0.23
Sample year *(2008)*	 -103.144 Hz	
*Age*2008*	1.816 Hz	0.021
Environment		2x10^{-16}

Table 53: KIT F2 in spontaneous speech in 2008 and 2016. Reference levels: females, college educated, 2016, /p/. Random effects: *speaker, word*. n= 6751

The developments in the frontness of spontaneous KIT in 2008 suggest the adoption of the NCS in spontaneous speech. The 2008 wordlist data, on the other hand, suggest an apparent-time *fronting* of KIT, as shown in Figure 93 below, which counters its NCS trajectory. This also suggests a good amount of style shifting in the frontness of KIT in 2008, which is confirmed by the F2 style-shifting slope in Figure 94 below. This suggests that at least some of the 2008 speakers who were backing KIT in spontaneous speech realize it notably fronter in more careful speech, indicating a shift *away* from the NCS in more careful speech.

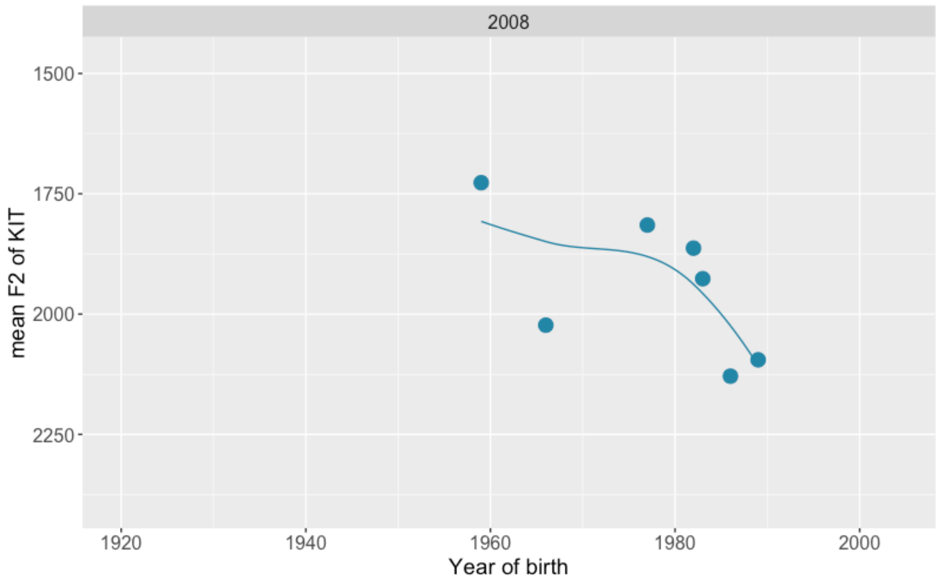


Figure 93: KIT F2 means in wordlist style in 2008.

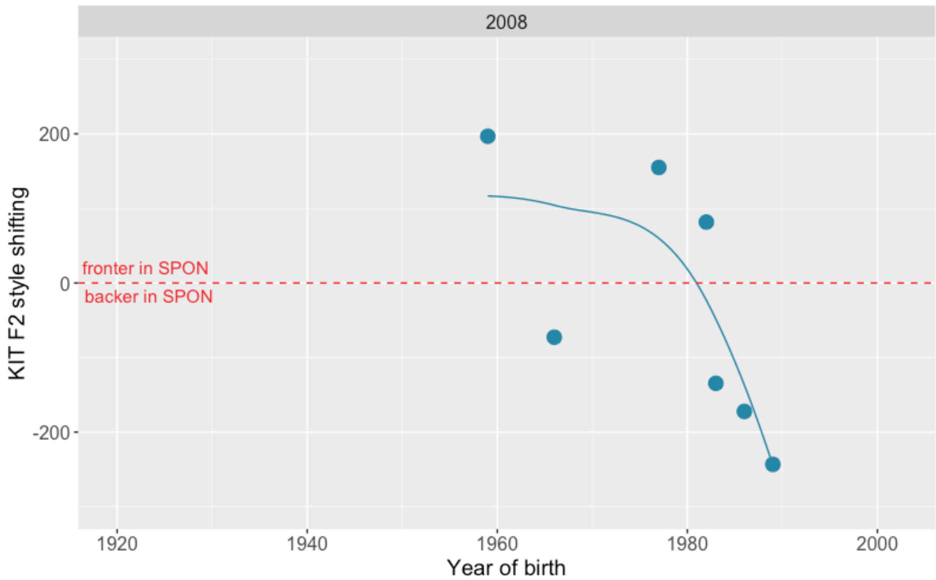


Figure 94: F2 difference between spontaneous and wordlist KIT means in 2008. A positive value indicates that the vowel is more backer in wordlist style than in spontaneous speech.

5.4 Summary: STRUT and KIT in Production

The analyses of STRUT and KIT showed that, overall, there is only slight, but very interesting variation in the realizations of both vowels, both in apparent and real time.

In spontaneous speech, some inter-speaker variation can be observed in the realizations of STRUT and KIT, particularly in real time. Both STRUT and KIT appear to have undergone momentary backing in the 2008 data, though neither of these two trends are statistically significant. Nevertheless, younger 2016 speakers realize a notably fronter spontaneous STRUT and KIT than their 2008 peers, as can be seen in Figures 95 and 96.

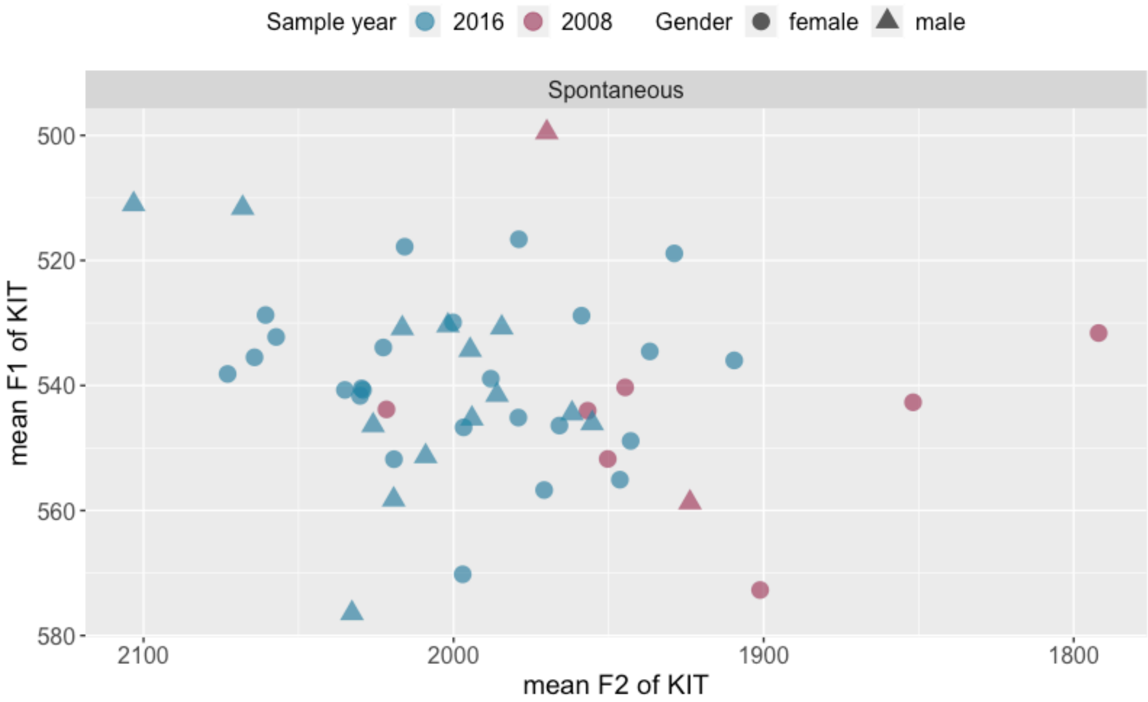


Figure 95: KIT F1 and F2 means in spontaneous speech in 2008 and 2016 by *gender*.

Furthermore, the frontness of spontaneous STRUT relative to LOT and the resulting participation in the UD criterion was found to differ between the two samples. While all but one of the 2008 speakers meet this criterion and are increasing the F2 difference between STRUT and LOT in apparent time (i.e. more retracted STRUT and fronter LOT), 2016 speakers participate in this criterion proportionally less. In fact, younger 2016 speakers are re-reversing the relative positions of STRUT and LOT, i.e. realizing STRUT in a position that is fronter than LOT. These differences between the 2008 and 2016 data are, however, primarily due to developments of LOT in spontaneous speech, though spontaneous STRUT also appears to be fronting in apparent time among college educated speakers in 2016.

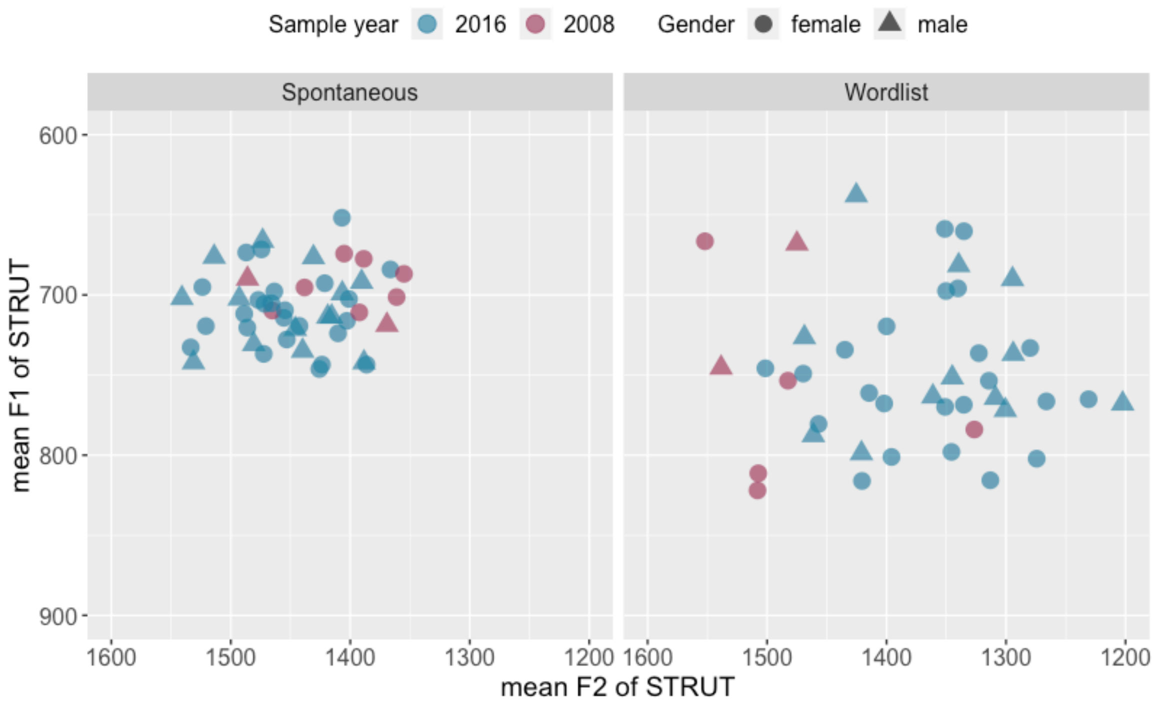


Figure 96: STRUT F1 and F2 means in 2008 and 2016 across *speech styles* by *gender*.

In wordlist style, the height of STRUT appears to be stable, while in terms of frontness, real-time differences suggest backing, and apparent-time trend suggest fronting of STRUT. Although the 2008 data suggest significant lowering of STRUT in apparent time, the 2016 data does not corroborate this trend, suggesting that the low starting point in 2008 was exceptional. Real-time differences do appear to exist on the front-back dimension, however. As can be seen in Figure 96 above, 2008 speakers produced a notably fronter STRUT than 2016 speakers in wordlist style. Interestingly, there are no apparent-time trends in either of the two data sets that would indicate a transition toward backer STRUT. In fact, college educated speakers in 2016 seem to be *fronting* wordlist STRUT. Compared to spontaneous speech, wordlist STRUT is produced lower and backer by the majority of 2016 speakers, while 2008 speakers shift to a fronter variant of STRUT in more careful speech. In other words, while 2008 speakers appear to have moved toward NCS STRUT in spontaneous speech, they were moving away from this pattern in more careful speech. A similar trend was observed for KIT in the 2008 data: In spontaneous speech, 2008 speakers seem to have retracted KIT in apparent time, indicating adoption of the NCS, while in wordlist style, they shifted toward a fronter realization of KIT, thus moving away from NCS patterns.

Despite the minimal real-time differences in the production of STRUT on the front-back dimension, there are significant differences in the realization of STRUT in relation to LOT. While all but one of the 2008 speakers meet the UD criterion, i.e. realize STRUT in a position backer than LOT, in both speech styles, only half of the 2016 speakers do so. Additionally, 2008 speakers seem to be increasing the distance between STRUT and LOT in apparent time and in more careful speech, with STRUT being produced in an increasingly backer position than LOT. Speakers in the 2016 sample, on the other hand, seem to be decreasing the distance between STRUT and LOT in apparent time in both speech styles, to the extent that many of the younger speakers no longer produce STRUT in a position backer than LOT. Many of these developments are, however, results of significant real-time differences in the frontness of LOT (see Chapter 6.2.3.1.1).

5.5 Discussion

The analysis in this chapter has shown that STRUT and KIT seem to have undergone the shifts that would have been expected from a community participating in the NCS; however, they did so only momentarily. KIT is now again realized in position well front of center and does not appear to be undergoing any apparent-time changes in any direction, while STRUT continues to be realized just back of center, though without strong indications of backing in apparent time. In fact, STRUT appears to be fronting for a certain group of speakers in apparent time. These findings for both KIT and STRUT contradict early reports of New York and the rest of the Inland North, where both have been found to shift along their respective NCS trajectories, i.e. backward. Real-time fronting of KIT also contradicts expected Elsewhere developments, as KIT would have been expected to centralize as part of this shift. STRUT fronting, however, is in line with changes of this vowel in the Elsewhere Shift.

In both cases, it seems that retracted variants were targeted for a short amount of time in spontaneous speech in 2008, but were simultaneously rejected in more careful speech, and later in spontaneous speech by the entire community. This is the same pattern that was identified for TRAP in Chapter 3.2. Contrary to TRAP, however, there is no evidence that the rejection of backed KIT and STRUT is due to negative evaluations of these variants. No one has, as of yet, reported any kind of social evaluation for either of the two vowels; likewise, the matched guise experiment in the present study returned no noteworthy results for STRUT (KIT was not included in the experiment). Thus, the motivation for the sudden rejection of backed KIT and STRUT may be of a different nature than that for raised TRAP.

In the case of STRUT, it is possible that backing may simply have reached its limits in the community, potentially because a different phoneme that occupies the phonetic space back of STRUT is preventing it from backing any further. Figure 97 suggests that this might be the case, as THOUGHT is produced just back of STRUT, with only small F2 differences between, and in fact, some overlap of the vowels. Only in wordlist style does THOUGHT appear to be moving out of the way by lowering in the vowel space for some speakers, thus making room for STRUT to move back. However, the analysis of THOUGHT in Chapter 6.2.3.1.2 will show that THOUGHT lowering is not a particularly significant trend in the community. Thus, it is possible that further backing of STRUT has been blocked by the presence of THOUGHT, especially in spontaneous speech. Why the community seems to allow more overlap of the vowels in wordlist style than in spontaneous remains a bit of a mystery.

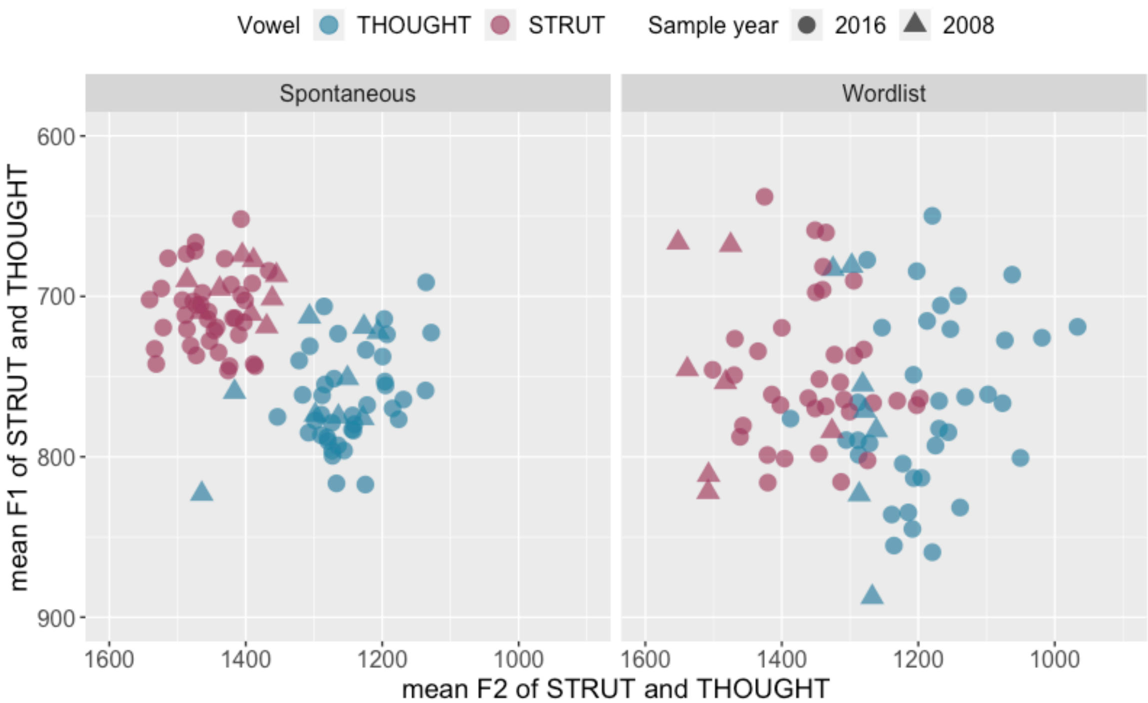


Figure 97: STRUT in relation to THOUGHT in 2008 and 2016 across *speech styles*.

The small F2 distance to THOUGHT suggests that, despite the rejection of the backed variant of STRUT of the younger 2008 speakers and the absence of apparent or real-time backing, STRUT *is* still realized in a relatively retracted position. According to E. R. Thomas (2001), STRUT is considered backed when it is realized back of center, which seems to always have been the case in Ogdensburg as Figure 97 above suggests. How STRUT came to be realized in this retracted position in the first place cannot be answered with any certainty in the present study, as there are no strong indications in the data that it was moved to this backed position at a certain point in time. The most plausible explanation therefore is that STRUT was backed at a time that precedes the years tracked in the present study. If this was the case, however, this would have happened independently of the NCS, especially considering that the backing of STRUT was, presumably, one of the latest changes in the NCS. Under this scenario, it is possible that early settlers brought retracted STRUT with them from New England, where retracted realizations of STRUT are commonly found, or that it has diffused from there to Ogdensburg at a point in time that precedes the current sample of speakers.

A last observation worth discussion is that, similar to the patterns reported for DRESS, students as a group differ slightly from the rest of the sample, producing higher KIT and STRUT than most other speakers in spontaneous speech. For neither KIT nor STRUT, however, is higher realization a contra-movement to an apparent-time lowering, nor does it appear to be an orientation toward older norms, as was found to be the case with DRESS. Thus, age-grading or influence from the students' parents are unlikely sources for their diverging behavior. The most plausible explanation, in the case of KIT and STRUT (and probably DRESS), is that these differences are simply not socially meaningful, especially considering that the F1 differences between students and other participants are very small.

Chapter 6: The Merger of LOT and THOUGHT

6.1 The Merger of LOT and THOUGHT – An Introduction

The merger of LOT and THOUGHT, i.e. COT-CAUGHT or low back merger, has been observed in varieties of American English since the early 1900s, most notably in Eastern New England, Eastern Pennsylvania and Canada, and today, in more than half of the continent. Even in the Inland North, where the presence of the NCS was assumed to offer "stable resistance" to the merger owing to LOT being fronted away from THOUGHT, recent studies have reported progress toward the merger.

One of the reasons that this merger is so common across many varieties of English presumably lies in the unstable relationship between LOT and THOUGHT that has characterized these low back vowels throughout the history of English. In order to get a better understanding of the ever-changing relationship between LOT and THOUGHT, their most important diachronic developments, starting in Middle English, will be described in the following subchapter, based on descriptions by Dobson (1957) and Wright (1905) as summarized by Johnson (2007), as well as on accounts provided by Wells (1982), Labov (1994) and *ANAE*. Many of these developments involved splits and mergers of various kinds, which often were somewhat messy rather than phonologically straightforward. Thus, the overview below is by no means a complete history of the developments of these two vowel classes, as many exceptions have played into the general rules. However, the outlined developments should provide an understanding of the formation of these two vowel classes that is sufficient for the purposes of the present study.

6.1.1 The History of LOT and THOUGHT

Today's LOT developed mainly out of Middle English [ɔ]. During the 16th and 17th century, [ɔ] lowered to [ɒ] in most phonological environments, e.g. in *pot* and *honest*, where, for the most part, it remained in most British varieties of English. In most American varieties, those that did not undergo immediate merger with [ɔ:], as well as in e.g. East Anglia, [ɒ] lowered and unrounded to today's [ɑ] in the 19th century and was joined by a backed variant of [a:] in pre-r environments (e.g. *far*) and the PALM class (e.g. *calm*). These developments are illustrated in Figure 98 below.

Today's THOUGHT is primarily the result of backing and monophthongization of Middle English /aʊ/ (e.g. *law*, *ball, taught*) and lengthening of Middle English [ɔ] as shown in Figure 98 below. In the 16th and 17th century, /aʊ/ monophthongized to [ɔː] in most environments. Simultaneously, [ɔ] split into [ɒ] as mentioned above, and a lengthened variant [ɔː]. The latter outcome, however, was limited to certain phonological environments. In British English, it only occurred before tautosyllabic /r, f, s, θ/, as in *north*, *off*, *loss*, *cough* in Southern England (NORTH and CLOTH), though it was maintained only in pre-/r/ environments, whereas in environments preceding fricatives it is now obsolescent, and CLOTH merged with LOT instead. In American English, lengthened [ɔː] in positions before both /r/ (NORTH) and voiceless fricatives (CLOTH) and before /ŋ/ collapsed with the monophthongized [ɔː] that had developed out of /aʊ/. Thus, CLOTH words such as *loss* and *cough* now commonly have THOUGHT in US varieties, and LOT in many British varieties of English.

This new [ɔː] contrasted from the newly developed [ɒ] only by length until [ɒ] was unrounded to [ɑ] in some varieties of English, including most US varieties. In Scotland, these two classes seem to have collapsed immediately, merging also with [a], with evidence dating back to the 17th century. In West England, on the other hand, [ɔː] was present, but [a] merged with [ɒ] at a later point. Many British varieties, however, maintained the three-way distinction between [a], [ɒ] and [ɔː], sometimes with alterations, commonly raising [ɔː] toward [o] (E. R. Thomas, 2001), or unrounding of [ɒ] (and then rounding again) (Trudgill, 1986). The same configuration of three distinct low vowels was also maintained in varieties of English spoken in the southern hemisphere (Wells, 1982). Figure 98 below summarizes these developments.

Owing to its historic formation, [ɔː] was constrained to relatively few phonological environments. Its distribution was skewed further in the 19th century by the migration of certain [ɒ]/[ɑ] words to the [ɔː] class by means of phonologically conditioned lexical transfer. This lexical diffusion from [ɒ]/[ɑ] to [ɔː] occurred before /θ, s, ŋ, ɡ/ (e.g. *broth, loss, long, dog*). Because this process was rather irregular, unpredictable lexical variation in the occurrence of [ɒ]/[ɑ] and [ɔː] is very common in these environments. Generally, uncommon and polysyllabic words remain unaffected by this transfer, i.e. maintaining [ɒ]/[ɑ], while high-frequency words often have [ɔː] instead of [ɒ]/[ɑ]. This transfer occurred in Southern England as well as in American varieties, however, it generally reversed in RP.

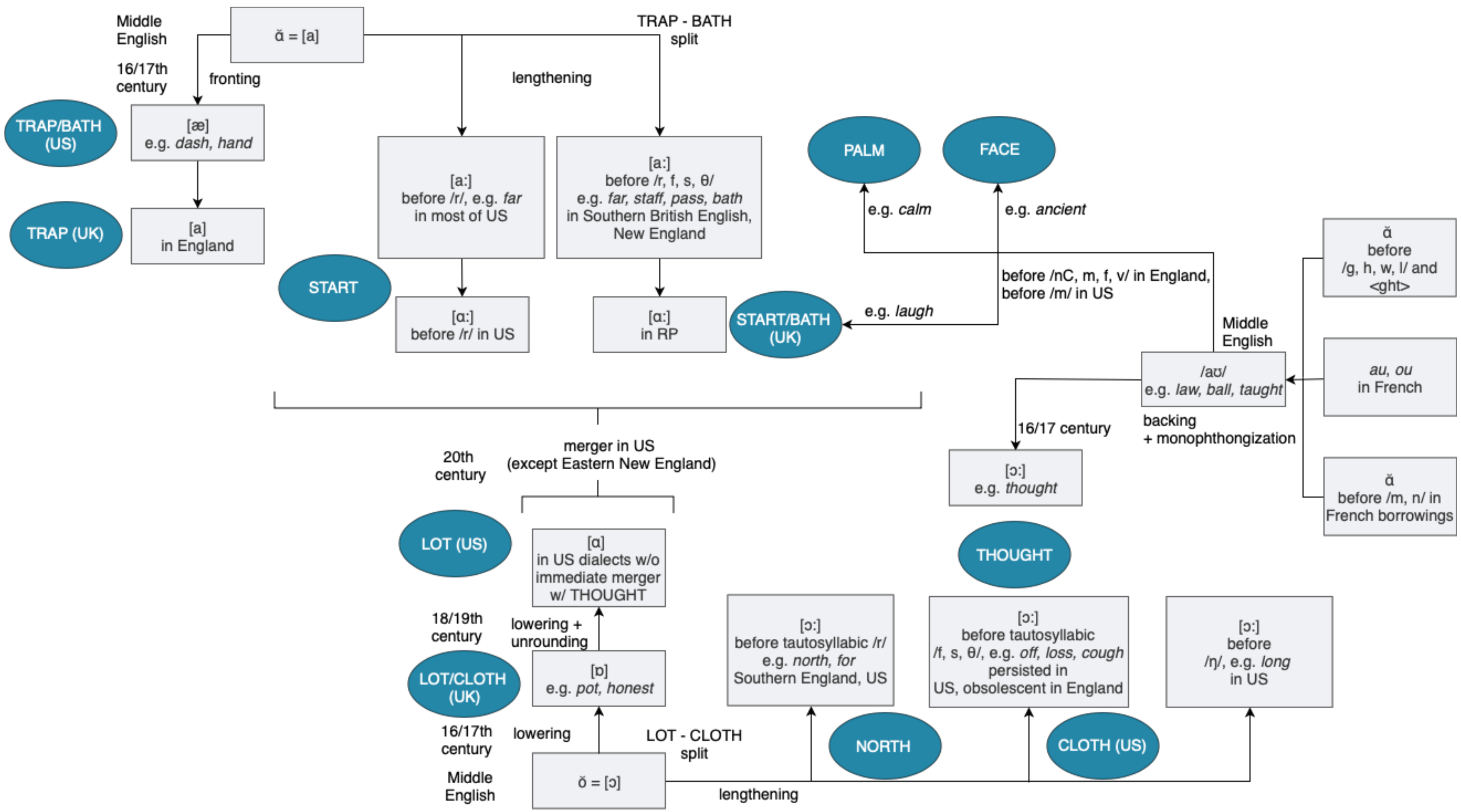


Figure 98: Historic development of LOT and THOUGHT.

In American English, the unstable relation between [ɑ] and [ɔː], now known as LOT and THOUGHT, continued and developed in multiple directions. For the vast majority of speakers, [ɑ] fronted to merge with [a] (e.g. *father*, *pajamas*), probably in the early 20th century, the only exceptions being Eastern New England and some coastal Southern areas, where they remained distinct (*LYS*). Frequently, but not always, this joined class later merged with THOUGHT, creating a three-way merger that today characterizes the speech of the majority of speakers as either a rounded or unrounded phoneme [ɑ, ɒ]. In some cases, however, LOT and THOUGHT merged without including [a], as for example in Eastern New England. Other dialects took measures to increase qualitative differences between LOT and THOUGHT. In New York City, Philadelphia and the Mid-Atlantic States, THOUGHT raised to mid or high back position, sometimes with the addition of an inglide, and in the South, THOUGHT developed a back upglide. In the Inland North, the stabilization of phonetic contrast between LOT and THOUGHT was achieved through the fronting of LOT to low central position, which, presumably, had been brought to the area by the early settlers from New England. As a result, both vowels have been clearly distinct in the North (e.g. Clopper et al., 2005).

6.1.1.1 LOT in the Inland North

LOT fronting is a relatively rare process in American English, and outside of the Inland North, fronted variants have only been reported for parts of the Midland, Southwestern New England, and St. John's in Newfoundland.

The fronting of LOT in the Inland North was first noted by Fasold (1969) in Detroit, however, it is believed that the early settlers of this area brought fronted LOT with them from New England rather than it being a new development unique to the Inland North. The NCS trajectory of LOT has mostly been described as unidirectional, moving LOT toward the front of the vowel space (e.g. Eckert, 1988; Labov, 1994; *LYS*), though fronting was sometimes found to be accompanied by a certain degree of lowering (e.g. Gordon, 2001). In this movement, LOT approached the qualities of TRAP, and in extreme cases resulted in a TRAP-like sound unless it was simultaneously lowered. In relation to THOUGHT, LOT in the Inland North has been described as being clearly distinct without any indication of approximation (McCarthy, 2007), the only exception being in pre-/r/ environment, where even speakers in the Inland North are fully merged (Labov, 1994). The reason behind this, of course, is that LOT fronting in the NCS moves LOT out of its low back position and away from THOUGHT.

Social Distribution of *LOT*

Regarding the social distribution of fronted LOT in the Inland North, earlier studies have come to different conclusions. Only one of them reported an ongoing change in apparent time in the fronting of LOT (*ANAE*). Some of the studies found that middle class speakers, or those who oriented toward the middle class, were more advanced in LOT fronting than lower middle-class speakers or those orienting toward the working class (Gordon, 2001; Herndobler, 1993). Similarly, *ANAE* found that higher education levels are correlated with increased LOT fronting. The findings regarding gender are even more diverse. While some found the generally assumed female lead to apply to LOT fronting (Eckert, 1989; Gordon, 2001), some found no significant differences between male and female speakers (e.g. Fasold, 1969; Herndobler, 1993), and others found males to be leading this change (*ANAE*), although it was also reported that the gender differences in these cases were not significant. In terms of style, the only study to report differences between speech styles found that wordlist tokens were more advanced in the fronting of LOT (McCarthy, 2007).

Recent Developments of *LOT*

More recent studies examining LOT fronting in the Inland North have observed a retreat from this feature (e.g. Fox, 2014, 2016; McCarthy, 2010; Morgan et al., 2017; Nesbitt, 2016; Wagner et al., 2016). Savage et al. (2016) suggest that change in progress toward more retracted LOT is led by young, college educated speakers; however, they do not provide any evidence or references for this. Wagner et al. (2016) have described the starting point for LOT retraction to lie among speakers born in the early 20th century, and add that some of the speakers in their sample have retracted LOT so far as to have merged with THOUGHT, a development that has long been thought impossible in NCS communities.

The reason for the retraction of LOT appear to be negative social meanings that have become attached to its fronted variant. Studies have shown that fronted LOT has reached the level of conscious awareness at least for some speakers (Bigham, 2010), and that it is evaluated negatively as “annoying” and “accented” (Savage et al., 2016).

6.1.1.2 THOUGHT *in the Inland North*

Lowering of THOUGHT is not uncommon in North American English. Most dialects that have undergone the COT-CAUGHT merger, even if merged on THOUGHT, show some degree of THOUGHT lowering. It has, however, been argued that the lowering and unrounding of THOUGHT for the merger is not historically related to the shifting of THOUGHT as part of the NCS (Stockwell & Minkova, 1997).

The shifting of THOUGHT in the Inland North was first observed in 1969 (Fasold) and described as part of the NCS by *LYS*. However, other than LOT, this particular change does not appear to be as tightly integrated into the NCS (*ANAE*). In the shifting of THOUGHT, THOUGHT loses some of its qualities that distinguish it from (non-NCS) LOT, i.e. height, frontness and rounding, as this shift has mostly been described as moving THOUGHT down and forward toward /ɒ/, or, if accompanied by unrounding, toward /ɑ/. The degree to which each of these processes occur, however, can vary within a single community (Gordon, 2001). Some studies report lowering of THOUGHT without fronting (*ANAE*; McCarthy, 2007), while others found a combination of both (e.g. Gordon, 2001). In a few cases, the shifting of THOUGHT and/or STRUT have resulted in phonetic overlap of these two vowels, either by means of STRUT backing toward THOUGHT (McCarthy, 2011) or THOUGHT fronting and unrounding, without lowering, toward STRUT (Gordon, 2001).

Social Distribution of THOUGHT

Fronting and lowering of THOUGHT was found to be socially conditioned to a certain extent. The majority of earlier studies found NCS-shifting of THOUGHT to be an ongoing process, often led by females (*ANAE*; Eckert, 1998; Gordon, 2001; McCarthy, 2007). In terms of social class, Fasold (1969) found that the upper working class and lower middle class were more advanced in shifting THOUGHT than speakers of other classes. This is mirrored by *ANAE*'s findings of speakers with higher levels of education shifting THOUGHT more frequently than those with lower levels of education. Eckert (e.g. 1988, 1989), on the other hand, found that working class-oriented students led over middle class-oriented students in this particular shift. McCarthy (2007) found no significant effect of social class on the shifting of THOUGHT in Chicago. No style shifting in the realization of THOUGHT has been reported as of yet.

Recent Developments of THOUGHT

Few of the more recent studies have taken developments in the realization of THOUGHT into account. Those that have, have come to contradictory conclusions. Wagner et al. (2016) do not report developments in THOUGHT specifically; however, they do point out that some of their younger speakers have adopted the COT-CAUGHT merger. Although this seems to be mainly due to the retraction of LOT, it suggests that THOUGHT is at least stable, if not approaching LOT in phonetic space. King (2017) reports apparent-time backing of THOUGHT in Rochester, while Durian and Cameron (2018) found THOUGHT to be lowering in Chicago, a change led by men. However, they also found that THOUGHT was not correlated to any other NCS features, indicating that THOUGHT was not an integral part of the NCS. None of these studies have considered potential positive or negative evaluations of different variants of THOUGHT.

6.1.2 LOT and THOUGHT in New York

Unrounding and centralization of LOT in New York State was observed as early as the 1830s (Labov, 2010, p. 162 referencing Barton, 1830), and NCS-like patterns were identified shortly after: In one of the earliest descriptions of the low back vowels in New York, Emerson (1891) reported some degree of THOUGHT fronting/lowering in Central New York in speakers most likely born between 1811 and 1841.[72] However, it has also been reported that LOT was frequently favored over THOUGHT in most words by most participants (Monroe, 1896), especially in the Upstate area. The only exception to this was St. Lawrence County, the home of Ogdensburg, where THOUGHT seemed to have been used considerably more frequently than in most other counties (C. K. Thomas, 1935-1937). Overall, C. K. Thomas pointed out, three distinct low back vowels were in use in the State of New York at that time: [ɔ], [ɒ] and [ɑ]. He further explained that speakers at that time did not appear to distinguish between [ɔ] and [ɒ], while [ɒ] were [ɑ] are clearly distinct (p. 68). For LOT, C. K. Thomas observed that the only identifiable variation, in accordance with the NCS, was fronting of the vowel. However, fronting of LOT in New York was found to occur “less frequently and less noticeably than in New England; … [and] is therefore more satisfactorily recorded as [ɑ˧] than as [a]” (p. 68). Additionally, C. K. Thomas noted a regional differentiation in the usage of fronted LOT when preceding /r/, which appeared

[72] Emerson (1891) provides, if any, only estimated ages or birth years for his participants.

to be more common among speakers from the central and western parts of New York, i.e. the birthplace of the NCS, than in the northern and eastern parts.

More recently, it has been found that LOT and THOUGHT are either discontinuing or reversing their movement along their NCS trajectories in the New York part of the Inland North. LOT has been found to be retracting quite rapidly throughout the State of New York in what appears to be a female led change (Dinkin, 2009, 2013; Driscoll & Lape, 2015; King, 2017). While most studies report a gradual change for LOT, this process has been found to be rather abrupt in Northern New York, the cut-off year being 1960 (Dinkin, 2009, 2011). THOUGHT, on the other hand, has been found to be stable (Dinkin, 2009, 2013) or backing (King, 2017) in apparent time. In cases of THOUGHT backing, LOT and THOUGHT are reported to remain securely distinct, with no encroaching merger (King, 2017). In the case of steadiness, on the other hand, the merger is reported as beginning to affect speakers' self-judgment. This effect, however, has been found to be relatively weak and sporadic, affecting only speakers born in 1982 or later (Dinkin, 2009).

The only exception to these observations in the State of New York is Ogdensburg. Ogdensburg was the only city in the Inland North Fringe where THOUGHT was still in the process of lowering in apparent time, while LOT showed only insignificant backing in apparent time, thus remaining in stable, low position (Dinkin, 2009). Nevertheless, Dinkin found the distance between both phonemes to be decreasing in apparent time as a result of THOUGHT lowering. In fact, out of all Inland North Fringe communities, evidence of an encroaching merger was strongest in Ogdensburg, with three out of nine speakers having transitional judgments, and one speaker, though distinct in judgments, had only 89 Hz in Cartesian distance between LOT and THOUGHT. How the relation between LOT and THOUGHT has developed more recently in Ogdensburg will be explored in the subsequent chapters.

6.2 Results: Merger in Production

In this chapter I will describe the status and progress of the merger of LOT and THOUGHT in production in Ogdensburg. First, the results for the two main measures employed to assess the degree of merger in production will be presented: Euclidean distance, and Bhattacharyya scores (see Chapter 2.6.3). For both measures, the 2016 data will be analyzed in detail first, looking into potential social effects of *speaker age*, *gender* and *education*, as well as differences between the three different speech styles (spontaneous speech, wordlist style, minimal pairs), before the results from the 2016 sample will be compared to speakers recorded in 2008. In a second step, complementary measures will be taken into consideration in order to determine the underlying mechanism behind the merger. These include developments in F1 and F2 of LOT and THOUGHT and their respective standard deviations. Additionally, vowel length for both LOT and THOUGHT will be analyzed as a potential contrasting factor in the participants' production of LOT and THOUGHT.

6.2.1 Euclidean Distance Between LOT and THOUGHT

The Euclidean distance between LOT and THOUGHT appears to differ considerably between speakers of different ages and educational levels in the 2016 sample. Speakers without a college degree produce LOT and THOUGHT with a slightly greater distance than speakers with a college degree. Students, on the other hand, produce both phonemes with a significantly smaller distance than college educated speakers. In keeping with this pattern, the distance between LOT and THOUGHT appears to be decreasing in apparent time. These observations, however, appear to depend to a great extent on the speech style. Thus, in the following paragraphs, Euclidean distance will be analyzed in more detail with regard to *speaker age* and *education* for each speech style separately.

In spontaneous speech, most of the 2016 speakers with a college education appear to produce both phonemes close to the sample average of about 200 Hz. Most of the participants without a college degree in this sample seem to have a somewhat greater distance between both phonemes, so that the average for this group increases slightly to about 270 Hz. The difference between these two groups is particularly pronounced for speakers born after 1980, as shown in Figure 99 below. The majority of students have smaller distances than older speakers in spontaneous speech, with an average of about 160 Hz.

The increasing divide between speakers born after 1980 in the 2016 data could also be *gender* related. As can be seen in Figure 99 below, younger speakers with a smaller distance between LOT and THOUGHT are female, while the two younger speakers with greater distances are male. However, no *gender* differences appear in the student group, where both males and females have comparable distances. In fact, most male students have slightly lower distances than the two females in this group. Thus, *gender* does not appear to be a determining factor in the LOT-THOUGHT distance in this speech style.

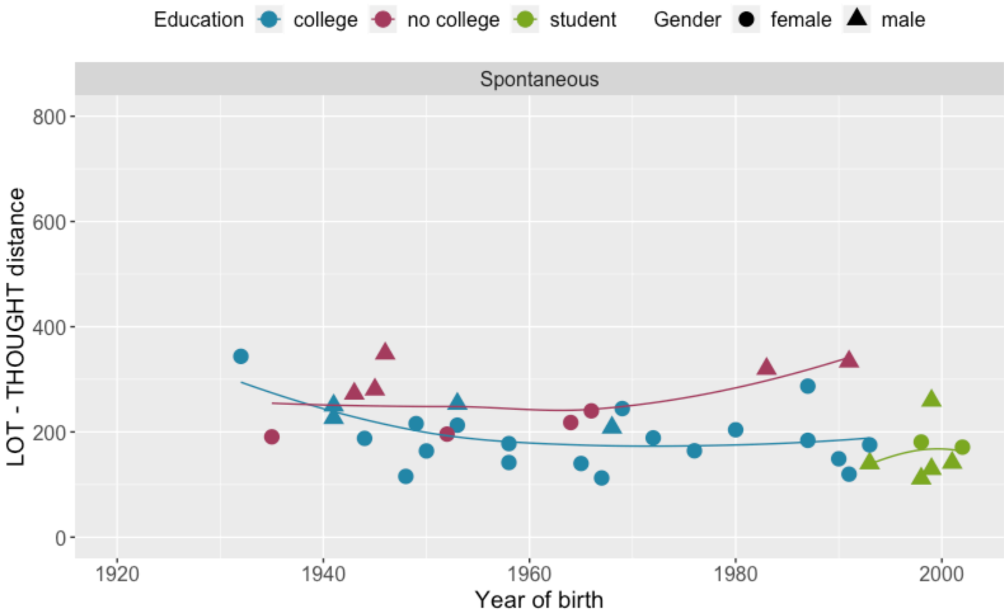


Figure 99: Euclidean distance between LOT and THOUGHT in spontaneous speech in 2016 by *education* and *gender*.

Across time, the Euclidean distance between LOT and THOUGHT does not appear to be changing drastically in spontaneous speech in 2016. However, since the students in the sample are simultaneously the youngest speakers, the observation that they collectively produce LOT and THOUGHT with a smaller distance may indicate a slight apparent-time trend toward the merger. For speakers without a college degree, the greater distances of the two younger speakers might be an indication of an apparent-time increase in distance. However, this observation is based on only two speakers, and thus cannot be concluded with any certainty.

In wordlist style, the Euclidean distance between LOT and THOUGHT shows significantly more variation than in spontaneous speech, but again appears to be conditioned by a combination of the speakers' *level of education* and *age* in the 2016 data. As Figure 100 below shows, most of the 2016 speakers produce wordlist LOT and THOUGHT with a distance between 100 and 600 Hz (mean= 301 Hz). Although this is a relatively wide range, there are speakers who fall short of it or exceed it.

The one speaker who has a LOT-THOUGHT distance that is greater than 600 Hz is a speaker without a college degree. On the other hand, those two who fall short of the 100 Hz mark are college educated speakers. This is indicative of the overall pattern. Although speakers from both educational groups spread across the entire distance range, speakers without a college degree tend to have a somewhat greater distance between wordlist LOT and THOUGHT (mean= 394 Hz) than speakers with a college degree (mean= 297 Hz). As can be seen in Figure 100, this difference is particularly pronounced among the younger speakers in these two educational groups. This appears to be due to an apparent time decrease in the Euclidean distance between wordlist LOT and THOUGHT among college educated speakers. Speakers with a college degree born after 1975 have distances of less than 300 Hz between both phonemes (mean= 233 Hz). While there are older speakers in this educational group who have comparably low distances, others in this age group have distances that are notably higher (mean= 330 Hz). This impression of an apparent-time decrease in the LOT-THOUGHT distance in wordlist style is strengthened by the observation that, as a group, students produce wordlist LOT and THOUGHT at a notably lower distance than any other speakers. Out of the seven students, five have a distance of 200 Hz or less between the two phonemes in this speech style (mean= 204).

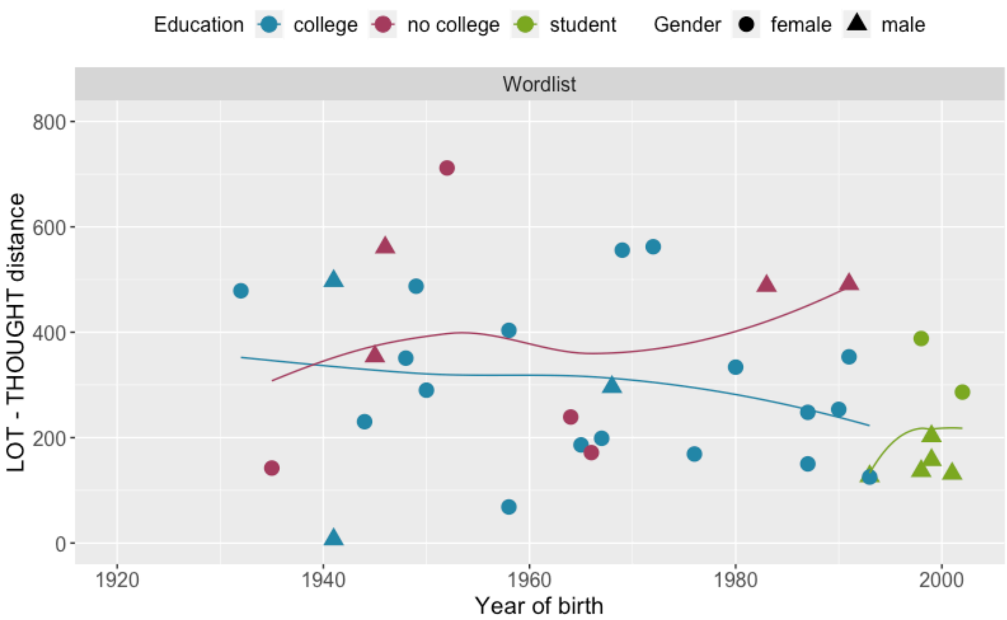


Figure 100: Euclidean distance between LOT and THOUGHT in wordlist style in 2016 by *education* and *gender*.

A second potential explanation for the increasing divide between speakers born after 1975 could be the *gender* pattern in both groups. As can be seen in Figure 101 below, younger speakers with a small distance between wordlist LOT and THOUGHT are all female, while the two younger speakers with greater distances are both male. Thus, it might be possible that the apparent-time trend toward a smaller distance between LOT and THOUGHT is a female-led change. However, the gender pattern in the student group is reversed.

Here, the five students with distances of 200 Hz or less are, in fact, all male, while the two female students have distances of nearly 300 and 400 Hz. Thus, it seems that *education* is the more relevant underlying factor in the apparent-time decrease in the LOT-THOUGHT distance in wordlist style.

In minimal pair production, the Euclidean distance between LOT and THOUGHT does not seem to be conditioned by the speakers' *level of education* but appears to be decreasing in apparent time across the entire 2016 sample. Figure 101 below shows that the majority of speakers fall within a distance range from 200 to 500 Hz (mean= 294 Hz). One exception to this, however, are the students in this sample. All students produce minimal pair LOT and THOUGHT with a distance of less than 200 Hz, and thus notably closer together than the majority of speakers in the other two educational groups. For speakers with and without a college degree, however, an apparent-time decrease in the distance between two phonemes emerges from the data. As can be seen in Figure 101, speakers in these two educational groups do not differ significantly in their distances between LOT and THOUGHT in minimal pair production, and in both groups, older speakers tend to have a greater distance between both vowels than younger speakers. However, only a few of the younger non-student speakers have crossed into the 200 Hz mark.

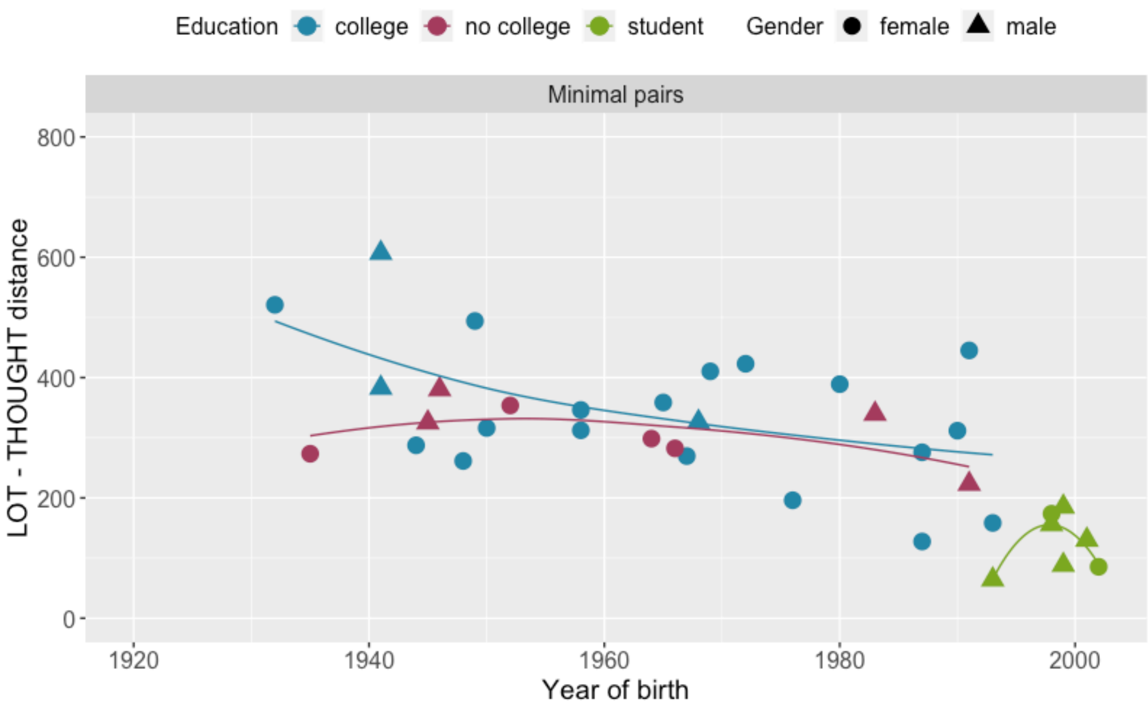


Figure 101: Euclidean distance between LOT and THOUGHT in minimal pairs in 2016 by *education* and *gender*.

Statistical testing, summarized in Table 54, supports the findings presented above. The model suggests that the main effect of *gender* is not significant, which supports the argument that the differences between younger speakers in spontaneous speech and wordlist style are not related to *gender*. Instead, the significant main effect of *education* suggests that these differences are due to the speakers' *educational backgrounds*, with speakers without a college degree having somewhat greater distances than speakers with a degree. Students have significantly lower distances than any other speakers, regardless of speech style. In fact, the low effect size of *age* suggests that inter-speaker differences are *only* determined by the speakers' *level of education*. The interaction between *age* and *style* suggests that changes over time differ significantly across the three speech styles. In minimal pair production, the distance between LOT and THOUGHT is decreasing about three times faster than in the other two styles. The difference between wordlist style and spontaneous speech, on the other hand, is relatively small.

Predictor	Coefficients	*p*
(Intercept)	214.031 Hz	
Age	-0.2065 Hz	
Style		
(Wordlist)	50.0482 Hz	
(Minimal pairs)	-44.0236 Hz	
Gender (Male)	29.0773 Hz	0.313
Education		
(No college)	30.4716 Hz	0.042
(Student)	-92.4096 Hz	
*Age*style*		
(Wordlist)	1.0565 Hz	0.019
(Minimal pairs)	2.9669 Hz	

Table 54: Euclidean distance between LOT and THOUGHT in 2016. Reference levels: spontaneous speech, females, college educated. Random effect: *speaker*. *n*= 111

In the analyses of the Euclidean distance between LOT and THOUGHT presented above, notable differences emerged between the three different speech styles. The extent of style shifting in the LOT-THOUGHT distance appears to depend on the speakers' educational background in the 2016 data, with college educated speakers shifting the most, and students the least.

College educated speakers tend to shift toward a greater distance between LOT and THOUGHT the more attention they pay to their speech. While in spontaneous speech, the average distance between the two phonemes for college speakers is only 194 Hz, it increases to 297 Hz in wordlist style, and to 344 Hz in minimal pairs. Thus, they shift notably from spontaneous speech to wordlist style and minimal pairs, but do not make a notable difference between the two more careful styles. However, the extent of style shifting appears to depend on the speakers' *age* in this educational group, since apparent-time decreases in the LOT-THOUGHT distance can be observed in the two more careful speech styles, while no changes are evident in spontaneous speech. As shown in Figure 102, older speakers shift to a significantly greater distance, especially from spontaneous speech to minimal pairs than younger speakers. The shift from spontaneous speech to wordlist style, on the other hand, does not differ notably between older and younger speakers.

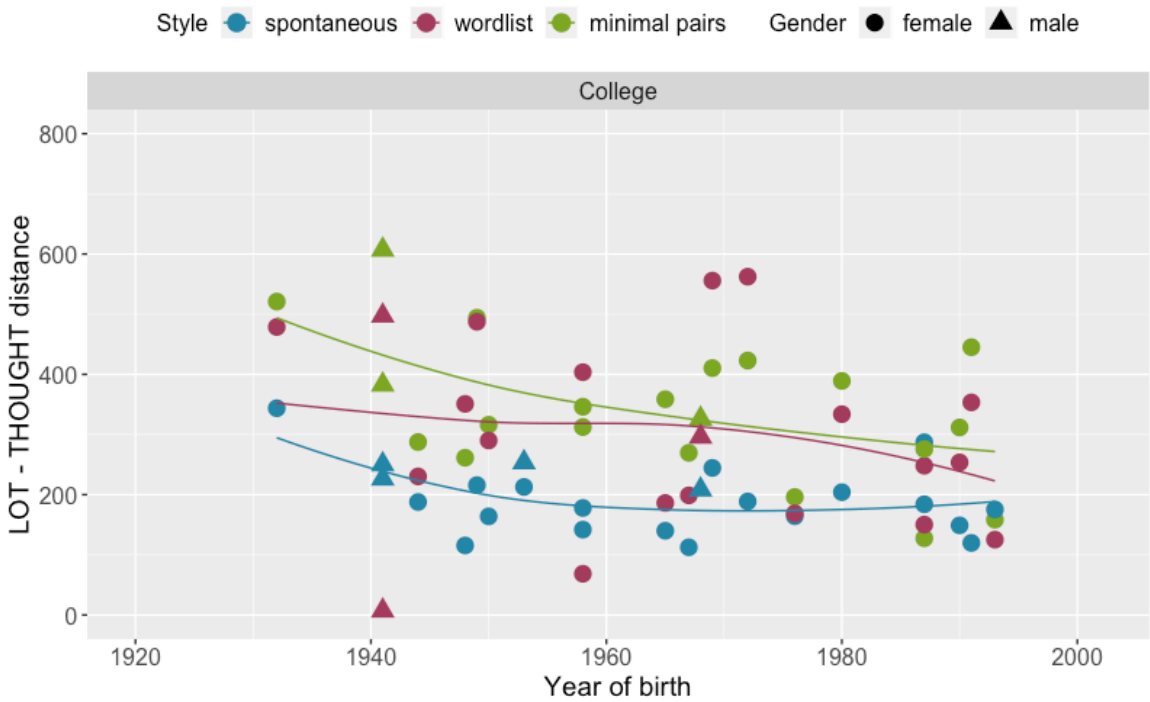


Figure 102: Effect of *style* on the LOT-THOUGHT distance for college educated speakers in 2016 by *gender*.

These observations are corroborated by statistics and summed up in Table 55 below. The results indicate that college educated speakers born before 1970 shift to a somewhat greater distance in wordlist style than speakers born after 1970. In the shift from spontaneous speech to minimal pairs, this difference between the two age groups is even more pronounced. Older college educated speakers also increase the extent of their shift the more attention they pay to speech, so that the shift from spontaneous speech to minimal pairs is greater than that from spontaneous speech to wordlist style. Younger speakers in this educational group, on the other hand, do not appear to treat wordlist style and minimal pairs differently, so that the shift from spontaneous speech to wordlist style and minimal pairs is about the same.

Predictor	yob < 1970		yob > 1970	
	Coefficient	*p*	Coefficient	*p*
(Intercept)	26.167 Hz		19.43 Hz	
Age	2.685 Hz	0.156	5.226 Hz	0.142
Style				
(Wordlist)	113.403 Hz	0.0002	90.391 Hz	0.066
(Minimal pairs)	178.182 Hz		106.788 Hz	
Gender (Male)	16.976 Hz	0.719	---	---

Table 55: Effect of *style* on the LOT-THOUGHT distance for college educated speakers in 2016. Reference level: spontaneous speech. Random effect: *speaker*. *n*= 41 (yob < 1970), *n*= 24 (yob > 1970)

A very different style-shifting pattern emerges for speakers without a college degree. Speakers in this group average around a 267 Hz distance in spontaneous speech and tend to shift toward a somewhat greater LOT-THOUGHT distance in wordlist style (mean= 395 Hz) and minimal pairs (mean= 310 Hz). However, this pattern appears to change for the youngest speakers in this group. Speakers born after 1980 produce LOT and THOUGHT much further apart in wordlist style than they do in minimal pairs and spontaneous speech, as can be seen in Figure 103.

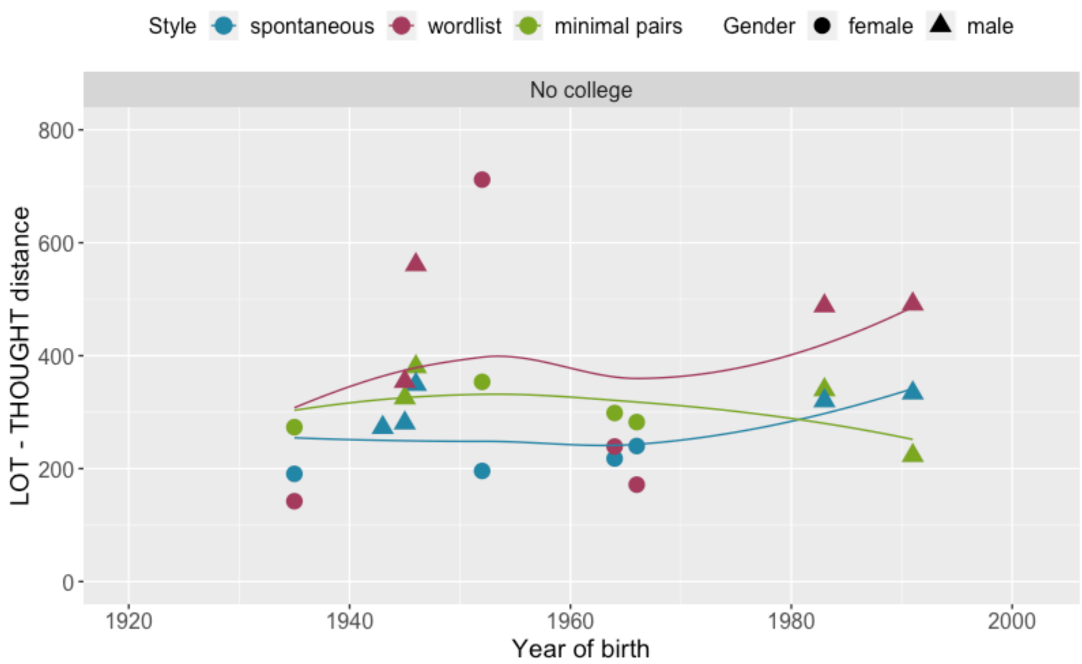


Figure 103: Effect of style on the LOT-THOUGHT distance for speakers without a college degree in 2016 by *gender*.

The regression model summed up in Table 56 confirms these observations. Only for younger speakers are style-shifting patterns found to be statistically significant, however.

Predictor	yob < 1980		yob > 1980	
	Coefficient	*p*	Coefficient	*p*
(Intercept)	252.751 Hz		207.025 Hz	
Age	-0.691 Hz	0.824	4.134 Hz	0.219
Style				
(Wordlist)	121.8 Hz	0.16	162.825 Hz	0.001
(Minimal pairs)	77.344 Hz		-45.576 Hz	
Gender (Male)	98.84 Hz	0.178	---	---

Table 56: Effect of *style* on the LOT-THOUGHT distance for speakers without a college degree in 2016. Reference level: spontaneous speech. Random effect: *speaker*. *n*= 19 (yob < 1980), *n*= 6 (yob > 1980)

The students in the 2016 sample appear to style shift the Euclidean distance between LOT and THOUGHT only to a minimal extent. As shown in Figure 104, the majority of students produce LOT and THOUGHT with a distance between 100 and 200 Hz in all three styles. Only in wordlist style do two female students shift to a somewhat greater distance between both phonemes. Otherwise, the mean distances do not differ notably from each other between the three speech styles for this group of speakers. It is, however, noteworthy that for the majority of students, LOT and THOUGHT in minimal pairs are produced with a slightly lower distance than they are in the other two speech styles.

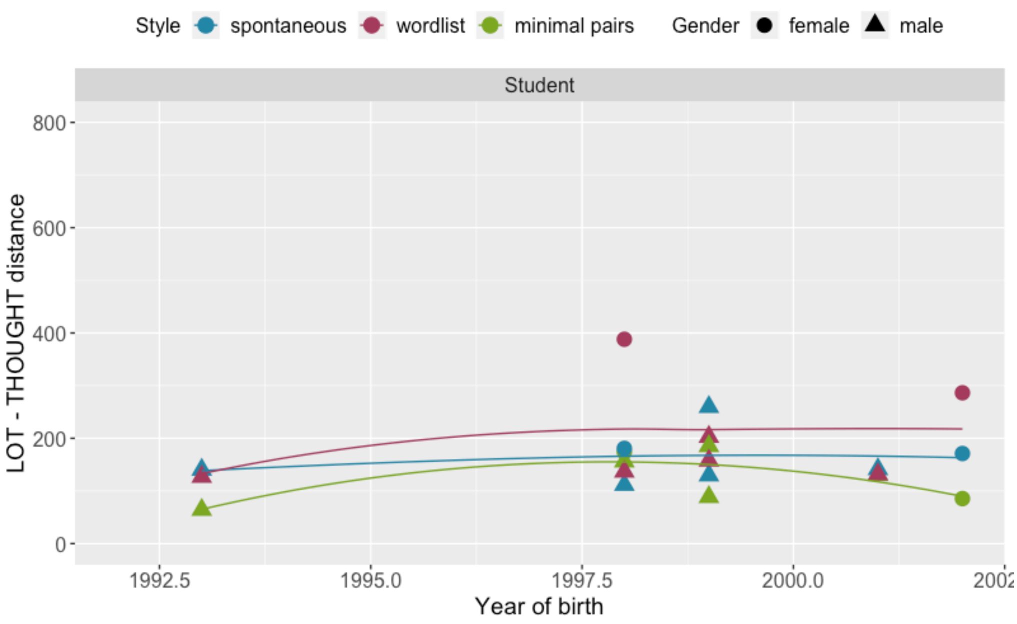


Figure 104: Effect of style on the LOT-THOUGHT distance for students in 2016 by *gender*.

Statistics corroborate these observations, as summed up in Table 57. The interaction between *gender* and *style* suggests that differences between male and female students are particularly pronounced in wordlist style. Here, male students produce LOT and THOUGHT significantly closer together than female students, i.e. female students style shift quite significantly in this speech style. In minimal pair production and spontaneous speech, on the other hand, the differences are negligible. However, because there are only two female students in the sample, the significance of these *gender* differences is questionable.

Predictor	Coefficient	*p*
(Intercept)	197.409 Hz	
Age	-1.339 Hz	0.754
Style		
(Wordlist)	161.27 Hz	
(Minimal pairs)	-46.387 Hz	
Gender (Male)	-16.841 Hz	
*Style*male*		
(Wordlist)	-166.609 Hz	0.002
(Minimal pairs)	14.704 Hz	

Table 57: Effect of *style* on the LOT-THOUGHT distance for students in 2016. Reference level: spontaneous speech. Random effect: *speaker*. *n*= 21

A comparison of speakers interviewed in 2008 and 2016 reveals that speakers from the two samples do not differ visibly from each other in their Euclidean distances between LOT and THOUGHT. As can be seen in Figure 105 below, this holds true for all three speech styles. However, to get a better sense of the developments regarding the distance between LOT and THOUGHT in the 2008 sample, speakers from both samples will be compared in more detail in each of the three speech styles in the following paragraphs.

In spontaneous speech, the majority of 2008 speakers produce LOT and THOUGHT with comparable distances as 2016 speakers. As can be seen in Figure 105 below, most of the 2008 speakers have a distance between 100 and 300 Hz in this speech style, which falls well within the range of 2016 speakers. There are, however, two speakers in the 2008 sample with a notably greater distance of 560 Hz and 730 Hz; these are the youngest speakers in the 2008 sample. Because of these two outliers in 2008, the average LOT-THOUGHT distance in spontaneous speech is notably higher in the 2008 sample (mean= 353 Hz) than in the 2016 sample (mean= 205 Hz). Also because of two young outliers, one might suspect a tendency toward greater distances in the 2008 sample. The regression model in Table 58 below supports these observations. With the absence of any major apparent-time developments in 2016 in spontaneous speech, the significant real-time differences that exist in the Euclidean distances between the two samples in this speech style result not from apparent-time decreases in 2016, but rather from a potentially increasing distance in 2008.

In wordlist style, the distances between LOT and THOUGHT of 2008 speakers fall within the range of 2016 speakers but tend to orient toward the lower end of this range. As mentioned above, the distances between the two vowels are relatively varied in wordlist style among 2016 speakers, ranging from about 100 to 600 Hz, with outliers who exceed or fall short of this range. As can be seen in Figure 105 below, 2008 speakers seem to be equally varied, and range in their wordlist distances from 50 to 400 Hz. Their mean distance of 203 Hz is somewhat lower than that of 2016 speakers (mean= 301 Hz) in this speech style. In apparent time, however, both samples appear to be progressing toward lower distances.

In minimal pair production, 2008 speakers appear to produce LOT and THOUGHT at about the same distance as 2016 speakers. As Figure 105 below illustrates, speakers in the 2008 sample range in their LOT-THOUGHT distances from 200 to 450 Hz (mean= 333 Hz), which is the same range that was identified for 2016 speakers in this speech style. Furthermore, an apparent-time decrease in the minimal pair distance between LOT and THOUGHT can be observed for 2008 speakers, similar to the trend observed in the 2016 data. The distances of the youngest 2008 speakers appear to be notably lower than those of the two middle aged speakers in the sample, although the oldest 2008 speaker has an equally small distance as the younger speakers.

One notable difference between the 2008 and 2016 data is the amount of style shifting in the LOT-THOUGHT distance. As described above, the majority of 2016 speakers appear to increase the distance the more attention they pay to their speech, with the exception of students, who style shift to a smaller extent and, in most cases, in the opposite direction. For 2008 speakers, on the other hand, no clear pattern of style shifting emerges from the data.

Figure 105: Euclidean distance between LOT and THOUGHT in 2008 and 2016 across *speech styles* by *gender*.

From spontaneous speech to wordlist style, it appears that some 2008 speakers shift to a somewhat greater distance, while others shift in the opposite direction. The latter is particularly notable for the two outliers with exceptionally great distances in spontaneous speech mentioned above. In this shift from spontaneous speech to wordlist style, these two speakers shift from distances of 560 Hz and 730 Hz in spontaneous speech to 270 Hz and 91 Hz in wordlist style, respectively. This is somewhat different from the 2016 sample, where the majority of speakers seem to shift to a greater distance from spontaneous speech to wordlist style. However, as mentioned above, the LOT-THOUGHT distances in wordlist style are quite varied in the 2016 sample, and there are speakers who shift to a smaller distance or who do not style shift at all in this sample as well.

From spontaneous speech to minimal pairs, 2008 speakers seem to follow the same pattern as most 2016 speakers, in that they shift to a greater distance in the most careful style. Again, the opposite is true for the two 2008 outliers, who shift to notably smaller distances.

Predictor	Coefficient	p
(Intercept)	174.293 Hz	
Age	0.585 Hz	
Style *(Wordlist)* *(Minimal pairs)*	 21.191 Hz -71.06 Hz	
Sample year (2008)	334.683 Hz	
Gender (Male)	9.326 Hz	0.708
*Age*style* *(Wordlist)* *(Minimal pairs)*	 1.675 Hz 3.545 Hz	0.006
*Age*2008*	-4.553 Hz	0.016
*2008*style* *(Wordlist)* *(Minimal pairs)*	 -240.789 Hz -88.524 Hz	0.001

Table 58: Euclidean distance between LOT and THOUGHT in 2008 and 2016. Reference levels: spontaneous speech, females, 2016. Random effect: *speaker*. *n*= 133

6.2.2 Bhattacharyya Scores

According to the second measure used to track the merger in production, the overlap of LOT and THOUGHT as measured by Bhattacharyya scores, the degree of merger appears to depend to a great extent on the speakers' *age* and/or *education* in the 2016 sample. Speakers with and without a college degree differ notably in their amount of overlap of LOT and THOUGHT. Likewise, students seem to produce both vowels with notably more overlap than speakers in the other two educational groups. Additionally, the amount of overlap appears to differ considerably between spontaneous speech and the two more careful styles. Thus, in the following paragraphs, Bhattacharyya scores will be analyzed in more depth for each speech style separately, focusing on the effect *age* and *education* on the amount of overlap of LOT and THOUGHT.

In spontaneous speech, the amount of overlap of LOT and THOUGHT ranges from 40 to 90% in spontaneous speech (mean= 67%) in 2016, and overall, no apparent-time trend toward higher or lower degrees are evident in the data, as can be seen in Figure 106 below. However, as the plot illustrates, a slightly higher degree of overlap of LOT and THOUGHT can be observed for college educated speakers (mean= 70%) and students (mean= 65%) than for speakers without a college degree (mean= 61%). This is particularly noticeable among younger speakers, as the two youngest speakers without a college degree have below-average amounts of overlap. This might be an indication of an apparent-time decrease in overlap of LOT and THOUGHT in this educational group. However, because this observation is based on only two speakers, the significance of this is questionable.

It is also possible that the differences between younger speakers with and without a college degree might be related to *gender* rather than education, as all younger speakers without a college degree are male, while those with a college degree are female. However, the *gender* pattern among students does not support this hypothesis. While two male students do have comparably low degrees of overlap, the remaining three are more similar to their female peers.

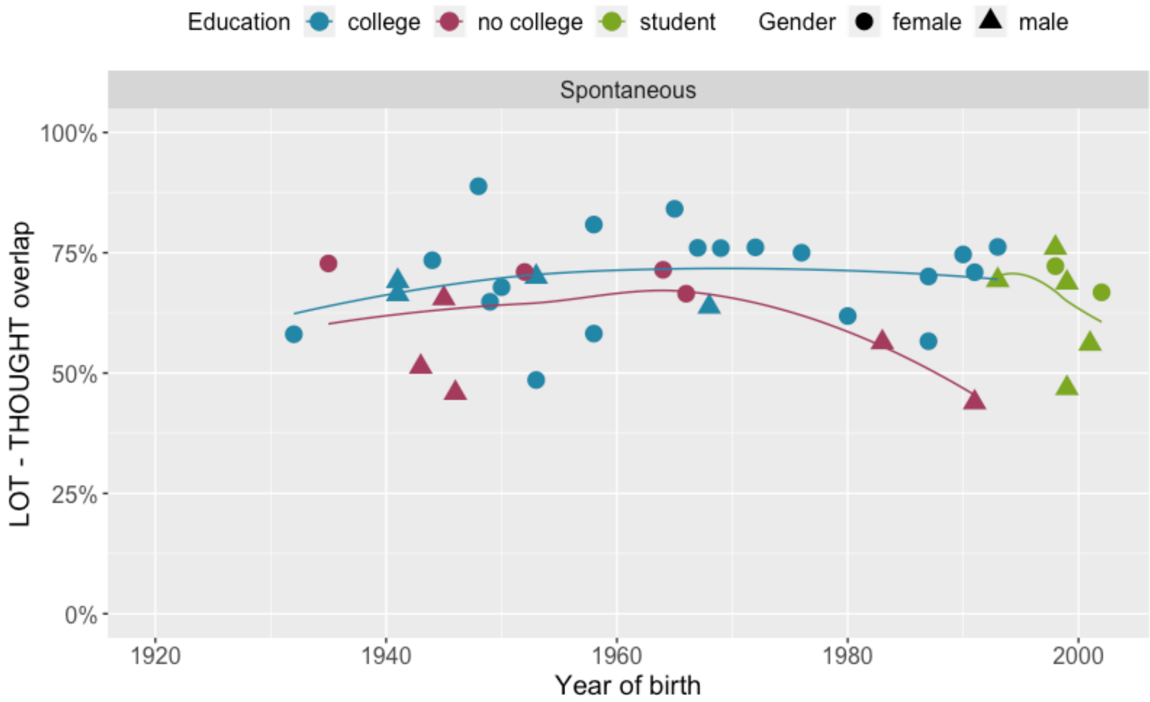


Figure 106: Overlap of LOT and THOUGHT in spontaneous speech in 2016 by *education* and *gender.*

In wordlist style, Figure 107 below suggests that the overall amount of overlap of LOT and THOUGHT varies notably, ranging from 0 to 65% (mean= 27%) in the 2016 sample. In this speech style, the speakers' *educational background* seems to be less relevant than in spontaneous speech. Although higher degrees of overlap can be observed for college educated speakers (mean= 28%) compared to speakers without a college degree (mean= 14%), this difference seems to be due to only one college educated female, Allison, born in 1993, who has a much higher than average degree of overlap (64%) of wordlist LOT and THOUGHT than her peers in this educational group. The remaining college educated speakers born after 1980 (mean= 16%) do not seem to differ notably from their peers without a college degree (mean= 7%). Speakers born before 1980 also do not seem to differ significantly based on their *educational background*.

Students, however, seem to have significantly more overlap of LOT and THOUGHT in wordlist style (mean= 39%) than most other speakers. Although the trend line in Figure 107 suggests that this might be the result of a gradual apparent-time development toward more overlap, this does not appear to be the case. Allison, the college educated speaker with 64% overlap born in 1993 mentioned above, was a recent college graduate at the time of interviewing and was therefore only just grouped with college educated speakers rather than with students. If she is excluded from the group of college educated speakers, the abrupt increase in overlap from adult speakers to students becomes much more pronounced. This suggests that the increased amount of overlap among students is not the result of a gradual apparent-time development, but rather a student-exclusive phenomenon. In fact, excluding Allison from the group of young college educated speakers (born after 1980) leads to a slight apparent-time *decrease* in overlap for this educational group, as the new combined average of 16% is somewhat lower than that of college educated speakers born before 1980 (mean= 29%). Similarly, the trend line for speakers without a college degree in Figure 107 suggests an apparent-time decrease in the amount of overlap of wordlist LOT and THOUGHT for this group of speakers.

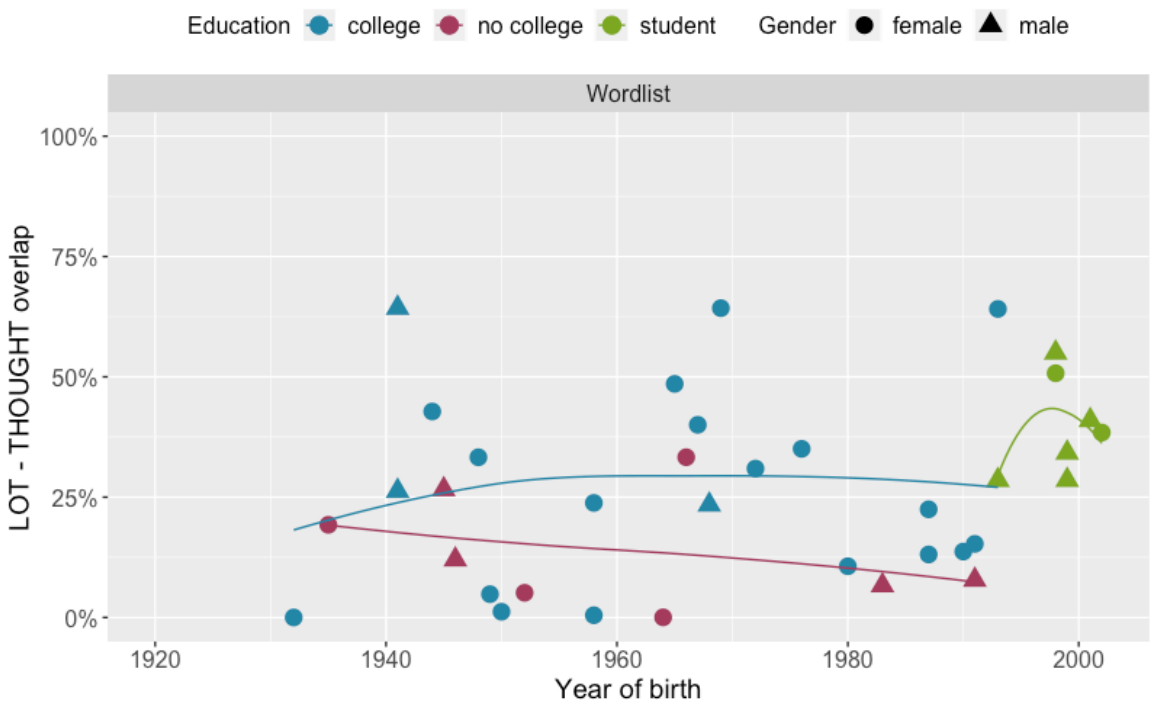


Figure 107: Overlap of LOT and THOUGHT in wordlist style in 2016 by *education* and *gender*.

In minimal pair production, the average amount of overlap of LOT and THOUGHT is 27% for 2016 speakers, with a potential apparent-time increase in overlap. As Figure 108 shows, speakers with and without a college degree do not seem to differ notably in their amount of overlap, however, students appear to have a much higher degree of overlap than most other speakers. As in wordlist style, this does not appear to be a gradual trend in apparent time, but rather a newly adopted behavior among the student group. Students have an average overlap of 63% in the production of LOT and THOUGHT in minimal pairs; on the other hand, the average overlap for college educated speakers in this speech style is 18%, and speakers without a college degree average at a comparable 19%. While a few of these speakers have relatively high percentages that are comparable to students, including some younger speakers born after 1980, they seem to be an exception, as most of their peers produce minimal pair LOT and THOUGHT with considerably less overlap.

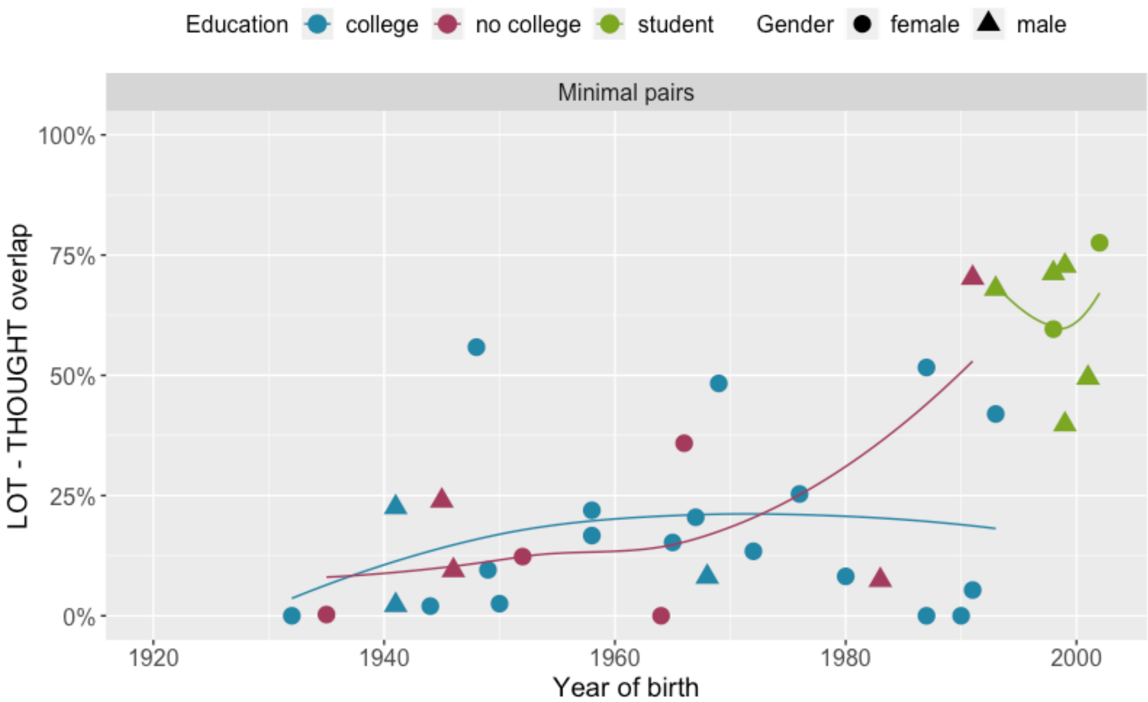


Figure 108: Overlap of LOT and THOUGHT in minimal pairs in 2016 by *education* and *gender*.

Statistical testing, summarized in Table 59 below, supports the findings presented for the 2016 sample above. The regression model predicts an insignificant main effect of *gender*, which supports the argument that the differences in overlap between younger speakers, especially in spontaneous speech, are not related to *gender*. Instead, the highly significant interaction between *style* and *education* suggests that these differences are due to the speakers' *educational backgrounds*. This supports the observation that, in wordlist style and minimal pairs, students have much more overlap of LOT and THOUGHT than any other speakers. In spontaneous speech, on the other hand, the predicted difference of about 8% between speakers with and without a college degree in spontaneous speech is quite small, and there is virtually no difference predicted between students and speakers without a college degree in this speech style.

Predictor	Coefficients	*p*
(Intercept)	75.406%	
Age	-0.101%	0.394
Style		
(Wordlist)	-42.755%	
(Minimal pairs)	-52.625%	
Gender (Male)	-1.625%	0.724
Education		
(No college)	-8.178%	
(Student)	-7.343%	
*Style*education*		
*No college*WL*	-4.461%	
*No college*MP*	11.49%	10^{-6}
*Student*WL*	17.099%	
*Student*MP*	50.097%	

Table 59: Overlap of LOT and THOUGHT in 2016. Reference levels: spontaneous speech, females, college educated. Random effect: *speaker*. *n*= 111

The regression model presented in Table 59 above suggests that, overall, there is less overlap of LOT and THOUGHT in the two more careful speech style than there is in spontaneous speech in the 2016 data. Overall, this appears to be the case for speakers with and without a college degree. Speakers from both groups appear to shift to a considerably smaller amount of overlap in the two more careful speech styles, while not making a significant difference between LOT and THOUGHT in wordlist style and minimal pair production, where they produce both vowels with comparable degrees of overlap. Students, on the other hand, tend to shift to less overlap of LOT and THOUGHT in wordlist style. In minimal pair production, on the other hand, they appear to produce both vowels with just as much overlap as they do in spontaneous speech. These style-shifting patterns will be explored in more detail in the following paragraphs for each educational group separately.

College educated speakers in 2016 appear to produce LOT and THOUGHT with considerably less overlap in wordlist style and minimal pairs than they do in spontaneous speech. As shown in Figure 109 below, the majority of speakers in this group have about

70% overlap of LOT and THOUGHT in their spontaneous speech production. This number drops down to 28% and 18% for wordlist style and minimal pair production, respectively. These numbers indicate that speakers with a college degree make a notable distinction between spontaneous speech and the two more careful styles, but they do not make a distinction between wordlist style and minimal pairs. These observations are corroborated by a regression model, summed up in Table 60. According to this model, college educated speakers have significantly less overlap in the two more careful styles than they do in spontaneous speech: about 43% less in wordlist style, and 53% less in minimal pairs. In neither of the three styles do any apparent-time changes emerge from the data for college educated speakers.

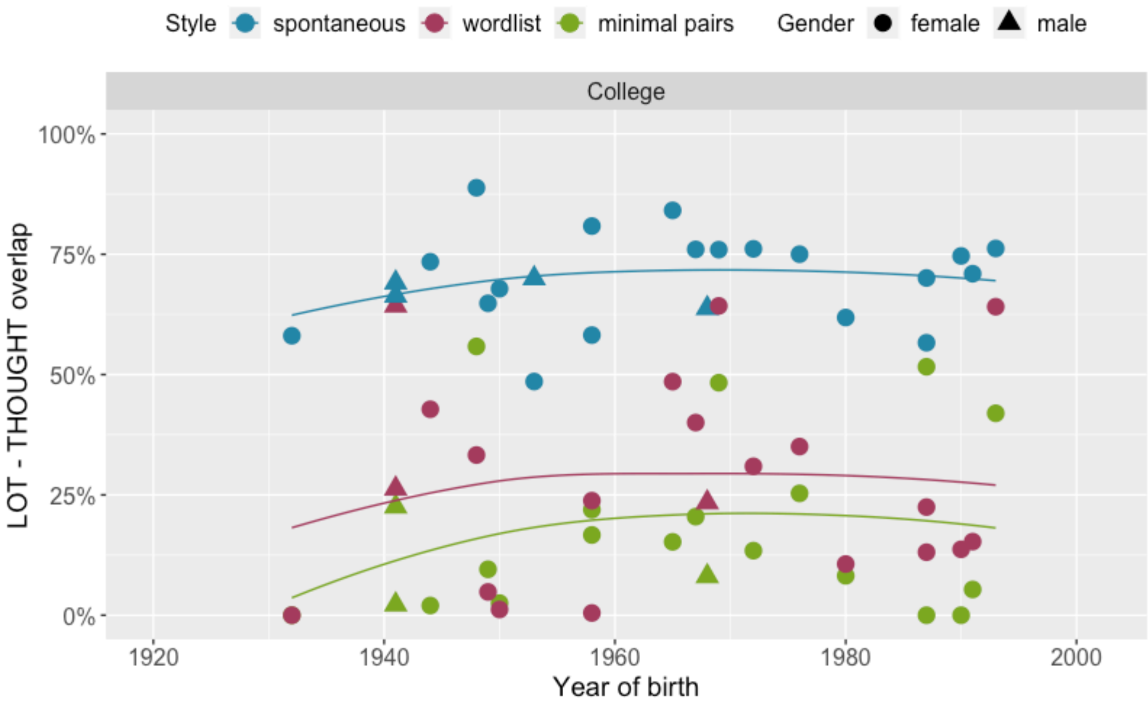


Figure 109: Effect of *style* on the LOT-THOUGHT overlap for college educated speakers in 2016 by *gender*.

Predictor	Coefficient	*p*
(Intercept)	76.19%	
Age	-12.96%	0.445
Style (*Wordlist*)	-42.81%	2×10^{-16}
(*Minimal pairs*)	-52.68%	
Gender (Male)	2.535%	0.757

Table 60: Effect of *style* on the overlap of LOT and THOUGHT for college educated speakers in 2016. Reference level: spontaneous speech. Random effect: *speaker*. *n*= 65

A similar style-shifting pattern in the overlap of LOT and THOUGHT can be observed for speakers without a college degree in the 2016 sample in Figure 110. In spontaneous speech, speakers in this group have an average overlap of 61%. In wordlist style, on the other hand, LOT and THOUGHT overlap in only 14% of all cases, and in minimal pairs in 20%. Thus, while they seem to differentiate between spontaneous speech and more careful speech, they treat wordlist style and minimal pair production relatively equally. The youngest speaker in this group, however, appears to shift to notably more overlap of both vowels in minimal pair production compared to both wordlist style and spontaneous speech, as can be seen in Figure 110.

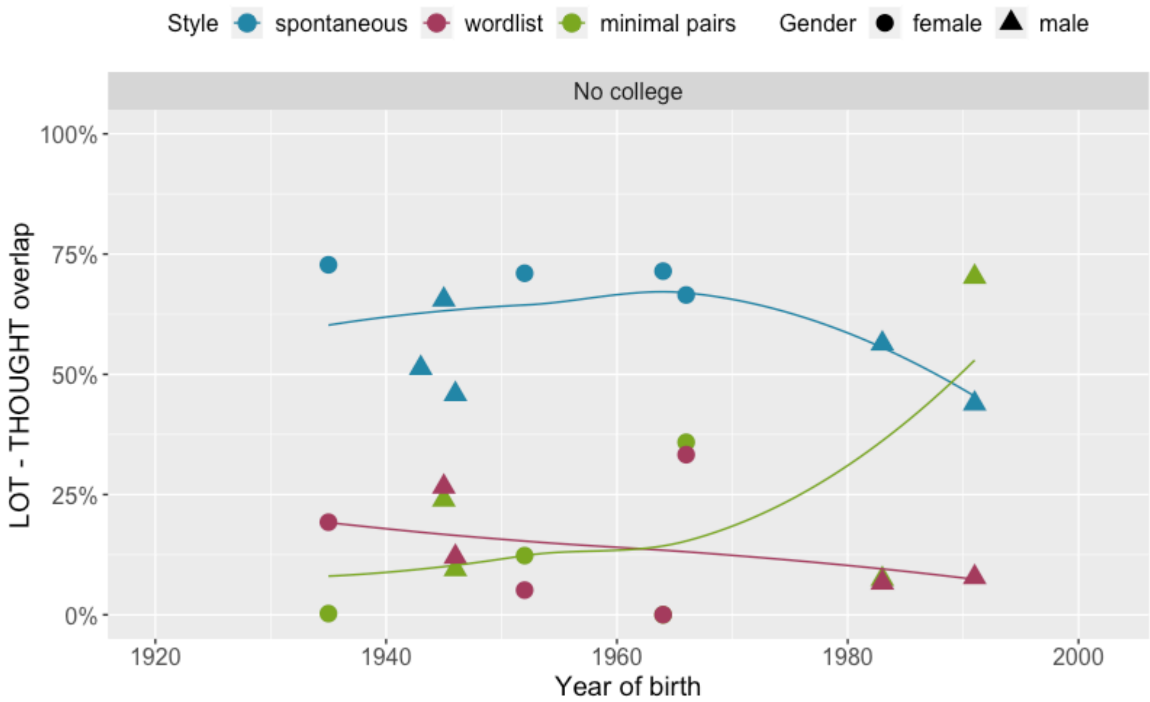


Figure 110: Effect of *style* on the LOT-THOUGHT overlap for speakers without a college degree in 2016 by *gender*.

Statistics support the general pattern that can be observed for this group of speakers. As summed up in Table 61 below, speakers without a college degree have about 47% less overlap in wordlist style, and about 41% less overlap in minimal pairs than in spontaneous speech. These differences are statistically significant and are comparable to the extent and direction of style shifting described above for college educated speakers. Similarly, no clear apparent-time trends toward more or less overlap emerge from the data.

Predictor	Coefficient	*p*
(Intercept)	68.87%	
Age	-0.117%	0.557
Style (*Wordlist*)	-47.056%	10^{-5}
(*Minimal pairs*)	-40.976%	
Gender (Male)	-2.918%	0.685

Table 61: Effect of *style* on the overlap of LOT and THOUGHT for speakers without a college degree in 2016. Reference level: spontaneous speech. Random effect: *speaker*. *n*= 25

The students in the 2016 sample behave very differently in terms of style shifting regarding the overlap of the two low back vowels. Figure 111 suggests that the only style shifting among the majority of students occurs in wordlist style, where they shift to a lower amount of overlap. In this speech style, they average at a relatively low 40% compared to 65% in spontaneous speech, and 63% in minimal pairs. In fact, some students realize LOT and THOUGHT with more overlap in minimal pairs than they do in spontaneous speech.

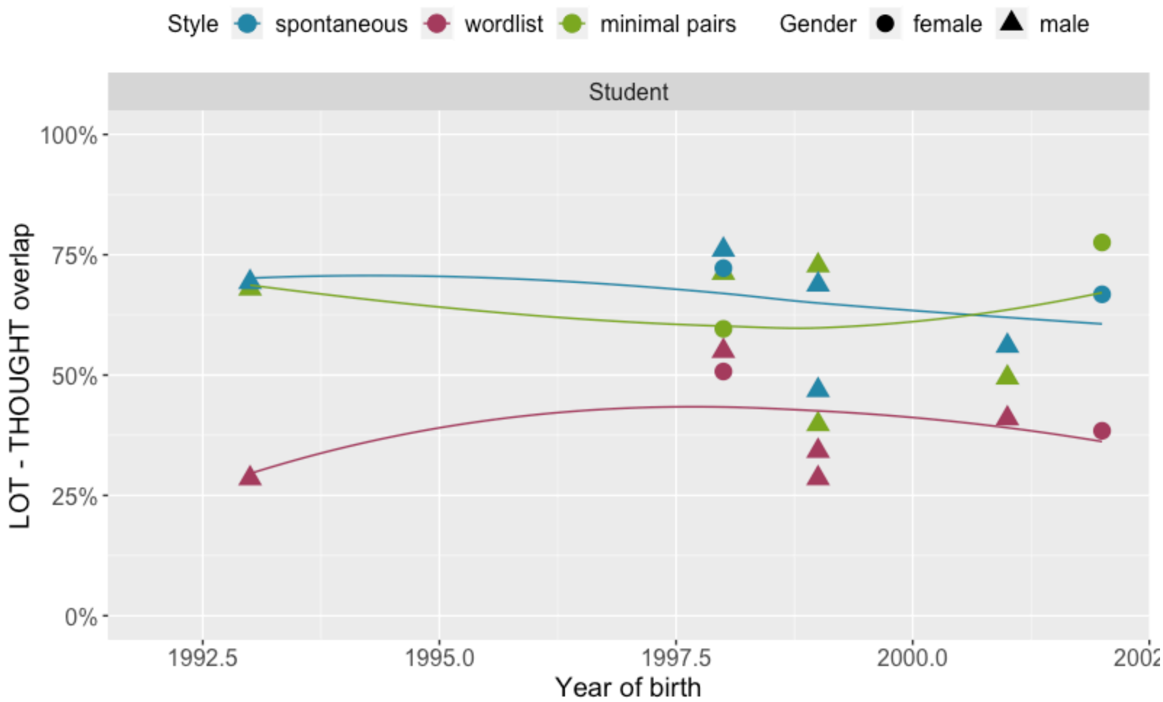


Figure 111: Effect of *style* on the LOT-THOUGHT overlap for students in 2016 by *gender*.

A regression model (Table 62 below) estimates the differences between spontaneous speech and wordlist style at about 26%, and at only 2.5% between spontaneous speech and minimal pairs, thus corroborating the observations outlined above. Students, then, appear to style shift significantly less than older speakers with and without a college degree.

Predictor	Coefficient	*p*
(Intercept)	58.407%	
Age	0.742%	0.64
Style (*Wordlist*)	-25.657%	0.0003
(*Minimal pairs*)	-2.527%	
Gender (Male)	-8.671%	0.373

Table 62: Effect of *style* on the overlap of LOT and THOUGHT for students in 2016. Reference level: spontaneous speech. Random effect: *speaker*. *n*= 21

In comparison to speakers interviewed in 2008, speakers in the 2016 sample appear to produce LOT and THOUGHT with a somewhat higher degree of overlap. However, as can be seen in Figure 112, the differences between the two samples seem to vary between the three speech styles. Thus, real-time differences in each style will be analyzed in more depth in the following paragraphs.

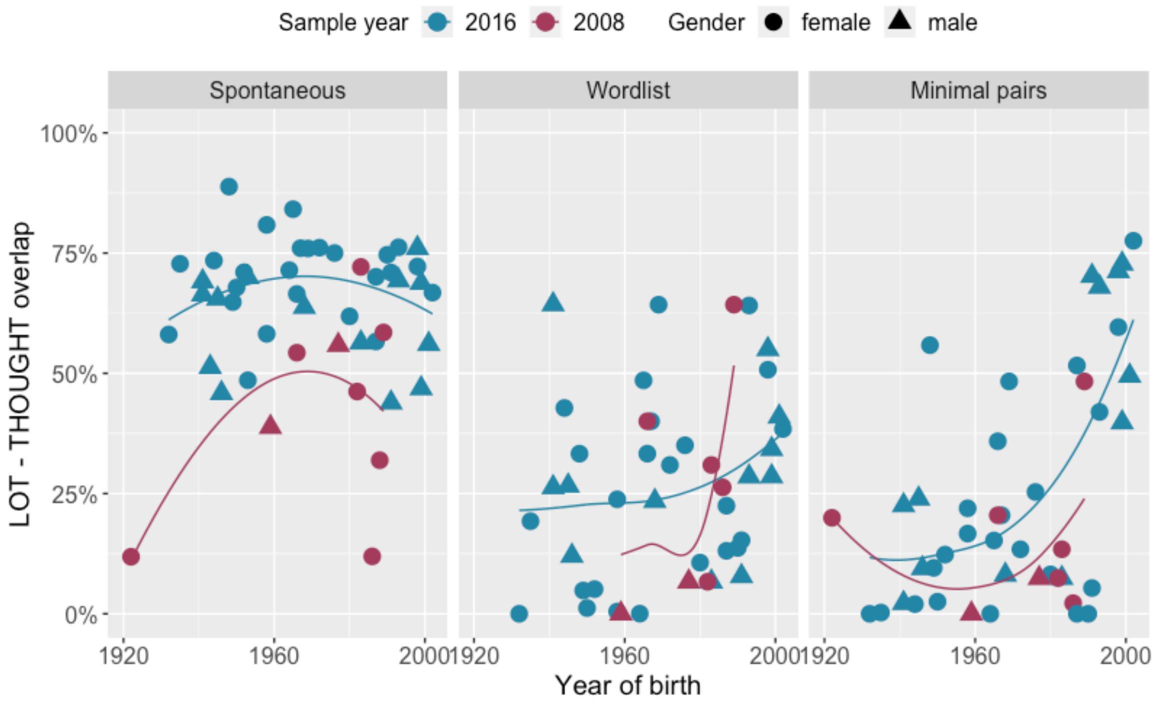


Figure 112: Overlap of LOT and THOUGHT in 2008 and 2016 across *speech styles* by *gender*.

In spontaneous speech, 2016 speakers appear to produce LOT and THOUGHT with a notably higher degree of overlap than 2008 speakers. Speakers in the 2016 sample range in overlap between 40 and 90% in spontaneous speech, with an average of 67%. Speakers in the 2008 sample, on the other hand, produce spontaneous LOT and THOUGHT with an average overlap of only 42%. This is even less overlap than speakers without a college degree in 2016, who are among the speakers with the least amount of overlap in the 2016 sample. In fact, three of the speakers in the 2008 sample have less overlap of both phonemes than any of the 2016 speakers, two of them notably so, as shown in Figure 112 above. Since these are two of the younger speakers in the 2008 sample, this may be an indication of an apparent-time trend toward a lower degree of overlap of LOT and THOUGHT. However, four speakers of similar ages have much higher overlap percentages, so that an apparent-time change in 2008seems unlikely.

In wordlist style, 2016 and 2008 speakers appear to produce the low back vowels with similar amounts of overlap. Similar to 2016 speakers, speakers in the 2008 sample range in their overlap of wordlist LOT and THOUGHT from 0 to 65%, with a mean of 25%, comparable to the 27% mean in 2016. As indicated in Figure 112 above, the 2008 speakers with the highest degree of overlap are generally among the younger ones, which might be an indication of an apparent-time increase in overlap in wordlist style in 2008, similar to the trends observed in 2016, although it is more likely that in the 2016 data, the increased amount of overlap is student-specific.

In the production of minimal pairs, speakers in the 2016 samples appear to produce LOT and THOUGHT with slightly more overlap than 2008 speakers. The range of overlap for the majority of 2016 speakers in minimal pairs is 0 to 75%, with a mean of 27%. Speakers in the 2008 sample fall into the same range, but none of them have an overlap of more than 50%. Thus, the average overlap in the 2008 sample is somewhat lower (15%) in this speech style than that of 2016 speakers. However, as Figure 112 above indicates, the overlap of LOT and THOUGHT in minimal pair production may be increasing in apparent time in the 2008 sample, as the youngest speaker has a notably higher degree of overlap than the rest of the speakers in the sample. This is similar to the 2016 data, where the youngest speakers, i.e. students, produce the low back vowels with considerably more overlap than most other speakers. The younger 2008 speaker with a notably higher amount of overlap belongs to the student group as well, however, the second student in this sample produces LOT and THOUGHT with as little overlap as the rest of the speakers.

For the 2008 sample, no clear style-shifting pattern in the LOT-THOUGHT overlap emerges from the data. The trend lines in Figure 112 above suggests that speakers in the 2008 sample follow the same general pattern as 2016 speakers (except students), i.e. producing LOT and THOUGHT in spontaneous speech with considerably more overlap than they do in more careful speech, without differentiating between wordlist style and minimal pairs. However, as can also be seen in Figure 112, not all 2008 speakers seem to follow this pattern, and the regression model in Table 63 predicts no significant style differences for 2008 speakers (-11% from spontaneous speech to wordlist style, and -3% from spontaneous speech to minimal pairs).

Predictor	Coefficient	*p*
(Intercept)	70.211%	
Age	-0.063%	
Style		
(Wordlist)	-31.903%	
(Minimal pairs)	-15.956%	
Sample year (2008)	-24.852%	
Gender (Male)	-1.047%	0.797
*Age*style*		0.002
(Wordlist)	-0.186%	
(Minimal pairs)	-0.535%	
*2008*style*		0.047
(Wordlist)	21.175%	
(Minimal pairs)	12.543%	

Table 63: Overlap of LOT and THOUGHT in 2008 and 2016. Reference levels: spontaneous speech, females, 2016. Random effect: *speaker*. *n*= 135

6.2.3 Mechanism of Merger

The results presented in the previous subchapters demonstrated that, while LOT and THOUGHT are still two distinct categories in Ogdensburg, there is definite progress toward the merger of the two vowels in apparent and real time, both according to the Euclidean distance as well as the Bhattacharyya score measure. In this chapter, the underlying mechanisms behind this progress will be examined.

6.2.3.1 Merger by Approximation

The observation that the Euclidean distance between LOT and THOUGHT is decreasing in the community suggests a merger by approximation. This suggests that LOT is moving back and up and/or that THOUGHT is moving down and forward. While lowering and fronting of THOUGHT would not be surprising, as this is not only the expected trajectory for the merger but also for the NCS, backing and raising of LOT would run counter to its NCS trajectory, which predicts fronting. Which of the potential movements are responsible for the observed approximation of the two vowels will be examined by analyzing spectral changes in LOT and THOUGHT in apparent and real time in the following subchapters. In the analysis of LOT, in addition to its role in the merger, I will also point out its involvement in the NCS, which, as the analysis will show, very limited.

6.2.3.1.1 LOT

Figure 113 suggests that there is a remarkable amount of variation in the realization of LOT in the 2016 sample. The vast majority of speakers in this sample produce LOT in low central position in spontaneous speech, with little evidence of apparent-time trends in either direction. In the two more careful styles, LOT is produced with significantly more variation, and is, overall, lower and backer than in spontaneous speech.

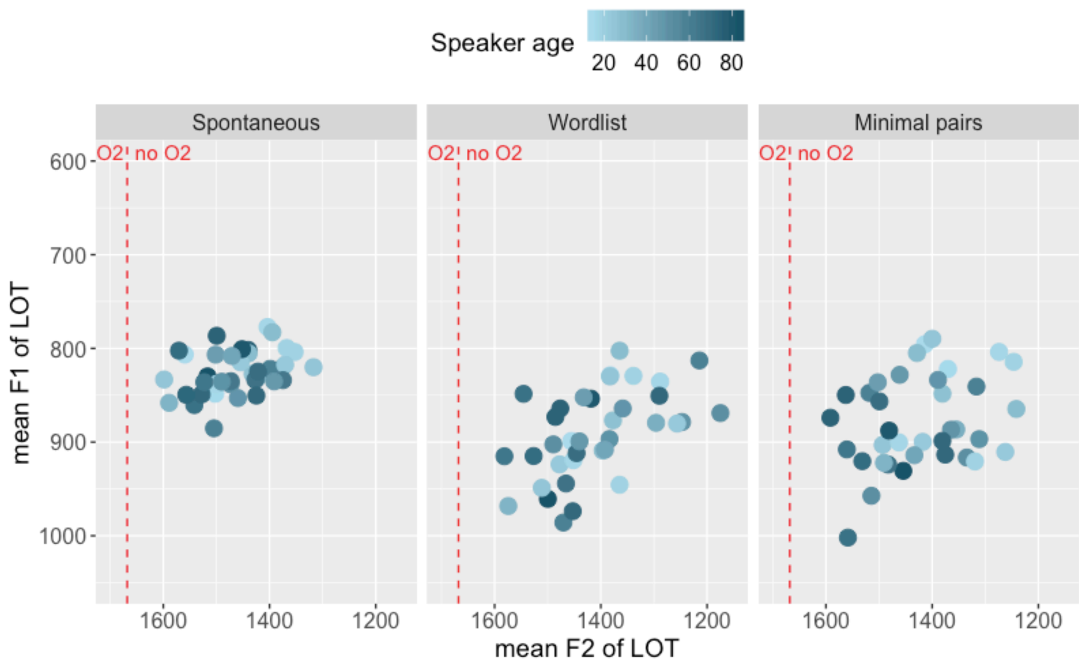


Figure 113: LOT F1 and F2 means in 2016 across *speech styles* by *age*. Lighter shades represent younger speakers, darker shades older speakers.

How low and *how* back, however, seems to depend on the speakers' *age*. While older speakers tend to produce LOT in a lower and fronter position, younger ones tend to orient toward a higher and backer LOT in wordlist style and minimal pairs. Furthermore, the height and front- or backness of LOT appear to depend significantly on the level of the speaker's education in each speech style in the 2016 data. However, as Figure 113 above shows, none of the speakers reach an F2 of 1668 Hz, which is the threshold that defines NCS-fronted LOT in the O2 criterion, in either speech style. Thus, LOT is not particularly fronted in the community, increasing the changes of a merger with THOUGHT.

LOT in Spontaneous Speech

In spontaneous speech, the height of LOT appears to be subject to very little variation, but seems to depend to a certain degree on the speakers' *age*, as well as their *level of education* in the 2016 sample. As Figure 114 below shows, the majority of 2016 participants produce spontaneous LOT at a similar height, ranging from about 800 to 850 Hz. Younger 2016 speakers, especially those with a college degree and students, seem to produce LOT with slightly smaller F1 means than older speakers, averaging around 800 Hz, while older speakers have F1 means closer to of 825 Hz. It seems that spontaneous LOT has undergone two opposing developments regarding its height in the 2016 data: LOT was lowering slightly among speakers born before 1960, regardless of their *educational level*; afterwards, speakers without a college degree have maintained the same height of spontaneous LOT, while college educated speakers started to raise it back up to a higher position. However, none of these apparent-time trends seem to be statistically significant according to the regression models presented in Table 64 below, and the lack of a significant interaction between *age* and *education* suggests that the apparent-time developments in the height of spontaneous LOT are not as drastic as Figure 114 may suggest. However, the model specific to speakers born after 1960 supports the observation that speakers without a college degree have a significantly lower LOT than college educated speakers, while no such difference is found for speakers born before 1960. The majority of students orient toward younger college educated speakers in their height of LOT in spontaneous speech.

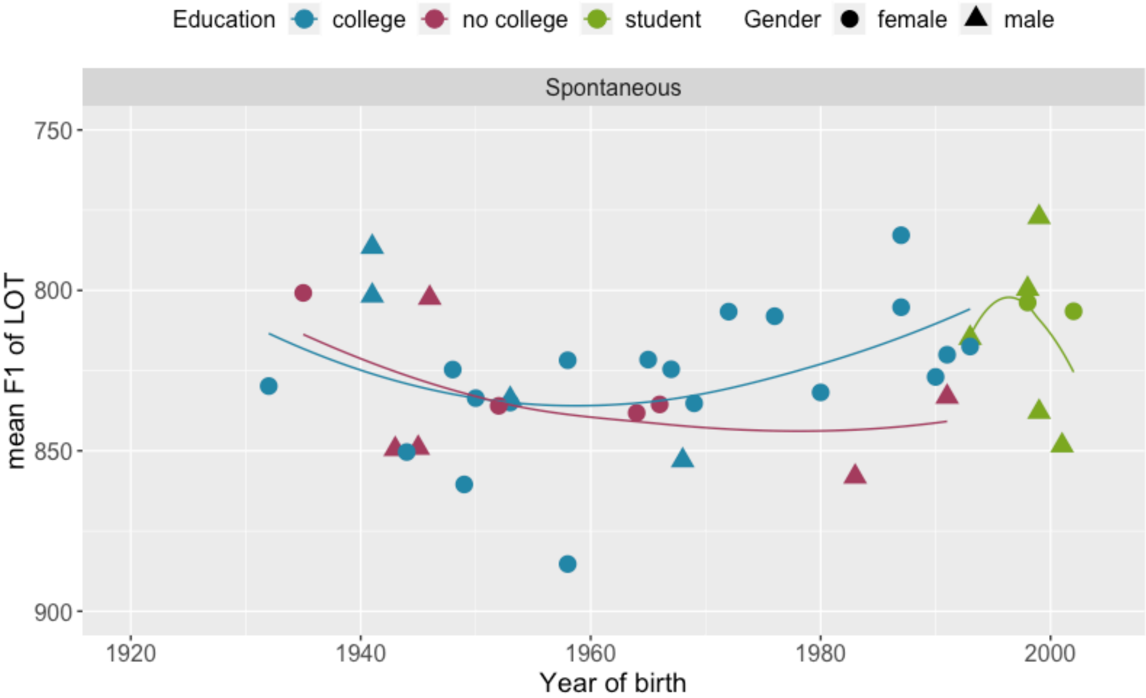


Figure 114: LOT F1 means in spontaneous speech in 2016 by *education* and *gender.*

Predictor	2016		Yob < 1960		Yob > 1960	
	Coefficient	*p*	Coefficient	*p*	Coefficient	*p*
(Intercept)	852.946 Hz		907.507 Hz		857.042 Hz	
Age	0.229 Hz	0.22	-0.413 Hz	0.539	-0.169 Hz	0.684
Education *(No college)* *(Student)*	14.545 Hz -14.653 Hz	0.048	5.853 Hz ---	0.573	26.63 Hz -15.811 Hz	0.018
Environment		2x10-16		2x10-16		2x10-16

Table 64: LOT F1 in spontaneous speech in 2016. Reference levels: college educated, /p/. Random effects: *speaker, word. n*=6403 (2016 overall)

The frontness of LOT in spontaneous speech also appears to depend on the speakers' *age* and *level of education* in the 2016 data. As can be seen in Figure 115 below, spontaneous LOT appears to be subject to two opposing apparent-time trends: It is backing among college educated speakers, while those without a college degree may be fronting it, leading to increasing differences between these two educational groups. The regression model presented in Table 65 below supports these observations partially, and predicts significant differences between speakers with and without a college degree, as well as nearly significant backing of LOT in apparent time for the sample as a whole. The lack of a significant interaction between *age* and *education* suggests that the apparent-time developments in the frontness of spontaneous LOT are not as severe as Figure 115 may suggest. However, regression models that test for the effect of *age* on the F2 of spontaneous LOT separately predict significant backing among college educated speakers (1.395 Hz, p= 0.013), while no apparent-time changes are predicted for speakers without a college degree (0.179 Hz, p= 0.683).

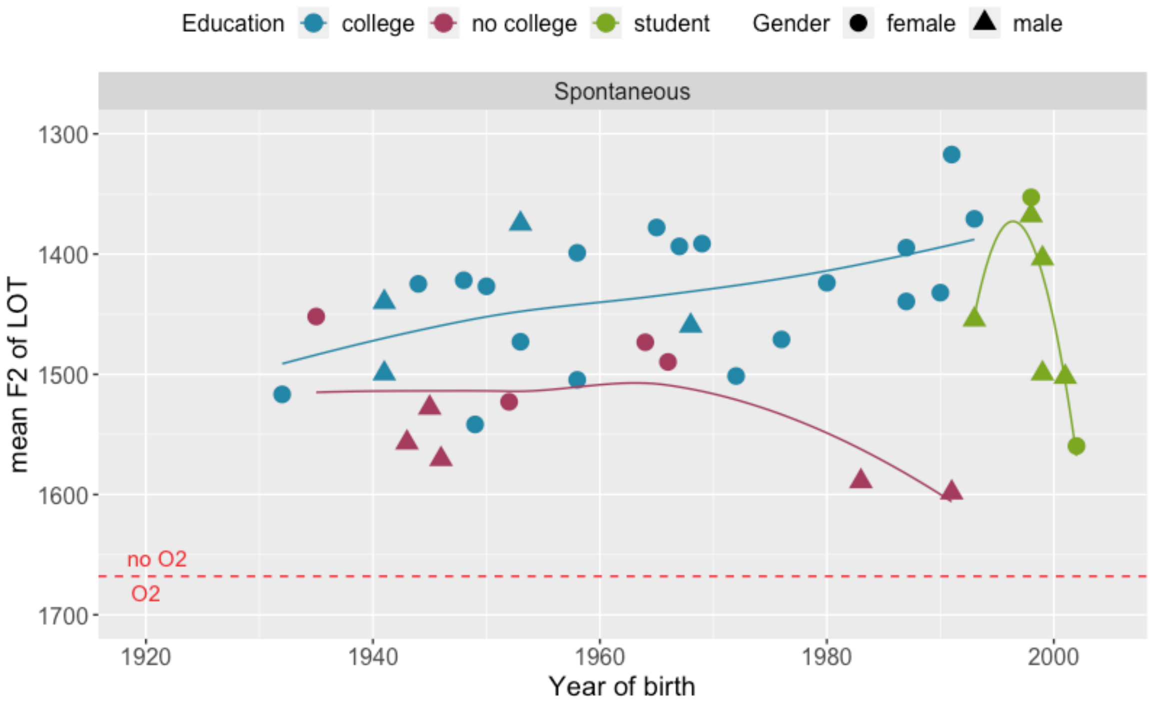


Figure 115: LOT F2 means in spontaneous speech in 2016 by *education* and *gender*.

As Figure 115 above shows, younger college educated speakers in the 2016 sample are exclusively female, while the two younger speakers without a college degree are male. This pattern may suggest that, rather than *education*, *gender* might be the determining factor in the frontness of spontaneous LOT. However, when taking students into account, *gender* appears to lose its divisiveness. Of the four students who have similar F2 means as younger college educated females, three are male. On the other hand, the student who produces LOT with an F2 that is closest to the two males without a college degree, is female. Thus, *education* rather than *gender* appears to be the most significant factor in the frontness of LOT in spontaneous speech. The regression model presented in Table 65 below does not predict *gender* to be of any significance in the frontness of spontaneous LOT.

Predictor	Coefficient	*p*
(Intercept)	1383.49 Hz	
Age	0.8014 Hz	0.057
Gender (Male)	24.9446 Hz	0.159
Education (No college)	62.7485 Hz	0.001
Environment		$5x10^{-5}$

Table 65: LOT F2 in spontaneous speech in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word*. *n*= 5697

In spontaneous speech, no significant differences in the height of LOT can be observed between 2008 and 2016 speakers, while slight real-time differences exist on the front-back dimension.

As can be seen in Figure 116, the majority of speakers from both samples produce LOT within a very narrow F1 range between 800 and 850 Hz in this speech style. Thus, they do not differ notably in their height of spontaneous LOT. Although the regression model presented in Table 66 below predicts a significant height difference between the two samples, the actual coefficient is quite small.

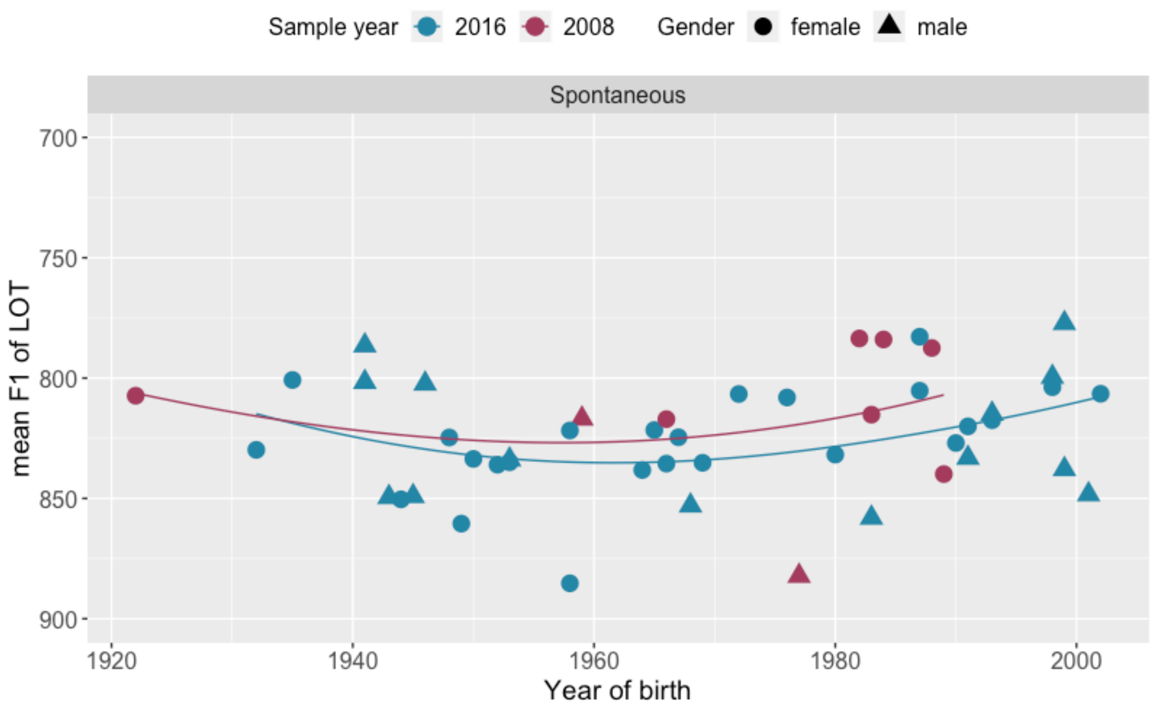


Figure 116: LOT F1 means in spontaneous speech in 2008 and 2016 by *gender.*

Furthermore, Figure 116 above suggests a slight raising of spontaneous LOT in apparent time in the 2008 data, similar to that detected among 2016 speakers. The regression model presented in Table 66 below supports the observation that, for both samples combined, the trend toward a slightly higher LOT in apparent time is significant, and the lack of a significant interaction between *age* and *sample year* confirms that the two samples do not differ notably in their apparent-time developments.

On the front-back dimension, the majority of speakers from both data sets produce spontaneous LOT with an F2 between 1400 and 1550 Hz, as shown in Figure 117. Thus, like 2016 speakers, none of the speakers from the 2008 sample meet the O2 NCS criterion in this speech style. The majority of 2008 speakers, however, fall within the higher end of this range, i.e. producing a somewhat fronter LOT than 2016 speakers. Additionally, while LOT seems to be backing slightly in apparent time among college educated speakers in the 2016 sample, no such trend emerges from the 2008 data. However, the lack of a significant interaction between *age* and *sample year* in the regression model in Table 67 below weakens the importance of this observation. Overall, however, it seems that the apparent-time trend toward backer LOT in spontaneous speech in 2016 is supported by real-time evidence.

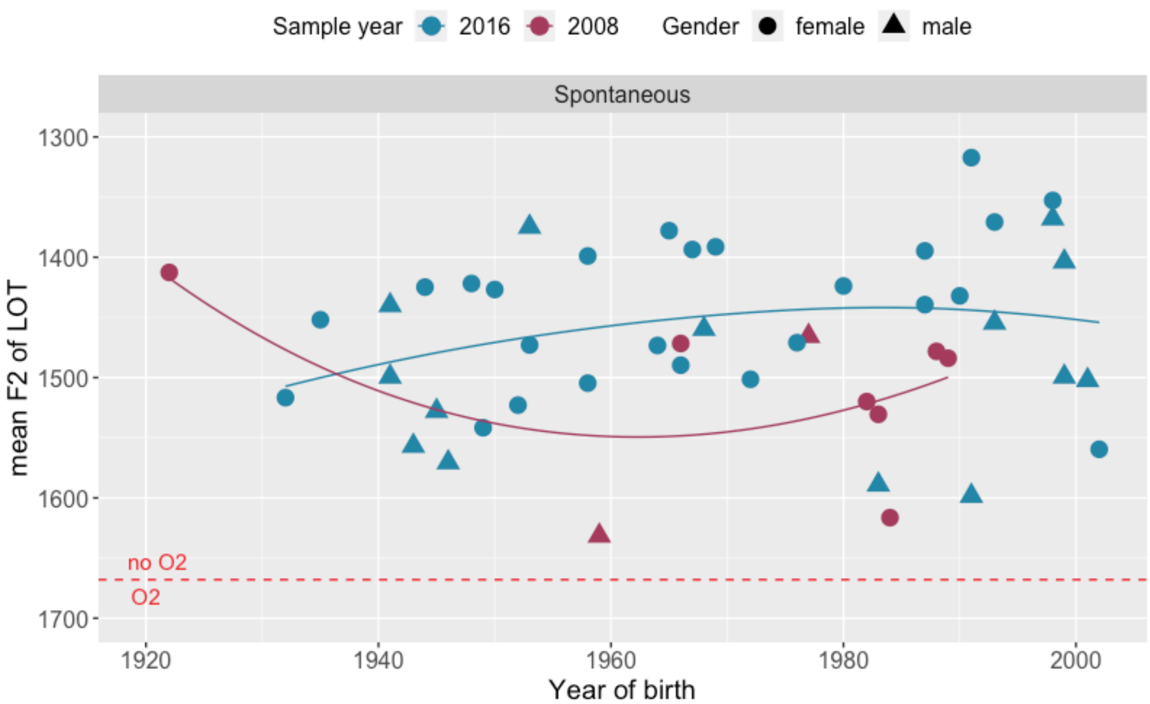


Figure 117: LOT F2 means in spontaneous speech in 2008 and 2016 by *gender.*

Predictor	Coefficient	*p*
(Intercept)	834.956 Hz	
Age	0.428 Hz	0.005
Gender (Male)	8.287 Hz	0.202
Sample year (2008)	-22.608 Hz	0.017
Environment		2x10-16

Table 66: LOT F1 in spontaneous speech in 2008 and 2016. Reference levels: females, 2016, /p/. Random effects: *speaker, word*. *n*= 6634

Predictor	Coefficient	*p*
(Intercept)	1402.958 Hz	
Age	0.789 Hz	0.041
Gender (Male)	35.525 Hz	0.04
Sample year (2008)	69.91 Hz	0.002
Environment		2x10-5

Table 67: LOT F2 in spontaneous speech in 2008 and 2016. Reference levels: females, 2016, /p/. Random effects: *speaker, word*. *n*= 6634

LOT in Wordlist Style

In wordlist style, LOT is produced within an F1 range between 850 Hz and 950 Hz by the majority of speakers in the 2016 sample. Within this range, the most determining factors for the height of LOT, again, appear to be *speaker age* and *level of education*. As can be seen in Figure 118 speakers with and without a college degree are continuously diverging from each other in their height of LOT, with college educated speakers moving toward a higher LOT, while speakers without a degree seem to be moving toward lower LOT. However, neither the apparent-time trends nor the differences between speakers of different educational backgrounds appear to be statistically significant (Table 68 below).

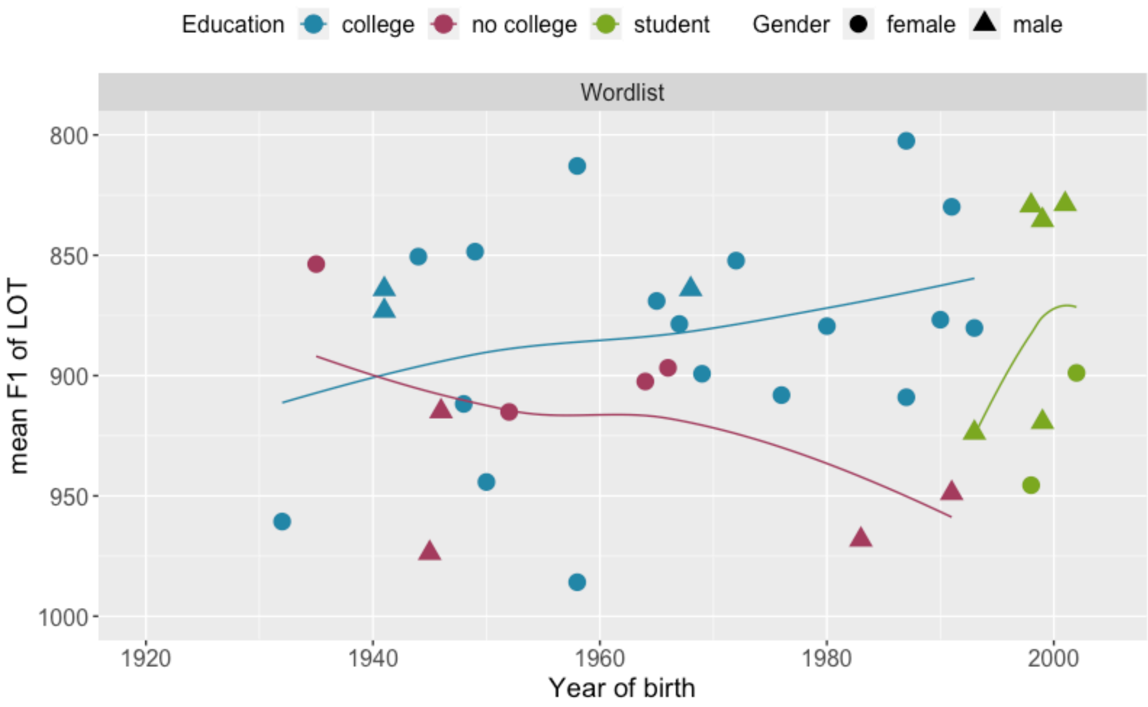


Figure 118: LOT F1 means in wordlist style in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	875.762 Hz	
Age	0.141 Hz	0.735
Gender (Male)	20.717 Hz	0.287
Education (No college)	16.81 Hz	0.363
Environment		0.016

Table 68: LOT F1 in wordlist style in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word*. *n*= 337

Again, the *gender* differences among younger speakers might suggest that the height difference is based on *gender*, not *education*. The *gender* distribution among students, however, does not supports this. In fact, the three students with the highest wordlist LOT are all male, while the student with the lowest LOT in this speech style is female. As with *education*, however, the regression model in Table 68 above finds no significant differences between male and female speakers.

In wordlist style, the speakers' *age* and *level of education* appear to have a notable impact on the frontness of LOT in the 2016 data. As can be seen in Figure 119 below, the oldest speakers in 2016, regardless of their educational background, produce wordlist LOT with an F2 of about 1500 Hz, thereby falling short of the 1668 Hz threshold of the O2 criterion. Younger speakers of the two educational groups, on the other hand, differ by up to 200 Hz in the frontness of LOT, owing to continuous fronting among speakers without a college degree, and retraction of LOT among college educated speakers. It should, however, be noted that this retraction appears to have halted. While LOT seems to have been retracted among college educated speakers until about 1970, it has remained steady at about 1350 Hz since then. Despite continuous fronting among speakers without a college degree, none of them reach the 1668 Hz mark of the O2 criterion. Again, the *gender* patterning of the students suggests that the differences in F2 of LOT among younger speakers with and without a college degree are not *gender*-based, as three out of five male students orient toward backer LOT at the same level as young females.

The regression model in Table 69 below confirms that the differences in the frontness of LOT between the two educational groups in wordlist style are statistically significant, while the lack of a significant interaction between *age* and *education* suggests that the apparent-time developments do not differ drastically between these two groups. However, regression models that test for the effect of *age* without *gender* for speakers with and without a college education separately find significant apparent-time LOT backing for college educated speakers (2.033 Hz, *p*= 0.044), but no significant changes for speakers without a college degree (-0.463 Hz, *p*= 0.717).

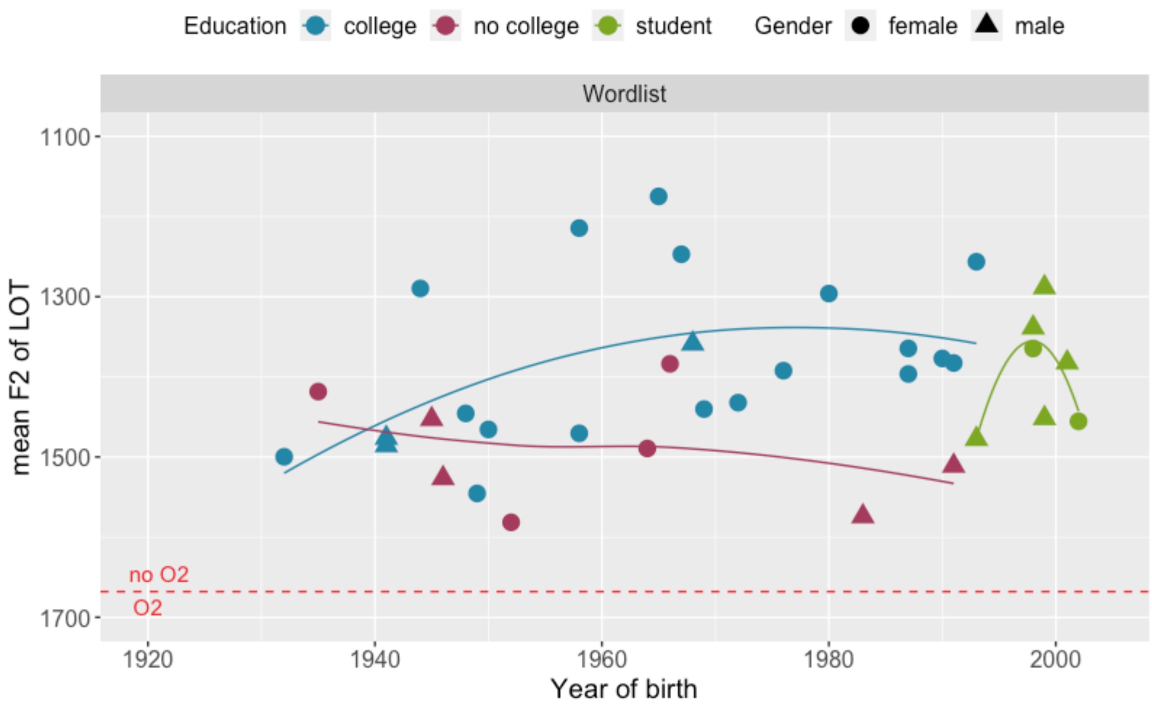


Figure 119: LOT F2 means in wordlist style in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	1335.494 Hz	
Age	1.366 Hz	0.087
Education (*No college*)	87.003 Hz	0.011
Environment		0.036

Table 69: LOT F2 in wordlist style in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word*. *n*= 337

In wordlist style, speakers in the 2008 sample appear to produce a notably higher and fronter LOT than 2016 speakers.

As described above, 2016 speakers produce wordlist LOT with an F1 between 850 and 950 Hz. All but one of the 2008 speakers, on the other hand, have F1 means of 800 Hz or less, as can be seen in Figure 120 below. However, the regression model presented in Table 70 below does not predict this difference between the speakers of the two samples to be of any significance. Contrary to the raising of LOT in apparent time in 2016, no clear apparent-time trends in either direction emerge from the 2008 data in Figure 120 below.

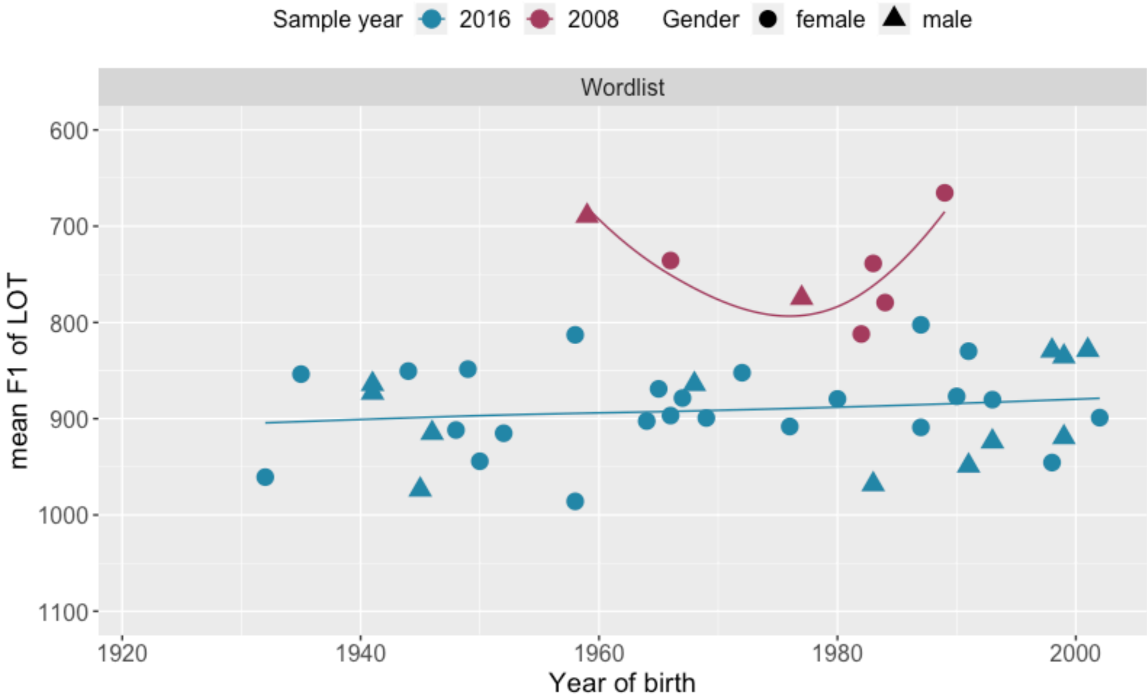


Figure 120: LOT F1 in wordlist style in 2008 and 2016 by *gender*.

On the front-back dimension most 2016 speakers range in F2 means between 1400 and 1500 Hz in wordlist style; on the other hand, 2008 speakers have an F2 range from 1600 to 1900 Hz, i.e. their wordlist LOT is 100 to 500 Hz fronter than that of 2016 speakers, as shown in Figure 121 below. However, the difference estimated by the regression model in Table 71 below is much smaller, and not statistically significant. Nevertheless, the difference is great enough to divide the two samples in terms of their O2 participation in this speech style: In the 2008 sample, five of the seven speakers produce LOT front enough to meet the O2 criterion, while none of the 2016 speakers do. Although no clear apparent-time trend emerges from the 2008 data, the regression model in Table 71 below predicts significant backing of wordlist LOT in the combined 2008 and 2016 data, while this trend was only significant for college educated speakers in 2016 alone.

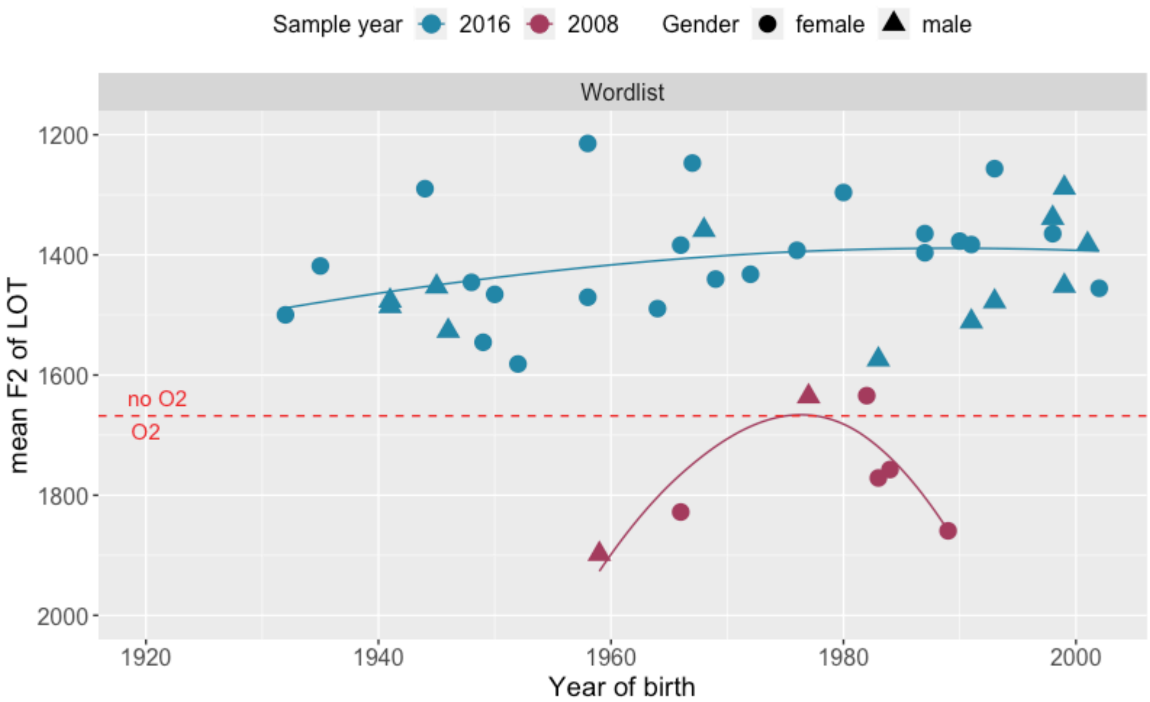


Figure 121: LOT F2 in wordlist style in 2008 and 2016 by *gender*.

Predictor	Coefficient	*p*
(Intercept)	875.618 Hz	
Age	0.38 Hz	0.234
Gender (Male)	-1.372 Hz	0.921
Sample year (2008)	-21.409 Hz	0.495
Environment		0.119

Table 70: LOT F1 in wordlist style in 2008 and 2016. Reference levels: females, 2016, /p/. Random effects: *speaker, word*. *n*= 464

Predictor	Coefficient	*p*
(Intercept)	1355.113 Hz	
Age	1.38 Hz	0.028
Gender (Male)	35.495 Hz	0.186
Sample year (2008)	64.561 Hz	0.266
Environment		0.208

Table 71: LOT F2 in wordlist style in 2008 and 2016. Reference levels: females, 2016, /p/. Random effects: *speaker, word*. *n*= 464

LOT in Minimal Pairs

In the most careful speech style, LOT appears to be undergoing a slight apparent-time raising in the 2016 data, depending on the speakers' *educational background*. Figure 122 below shows that the majority of speakers in this sample produce minimal pair LOT within an F1 range between 800 and 950 Hz. Students and younger college educated speakers appear to orient toward the lower end of this range, suggesting that college educated speakers are raising LOT in apparent time in this speech style. Speakers without a college degree do not seem to participate in this trend. As a result, an increasing height difference can be observed between these two educational groups. Speakers from both groups born before 1960 produce minimal pair LOT with a mean F1 of about 900 Hz regardless of their *educational background*. Speakers from the two groups born after 1960, on the other hand, have an average F1 difference of about 60 Hz. The students in the 2016 sample appear to fall in between the F1 ranges of younger speakers with and without a college degree. While some of their means are closer to those of speakers without a college degree, others appear to follow the apparent-time raising of college educated speakers. In fact, four of the seven students in the sample are among the six speakers with the lowest F1 means in this speech style in the entire sample. Because this group of four includes both male and female students, *gender* does not appear to be a significant factor in the height of minimal pair LOT, although the *gender* distribution of younger speakers with and without a college degree may suggest otherwise.

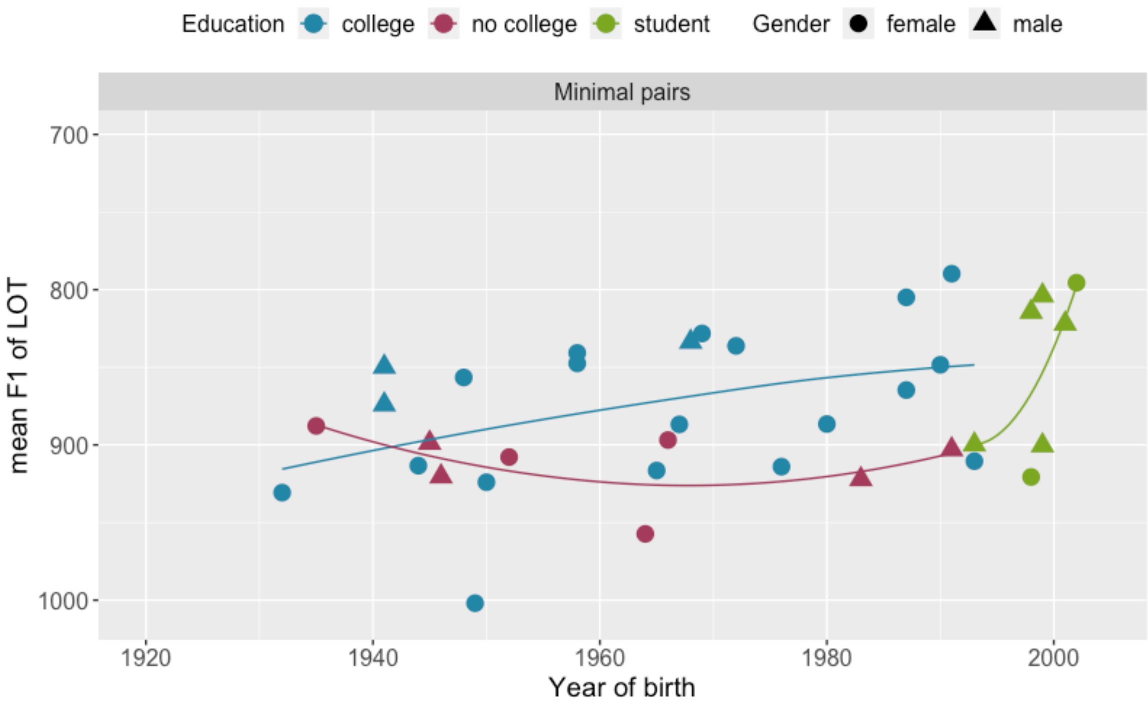


Figure 122: LOT F1 means in minimal pairs in 2016 by *education* and *gender*.

The regression model in Table 72 below supports the observation that significant differences exist between speakers with and without a college degree, while the effect of *gender* in the height of LOT in minimal pair production is found to be insignificant. The effect of *age* does not quite reach the level of statistical significance in this model, but does reach this level in a model that considers only college educated speakers (1.429 Hz, p= 0.02).

Predictor	Coefficient	p
(Intercept)	774.851 Hz	
Age	0.8 Hz	0.075
Gender (Male)	-20.38 Hz	0.313
Education (No college)	41.078 Hz	0.04
Environment		0.002

Table 72: LOT F1 in minimal pairs in 2016. Reference levels: females, college educated, /n/. Random effects: *speaker, word*. *n*= 571

F2 of LOT also seems to depend on *speaker age* and *education* in the production of minimal pairs in 2016. Figure 123 below suggests that college educated speakers are significantly backing LOT in apparent time, while for speakers without a college degree the frontness of LOT remains stable. As a result, 2016 speakers with and without a college degree produce minimal pair LOT with an increasingly different F2. Younger speakers with a college degree average around 1300 Hz, while those without have F2 means around 1500 Hz. No such difference can be observed for older speakers in the sample. The students in this data set have LOT F2 means that range from 1280 to 1480 Hz in minimal pair production, and thus fall in between speakers with and without a college degree, although they seem to lean more toward backer LOT in this speech style. This tendency can be observed for both male and female students. Thus, *gender* does not appear to be as deciding a factor as *education* in the frontness of minimal pair LOT, although the *gender* distribution among younger speakers with and without a college degree may suggest otherwise. None of the 2016 speakers meet the O2 criterion in minimal pairs.

The regression model presented in Table 73 below supports these observations only partially. While it does corroborate the impression of a significant apparent-time backing of LOT in minimal pairs, it does not predict any significant differences in the frontness of LOT between speakers with and without a college degree, nor does it predict the apparent-time developments in these two groups to differ significantly, as indicated by the lack of a significant interaction between *age* and *education*. Tested separately, however, significant backing in apparent time is predicted only for college educated speakers (2.93 Hz, *p*= 0.006), but not for speakers without a college degree (0.06 Hz, *p*= 0.976).

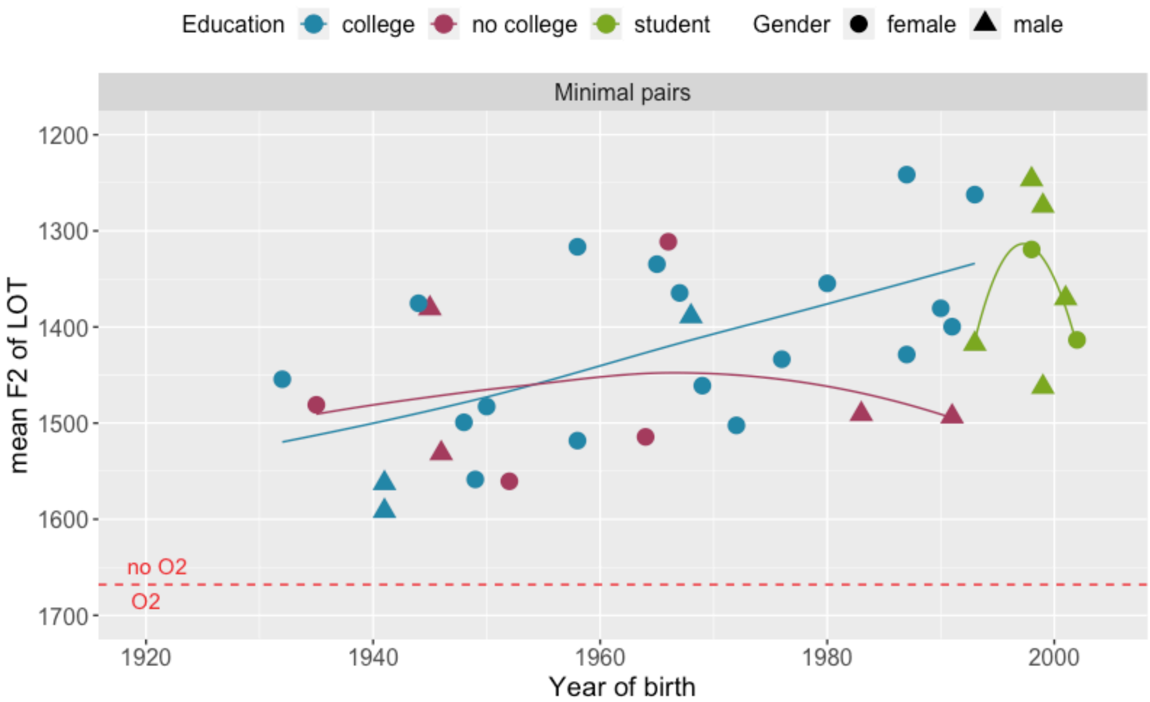


Figure 123: LOT F2 means in minimal pairs in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	1239.505 Hz	
Age	2.33 Hz	0.008
Gender (Male)	22.723 Hz	0.531
Education (No college)	46.356 Hz	0.227
Environment		0.057

Table 73: LOT F2 in minimal pairs in 2016. Reference levels: females, college educated, /n/. Random effects: *speaker, word*. *n*= 571

In minimal pair production, speakers from the 2008 sample seem to produce LOT at a comparable height, but notably fronter in the vowel space than 2016 speakers.

As can be seen in Figure 124 below, the majority of speakers from both samples produce minimal pair LOT with F1 means between 850 and 950 Hz. Thus, no notable differences in the height of LOT can be observed in this speech style for the majority of speakers. However, as Figure 124 indicates, 2008 speakers appear to be undergoing a different apparent-time trend than 2016 speakers, as they seem to be lowering minimal pair LOT slightly, while the majority of 2016 speakers is raising it. The regression model in Table 74 below confirms that this difference is statistically significant, and predicts that 2008 speakers are lowering LOT at about the same pace as 2016 speakers are raising it.

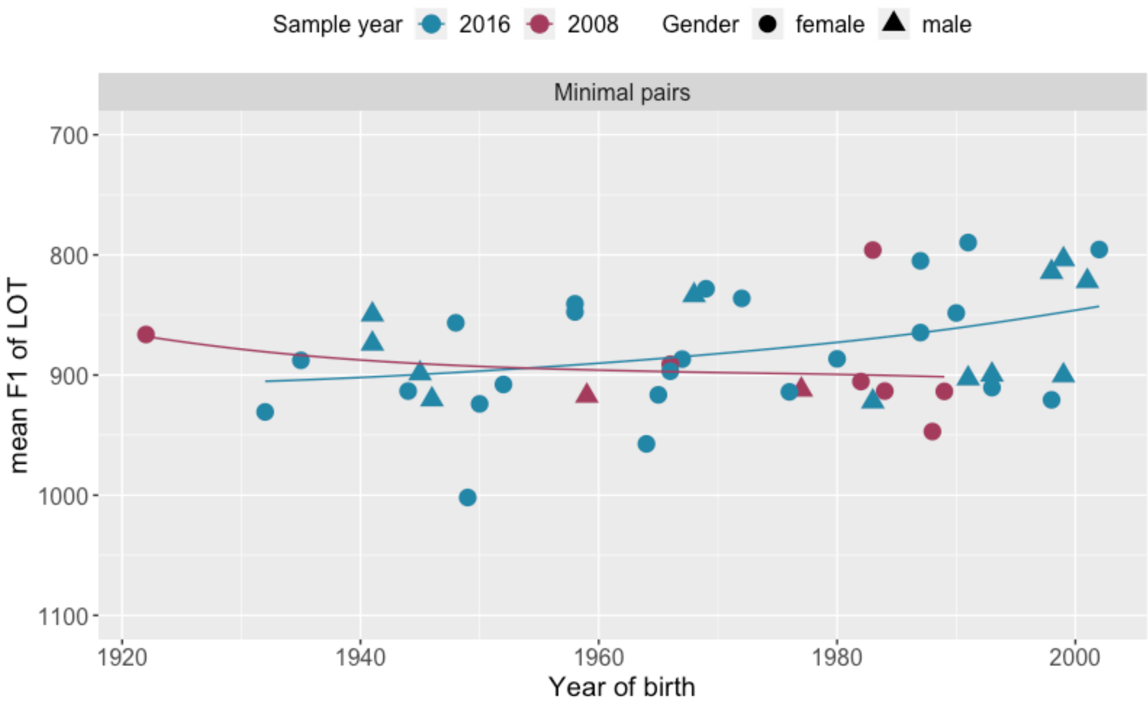


Figure 124: LOT F1 in minimal pairs in 2008 and 2016 by *gender.*

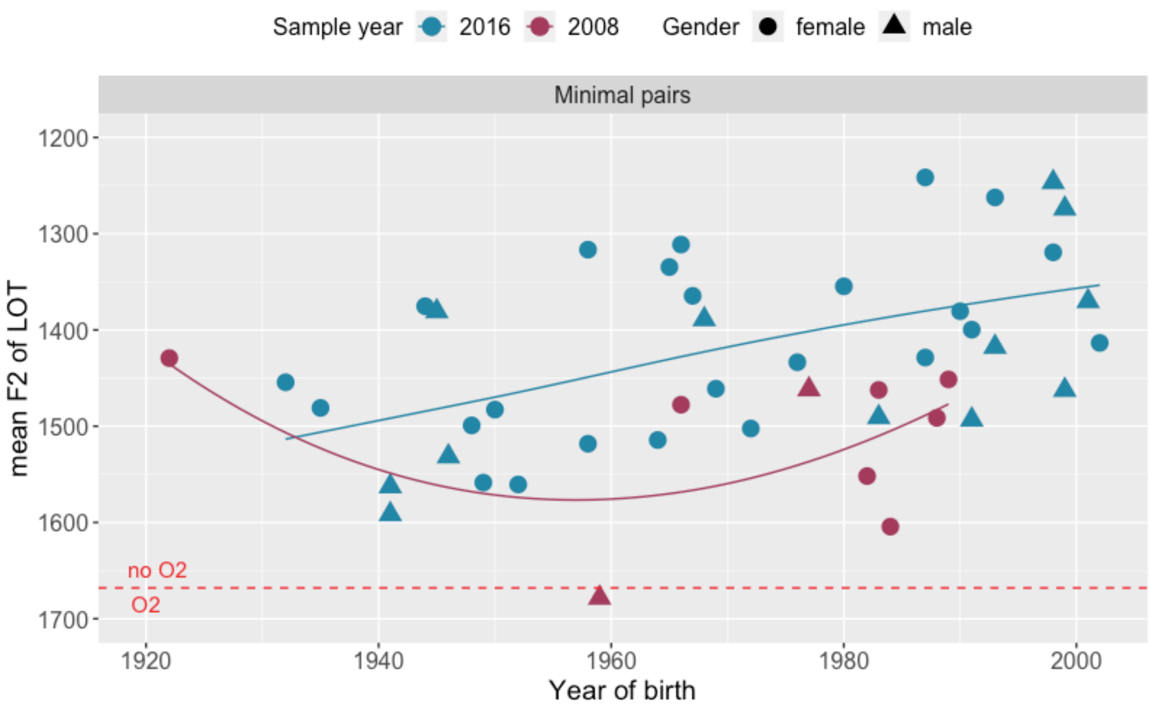


Figure 125: LOT F2 in minimal pairs in 2008 and 2016 by *gender*.

Regarding F2, Figure 125 above shows that 2008 speakers are producing LOT in minimal pairs in a fronter position than 2016 speakers. The F2 mean range for 2016 speakers is 1350 to 1500 Hz. While many of the 2008 speakers fall within this range, two of them have F2 means that are greater than 1500 Hz. The two samples also appear to differ in apparent-time developments: Most speakers in 2016 are significantly retracting LOT to a backer position, while in the 2008 data, no clear trend can be observed. The regression model in Table 75 confirms the observation that developments over time differ significantly between the two samples. Despite these inter-set differences, O2 participation in minimal pairs is just as weak among 2008 speakers as it is in the 2016 data, as only one 2008 speaker reaches the 1668 Hz threshold in this style, though he does so just barely, as can be seen in Figure 125 above.

Predictor	Coefficient	*p*
(Intercept)	783.092 Hz	
Age	0.925 Hz	
Gender (Male)	-6.646 Hz	0.657
Sample year (2008)	139.998 Hz	
*Age*2008*	-2.101 Hz	0.031
Environment		0.0001

Table 74: LOT F1 in minimal pairs in 2008 and 2016. Reference levels: females, 2016, /n/. Random effects: *speaker, word*. *n*= 748

Predictor	Coefficient	*p*
(Intercept)	1270.016 Hz	
Age	2.505 Hz	
Gender (Male)	33.692 Hz	0.198
Sample year (2008)	278.492 Hz	
*Age*2008*	-3.579 Hz	0.025
Environment		0.059

Table 75: LOT F2 in minimal pairs in 2008 and 2016. Reference levels: females, 2016, /n/. Random effects: *speaker, word*. *n*= 748

Style shifting LOT

The height and frontness at which speakers in the 2016 sample produce LOT appears to depend on *speech style.* The vast majority of speakers in this sample appear to produce a considerably higher and fronter LOT in spontaneous speech than they do in the more careful styles, as shown in Figures 126 and 127 below.

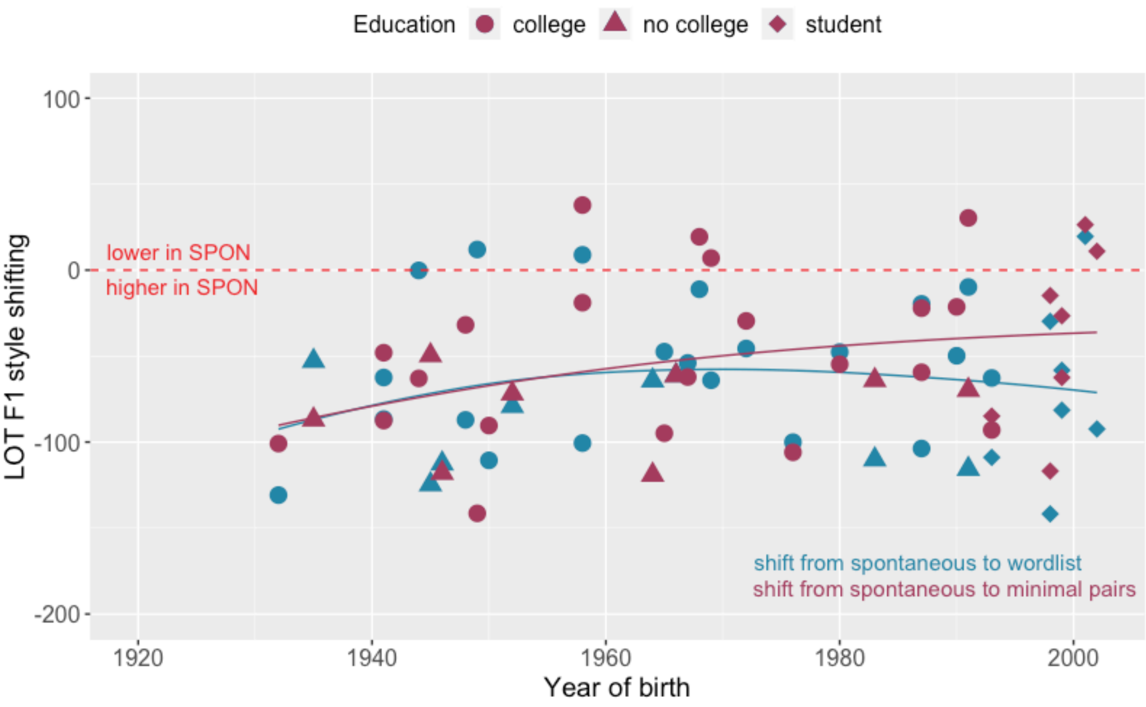


Figure 126: F1 difference between spontaneous, wordlist and minimal pair LOT in 2016 by *education*. A positive value indicates that the vowel is lower in spontaneous speech. Blue represents the shift from spontaneous speech to wordlist style, wine-red the shift from spontaneous speech to minimal pairs.

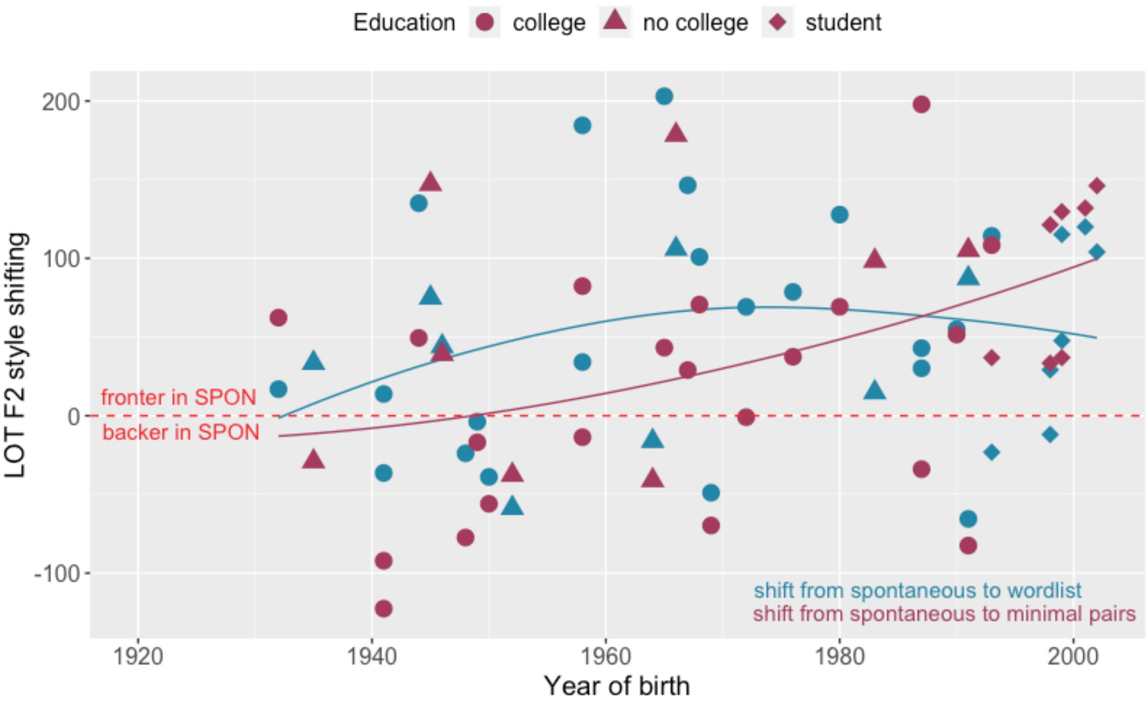


Figure 127: F2 difference between spontaneous, wordlist and minimal pair LOT in 2016 by *education*. A positive value indicates that the vowel is fronter in spontaneous speech. Blue represents the shift from spontaneous speech to wordlist style, wine-red the shift from spontaneous speech to minimal pairs.

While the shift from spontaneous speech to wordlist style appears to remain relatively consistent, the range and direction of style shifting from spontaneous speech to minimal pairs seem to be change in apparent time. As Figures 126 and 127 above show, older speakers in the 2016 sample shift to a notably lower and somewhat fronter LOT in minimal pairs. The youngest speakers in the sample, on the other hand, reduce the height differences in their shift to minimal pairs, while simultaneously increasing the differences in frontness in the opposite direction than older speakers, i.e. they shift to a notably *backer* LOT in minimal pairs while maintaining the same height. Figure 127 suggests that the effect of *style* on the frontness of LOT switches from negative to positive between 1950 and 1960 in apparent time. Afterwards, only a small minority of speakers shift in the opposite direction, i.e. to a fronter LOT in minimal pairs.

The regression model in Table 76 confirms that younger speakers make only minimal style differences between spontaneous and minimal pair LOT, and that these style differences are greater among older speakers.

Predictor	Coefficient	*p*
(Intercept)	849.599 Hz	
Age	0.249 Hz	
Gender (Male)	2.85 Hz	0.695
Education		
(No college)	16.696 Hz	0.03
(Students)	-14.53 Hz	
Style		
(Wordlist)	27.449 Hz	
(Minimal pairs)	13.288 Hz	
*Age*style*		
(Wordlist)	-0.059 Hz	0.012
(Minimal pairs)	0.451 Hz	
Environment		$2x10^{-16}$

Table 76: Effect of style on F1 of LOT in 2016. Reference levels: females, college educated, spontaneous speech, /p/. Random effects: *speaker, word*. n= 7537

Predictor	Coefficient	*p*
(Intercept)	1391.551 Hz	
Age	0.71 Hz	
Gender (Male)	23.099 Hz	0.207
Education		
(No college)	66.273 Hz	
(Students)	-0.809 Hz	
Style		
(Wordlist)	-26.026 Hz	
(Minimal pairs)	-90.23 Hz	
*Age*style*		
(Wordlist)	0.511 Hz	$5x10^{-6}$
(Minimal pairs)	1.568 Hz	
*Style*education*		
*No college*WL*	16.824 Hz	
*No college*MP*	-34.493 Hz	
*Students*WL*	32.392 Hz	0.034
*Students*MP*	4.053 Hz	
Environment		$9x10^{-7}$

Table 77: Effect of style on F2 of LOT in 2016. Reference levels: females, college educated, spontaneous speech, /p/. Random effects: *speaker, word*. n= 7537

Likewise, the regression model presented in Table 77 above supports the observation that younger speakers in the 2016 sample produce a backer LOT in minimal pairs than they do in spontaneous speech, and that these style differences are largely absent among older speakers in this sample. Furthermore, this results of this regression model suggest a significant difference in the effect of *style* across different educational groups in the 2016 data: The range of style shifting among speakers without a college education in this sample is smaller than that of speakers with a college degree, i.e. their wordlist and minimal pair LOT is more similar to their spontaneous LOT than it is for college educated speakers. The estimated difference in range of shifting, however, is relatively small considering that it concerns F2.

Speakers from the 2008 and 2016 samples appear to differ in their amount and direction of style shifting in the realization of LOT. As can be seen in Figures 130 and 131 below, a clear difference appears to exist between all three speech styles on the height dimension for 2008 speakers: They produce LOT in wordlist style the highest up and the furthest forward in the vowel space compared to the other two speech styles. This is the opposite pattern of what was observed for speakers in the 2016 sample, where wordlist LOT is produced lower and further back than both spontaneous and minimal pair LOT by most speakers.

The shift from spontaneous speech to minimal pairs LOT in the 2008 sample resembles that of 2016 speakers for the most part, with a clear shift to a lower LOT in minimal pairs, but little difference on the front-back dimension. Only among younger speakers do inter-set differences become somewhat more notable for these two speech styles, especially on the height dimension, owing to the apparent-time raising of LOT in minimal pairs in the 2016 sample, and the absence of such trend in the 2008 data. Thus, younger 2016 speakers produce minimal pair and spontaneous LOT with smaller height differences than younger 2008 speakers. As a result of the style patterns in 2008, some of the speakers from this sample meet the O2 NCS criterion in wordlist style and minimal pairs, while none of them do in spontaneous speech.

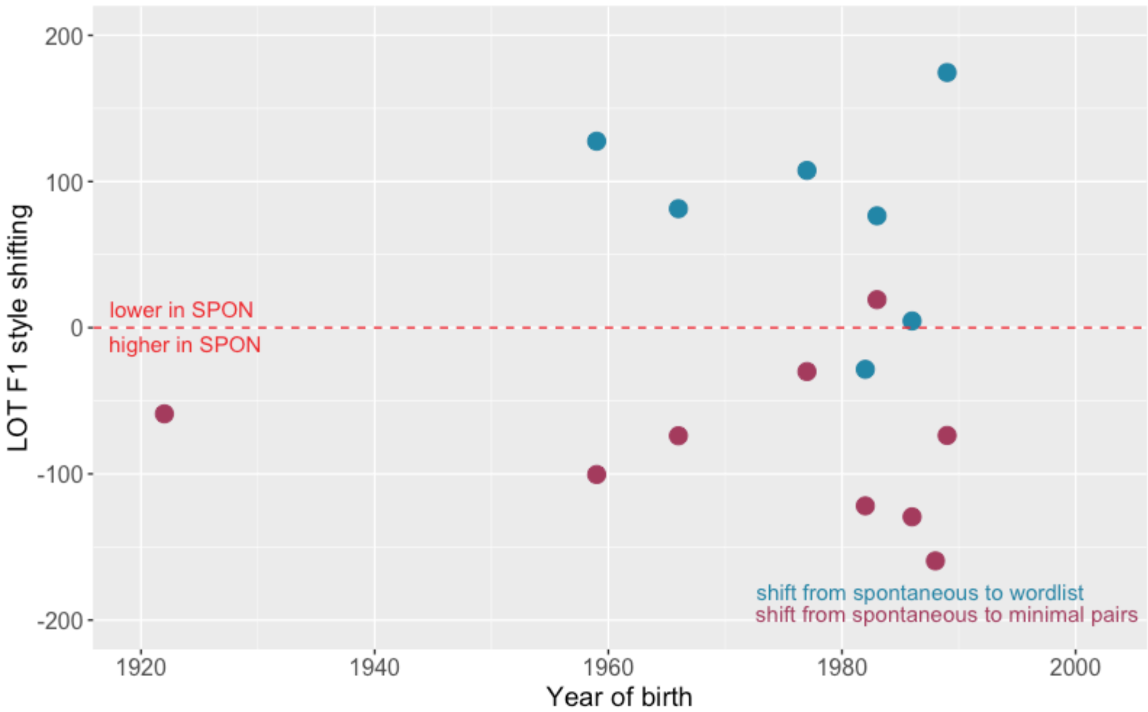


Figure 128: F1 difference between spontaneous, wordlist and minimal pair LOT in 2008. A positive value indicates that the vowel is lower in spontaneous speech. Blue represents the shift from spontaneous speech to wordlist style, wine-red the shift from spontaneous speech to minimal pairs.

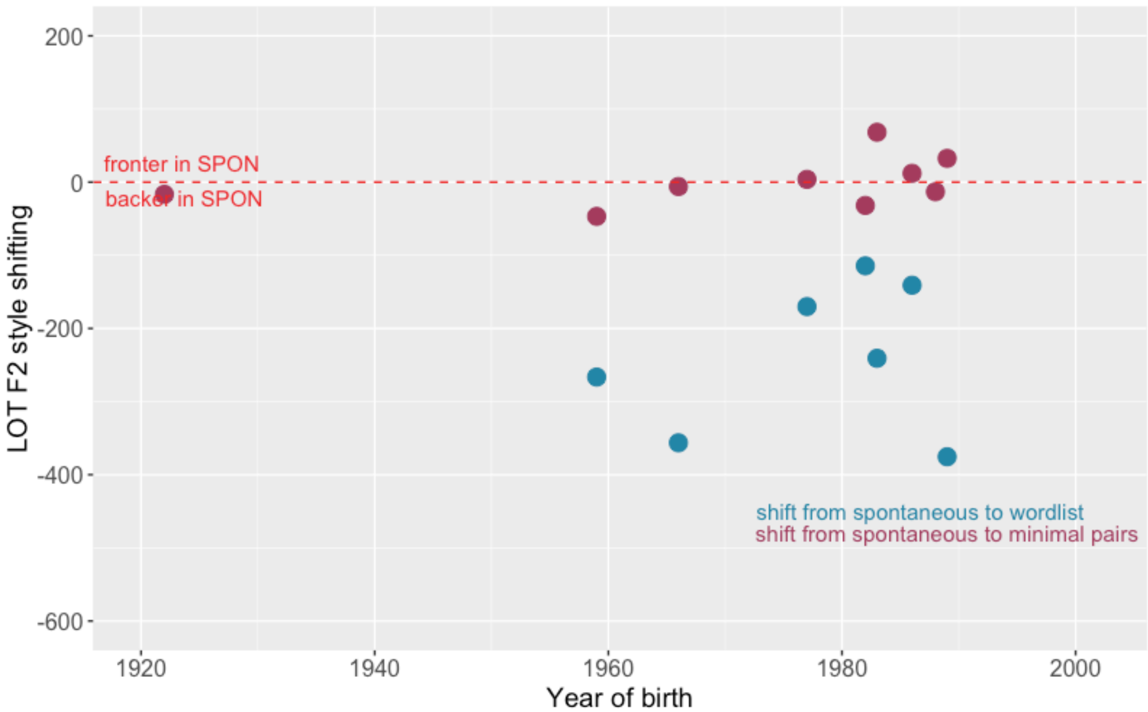


Figure 129: F2 difference between spontaneous, wordlist and minimal pair LOT in 2008. A positive value indicates that the vowel is fronter in spontaneous speech. Blue represents the shift from spontaneous speech to wordlist style, wine-red the shift from spontaneous speech to minimal pairs.

Interim Summary: LOT in Production

The analysis of LOT production has shown that the vowel is undergoing significant changes in apparent and real time in the community. In all three speech styles, LOT is undergoing retraction in the 2016 data, though the pace of this differs from style to style, and only college educated speakers seem to participate in this trend.

In spontaneous speech, LOT retraction is progressing quite slowly in the 2016 data (5 Hz per 10 years), but is supported by real-time evidence from the 2008 data as shown in Figure 130 below. Retraction is accompanied by a slight raising of LOT in this speech style, which can be observed in both apparent and real time as well. All of these changes, however, appear to depend on the *educational background* of the speakers, as speakers without a college degree tend to not participate in either of the two trends. Compared to the other speech styles, spontaneous LOT is produced the highest and frontest for the majority of 2016 speakers.

The changes in wordlist LOT are proceeding at an intermediate pace in the 2016 data. While it seems to be maintaining its height in the 2016 sample, it is backing by about 14 Hz per 10 years in apparent time. This, however, is again dependent on the speaker's *level of education*, as this trend can only be observed for college educated speakers. Compared to spontaneous speech, wordlist LOT is produced notably lower and backer by the majority of 2016 speakers. In the 2008 data, on the other hand, wordlist LOT is significantly higher and fronter than it is in spontaneous speech.

The highest degree of change in the realization of LOT can be observed in minimal pairs. In the 2016 data, minimal pair LOT is raising by 9, and backing by 23 Hz per 10 years. Both of these apparent-time trends are supported by real-time evidence, as can be seen in Figure 130 below. While minimal pair LOT is produced increasingly backer than spontaneous LOT in the 2016 data, it is notably *lower* for the majority of speakers. This F1 style difference, however, is decreasing and, in extreme cases, has reversed among the younger speakers in the 2016 sample.

As a result of these developments, participation in NCS-fronted LOT as defined by the O2 criterion is decreasing in the community in real time. While the majority of 2008 speakers participate in this feature in wordlist style, and show little to no indication of a reversal, none of the 2016 speakers participate in this criterion in this speech style. Similarly, in minimal pairs, one of the 2008 speakers meets the O2 criterion, while none of the speakers from the 2016 sample reach the 1668 Hz threshold for this criterion.

Interestingly, there is no indication of LOT fronting at any point in time in the data at all. This is quite surprising, as evidence of LOT fronting would have been expected around the time the NCS developed in the Inland North. Given the absence of any such trend, the lack of O2 participation is not surprising.

On the other hand, the results presented here indicate definite progress toward a merger with THOUGHT by mechanism of approximation, as LOT is backing and raising, thereby approaching THOUGHT. Whether a similar approximation of THOUGHT toward LOT contributes to the progress toward the merger will be explored in the following subchapter.

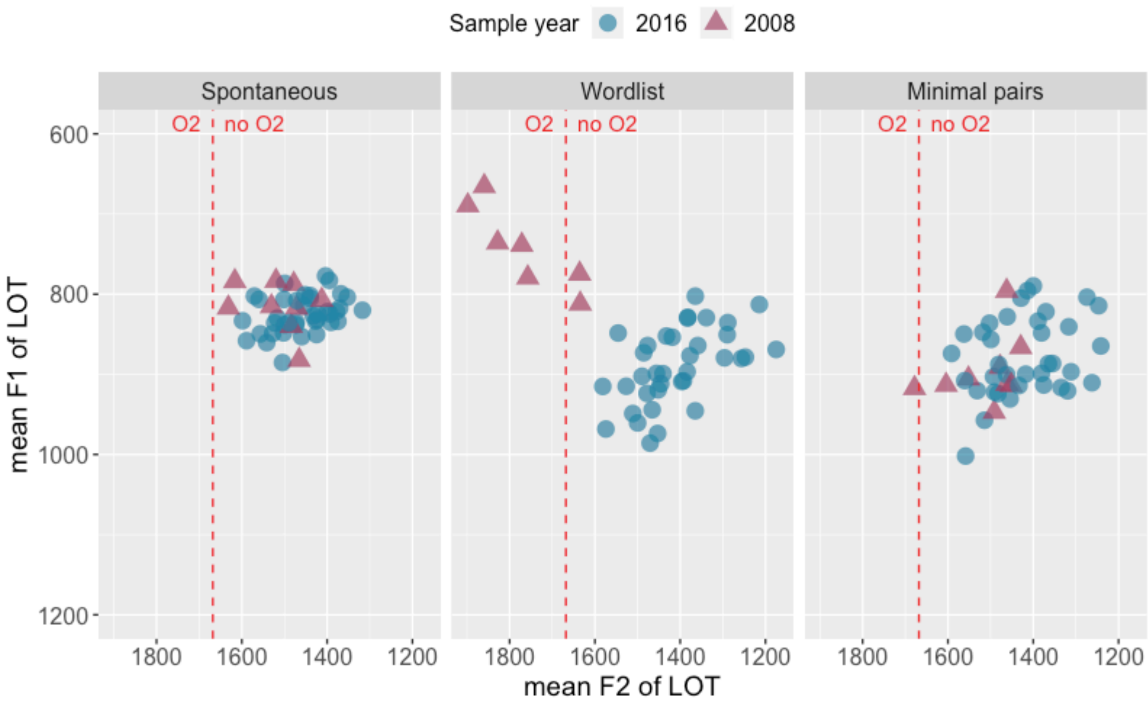


Figure 130: LOT F1 and F2 means in 2008 and 2016 across *speech styles.*

6.2.3.1.2 THOUGHT

The height and frontness of THOUGHT appear to slightly depend on the speaker's *age* in the 2016 sample, especially in the two more careful speech styles. Figure 131 below suggests that the vast majority of speakers in this sample produce THOUGHT relatively high in the vowel space and back of center in spontaneous speech, with little evidence of apparent-time trends in either direction. In the two more careful styles, notably more inter-speaker variation emerges in the production of THOUGHT in the 2016 sample. THOUGHT appears to be produced lower and backer in the vowel space by some speakers, while others appear to be moving in the opposite direction on the height dimension. The speaker's *age* seems to be an indication of *which* direction of change they participate in: While older speakers tend to produce THOUGHT in a higher position in the more careful styles, younger ones tend to orient toward a lower THOUGHT in these speech styles.

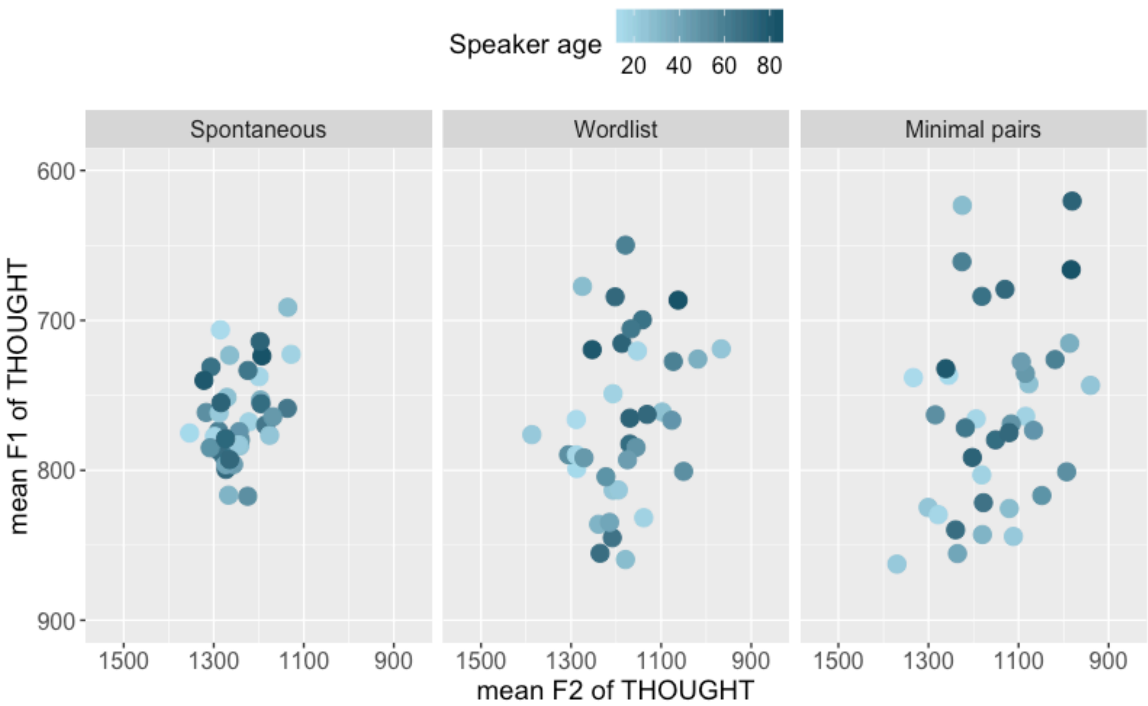


Figure 131: THOUGHT F1 and F2 means in 2016 across *speech styles* by *age*. Lighter shades represent younger speakers, darker shades older speakers.

THOUGHT in Spontaneous Speech

In terms of height, THOUGHT in spontaneous speech appears to be produced with relative consistency by 2016 speakers. Figure 132 below shows that the majority of speakers in this sample produce spontaneous THOUGHT with F1 means between 750 and 800 Hz. Only a few speakers have F1 means that are smaller than that, i.e. realize a somewhat higher THOUGHT, the majority of them being female. Thus, the height of THOUGHT may depend on the speakers' *gender* in this speech style in the 2016 data. The regression models presented in Table 78 below, however, predict only minor *gender* differences, that do not, or just barely, reach the level of statistical significance.

Figure 132 below also suggests apparent-time changes in the height of THOUGHT in the 2016 sample. The direction of change, however, appears to depend on the generation of speakers, as older speakers seem to have lowered THOUGHT, while younger speakers are raising it. Among younger speakers, the pace of THOUGHT raising appears to depend on the speaker's *gender*, with females raising it at a slightly faster pace than males. The regression models in Table 78 partially support these observations. Among speakers born before 1970, THOUGHT is estimated to lower, albeit slightly, at a statistically significant level. The model for younger speakers predicts THOUGHT raising that matches the pace of *lowering* among older speakers. In other words, younger speakers have raised THOUGHT back to its starting point. However, THOUGHT raising among younger speakers lacks statistical significance.

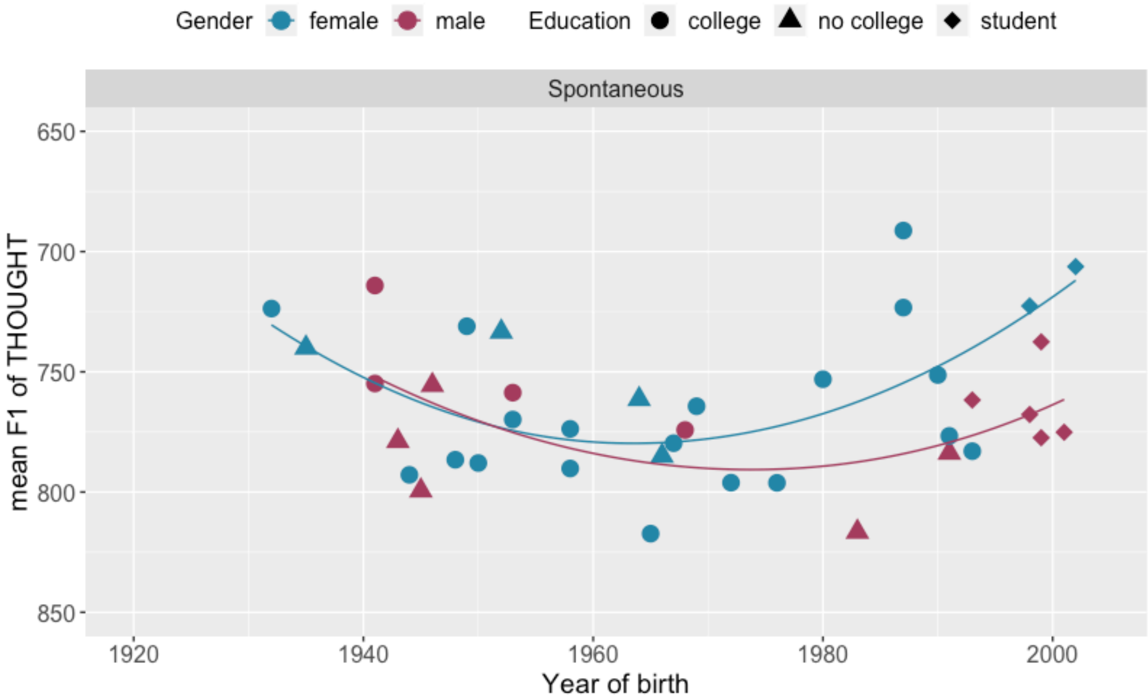


Figure 132: THOUGHT F1 means in spontaneous speech in 2016 by *gender* and *education.*

Predictor	2016		Yob < 1970		Yob > 1970	
	Coefficient	*p*	Coefficient	*p*	Coefficient	*p*
(Intercept)	782.789 Hz		824.463 Hz		753.678 Hz	
Age	-0.075 Hz	0.672	-0.709 Hz	0.04	0.76 Hz	0.401
Gender *(Male)*	13.407 Hz	0.078	15.702 Hz	0.048	27.208 Hz	0.1
Education *(No college)*	10.124 Hz	0.176	10.19 Hz	0.176	---	---
Environment		2x10^{-16}		2x10^{-16}		2x10^{-16}

Table 78: THOUGHT F1 in spontaneous speech in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word. n=* 4471 (overall 2016)

In terms of frontness, spontaneous THOUGHT appears to be produced with relative consistency in the 2016 sample as well. As Figure 133 indicates, the majority of 2016 speakers produce THOUGHT in spontaneous speech with a mean F2 between 1200 and 1300 Hz, regardless of *gender* and *age*. However, the speaker's *level of education* does appear to determine the frontness of spontaneous THOUGHT to a certain extent. Overall, it seems that college educated speakers produce THOUGHT further back than speakers without a college degree. The regression model in Table 79 predicts this difference to be statistically significant, however, the estimated difference is very small, especially considering that it refers to the front-back dimension.

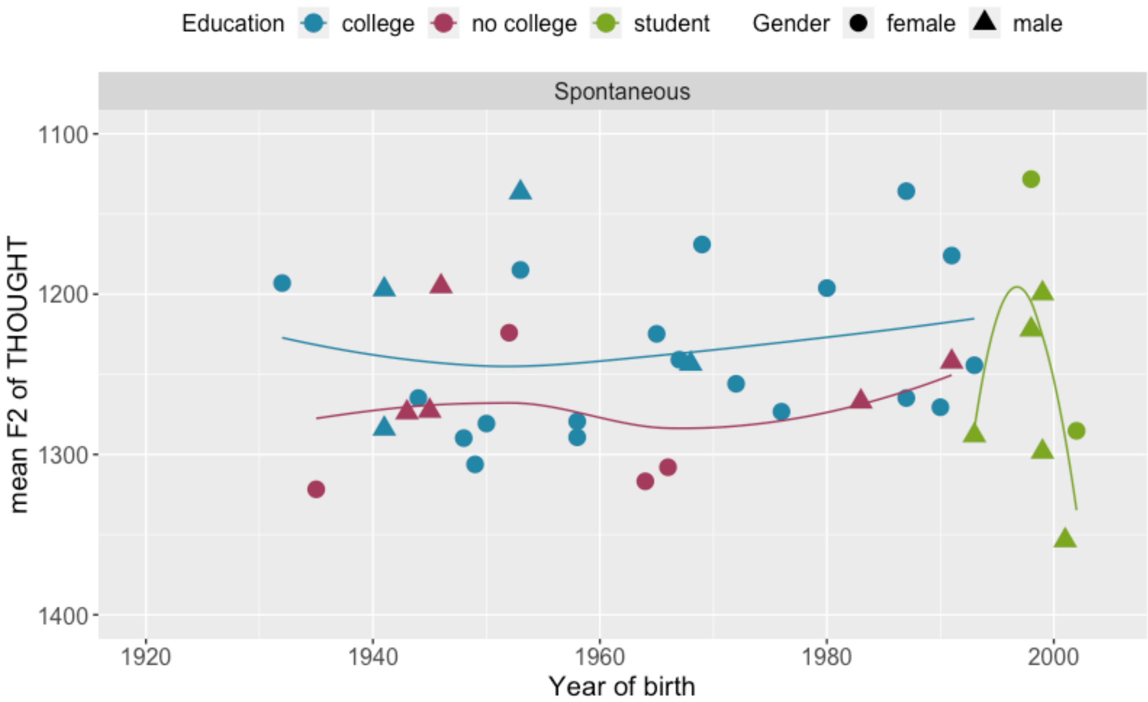


Figure 133: THOUGHT F2 means in spontaneous speech in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	1231.016 Hz	
Age	0.413 Hz	0.293
Gender (Male)	-3.153 Hz	0.847
Education (No college)	35.033 Hz	0.039
Environment		$2x10^{-16}$

Table 79: THOUGHT F2 in spontaneous speech in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word*. *n*= 4471

In real time, few differences emerge from the 2008 and 2016 data for both the height and frontness of THOUGHT in spontaneous speech.

The majority of speakers from both samples produce spontaneous THOUGHT within an F1 range between 700 and 800 Hz, the same range with which they produce LOT in this speech style. Although Figure 134 suggests that 2008 speakers might be producing a slightly higher THOUGHT than 2016 speakers, this difference is relatively small, and does not quite reach the level of statistical significance (Table 80 below). Furthermore, Figure 134 suggests a potential apparent-time lowering of spontaneous THOUGHT in apparent time in the 2008 sample, while no such trend can be detected for 2016 speakers.

The regression model presented in Table 80 below support this observation. It predicts no apparent-time changes in the 2016 data, but an increase in F1 of about 12 Hz per 10 years for 2008 speakers. The differences in apparent-time developments in the two samples are statistically significant, the apparent-time lowering in 2008 itself, however, is not (p= 0.216). The regression model in Table 80 also predicts a significant effect of *gender* on the height of THOUGHT, however, the predicted difference is so small that it can be dismissed.

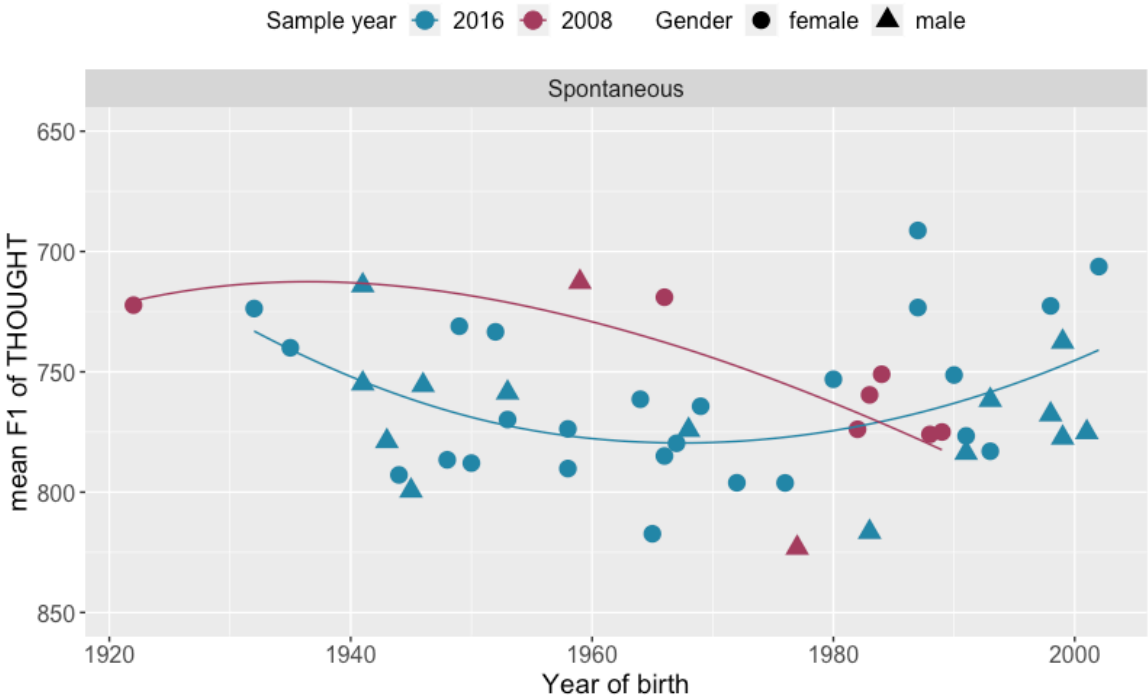


Figure 134: THOUGHT F1 means in spontaneous speech in 2008 and 2016 by *gender*.

The frontness of THOUGHT does not differ to a great extent between the 2008 and 2016 sample in spontaneous speech, with the potential exception of speakers without a college degree. As Figure 135 shows, the majority of speakers from both samples produce spontaneous THOUGHT within the same F2 range, from about 1200 to 1300 Hz. However, two of the 2008 speakers clearly stand out in Figure 135: Two younger speakers without a college degree produce spontaneous THOUGHT notably fronter than any other speakers, with F2 means of 1465 Hz (male speaker) and 1416 Hz (female speaker).

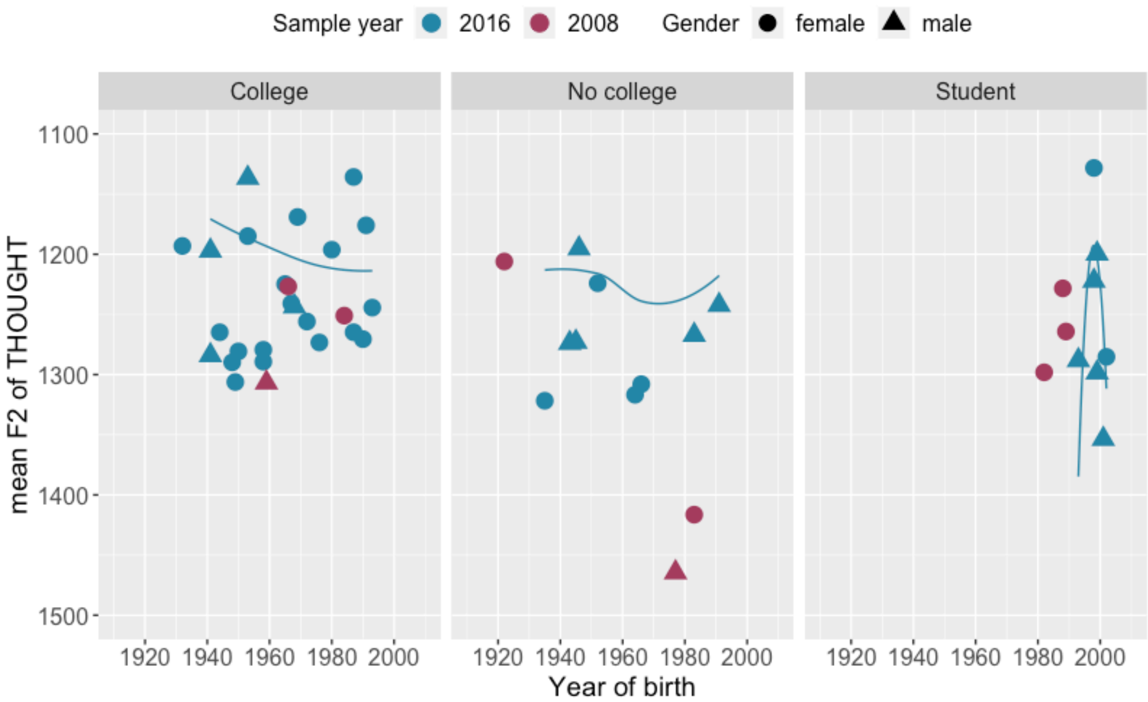


Figure 135: THOUGHT F2 means in spontaneous speech in 2008 and 2016 across *educational levels* by *gender*.

Presumably because of these two outliers, the regression model in Table 81 below predicts significant inter-set differences for the frontness of THOUGHT in spontaneous speech, as well as significant differences in apparent-time trends in the two sets. Speakers in the 2008 sample are estimated a have notably fronter THOUGHT in this speech style, and to be significantly fronting it in apparent time (p= 7x10^{-5}), a trend that is absent in the 2016 data and that is.

Predictor	Coefficient	*p*
(Intercept)	762.765 Hz	
Age	0.246 Hz	
Gender *(Male)*	12.412 Hz	0.045
Sample year *(2008)*	43.764 Hz	
*Age*2008*	-1.234 Hz	0.008
Environment		$2x10^{-16}$

Table 80: THOUGHT F1 in spontaneous speech in 2008 and 2016. Reference levels: females, 2016, /p/. Random effects: *speaker, word. n*= 5137

Predictor	Coefficient	*p*
(Intercept)	1230.496 Hz	
Age	0.384 Hz	
Gender *(Male)*	2.584 Hz	0.866
Education *(No college)*	36.865 Hz	0.02
Sample year *(2008)*	227.479 Hz	
*Age*2008*	-3.336 Hz	0.003
Environment		$2x10^{-16}$

Table 81: THOUGHT F2 in spontaneous speech in 2008 and 2016. Reference levels: females, college educated, 2016, /p/. Random effects: *speaker, word. n*= 4547

THOUGHT in Wordlist Style

In terms of height, THOUGHT in wordlist style shows somewhat more inter-speaker variation than in spontaneous speech. In this speech style, THOUGHT is produced with F1 means between 700 and 850 Hz by the vast majority of 2016 speakers. However, as Figure 136 indicates, the height of wordlist THOUGHT appears to be deceasing, i.e. THOUGHT is lowering in apparent time. More precisely, lowering may have occurred between 1930 and 1970. Since then, the height of wordlist THOUGHT has been kept more or less consistent, averaging at about 800 Hz. The regression models in Table 82 below, however, do not predict any significant changes in the height of THOUGHT in spontaneous speech, neither for the entire 2016 sample, nor for speakers born before or after 1970.

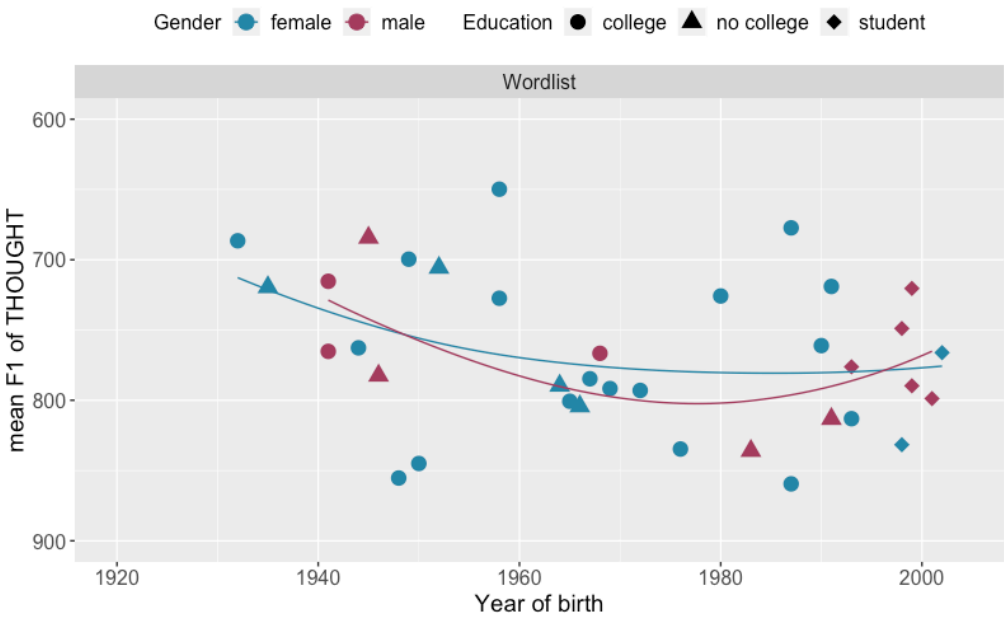


Figure 136: THOUGHT F1 means in wordlist style in 2016 by *gender* and *education*.

Predictor	2016		Yob < 1970		Yob > 1970	
	Coefficient	*p*	Coefficient	*p*	Coefficient	*p*
(Intercept)	843.457 Hz		895.688 Hz		743.794 Hz	
Age	-0.898 Hz	0.112	-1.49 Hz	0.19	1.551 Hz	0.589
Gender *(Male)*	9.217 Hz	0.716	-0.152 Hz	0.996	52.693 Hz	0.29
Education *(No college)*	7.257 Hz	0.764	-3.951 Hz	0.881	---	---
Environment		0.012		0.028		$2x10^{-16}$

Table 82: THOUGHT F1 in wordlist style in 2016. Reference levels: females, college educated, /s/. Random effects: *speaker, word*. *n*= 312 (overall 2016)

As Figure 137 shows, THOUGHT in wordlist style appears to be produced with relatively consistent F2 means by the majority of speakers in the 2016 sample, averaging between 1100 and 1300 Hz. However, the frontness of THOUGHT seems to be subject to some social variation depending on the *educational level* of the speakers. Speakers with a college degree seem to produce a slightly fronter THOUGHT than speakers without a degree, and four of the five students in the 2016 sample seem to produce THOUGHT fronter than most other speakers in this speech style, as can be seen in Figure 137. The regression model presented in Table 83 below finds these differences between the three educational groups to be statistically significant.

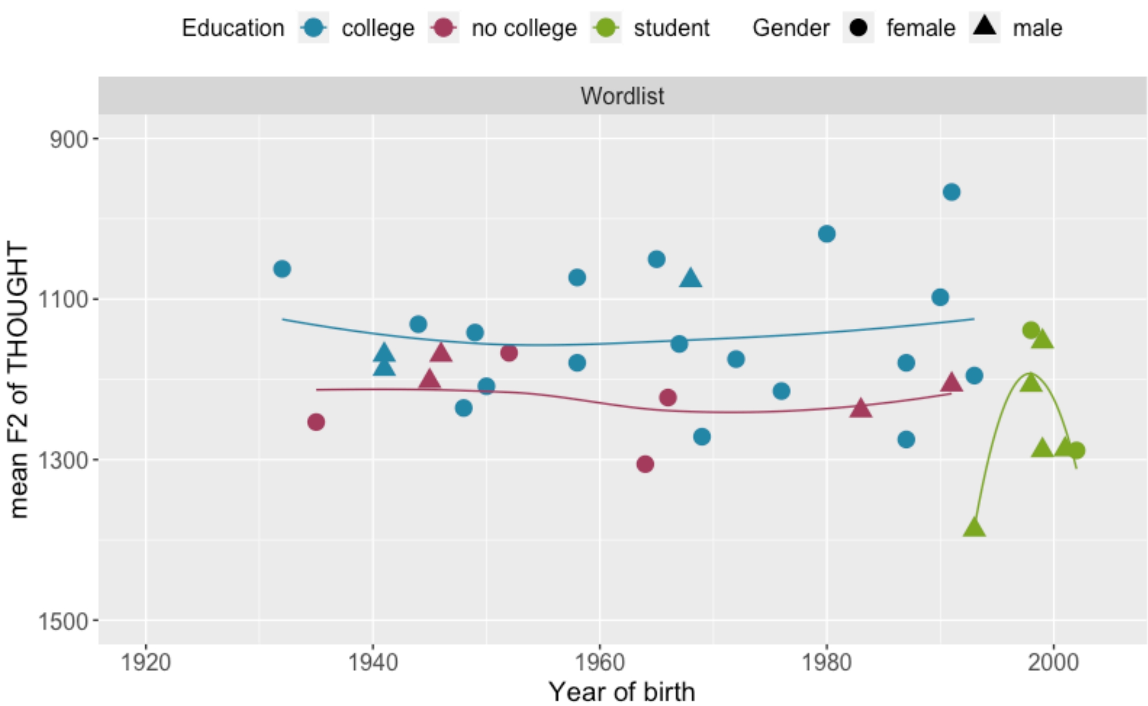


Figure 137: THOUGHT F2 means in wordlist style in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	1136.834 Hz	
Age	0.381 Hz	0.626
Gender (Male)	0.139 Hz	0.996
Education *(No college)* *(Student)*	 72.162 Hz 128.27 Hz	0.014
Environment		0.034

Table 83: THOUGHT F2 in wordlist style in 2016. Reference levels: females, college educated, /s/. Random effects: *speaker, word*. *n*= 312

In wordlist style, differences between speakers from the 2008 and 2016 samples can only be observed on the front-back dimension.

As Figure 139 shows, speakers from both data sets produce wordlist THOUGHT with similar F1 means, i.e. at a similar height, and the regression in Table 84 below does not predict any significant effects for any of the social factors included in the model.

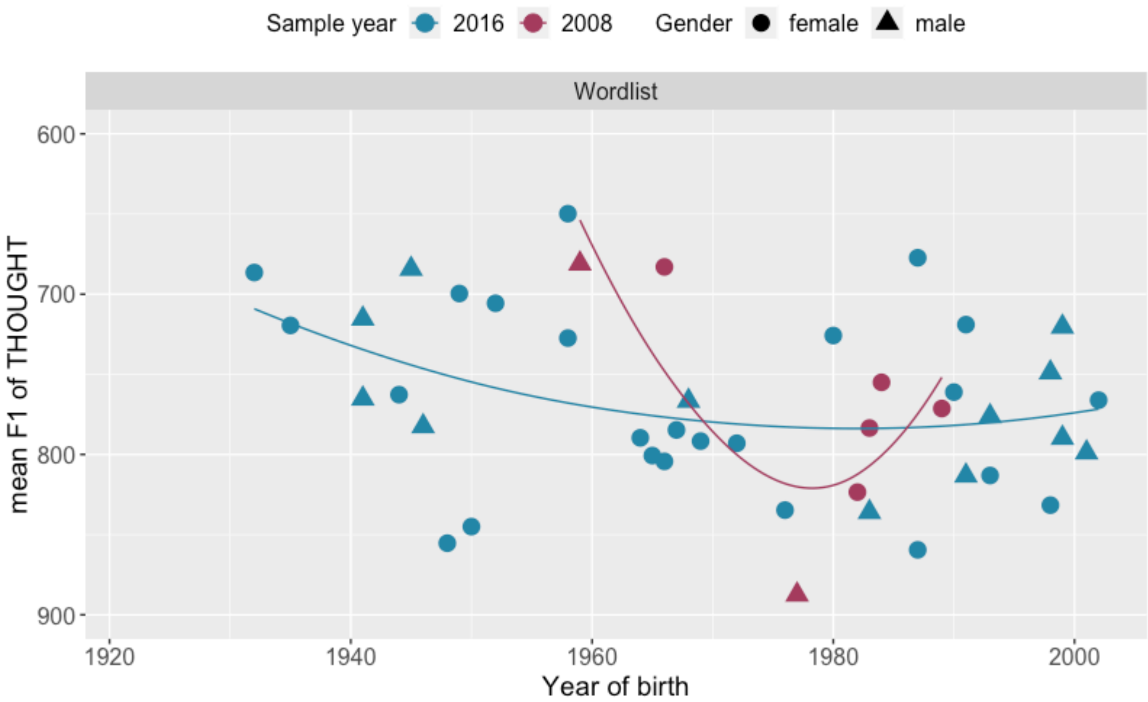


Figure 138: THOUGHT F1 means in wordlist style in 2008 and 2016 by *gender*.

In terms of frontness, 2008 speakers appear to produce a considerably fronter wordlist THOUGHT than speakers interviewed in 2016. As Figure 139 below illustrates, this difference is particularly pronounced among speakers with a college degree. In this group, 2008 speakers have an average F2 of 1300 Hz, while most 2016 speakers produce it with

F2 means of 1200 Hz or less. The regression model presented in Table 85 predicts that *education* does have a significant effect on the frontness of wordlist THOUGHT in the combined 2008 and 2016 sample. However, this effect seems to be based primarily on intra-set differences in the 2016 data, as Figure 139 suggests no differences between speakers with different *education backgrounds* in 2008. Regression models that consider each educational group separately support the observation that real-time differences exist between college educated speakers (142 Hz, *p*= 0.033), but not speakers without a college degree (4 Hz, *p*= 0.956) or students (13 Hz, *p*= 0.917).

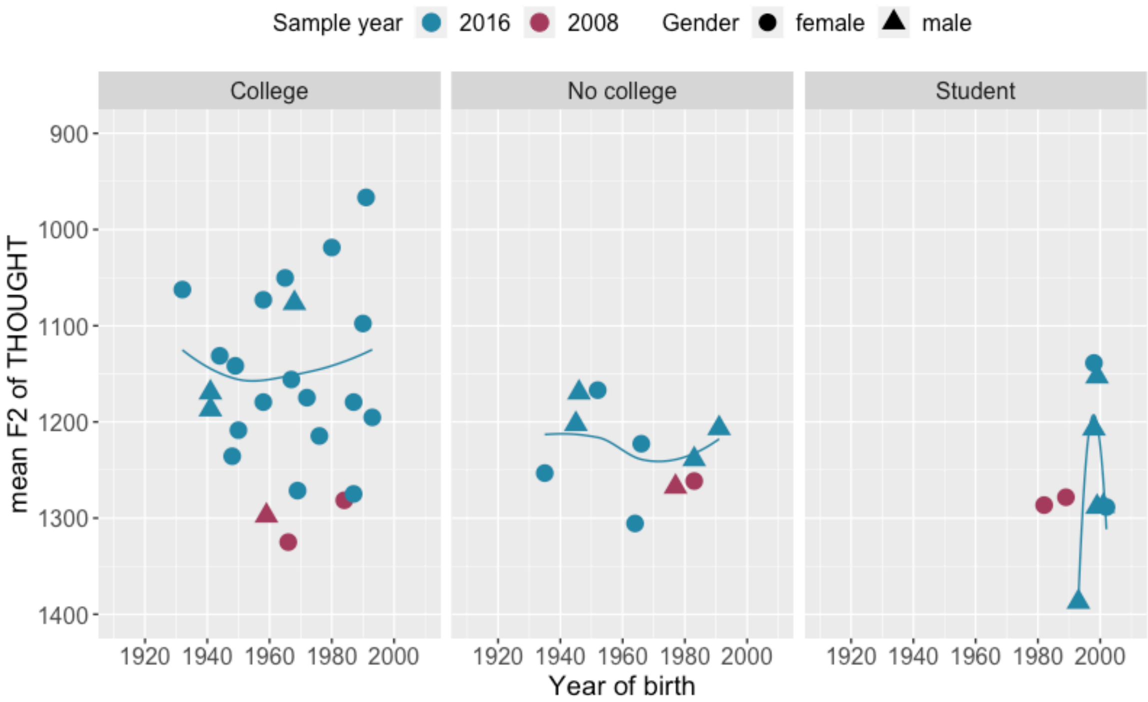


Figure 139: THOUGHT F2 means in wordlist style in 2008 and 2016 across *educational levels* by *gender*.

Predictor	Coefficient	*p*
(Intercept)	831.719 Hz	
Age	-0.721 Hz	0.061
Gender (Male)	1.697 Hz	0.918
Sample year (2008)	12.207 Hz	0.675
Environment		0.002

Table 84: THOUGHT F1 in wordlist style in 2008 and 2016. Reference levels: females, 2016, /s/. Random effects: *speaker, word*. *n*= 408

Predictor	Coefficient	*p*
(Intercept)	1134.981 Hz	
Age	0.457 Hz	0.53
Gender (Male)	-15.766 Hz	0.617
Education (No college)	63.953 Hz	0.039
Sample year (2008)	95.932 Hz	0.203
Environment		0.137

Table 85: THOUGHT F2 in wordlist style in 2008 and 2016. Reference levels: females, college educated, 2016, /s/. Random effects: *speaker, word*. *n*= 327

THOUGHT in Minimal Pairs

In minimal pairs, THOUGHT is produced with F1 means between 700 and 850 Hz by the majority of 2016 speakers, as can be seen in Figure 140. The plot suggests that minimal pair THOUGHT is undergoing a slight lowering in apparent time; however, according to the estimates of the regression model presented in Table 86, this is a relatively minor change that does not reach the level of statistical significance.

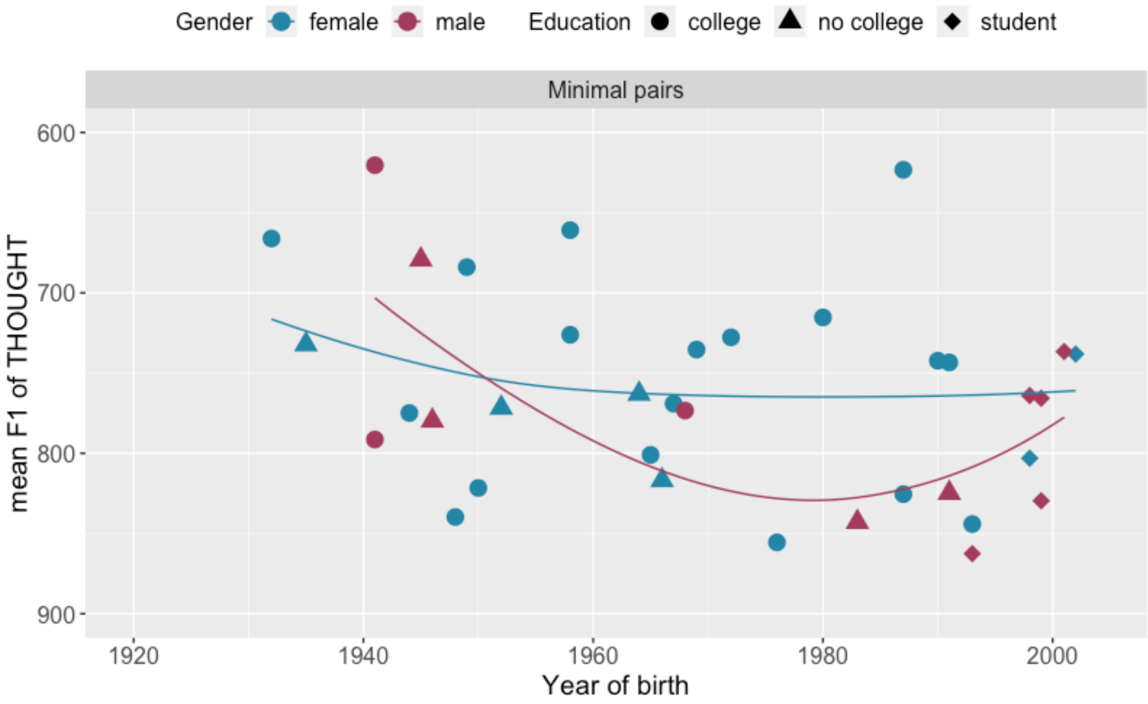


Figure 140: THOUGHT F1 means in minimal pairs in 2016 by *gender* and *education*.

Predictor	Coefficient	*p*
(Intercept)	763.616 Hz	
Age	-1.107 Hz	0.092
Gender *(Male)*	0.986 Hz	0.973
Education *(No college)*	34.571 Hz	0.224
Environment		9x10^{-5}

Table 86: THOUGHT F1 in minimal pairs in 2016.
Reference levels: females, college educated, /n/.
Random effects: *speaker, word. n*= 565

The frontness of THOUGHT shows some inter-speaker variation in the 2016 sample in minimal pairs. As Figure 141 below shows, the F2 range with which THOUGHT in minimal pairs is produced is relatively wide, starting as far back as less than 1000 Hz, and reaching as far front as almost 1400 Hz. Figure 141 suggest that college educated speakers in the sample orient more commonly toward the lower end of this range, i.e. producing THOUGHT further back in the vowel space. Speakers without a college degree, on the other hand, lean somewhat more toward producing a fronter THOUGHT. The regression model in Table 87 supports this observation and predicts a significant F2 difference between the two educational groups in the 2016 sample. As Figure 141 shows, the majority of students' F2 means are similar to those of speakers without a college degree, i.e. they produce minimal pair THOUGHT in a somewhat fronted position. Two of the students produce THOUGHT fronter than any other speakers in the 2016 sample.

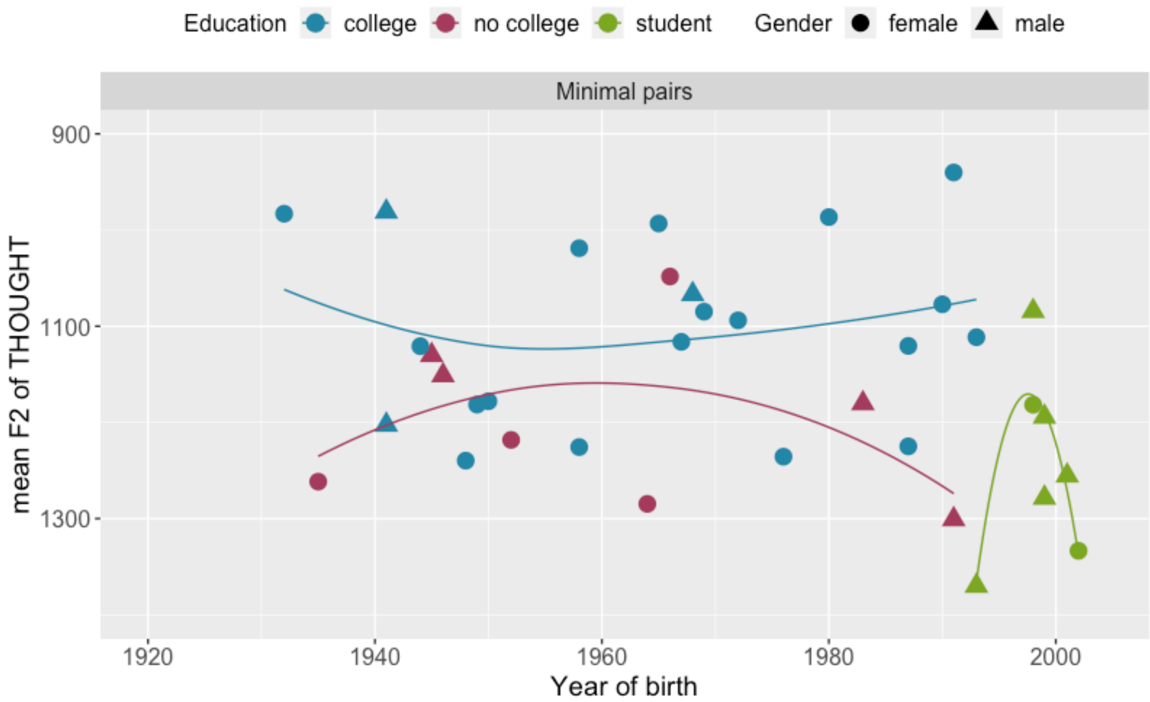


Figure 141: THOUGHT F2 means in minimal pairs in 2016 by *education* and *gender*.

Predictor	Coefficient	*p*
(Intercept)	1011.755 Hz	
Age	0.245 Hz	0.794
Gender *(Male)*	-22.194 Hz	0.608
Education *(No college)*	102.697 Hz	0.019
Environment		0.0005

Table 87: THOUGHT F2 in minimal pairs in 2016.
Reference levels: females, college educated, /n/.
Random effects: *speaker, word*. *n*= 565

Real-time differences between speakers from the 2008 and 2016 samples in the production of THOUGHT in minimal pairs can only be observed in the frontness of the vowel.

No significant height differences can be observed in the production of minimal pair THOUGHT between 2008 and 2016 speakers. As can be seen in Figure 142, F1 means for minimal pair THOUGHT do not seem to differ notably for speakers in the two samples. Apparent-time trends toward lower THOUGHT are also very similar in both data sets. The regression model in Table 88 below finds that apparent-time lowering for the combined data is statistically significant and estimates this change to progress by about 10 Hz per 10 years.

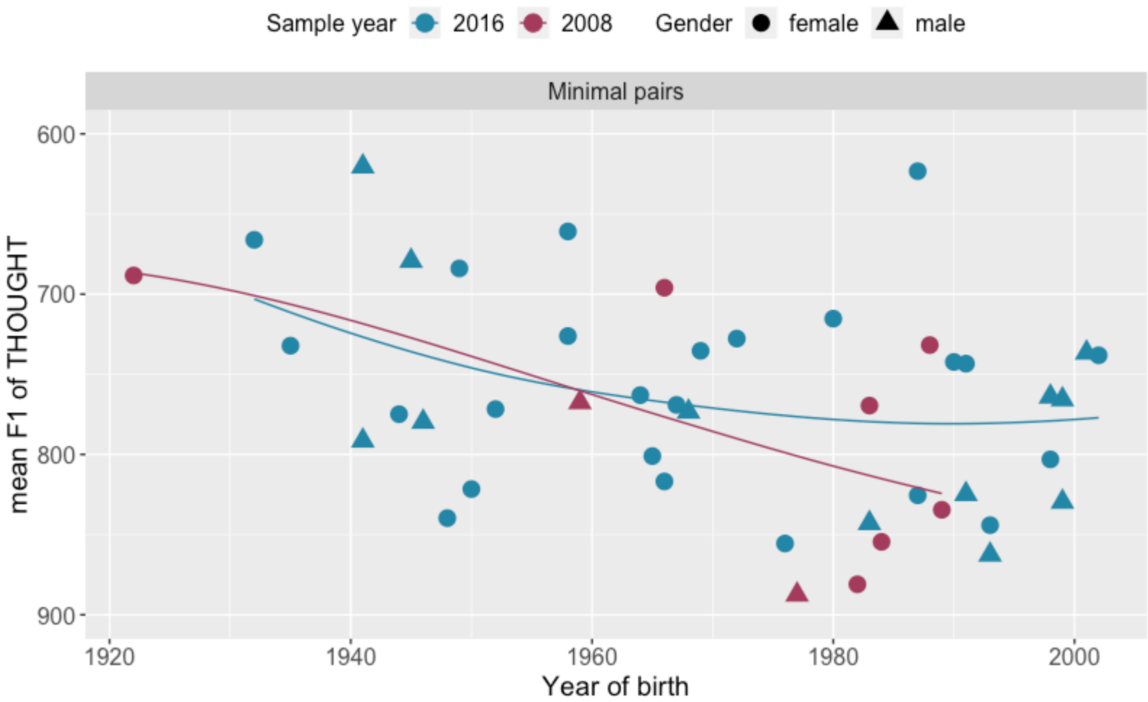


Figure 142: THOUGHT F1 means in minimal pairs in 2008 and 2016 by *gender*.

In terms of frontness, 2008 speakers appear to produce a considerably fronter minimal pair THOUGHT than speakers interviewed in 2016, as can be seen in Figure 143 below. The regression model in Table 89 below predicts that this difference is statistically significant. There are, however, no indications of any apparent-time trends that have led to this real-time difference.

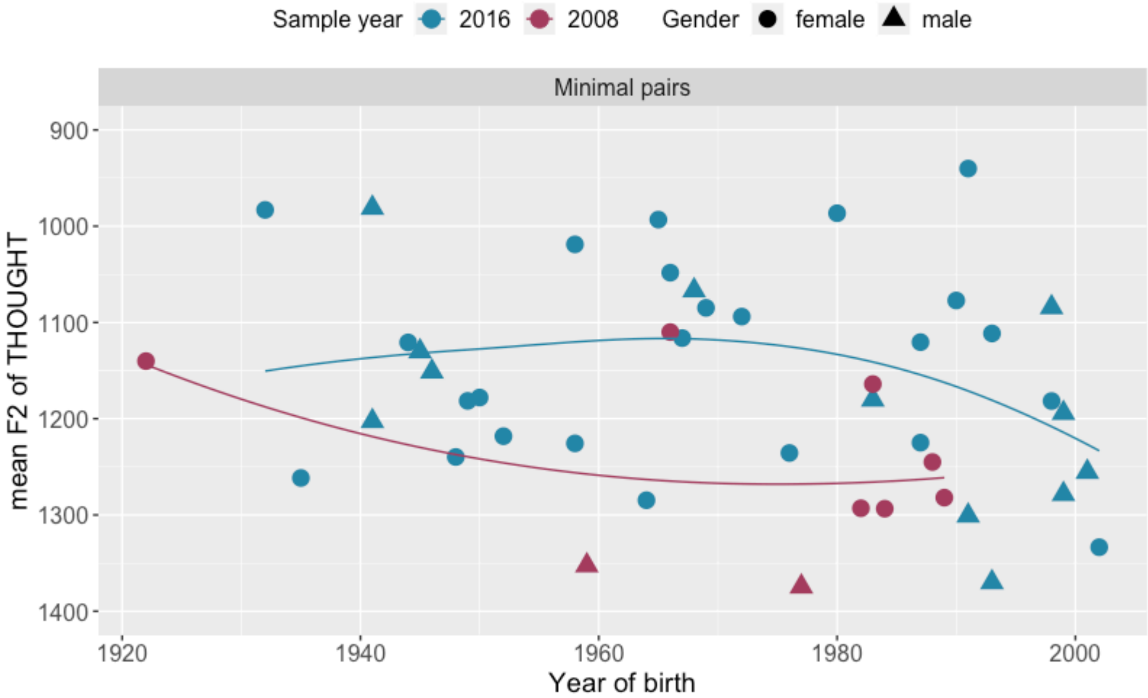


Figure 143: THOUGHT F2 means in minimal pairs in 2008 and 2016 by *gender*.

Predictor	Coefficient	*p*
(Intercept)	765.451 Hz	
Age	-0.963 Hz	0.026
Gender (*Male*)	13.652 Hz	0.488
Sample year (*2008*)	41.473 Hz	0.129
Environment		$2x10^{-5}$

Table 88: THOUGHT F1 in minimal pairs in 2008 and 2016. Reference levels: females, 2016, /n/. Random effects: *speaker, word*. *n*= 727

Predictor	Coefficient	*p*
(Intercept)	1118.913 Hz	
Age	-1.022 Hz	0.15
Gender (*Male*)	56.787 Hz	0.088
Sample year (*2008*)	129.559 Hz	0.004
Environment		0.0003

Table 89: THOUGHT F2 in minimal pairs in 2008 and 2016. Reference levels: females, 2016, /n/. Random effects: *speaker, word*. *n*= 727

Style shifting THOUGHT

The range and direction of style shifting THOUGHT in the 2016 sample appears to depend on the speakers' *age*. As shown in Figures 144 and 145 below, the majority of older speakers in the 2016 sample shift toward a slightly higher and backer THOUGHT in careful speech, while younger speakers in this sample realize a lower THOUGHT in more careful speech. These generational differences indicate an apparent-time change in progress affecting the direction of style shifting: The community is changing toward lower THOUGHT in more careful speech. The main effect of *style* in the regression model in Table 90 below confirms that younger speakers have THOUGHT lower in wordlist style and minimal pairs than in spontaneous speech, while the interaction between *age* and *style* indicates that older speakers have THOUGHT higher in the more careful styles. The effect of *style* switches from positive to negative between 1960 and 1970 in apparent time, as shown in Figure 144. These differences in intra-speaker variation between older and younger speakers in 2016 are due to differently paced apparent-time changes in each of the three styles. On the height dimension, THOUGHT is lowering significantly faster in wordlist style and minimal pairs than in spontaneous speech (Table 90 below).

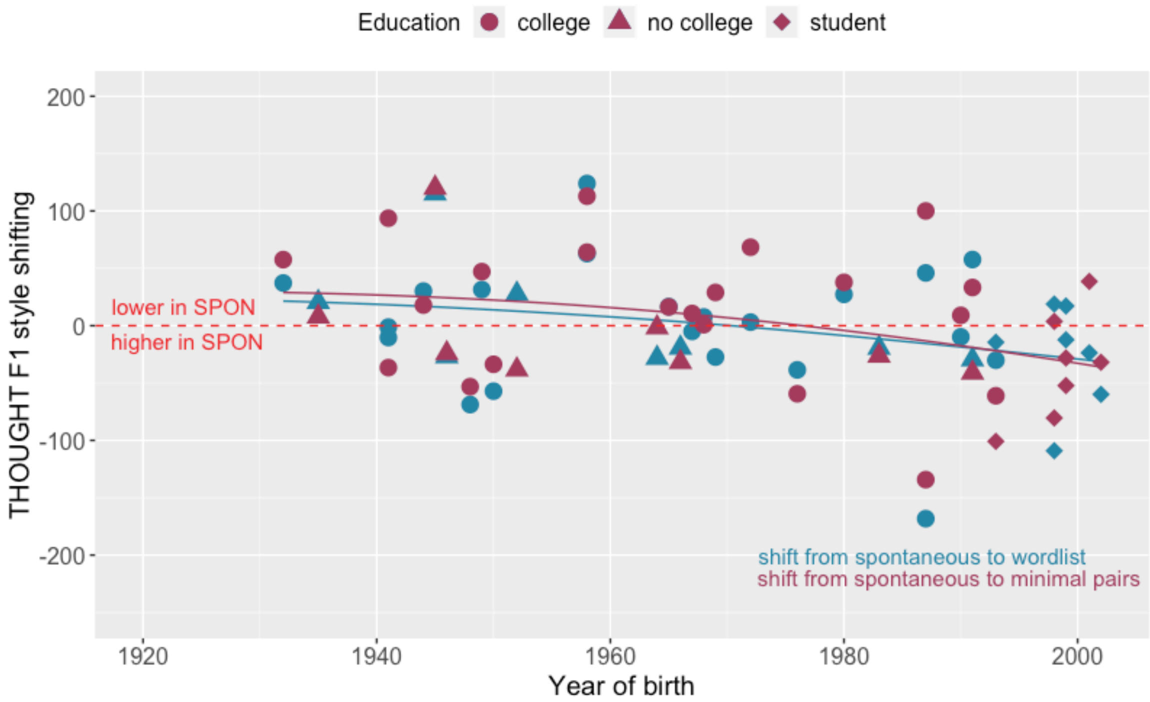


Figure 144: F1 difference between spontaneous, wordlist and minimal pair THOUGHT in 2016 by *education*. A positive value indicates that the vowel is lower spontaneous speech. Blue represents the shift from spontaneous speech to wordlist style, wine-red the shift from spontaneous speech to minimal pairs.

Regarding F2, Figure 145 below suggests a difference in the style shifting of the frontness of THOUGHT between speakers of different educational backgrounds. While college educated speakers tend to shift toward a notably more retracted THOUGHT in wordlist style, speakers without a college degree and students shift to a less notable extent, or, in fact, in the opposite direction. The main effect of *style* in combination with the interaction between *education* and *style* in the regression model presented in Table 91 below confirms the significance of these differences in style shifting. Because students are simultaneously the youngest speakers in the sample, their minimized range of style shifting suggests an apparent-time change in progress regarding the direction of style shifting in Figure 145. However, the regression model in Table 91 predicts no significant effect of *age*, nor does it find a significant interaction between *age* and *style*. Thus, it seems that, on the front-back dimension, minimized style differences among younger speakers are not the result of gradual apparent-time change, but rather recent developments that are more or less exclusive to the students in the 2016 data.

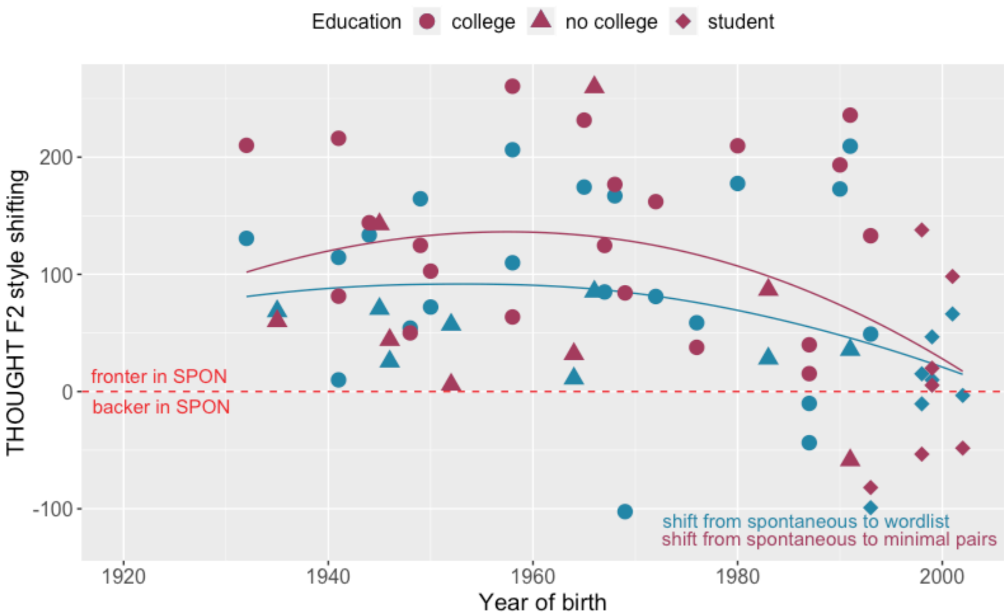


Figure 145: F2 difference between spontaneous, wordlist and minimal pair THOUGHT in 2016 by *education*. A positive value indicates that the vowel is fronter in spontaneous speech. Blue represents the shift from spontaneous speech to wordlist style, wine-red the shift from spontaneous speech to minimal pairs.

Predictor	Coefficient	*p*
(Intercept)	778.233 Hz	
Age	-0.035 Hz	
Gender (Male)	16.023 Hz	0.067
Education *(No college)* *(Students)*	 11.565 Hz -23.189 Hz	0.069
Style *(Wordlist)* *(Minimal pairs)*	 43.05 Hz 42.662 Hz	
*Age*style* *(Wordlist)* *(Minimal pairs)*	 -0.907 Hz -1.113 Hz	10^{-15}
Environment		$2x10^{-16}$

Table 90: Effect of style on F1 of THOUGHT in 2016. Reference levels: females, college educated, spontaneous speech, /p/. Random effects: *speaker, word*. *n*= 6111

Predictor	Coefficient	*p*
(Intercept)	1233.953 Hz	
Age	0.365 Hz	0.453
Gender (Male)	-0.503 Hz	0.978
Education *(No college)* *(Students)*	 33.753 Hz 77.692 Hz	
Style *(Wordlist)* *(Minimal pairs)*	 -47.402 Hz -95.694 Hz	
*Style*education* *No college*WL* *No college*MP* *Students*WL* *Students*MP*	 38.03 Hz 53.513 Hz 54.377 Hz 83.115 Hz	$5x10^{-8}$
Environment		$2x10^{-16}$

Table 91: Effect of style on F2 of THOUGHT in 2016. Reference levels: females, college educated, spontaneous speech, /p/. Random effects: *speaker, word*. *n*= 6111

Speakers in the 2008 sample appear to follow similar style-shifting patterns for THOUGHT as 2016 speakers in terms of frontness but differ notably in their style differences on the height dimension.

In terms of height, 2008 speakers differ from 2016 speakers in their shift from spontaneous speech to minimal pairs. As can be seen in Figure 146, most of the 2008 speakers realize a lower THOUGHT in minimal pairs than they do in wordlist style, while in 2016 this pattern could only be observed for speakers born after 1970. Furthermore, the range of shifting is greater for 2008 speakers, i.e. their minimal pair THOUGHT is notably lower than it is for those 2016 speakers who shift in this direction. This real-time difference contradicts the apparent-time trends observed in the style shifting of the height of THOUGHT in 2016: While, in apparent time, the effect of *style* switches from positive to negative, in real time, it switches from negative to positive or to a lower negative.

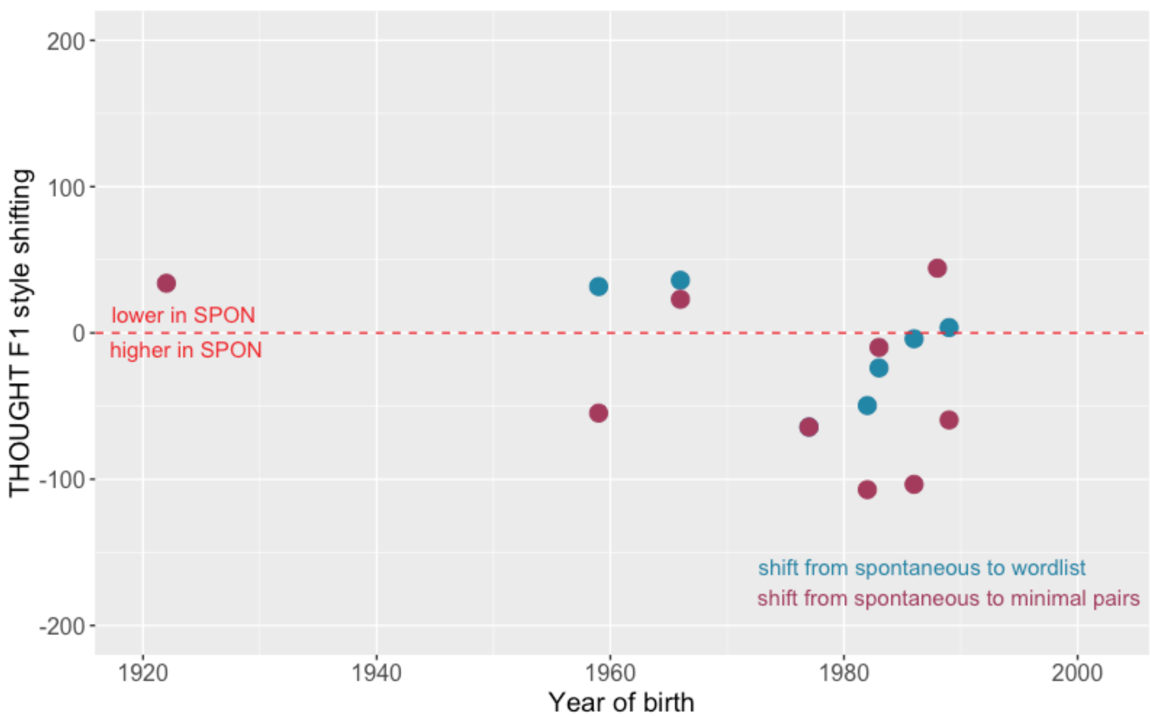


Figure 146: F1 difference between spontaneous, wordlist and minimal pair THOUGHT in 2008. A positive value indicates that the vowel is lower in spontaneous speech. Blue represents the shift from spontaneous speech to wordlist style, wine-red the shift from spontaneous speech to minimal pairs.

Figure 147 below suggests that 2008 speakers produce THOUGHT in minimal pairs and wordlist style the furthest back in the vowel space compared to spontaneous THOUGHT, which was observed for 2016 speakers as well. Two of the youngest speakers in the 2008 sample seem to have minimized this pattern, and two of the older speakers show a reversed effect of style.

Figure 147: F2 difference between spontaneous, wordlist and minimal pair THOUGHT in 2008. A positive value indicates that the vowel is fronter in spontaneous speech. Blue represents the shift from spontaneous speech to wordlist style, wine-red the shift from spontaneous speech to minimal pairs.

Interim Summary: THOUGHT in Production

The analysis of the production of THOUGHT has shown that the vowel is subject to some variation in the community. This variation appears to stem from differences between speakers of different *ages* and *educational levels*, as well as from the effects of *speech style*.

In spontaneous speech, significant inter-speaker variation can be observed between speakers of different *educational levels*. Speakers without a college education produce spontaneous THOUGHT notably fronter than most other speakers. This is true for both 2008 and 2016 speakers, but the difference is much more prominent for two of the three 2008 speakers in this educational group. This appears to be a case of stable variation, and the only significant apparent-time development that can be observed in this speech style is that THOUGHT was lowered slightly by 2016 speakers born before 1970. However, this lowering has since been reversed by younger speakers. Nevertheless, for most speakers, spontaneous THOUGHT is produced the lowest and furthest forward in the vowel space compared to the other two speech styles, only slightly higher, but still a good bit backer than spontaneous LOT.

Variation in wordlist THOUGHT is limited to speakers of different *educational levels* and to a notable real-time difference within the group of college educated speakers. College educated speakers in the 2016 sample produce wordlist THOUGHT notably backer than most other speakers, especially speakers with the same level of education in the 2008 sample, leading to a significant real-time difference in the frontness of wordlist THOUGHT for speakers in this educational group. Overall, the majority of speakers produce wordlist THOUGHT notably backer and slightly higher than THOUGHT in spontaneous speech. For younger speakers, however, this pattern reverses on the height dimension, and minimizes on the front-back dimension, as students in the 2016 sample appear to produce a somewhat lowered and fronted THOUGHT in this speech style. THOUGHT in wordlist style is produced only slightly higher, but somewhat backer in the vowel space than LOT in this speech style.

In minimal pairs, THOUGHT appears to be lowering slightly in apparent time, and backing in real time, which can be seen in Figure 148 below. However, there is no real-time evidence for THOUGHT lowering, and no apparent-time evidence for backing. Instead, notably fronted THOUGHT can be observed for students in the 2016 sample, suggesting that THOUGHT might be undergoing fronting rather than backing in apparent time. This prediction is strengthened by the decreasing style differences in the 2016 sample on the front-back dimension, as the youngest speakers produce THOUGHT in minimal pairs with very similar F2 values as they do in spontaneous speech, while for most older speakers, THOUGHT is much backer in minimal pair production. Furthermore, the youngest speakers have reversed the style differences on the height dimension, producing minimal pair THOUGHT in a slightly lower position than spontaneous THOUGHT. For the majority of speakers, THOUGHT in minimal pairs is slightly higher in the vowel space than LOT in this speech style, though there is some overlap on the height dimension for some speakers. On the front-back dimension, on the other hand, there is still a good distance between the two vowels for most speakers.

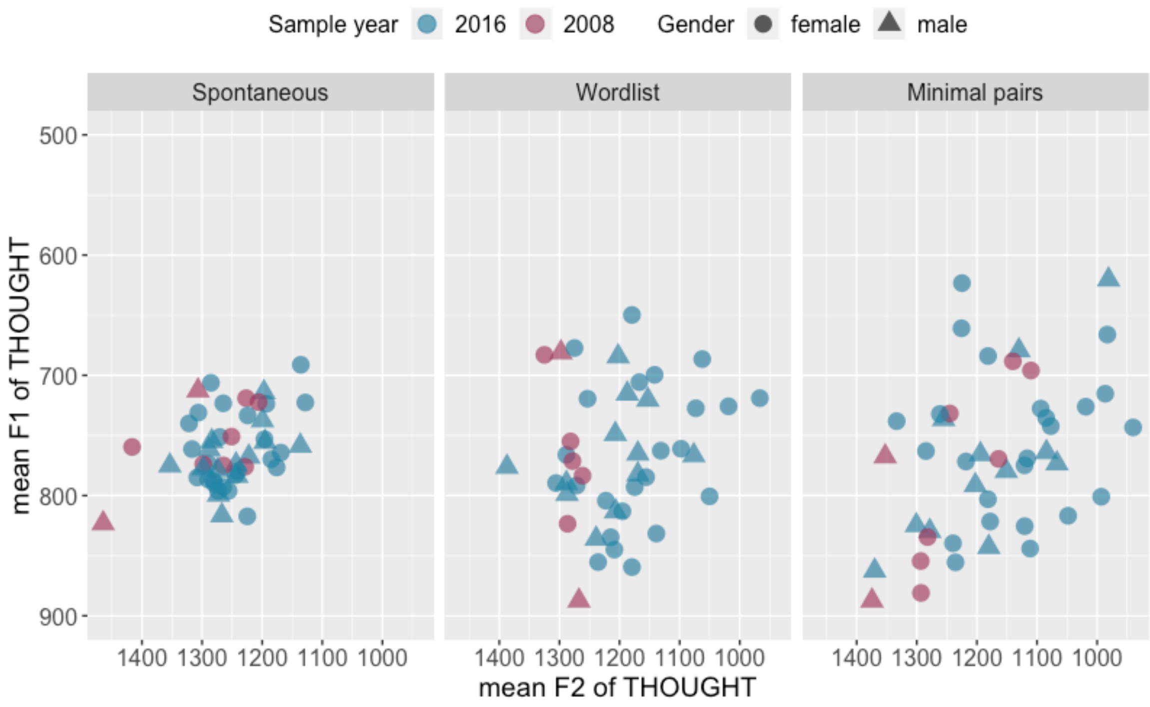


Figure 148: THOUGHT F1 and F2 means in 2008 and 2016 across *speech styles* by *gender*.

The results of the analysis of THOUGHT presented in this chapter are similar to those presented for LOT, in that there is very little evidence of speakers in Ogdensburg having participated in the lowering and fronting of THOUGHT that would have been expected around the time the NCS developed in the Inland North. Only in spontaneous speech is there slight evidence for some THOUGHT lowering among older speakers in the data, and some real-time evidence suggests that some 2008 speakers may have had fronter THOUGHT than 2016 speakers in spontaneous speech and minimal pair production.

On the other hand, the results presented here indicate slight progress toward a merger with LOT. THOUGHT is lowering and fronting, thereby moving closer to LOT. These developments are particularly noticeable in the two more careful styles, with students producing a notably fronter THOUGHT than most other speakers. However, these developments do not appear to be as robust as those observed for LOT, so that it seems that the majority of progress toward the merger by mechanism of approximation is accounted for by the backing and raising of LOT.

6.2.3.2 Merger by Expansion

The results regarding the overlap of LOT and THOUGHT in Chapter 6.2.2 suggested increasing overlap of the two vowels in apparent time. This increase is likely the result of the phonetic approximation of LOT toward THOUGHT and vice versa, which was found to be progressing in apparent time as well in Chapter 6.2.3.1 above. However, it is also possible that the overlap of LOT and THOUGHT has increased as a result of each of the two vowels expanding their phonetic range, thereby starting to occupy parts of the phonetic space that has, traditionally, been occupied by the other phoneme. In this case, expansion might be an additional mechanism that is advancing the COT-CAUGHT merger in Ogdensburg.

To test whether or not a merger by expansion is at play in the community, standard deviations will be analyzed for F1 and F2 of both LOT and THOUGHT. Standard deviations measure the dispersion of data points relative to their mean value. Thus, the higher the standard deviation for F1 and F2 for a particular vowel, the more phonetic space it takes up. Because merger by expansion has been found to be an abrupt rather than a continuous process, gradual apparent-time progress toward higher standard deviations for either of the two vowels is not expected. There may, however, be real-time or generational differences that could indicate merger by expansion.

Figures 149 and 150 below illustrate the standard deviations for F1 and F2 of LOT and THOUGHT for both the 2008 and 2016 samples in spontaneous speech, wordlist style and minimal pairs. The plots show that, for the most part, both vowels have very similar standard deviations on both dimensions, suggesting that they do not differ notably in terms of their phonetic ranges in the vowel space.

As expected, no apparent-time trends toward greater ranges can be observed for either of the two vowels on either dimension. In apparent time, standard deviations appear to remain steady in both samples. There is also no evidence of generational change; younger speakers do not appear to suddenly produce either vowels with notably higher standard deviations in any speech style than older speakers do. However, as Figures 149 and 150 below suggest, there are a few real-time differences, which will be explored for each speech style separately below.

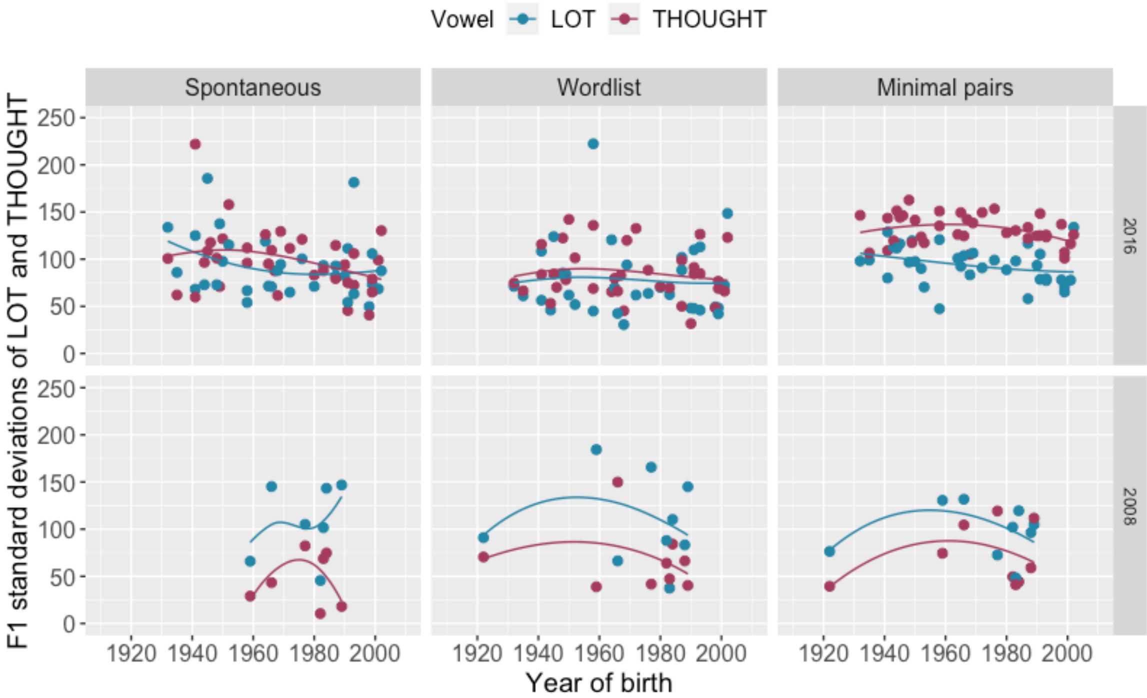


Figure 149: Standard deviations for F1 of LOT and THOUGHT in 2008 and 2016 across *speech styles*.

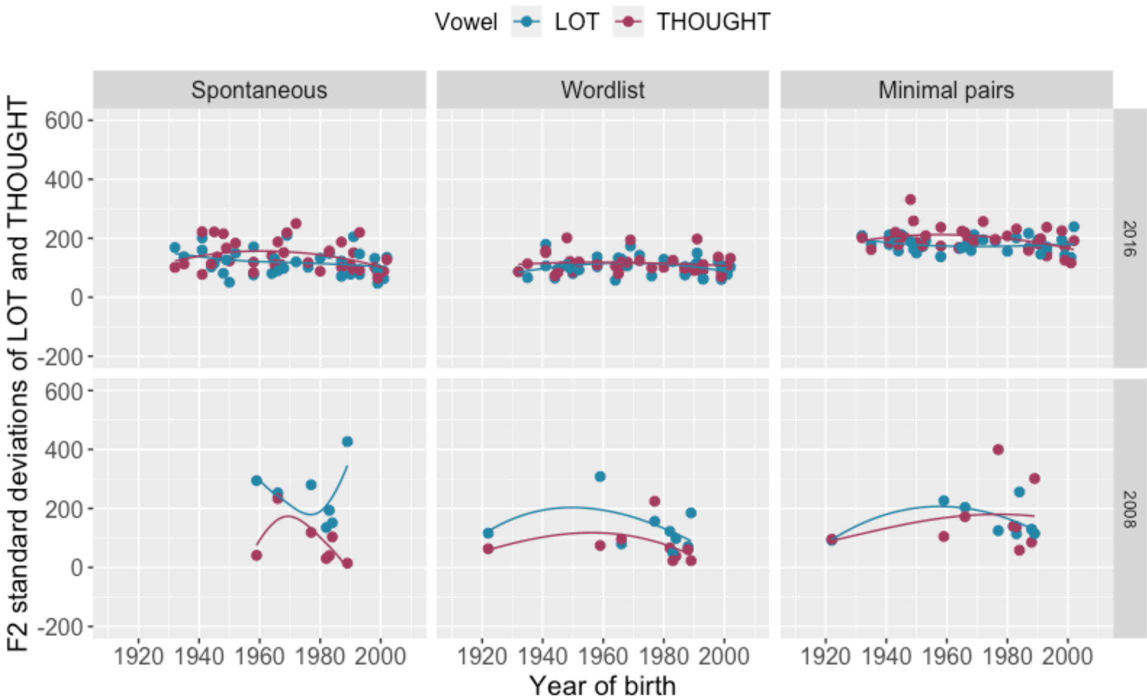


Figure 150: Standard deviations for F2 of LOT and THOUGHT in 2008 and 2016 across *speech styles*.

In spontaneous speech, there is some evidence of an increased phonetic range for THOUGHT. As shown in Figures 149 and 150 above, 2016 speakers appear to employ a wider range of phonetic space in the height and frontness of THOUGHT than 2008 speakers. The regression models presented in Tables 92 and 93 find the difference on both dimensions to be statistically significant.

For spontaneous LOT, on the other hand, *2008 speakers* appear to employ a wider range on both dimensions than 2016 speakers do, i.e. there is, if anything, a real-time decrease in phonetic ranges of the height and frontness of LOT. The regression models in Tables 92 and 93 support these observations, though only for F2 is the predicted real-time difference noteworthy.

Interestingly, students, who tend to be the most merged of all speakers, appear to have smaller overall ranges for LOT and THOUGHT than most other speakers. Although *education* is not found to have a significant effect on F1 or F2 standard deviations, this observation suggests that expansion may not be the driving force behind the merger.

Predictor	Coefficient	*p*
(Intercept)	75.272 Hz	
Age	0.319 Hz	0.205
Gender (Male)	-0.054 Hz	0.995
Education *(No college)* *(Students)*	11.83 Hz -0.952 Hz	0.48
Vowel (THOUGHT)	5.716 Hz	
Sample year (2008)	16.9 Hz	
2008 THOUGHT*	-66.72 Hz	0.0005

Table 92: Standard deviations for F1 of LOT and THOUGHT in 2008 and 2016 in spontaneous speech. Reference levels: females, college educated, LOT, 2016. Random effect: *speaker*. *n*= 82

Predictor	Coefficient	*p*
(Intercept)	115.237 Hz	
Age	0.218 Hz	0.582
Gender (Male)	1.923 Hz	0.9
Education *(No college)* *(Students)*	-5.532 Hz -34.05 Hz	0.302
Vowel (THOUGHT)	19.085 Hz	
Sample year (2008)	134.729 Hz	
2008 THOUGHT*	-183.693 Hz	$5x10^{-7}$

Table 93: Standard deviations for F2 of LOT and THOUGHT in 2008 and 2016 in spontaneous speech. Reference levels: females, college educated, LOT, 2016. Random effect: *speaker*. *n*= 82

In wordlist style, Figures 149 and 150 above suggest a slight real-time increase in the phonetic range of THOUGHT between younger speakers in the 2008 and 2016 samples. Speakers in the 2016 sample born after 1980 realize THOUGHT in wordlist style with somewhat wider ranges than their peers in 2008 both in terms of height and frontness. Although the differences appear to be relatively small in Figures 149 and 150, the regression models in Tables 94 and 95 below find them to be statistically significant on both dimensions for the overall samples. The estimated coefficient for F2, however, is quite small for the front-back dimension.

LOT in wordlist style appears to have occupied a wider range in 2008. Especially on the height dimension does LOT seem to have been produced with a wider by *2008 speakers* compared to 2016 speakers, as can be seen in Figure 149 above. According to the estimate provided by the regression model in Table 94, however, this real-time difference is relatively small. In any case, a wider F1 range in 2008 would not be an indication of a merger by expansion in real time.

Again, students, seem to employ slightly smaller ranges than other speakers, particularly on the front-back dimension. However, this difference does not reach the level of statistical significance.

The regression model in Table 95 also predicts a significant *gender* difference in the F2 standard deviations of LOT and THOUGHT, as males seem to employ a wider range on the front-back dimension. However, the predicted difference is so small that it does not require further analysis.

Predictor	Coefficient	*p*
(Intercept)	69.718 Hz	
Age	0.195 Hz	0.451
Gender (Male)	1.226 Hz	0.906
Education *(No college)* *(Students)*	 -9.3 Hz 0.565 Hz	0.687
Vowel (THOUGHT)	7.688 Hz	
Sample year (2008)	32.326 Hz	
2008 THOUGHT*	-48.611 Hz	0.005

Table 94: Standard deviations for F1 of LOT and THOUGHT in 2008 and 2016 in wordlist style. Reference levels: females, college educated, LOT, 2016. Random effect: *speaker*. *n*= 86

Predictor	Coefficient	*p*
(Intercept)	101.98 Hz	
Age	-0.003 Hz	0.991
Gender (Male)	32.96 Hz	0.006
Education *(No college)* *(Students)*	 -14.36 Hz -33.778 Hz	0.098
Vowel (THOUGHT)	10.306 Hz	
Sample year (2008)	38.696 Hz	
2008 THOUGHT*	-67.848 Hz	0.001

Table 95: Standard deviations for F2 of LOT and THOUGHT in 2008 and 2016 in wordlist style. Reference levels: females, college educated, LOT, 2016. Random effect: *speaker*. *n*= 86

In minimal pairs, real-time differences are particularly noticeable on the height dimension of THOUGHT. Figure 149 above shows that 2016 speakers collectively have much greater standard deviations than 2008 speakers, i.e. employ a wider range in their realizations of THOUGHT on the height dimension. The regression model presented in Table 96 below predicts this difference to be statistically significant. For F2, no noteworthy real-time differences emerge from the data (Table 97 below), however, students are again estimated to have slightly lower standard deviations than other speakers. For LOT, on the other hand, no noteworthy differences emerge from the data.

Predictor	Coefficient	*p*
(Intercept)	91.609 Hz	
Age	0.129 Hz	0.358
Gender (Male)	-2.308 Hz	0.67
Education *(No college)* *(Students)*	 -7.16 Hz -6.302 Hz	0.435
Vowel (THOUGHT)	37.085 Hz	
Sample year (2008)	5.831 Hz	
2008 THOUGHT*	-63.693 Hz	$7x10^{-8}$

Table 96: Standard deviations for F1 of LOT and THOUGHT in 2008 and 2016 in minimal pairs. Reference levels: females, college educated, LOT, 2016. Random effect: *speaker*. *n*= 90

Predictor	Coefficient	*p*
(Intercept)	187.494 Hz	
Age	-0.154 Hz	0.61
Gender (Male)	10.814 Hz	0.361
Education *(No college)* *(Students)*	 -6.21 Hz -31.306 Hz	0.198
Vowel (THOUGHT)	19.737 Hz	
Sample year (2008)	-15.141 Hz	
2008 THOUGHT*	-9.148 Hz	0.724

Table 97: Standard deviations for F2 of LOT and THOUGHT in 2008 and 2016 in minimal pairs. Reference levels: females, college educated, LOT, 2016. Random effect: *speaker*. *n*= 90

The results from the analysis of the standard deviations of LOT and THOUGHT suggest that 2016 speakers utilize a wider phonetic range in their production of THOUGHT than 2008 speakers do in all three speech styles. While wider ranges for LOT in 2008 might, to a certain extent, offset the effect that an increased range for THOUGHT may have on the merger, these real-time differences might be an indication of merger by expansion. However, standard deviations do not tell us anything about the *direction* of expansion. Thus, although 2016 speakers do have greater standard deviations for THOUGHT, they might produce some of their THOUGHT words *higher and backer* rather than lower and fronter. To determine in which direction THOUGHT is expanding, if there *is* any particular direction, its distribution in the vowel space for four of the most merged 2016 speakers can be compared.

Figure 151 below shows that, while THOUGHT in spontaneous speech does seem to take up a good amount of phonetic space for speakers with a relatively high degree of merger, it does not expand in the direction that would be expected for a COT-CAUGHT merger by expansion. For all four 2016 speakers, the range of THOUGHT expands up- and backwards relative to their means, i.e. in the opposite direction of what would be expected for the merger with LOT. The speakers presented here are two young males (top row), one middle-aged female (bottom left), and an older female (bottom right). Thus, this pattern of expansion does not appear to be *age* or *gender*-related.

For minimal pairs, Figure 152 below suggests a more evenly distributed range for THOUGHT. Four of the most merged speakers in minimal pairs, both male and female, appear to shift down- and forward just as much as they shift up- and backward in this speech style.

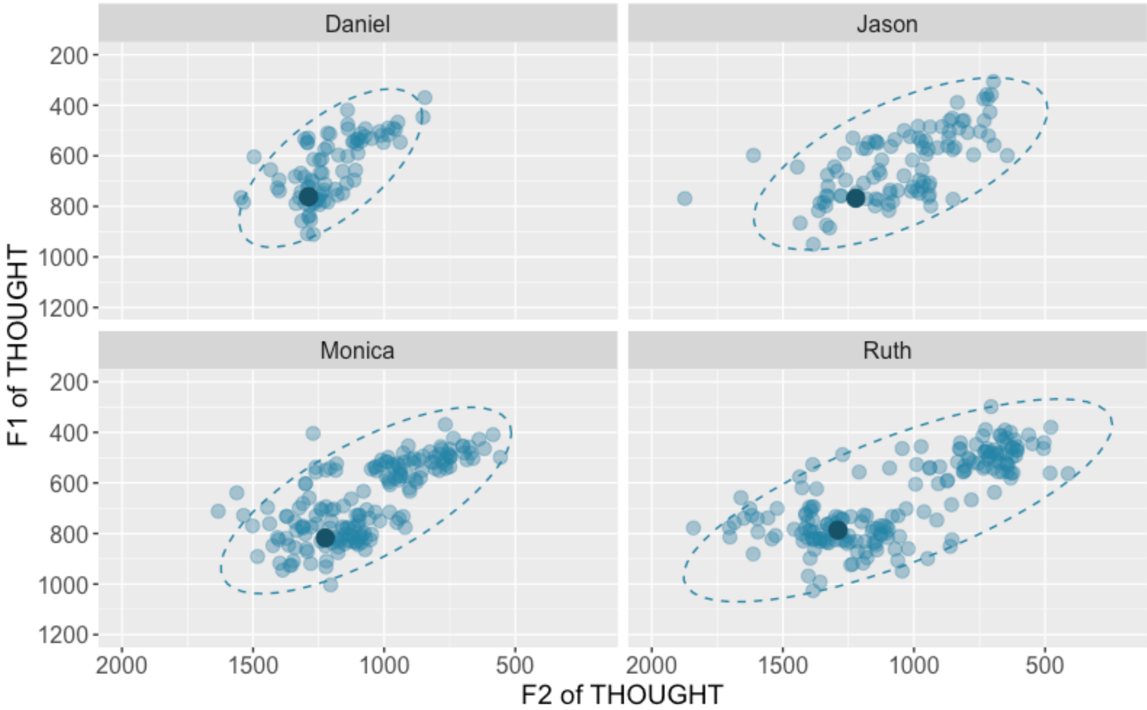


Figure 151: Expansion of THOUGHT in the vowel space in spontaneous speech. The small translucent dots represent the speakers' individual THOUGHT tokens, the bigger and solid dot represents their mean THOUGHT in this speech style.

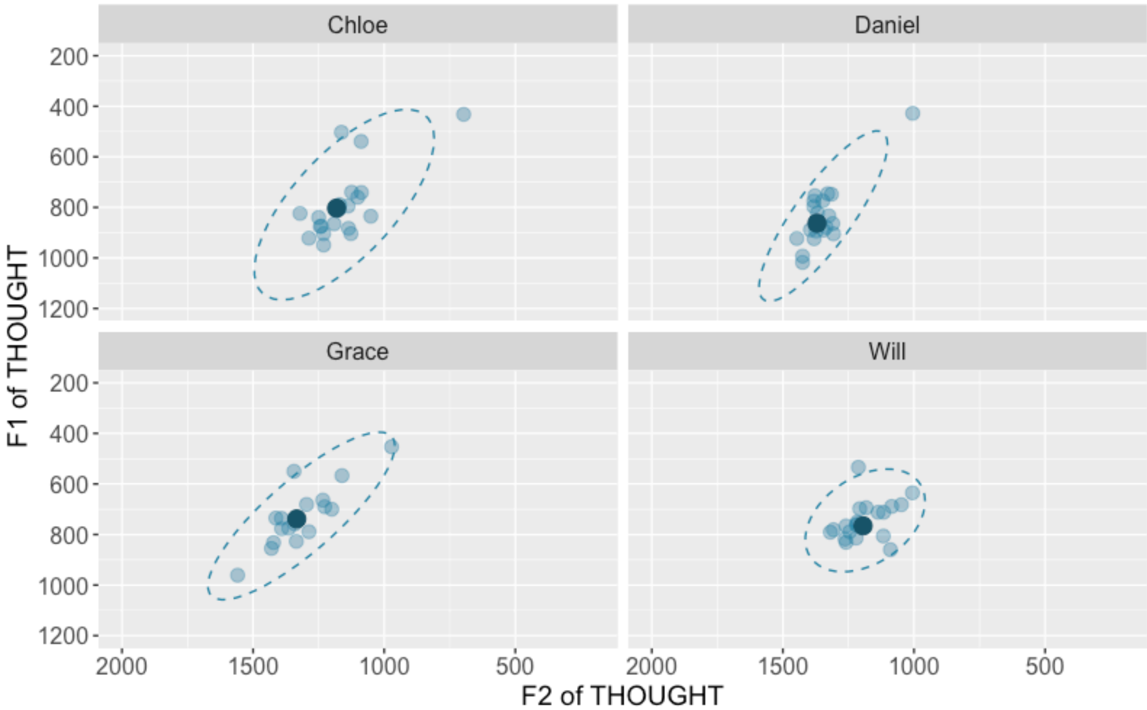


Figure 152: Expansion of THOUGHT in the vowel space in minimal pairs. The small translucent dots represent the speakers' individual THOUGHT tokens, the bigger and solid dot represents their mean THOUGHT in this speech style.

Overall, there is little evidence that the advancement of the COT-CAUGHT merger in Ogdensburg is driven by a merger by expansion. While the increasing overlap of LOT and THOUGHT might have suggested that the two vowels are expanding their phonetic ranges, this increase primarily seems to be the result of gradual approximation. Whether it is also advanced by means of phonological transfer will be explored in the following subchapter.

6.2.3.3 Merger by Phonological Transfer

The increasing overlap of LOT with THOUGHT may also be a result of phonological transfer from one category to the other. Dinkin (2009, 2016) found this to be the case in Northern New York in his study, including Ogdensburg. He found that, when preceding a cluster of /lf/ or/lv/, LOT was commonly transferred to THOUGHT. To test whether this is also the case in the 2016 sample, we can examine the distribution of /ɑlC/[73] words in the vowel space. This analysis is based on a selection of four speakers. Since /lC/ environments do not occur in any of the minimal pairs, this speech style will be excluded from the analysis.

6.2.3.3.1 Spontaneous Speech

Figures 153 to 155 below present a selection of LOT and THOUGHT tokens produced by four speakers, three from the 2016, one from the 2008 sample.[74]

For all four speakers, instances of LOT in an /lC/ cluster are among those LOT tokens that are produced the highest and furthest back in the vowel space. There are, however, LOT tokens in other environments that are produced with similar spectral qualities for all four speakers, so that there is no clear separation between /ɑlC/ and other LOT tokens.

Furthermore, there are only a few instances of /ɑlC/ tokens that are produced in the "core" of the phonetic space occupied by THOUGHT. This appears to be the case for Shelley in Figure 153, and Brian in Figure 154. However, for Shelley, this observation is only based on one /ɑlC/ token, and a few repetitive THOUGHT words. Brian, on the other hand, has four out of five /ɑlC/ tokens approaching central THOUGHT qualities. However, this may simply be the result of a good amount of overlap of his LOT and THOUGHT classes in this speech style (66%).

---

[73] C, in this case, stands for /f/ and /v/.

[74] Overlapping words were removed from these plots in order to ensure readability. There was, however, no control over *which* words were removed, so that I had little to no control over the selection of words that are shown in the plots.

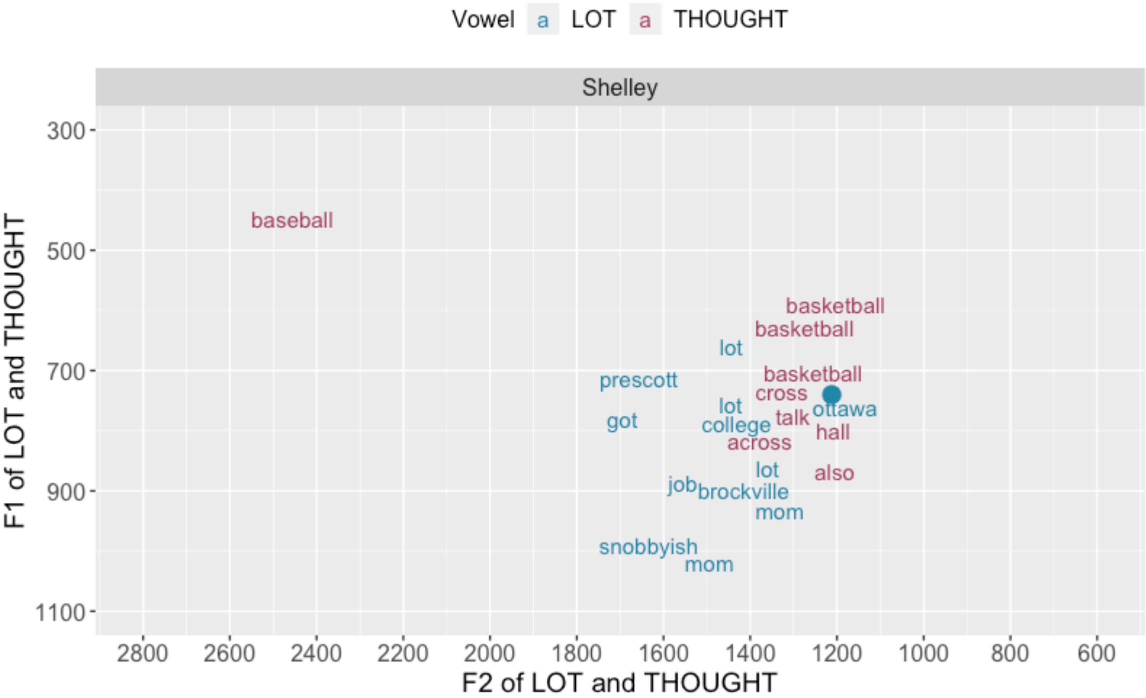


Figure 153: Distribution of /ɑlC/ clusters in relation to LOT and THOUGHT for Shelley, born in 1989. She was interviewed in 2008, and produced one token of *involved* in spontaneous speech, which is represented by the solid dot in the plot.

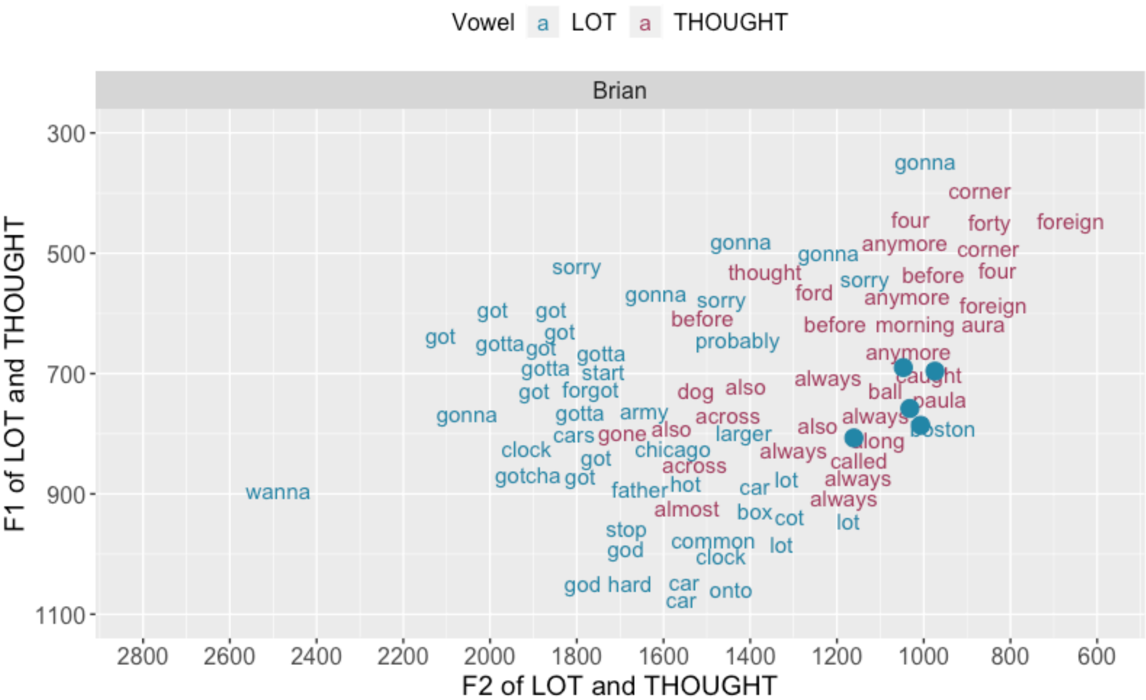


Figure 154: Distribution of /ɑlC/ clusters in relation to LOT and THOUGHT for Brian, born in 1945. He was interviewed in 2016, and produced three tokens of *involved*, and two tokens of *evolved* in spontaneous speech, which are represented by the solid dot in the plot.

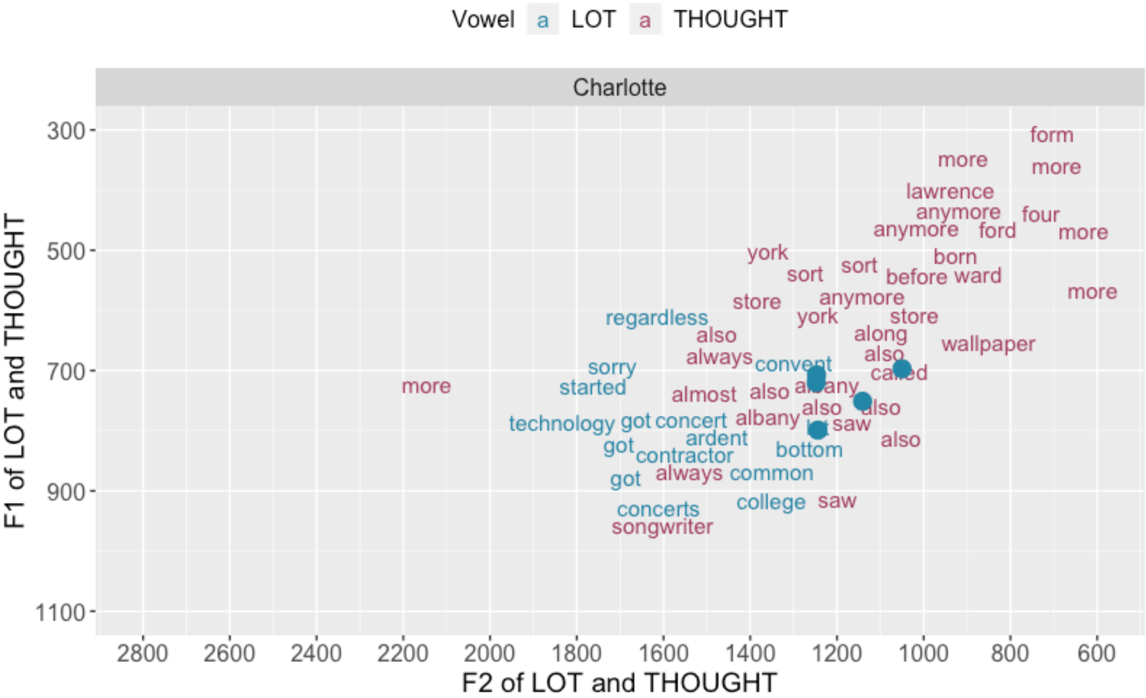


Figure 155: Distribution of /ɑlC/ clusters in relation to LOT and THOUGHT for Charlotte, born in 1958. She was interviewed in 2016, and produced five tokens of *involved* in spontaneous speech, which are represented by the solid dots in the plot.

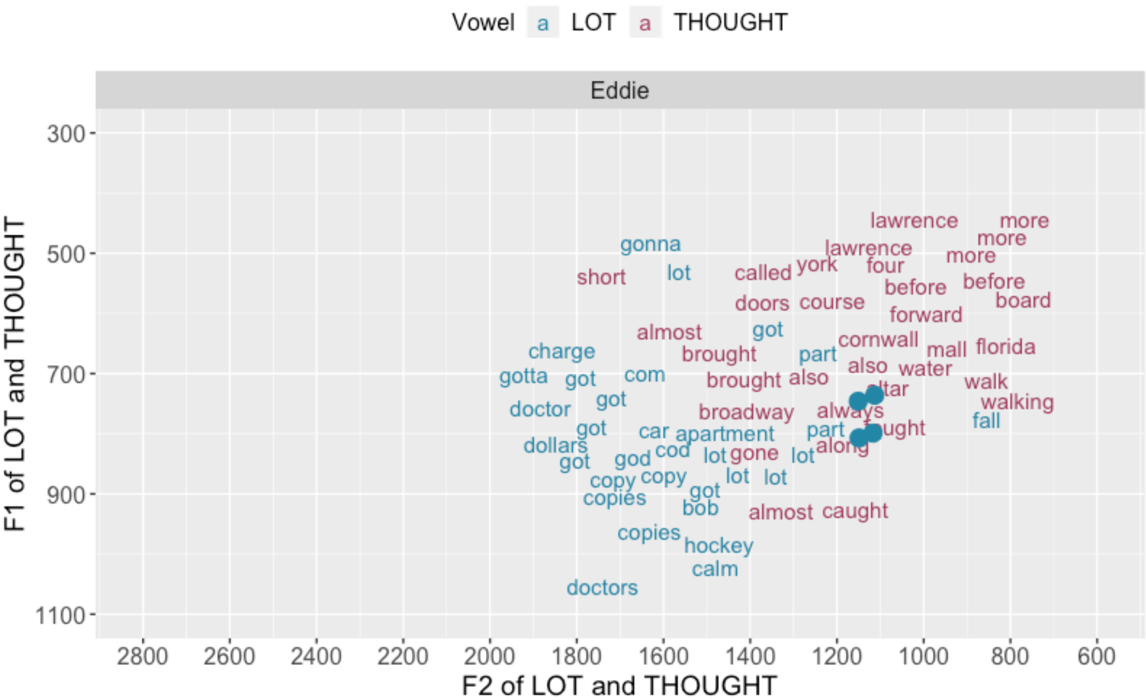


Figure 156: Distribution of /ɑlC/ clusters in relation to LOT and THOUGHT for Eddie, born in 1943. He was interviewed in 2016, and produced four tokens of *involved* in spontaneous speech, which are represented by the solid dot in the plot.

Neither of these four speakers therefore make very strong cases for a merger by phonological transfer, and this appears to be true for the rest of the speakers in the sample as well. The lexical items containing the /ɑlC/ cluster that occurred most commonly in the sample are variations of the word *involve*. It was produced a total of 24 times by 15 of the 2016 speakers and one 2008 speaker. For none of these 24 tokens does there appear to be a clear separation between /ɑlC/ tokens and LOT in other environments. They do, however, tend to be realized notably higher and backer than LOT in other environments, as confirmed by the regression models testing for the effect of following /lC/ clusters on F1 and F2 of LOT in spontaneous speech presented in Tables 98 and 99. Overall, it seems that, at least in spontaneous speech, there is no evidence of phonological transfer as a driving mechanism behind the COT-CAUGHT merger. Instead, it seems that LOT in /lC/ clusters merely has spectral qualities that are somewhat closer to THOUGHT than LOT in different environments for some of the speakers in the sample.

Predictor	Coefficient	*p*
(Intercept)	808.566 Hz	
Age	0.171 Hz	0.318
Gender (Male)	10 Hz	0.13
Education *(No college)* *(Student)*	 10.3 35 Hz -18.662 Hz	0.043
Preceding /lC/	-63.643 Hz	0.028

Table 98: Effect of /lC/ on F1 of LOT in spontaneous speech in the combined 2008 and 2016 sample. Reference levels: females, college educated, not preceding /lC/. Random effects: *speaker, word*. *n*= 6634

Predictor	Coefficient	*p*
(Intercept)	1397.316 Hz	
Age	0.484 Hz	0.337
Gender (Male)	20.006 Hz	0.289
Education *(No college)* *(Student)*	 46.376 Hz 7.637 Hz	0.097
Preceding /lC/	-285.069 Hz	$5x10^{-10}$

Table 99: Effect of /lC/ on F2 of LOT in spontaneous speech in the combined 2008 and 2016 sample. Reference levels: females, college educated, not preceding /lC/. Random effects: *speaker, word*. *n*= 6634

6.2.3.3.2 Wordlist Style

The wordlist that the participants in this study were asked to read included two words in which LOT occurred in an /lC/ environment: *golf* and *revolve*. Their distribution in the vowel space, relative to the community means for LOT and THOUGHT, are illustrated in Figure 157 below.[75] As can be seen in the plot, the majority of *golf* and *revolve* tokens are centered not only around, but in fact behind and above the mean for wordlist THOUGHT, far

[75] Again, this plot presents a random selection of /ɑlC/ words to avoid over-plotting. This selection is, however, representative of the overall distribution of these tokens.

away from mean LOT, indicating rather strongly that, in this speech style, /ɑlC/ may have been transferred to the THOUGHT class.

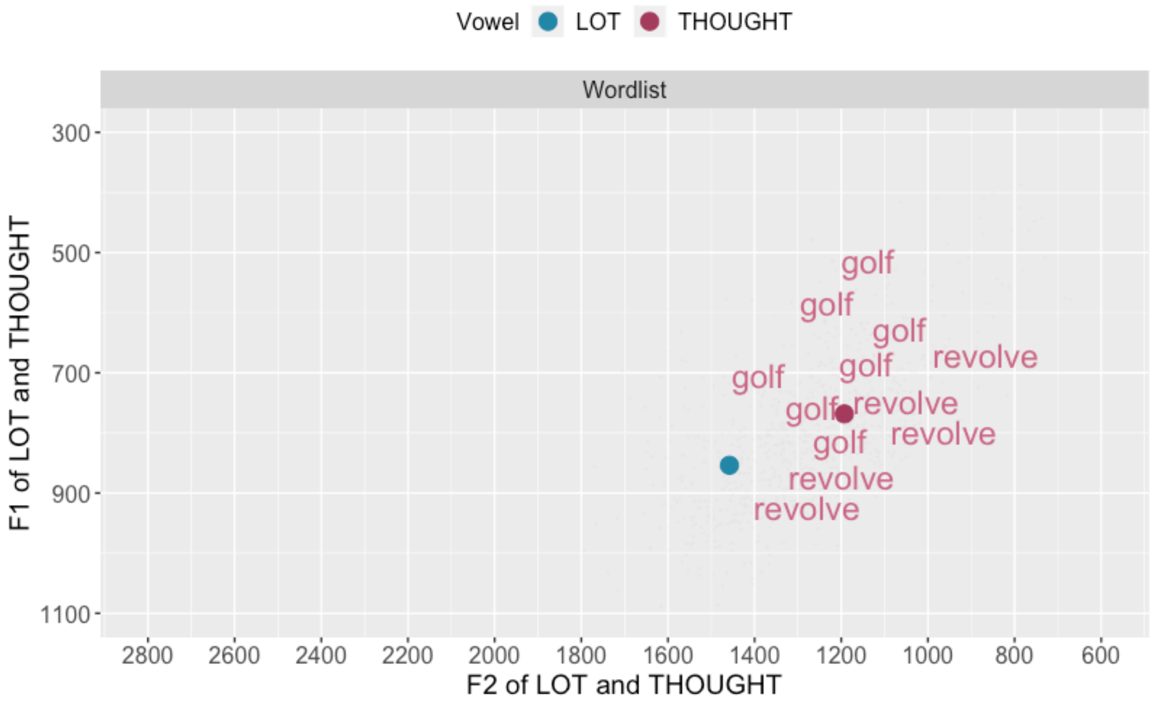


Figure 157: Distribution of LOT in /lC/ clusters in wordlist style. /ɑlC/ tokens are plotted in relation to the community means for LOT and THOUGHT, represented by the solid dots in the plot.

To test whether this holds true for individual speakers and their individual distribution of LOT and THOUGHT, four speakers will be analyzed individually: two who are clearly distinct in their production of the two vowels in wordlist style, and two who are relatively merged.

Figures 158 and 159 below illustrate the distribution of *revolve* and *golf* in the vowel spaces of Mark and Allison, two younger and relatively merged speakers from the 2016 sample. For both speakers, /ɑlC/ tokens are notably higher than the remaining LOT words, exceeding also the height of the majority of THOUGHT words. In three out of four cases, they are also notably backer than other LOT words, the only exception being Allison's *golf* token, which resembles most other LOT tokens in frontness, but is still notably higher. Overall, however, these two speakers provide strong evidence that /ɑlC/ has, in fact, been transferred to THOUGHT. The wordlist /ɑlC/ tokens *revolve* and *golf* are represented by the solid dots in the plots below.

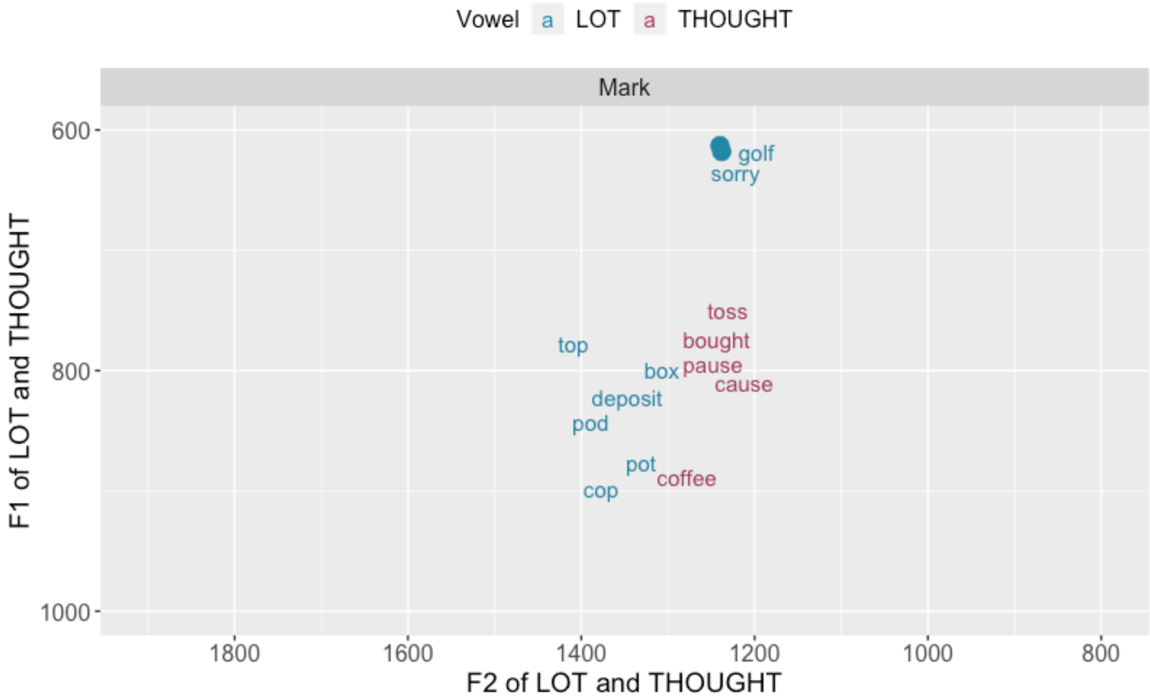


Figure 158: Distribution of LOT in /lC/ clusters in wordlist style for Mark, born in 2001.

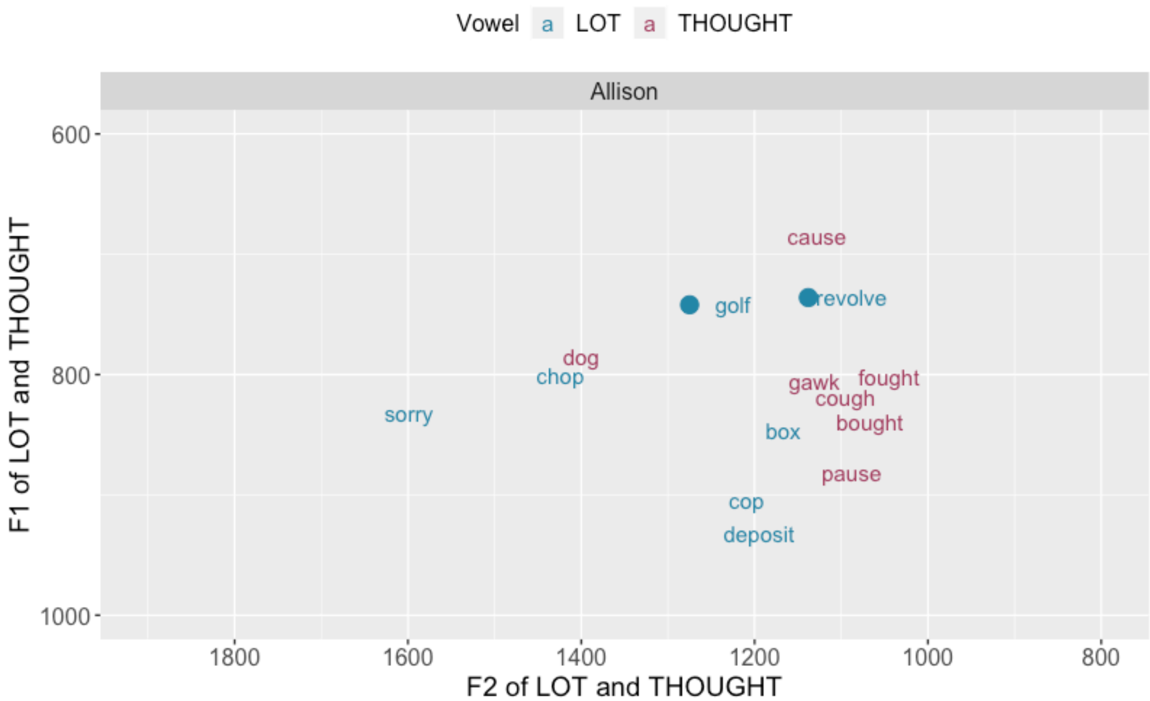


Figure 159: Distribution of LOT in /lC/ clusters in wordlist style for Allison, born in 1993.

Figures 160 and 161 below illustrate the distribution of *revolve* and *golf* in the vowel spaces of Scott and Helen, two older speakers from the 2016 sample. Although both of these speakers are clearly distinct in their production of LOT and THOUGHT in wordlist style, they show the same behavior in their production of /ɑlC/ tokens as the two more merged speakers in Figures 160 and 161 below. For both Scott and Helen *golf* is notably higher and further back than any other LOT words, and in fact than most THOUGHT words as well. *Revolve* differs to a less extreme extent from other LOT words, especially for Scott. It is, however, still one of the backest and highest instances of LOT for both speakers.

The comparison of the two younger, more merged and two older distinct speakers suggests that the transfer of /ɑlC/ to the THOUGHT class in wordlist style proceeded independent of the speaker's *age* and their participation in the merger. The regression models in Tables 100 and 101 support this observation. The model in Table 100 suggests that /ɑlC/ is realized notably higher than wordlist LOT in other environments. However, it also predicts that /ɑlC/ is raising significantly more in apparent time relative to LOT in other environments, though a separate model suggests that no significant apparent-time differences exist in the height of wordlist /ɑlC/ (0.761 Hz, p= 0.23). The same appears to be the case for F2, indicated by the significant effect of /lC/ in Table 101, as well as by the insignificant effect of *age* in a separate model (1.566 Hz, p= 0.16). The estimated F2 differences between wordlist LOT in /lC/ and other environments is substantial, and statistically significant. The same holds true for F1, as tested in a separate model without the interaction between *age* and /lC/ (-113.18 Hz, p= 0.045). Thus, it appears that speakers of all ages have participated in this transfer, regardless of whether or not they participate in the merger otherwise.

Predictor	Coefficient	*p*
(Intercept)	831.837 Hz	
Age	0.123 Hz	
Gender (Male)	-4.315 Hz	0.776
Education *(No college)* *(Student)*	 23.543 Hz -3.045 Hz	0.333
/lC/ cluster	-159.767 Hz	
Age/lC/*	1.148 Hz	0.018

Table 100: Effect of /lC/ on F1 of LOT in wordlist style in the combined 2008 and 2016 sample. Reference levels: females, college educated, not preceding /lC/. Random effects: *speaker, word*. *n*= 464

Predictor	Coefficient	*p*
(Intercept)	1384.539 Hz	
Age	1.37 Hz	0.076
Gender (Male)	18.026 Hz	0.528
Education *(No college)* *(Student)*	 55.642 Hz 29.009 Hz	0.218
/lC/ cluster	-271.658 Hz	0.019

Table 101: Effect of /lC/ on F2 of LOT in wordlist style in the combined 2008 and 2016 sample. Reference levels: females, college educated, not preceding /lC/. Random effects: *speaker, word*. *n*= 464

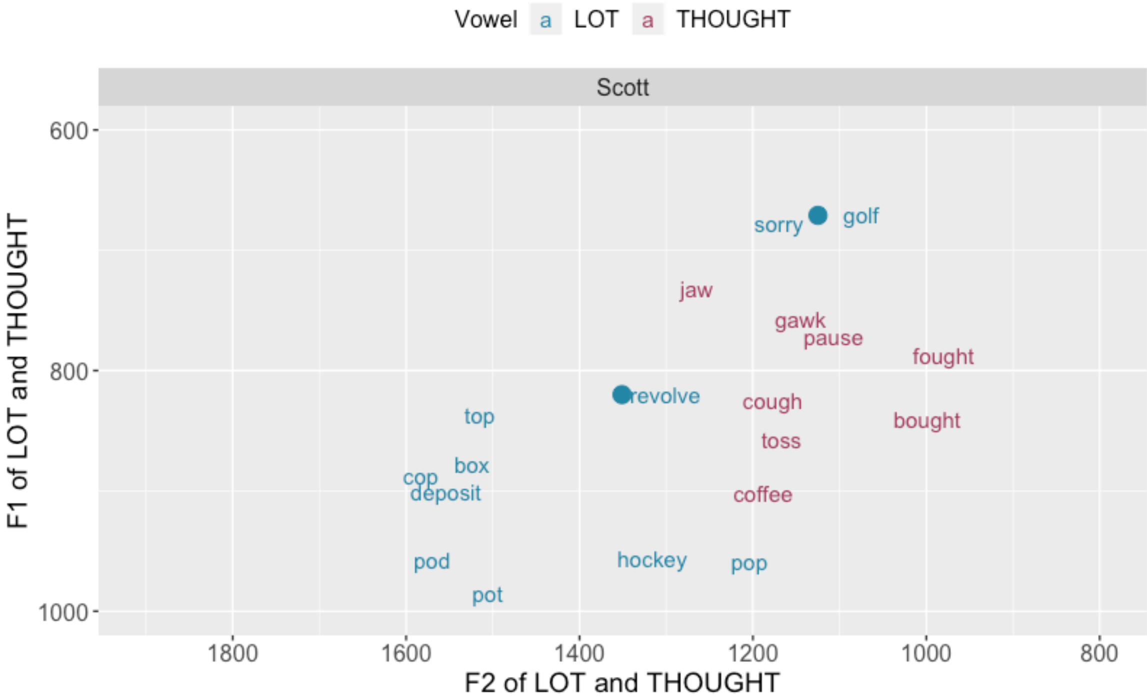


Figure 160: Distribution of LOT in /lC/ clusters in wordlist style for Scott, born in 1946.

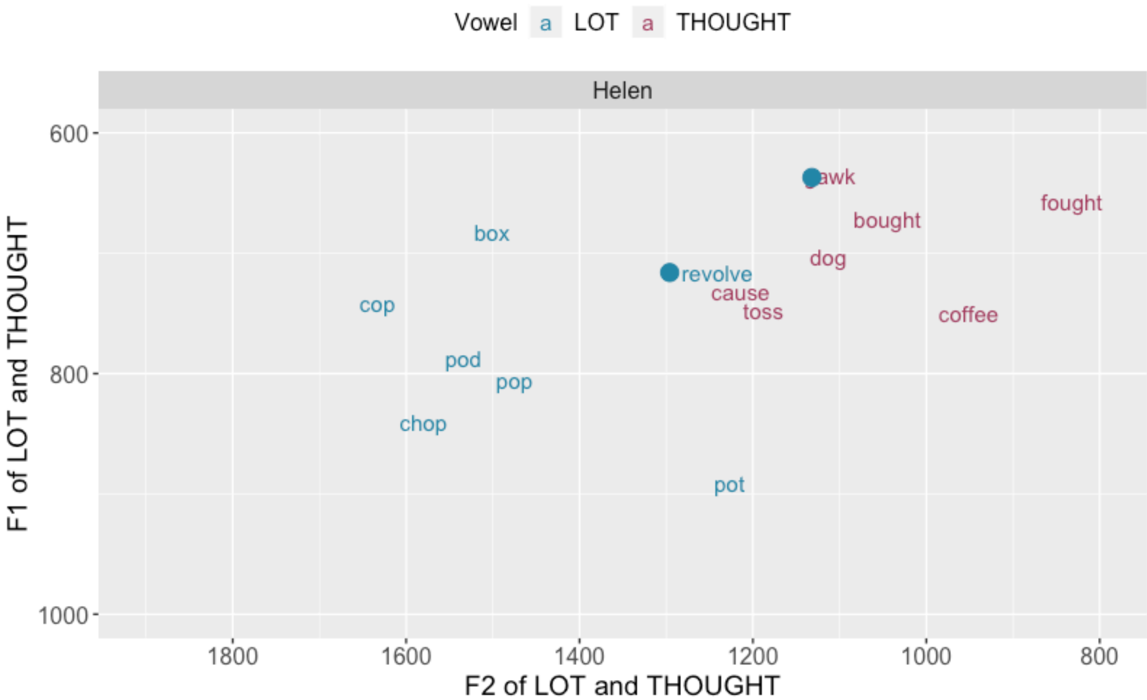


Figure 161: Distribution of LOT in /lC/ clusters in wordlist style for Helen, born in 1949.

6.2.4 The "New" LOT and THOUGHT

The phonetic qualities of the "new" phoneme that develops out of the merger of LOT and THOUGHT can vary depending on the underlying mechanism driving the merger (see Chapter 1.2.2). The analyses in Chapter 6.2.3 have shown that, with the exception of potential phonological transfer of /ɑlC/, the merger in Ogdensburg appears to be progressing primarily by means of approximation. The main actor in this process appears to be LOT, with significant raising and backing toward THOUGHT, while only slight lowering and fronting can be observed for THOUGHT. From these observations it follows that LOT and THOUGHT can be expected to merge in low back position, with intermediate spectral qualities that are somewhat closer to traditional THOUGHT than to traditional LOT.

To see whether this is the case, the vowel spaces of four of the most merged speakers can be assessed regarding the distribution of LOT and THOUGHT. Figures 162 to 165 below illustrate the distribution of LOT and THOUGHT in the vowel spaces of four speakers recorded in 2016. As can be seen in the plots, all of these four speakers appear to merge LOT and THOUGHT in low position just back of center, within an F1 range between 700 and 850 Hz, and an F2 range between 1200 and 1400 Hz. This appears to be the case regardless of *speech style*, as Grace's means in Figure 162 below represent LOT and THOUGHT in minimal pairs[76], while for the other three speakers, they represent spontaneous speech. In all four plots, the small translucent dots represent the speakers' individual tokens, the bigger and solid dots represent their LOT and THOUGHT means.

[76] Grace is most merged in minimal pair production, so that her LOT and THOUGHT means in Figure 162 are taken from this speech style. In order to illustrate their relative position, they were plotted against individual tokens for all other vowels in spontaneous speech, as minimal pair data is only available for LOT and THOUGHT.

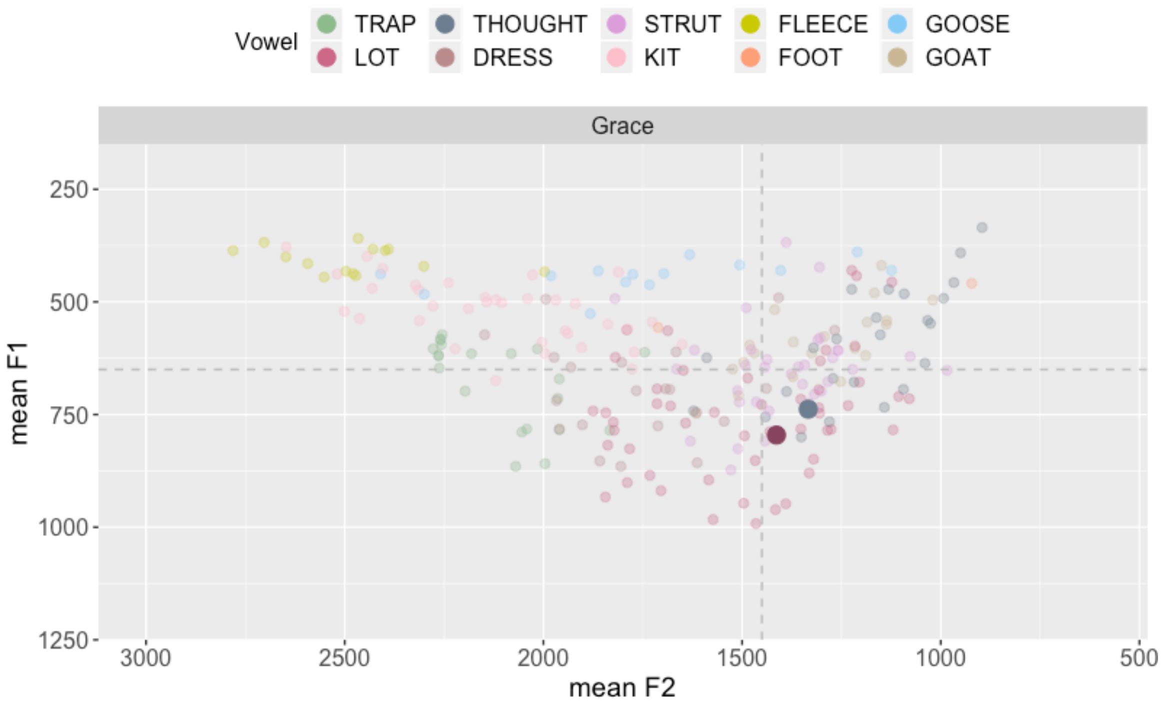


Figure 162: Distribution of LOT and THOUGHT in Grace's vowel space.

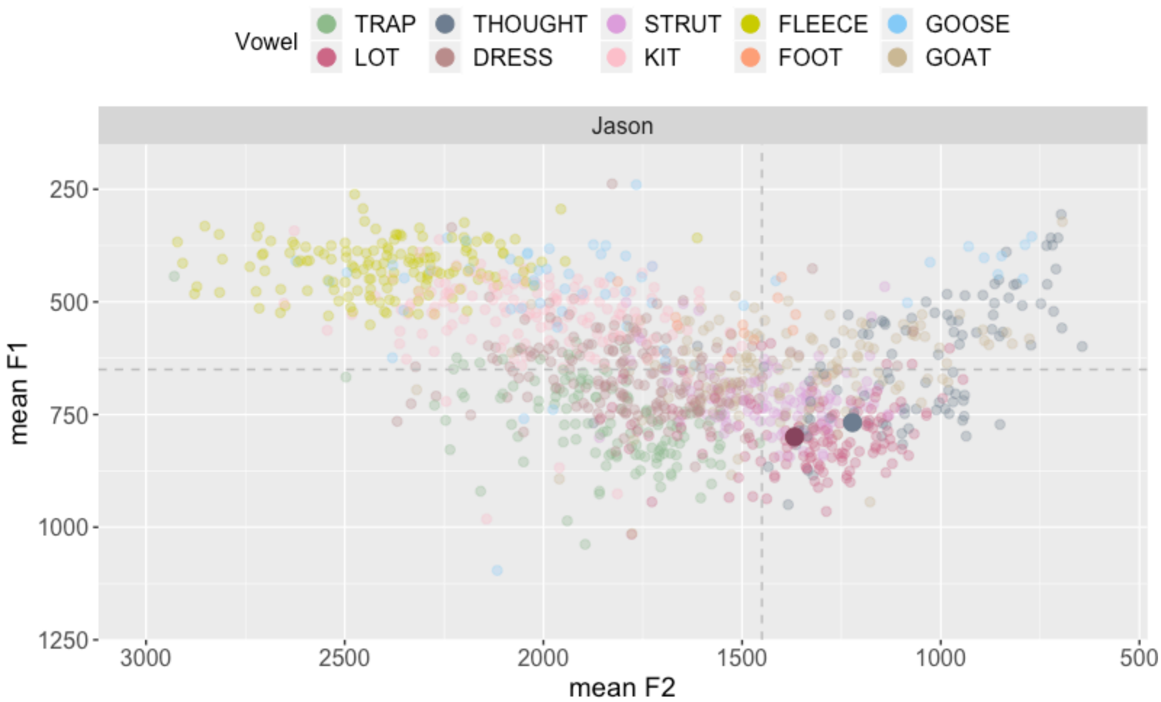


Figure 163: Distribution of LOT and THOUGHT in Jason's vowel space.

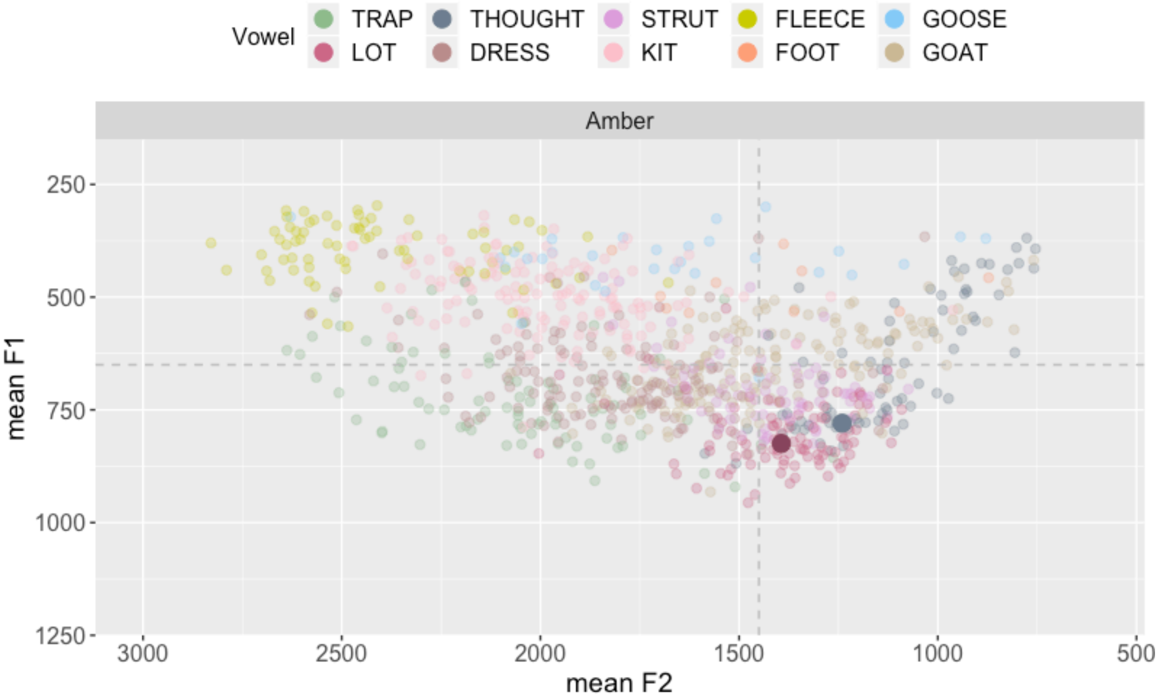


Figure 164: Distribution of LOT and THOUGHT in Amber's vowel space.

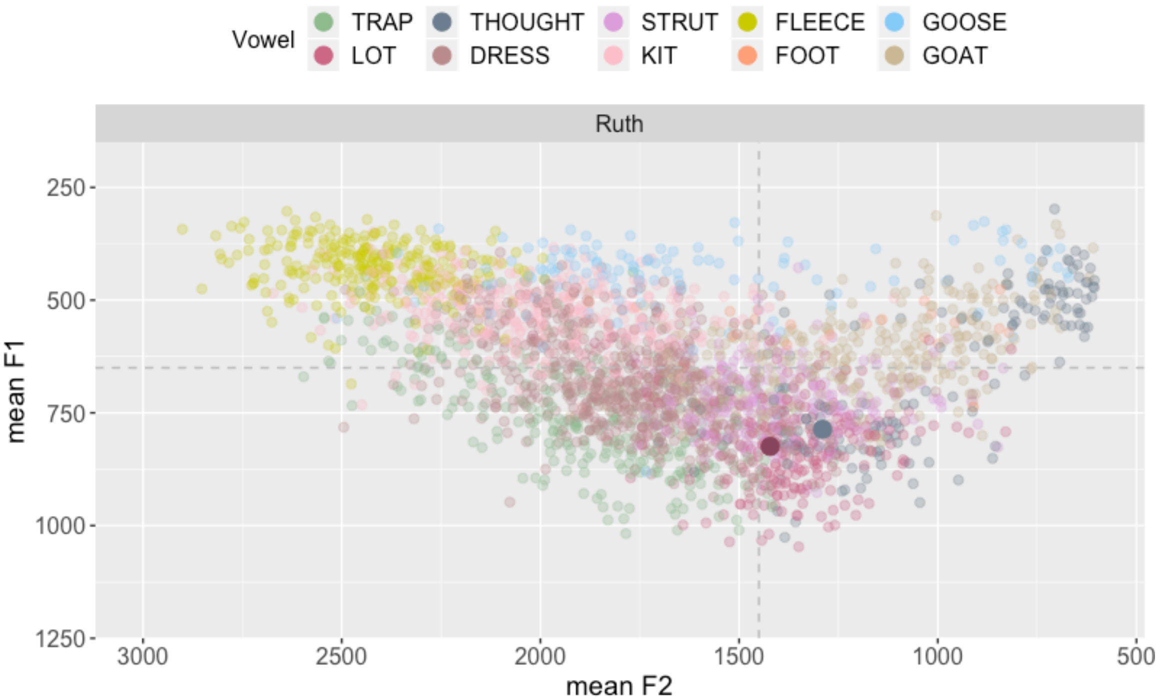


Figure 165: Distribution of LOT and THOUGHT in Ruth's vowel space.

6.2.5 Vowel Duration as a Counteracting Force

In Chapter 1.2.2, potential counteracting forces in the merger of vowel classes were introduced. These included acoustic parameters such as the presence of a glide on one of the members of the merger as well as differences in rounding, formant trajectories, creakiness/breathiness, nasality, or vowel duration. In this subchapter, I will examine the role of the latter in the merger of LOT and THOUGHT in Ogdensburg. Because vowel duration was not measured in the 2008 data, this chapter focuses on the 2016 sample only.

The differences in the duration of LOT and THOUGHT appear to depend the *speech style*, as can be seen in Figure 166 below. While virtually no duration differences appear to exist in minimal pair production, slight differences can be observed in spontaneous speech, and significant differences in wordlist style. The regression models presented in Table 102 below support these observations. The regression for minimal pairs predicts no significant difference in duration between the two vowels, and as Figure 166 below shows, only four speakers have differences of more than 25 ms, which has been established as the perceptual threshold. The regression models for spontaneous speech and wordlist style, on the other hand, do find the duration differences between LOT and THOUGHT to be significant in these speech styles. However, the estimated difference in spontaneous speech is only 8.5 ms, and as Figure 166 below shows, there are only a few speakers who cross the threshold of 25 ms, producing spontaneous THOUGHT with a notably longer duration than spontaneous LOT. In wordlist style, all but two speakers produce a much longer THOUGHT in comparison to LOT, with a significant estimated difference of 54.8 ms. In fact, wordlist reading is the only style in which all speakers produce THOUGHT with a notably longer duration than LOT. In the two other speech styles, no such consensus appears to exist in the community.

Interestingly, the two speech styles in which LOT and THOUGHT show the highest degree of merger or the most progress toward the merger, i.e. spontaneous speech and minimal pairs, are the two speech styles without any significant duration differences between the two vowels. Thus, it appears that duration is not used as a phonetic cue to maintain a distinction between LOT and THOUGHT as they are becoming more similar in terms of spectral qualities.

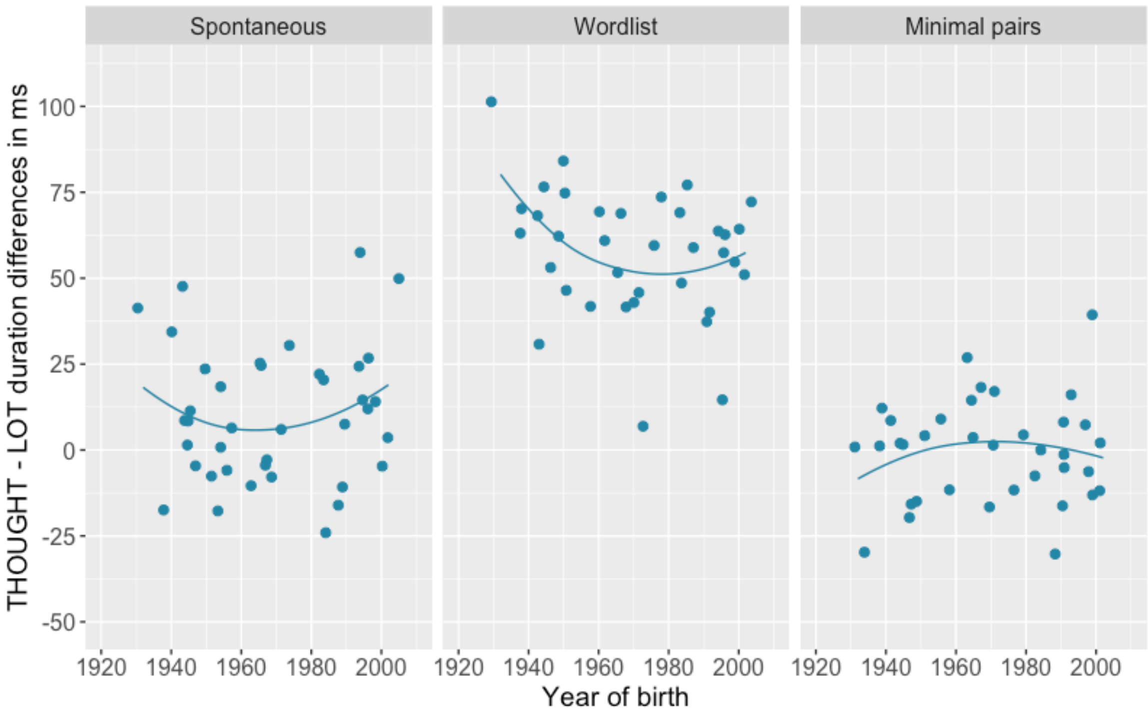


Figure 166: Duration difference between LOT and THOUGHT in 2016 across *speech styles*. A positive value indicates that THOUGHT is longer than LOT.

Predictor	Spontaneous speech		Wordlist style		Minimal pairs	
	Coefficient	*p*	Coefficient	*p*	Coefficient	*p*
(Intercept)	133.2 ms		121.9 ms		212.2 ms	
Age	0.3 ms	0.009	0.6 ms	0.031	0.5 ms	0.052
Gender (Male)	-3.9 ms	0.351	-7.1 ms	0.55	- 10.6 ms	0.392
Vowel (THOUGHT)	8.5 ms	0.003	54.8 ms	0.02	2.2 ms	0.67
Environment		2x10^{-16}		0.038		3x10^{-14}

Table 102: Duration of LOT and THOUGHT in 2016 across *speech styles*. Reference levels: female, LOT, /p/. Random effects: *speaker, word*. *n*= 11428 (spontaneous), *n*= 814 (wordlist), *n*= 1406 (minimal pairs)

It is of course possible that the most merged speakers in the data do not conform to these general observations. To test this, the duration differences for the most merged speakers in each speech style are summed up in Table 103 below.

As the numbers indicate, all of these speakers follow the same overall pattern that was presented above. All of them produce THOUGHT with notably longer durations than LOT in wordlist style, thus clearly differentiating the two vowels. In spontaneous speech, on the other hand, only three speakers, Daniel, Allison and Rachel, produce THOUGHT with a somewhat longer duration than LOT. Allison, however, remains below the perceptual boundary of 25 ms, and Daniel barely crosses it by 1.7 ms. In minimal pairs, a significant duration difference can only be observed for one speaker, Jason, whose minimal pair

THOUGHT is, on average, 39.4 ms longer than LOT in this speech style. Overall, however, it seems unlikely that duration is used as a differentiating acoustic parameter in the production of LOT and THOUGHT, even among the most merged speakers in the sample.

It is not clear why, in wordlist style, LOT and THOUGHT differ significantly more in length than they do in the other two speech styles. The most plausible explanation at this point is that this is simply a result of articulatory effort in the reading of isolated words. Presenting the participants with paired rather than isolated words in the minimal pair reading may have eliminated this effect.

Speaker (yob)	Spontaneous speech			Wordlist style			Minimal pairs		
	Distance in Hz	Overlap in %	Duration difference in ms	Distance in Hz	Overlap in %	Duration difference in ms	Distance in Hz	Overlap in %	Duration difference in ms
Mark (2001)	141.4	56.1	-4.7	131.7	41	51	130.2	49.5	-13.1
Ben (1999)	129.5	46.9	12	157.6	34.2	57.4	185.2	39.8	7.4
Jason (1998)	111.5	76	14.6	136.7	55	64.3	156	71.2	39.4
Daniel (1993)	140.4	69.3	26.7	126.8	28.5	62.7	64.1	68	16.1
Allison (1993)	175.4	76.2	24.4	125.2	64.1	63.7	158.4	42	-5.1
Rachel (1976)	164.2	75	30.4	168.9	35	59.5	196.1	25.3	4.4
Charlotte (1958)	141.9	80.9	6.4	68.5	23.8	41.8	312.3	16.7	-11.6

Table 103: Duration difference between LOT and THOUGHT for the most merged speakers in 2016 across *speech styles*.

6.2.6 Summary: Merger in Production

The analysis of the merger in production has shown that, at this point in time, LOT and THOUGHT appear to be two distinct phonological categories for the majority of speakers in this study. However, the data provide evidence of apparent and real-time transition toward merger in production. The pace of this progress depends significantly on the *speech style*, and, to some extent, on the speakers' *educational background*.

The highest degree of merger of LOT and THOUGHT for most speakers can be observed in spontaneous speech. Here, the majority of speakers have the smallest distances between, and the most overlap of LOT and THOUGHT. However, for the majority of speakers, the two categories are still clearly distinct in production in this speech style. There does appear to be a difference between speakers with a college education and those without, with the former group producing LOT and THOUGHT closer together and with more overlap than the latter. Overall, there is some evidence of a slight progress toward more merged production in apparent time in the 2016 data, however, speakers without a college degree do not seem to participate in this trend. Progress toward the merger results primarily from retraction and a slight raising of LOT among college educated speakers in apparent time. Real-time differences between the 2008 and 2016 data support these observations, especially in the overlap measure, where 2016 speakers show significantly more overlap than 2008 speakers, potentially because 2016 speakers produce a somewhat backer LOT. There is no significant evidence that differences in vowel duration are used as a contrasting feature in the production of LOT and THOUGHT in spontaneous speech.

Wordlist style appears to be the speech style with the lowest degree of merger of LOT and THOUGHT. Thus, speakers tend to shift toward less merged production when paying more attention to speech. In the 2016 data, speakers without a college degree produce wordlist LOT and THOUGHT with the greatest distance and the least amount of overlap compared to other speakers. In comparison, students produce both phonemes at about half the distance, and with more than twice as much overlap. College educated speakers appear to produce wordlist LOT and THOUGHT with distances and overlap that are intermediate compared to the other two groups. The 2016 data does suggest a potential apparent-time trend toward more merged production in both the distance and the overlap measure for college educated speakers, resulting from apparent-time LOT backing in this group of speakers. However, it is possible that the higher degree of merger among younger speakers is not a result of a gradual change, but a rather abrupt development that is specific to the student group in the sample. This is supported by the observation that students, as a group, produce THOUGHT in a fronter position than most other speakers. In any case, speakers without a college education seem to be excluded from this trend toward more merged production. The evidence for progress toward more merged production in wordlist style is notably weaker in the 2008 data and is not supported by real-time comparisons between the two data sets. Wordlist style is the only speech style where LOT and THOUGHT are clearly differentiated by length, as THOUGHT is notably longer than LOT.

The degree of merger in minimal pair production is relatively low in comparison to spontaneous speech for the majority of speakers. This indicates that most speakers shift toward less merged production in this speech style, at a level that is comparable to the shift from spontaneous speech to wordlist style. However, the most notable apparent-time trends toward a higher degree of merger can be observed in this speech style. In both the 2008 and 2016 sample, the distance between LOT and THOUGHT appears to be decreasing significantly in apparent time. The same trend can be observed in the amount of overlap, but it is much less pronounced for the 2016 sample in this measure. Real-time differences between the 2008 and 2016 samples corroborate these apparent-time trends. The primary underlying mechanism, again, appears to be merger by approximation, with LOT raising and backing significantly in apparent and real time, and some degree of THOUGHT lowering. Additionally, students in the 2016 sample appear to have a somewhat fronter THOUGHT than the rest of the speakers in this sample. Thus, students in the 2016 sample clearly stand out in their production of minimal pairs, with significantly smaller distances and more overlap than any other speakers, and a reversed style-shifting pattern, i.e. an *increased* degree of merger compared to the other two styles. Otherwise, no differences between educational groups emerge in this speech style. There is no evidence for a large-scale increase in duration differences between minimal pair LOT and THOUGHT that might serve to contrast the two vowels as they are assimilating in terms of spectral qualities.

Overall, the merger, and thus reversal of NCS-fronted LOT, appears to be a change that is led by college educates speakers, and that has been advanced quite rapidly by students in the more careful speech styles. Speakers who are the most advanced in merged production show a notable amount of overlap of LOT with THOUGHT BUT maintain a small but consistent distinction between LOT and THOUGHT, mainly on the front-back dimension.

6.3 Results: Merger in Perception

The results presented above provide evidence for an advancing merger of LOT and THOUGHT in production, which should entail, or in fact be preceded by, merger in perception. To determine to what extent speakers perceive LOT and THOUGHT as distinct linguistic categories, the data from the self-judgments and the commutation test will be analyzed in the following subchapters. Speakers from the 2008 sample are only included in the analysis of self-judgments, as commutation test scores are not available for this sample.

6.3.1 Self-Judgment of Minimal Pairs

The results of the self-judgment test suggest that the majority of speakers in both the 2008 and 2016 sample are clearly distinct in their perception of LOT and THOUGHT, but that the community appears to be transitioning toward merged perception. Overall, 28 of the 42 speakers whose self-judgments were recorded considered themselves as "distinct" in their production of LOT and THOUGHT, as shown in Figure 167. Only about half as many of the participants were either unsure in their response or thought that some minimal pairs sounded more alike than others and were therefore classified as "transitional". Only one speaker responded that she thought all minimal pairs sounded the same.

Figure 167: Self-judgment of minimal pair production.

Whether or not participants are distinct or transitional in their judgment appears to depend on their *level of education*, their *age* as well as their *gender*. As can be seen in Figure 167 above, all speakers without a college degree judge themselves as distinct in their production of LOT and THOUGHT minimal pairs. The same seems to be the case for the majority of speakers with a college education. However, a number of younger speakers from this educational group, as well as one of the older speakers, appear to be transitional in their judgments of minimal pair production. All of the college educated participants who are transitional in their judgments are female, while all male speakers with a college degree appear to consider their minimal pair production to be distinct. Students appear to be collectively transitional in their judgments. Only one of the speakers from this group responded that minimal pairs sounded different, and one judged them to sound the same.[77] Thus, it seems that there is an apparent-time trend toward merged perception in the community, led by young college educated females. However, there is no real-time evidence for this. Both samples have similar ratios of distinct and transitional speakers, with three out of seven transitional in 2008, and 10 or 11 out of 33 in 2016.

6.3.2 Commutation Test – Identifying Minimal Pairs

The commutation test scores suggest that, while the majority of speakers in the 2016 sample are clearly distinct in their perception of LOT and THOUGHT, the ability to identify LOT and THOUGHT is decreasing in apparent time. This is evident in both the identification of *cot* and *caught* in the clearly distinct line-up, as well as in the participants' identification of their own *cot-caught* production.

As can be seen in Figure 168 below, the majority of participants identified all clearly distinct instances of *cot* and *caught* in the commutation test correctly. Thus, neither of them seems to be completely merged in the perception of the two phonological categories LOT and THOUGHT. However, as shown in Figure 168, this ability seems to be diminishing in apparent time. The majority of participants born before 1990 did not seem to have any problem in identifying clearly distinct *cot* and *caught* correctly. In this age group, only three participants did not score 10 out of 10. Each of them, however, misidentified only one item in the line-up. Participants born after 1990 seem to have had more difficulty with the task. In this age group, three of the 10 participants did not reach the highest score of 10, and they confused notably more items in the line-up than older participants who did not reach the high score. The oldest of these three younger participants misidentified two words, the youngest of them six. The third one of them, born in 1998, scored zero in the identification of minimal pairs in this line-up. Thus, the speakers' *age* appears to be the most relevant factor in the ability to correctly identify LOT and THOUGHT. However, Figure 168 below indicates that *gender* and/or *education* may also

[77] It should be noted that the speaker with merged judgment was particularly shy and may just have responded with what she thought was the expected or correct answer. Thus, this judgment may not be entirely reliable. However, it can be assumed that, like other students, she was at least transitional in her judgment.

play a role in this ability, as the three older speakers who each identified one token incorrectly are all female, two of them college educated.

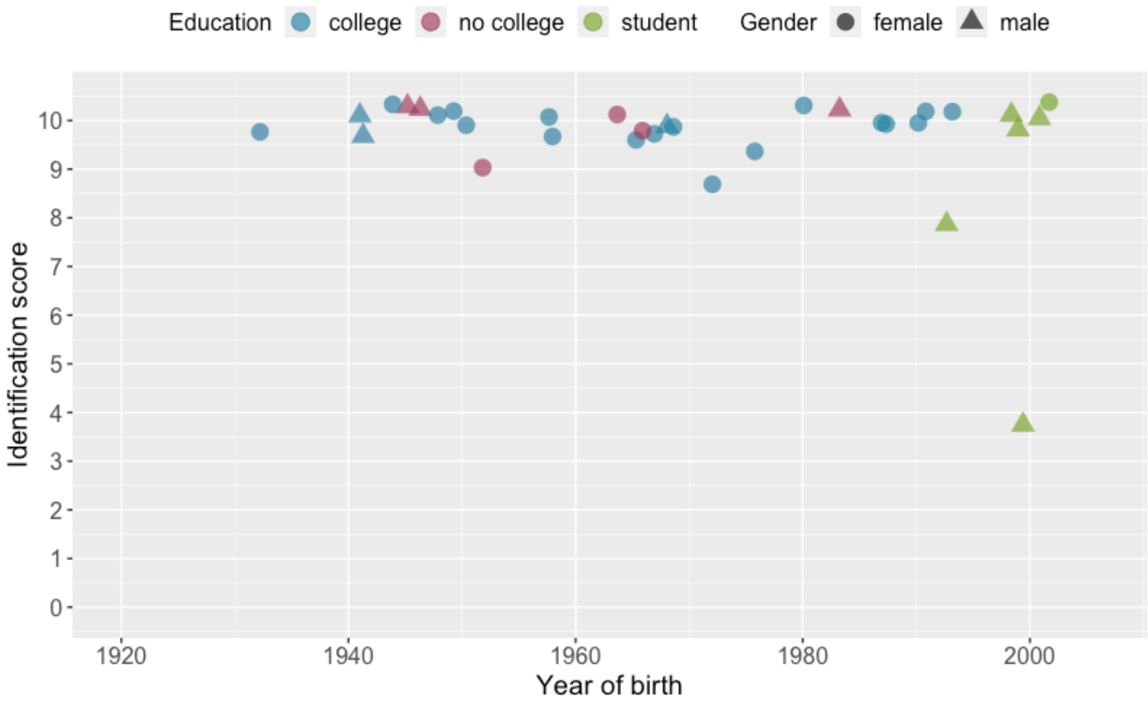


Figure 168: Commutation test scores for the clearly distinct line-up of *cot* and *caught* by *education* and *gender*.

Categorizing *cot* and *caught* in the self-commutation test posed a greater problem for many of the participants. Out of the 34 participants who completed the test, 14 did not reach the full score of 10 when they were asked to identify their own *cot* and *caught*. As shown in Figure 169 below, the number of misidentified words from each participant's own production ranges from one to seven. As also becomes evident from Figure 169 below, neither the speakers' *gender* nor *education* appear to be determining factors in the ability to identify *cot* and *caught* in their own speech production. Out of the 21 college educated speakers, five females misidentified at least one, and up to five of their words. Three of them are among the younger speakers in this group, born after 1970. In the group of six speakers without a college degree, half of the participants, both male and female, misidentified one to three of their words. For this group, however, *age* does not appear to have a significant influence on their ability to identify *cot* and *caught* correctly. One of the speakers was born after 1980, while the other two were born before 1960. In fact, the younger participant in this group misidentified fewer words than the two older speakers. Of the seven students in the sample, only one managed to correctly identify all 10 instances of *cot* and *caught*. Thus, it seems that the participants' ability to discriminate LOT and THOUGHT in their own minimal pair production depends to a great extent on their *age*; however, *age* does not appear to be the only relevant factor, as there are older speakers who misidentify some of their words as well. Out of these eight older speakers, two are male, six are female. Of the six females, five are college educated, and four of them misidentified four or more words; results that are very similar to those in the student group.

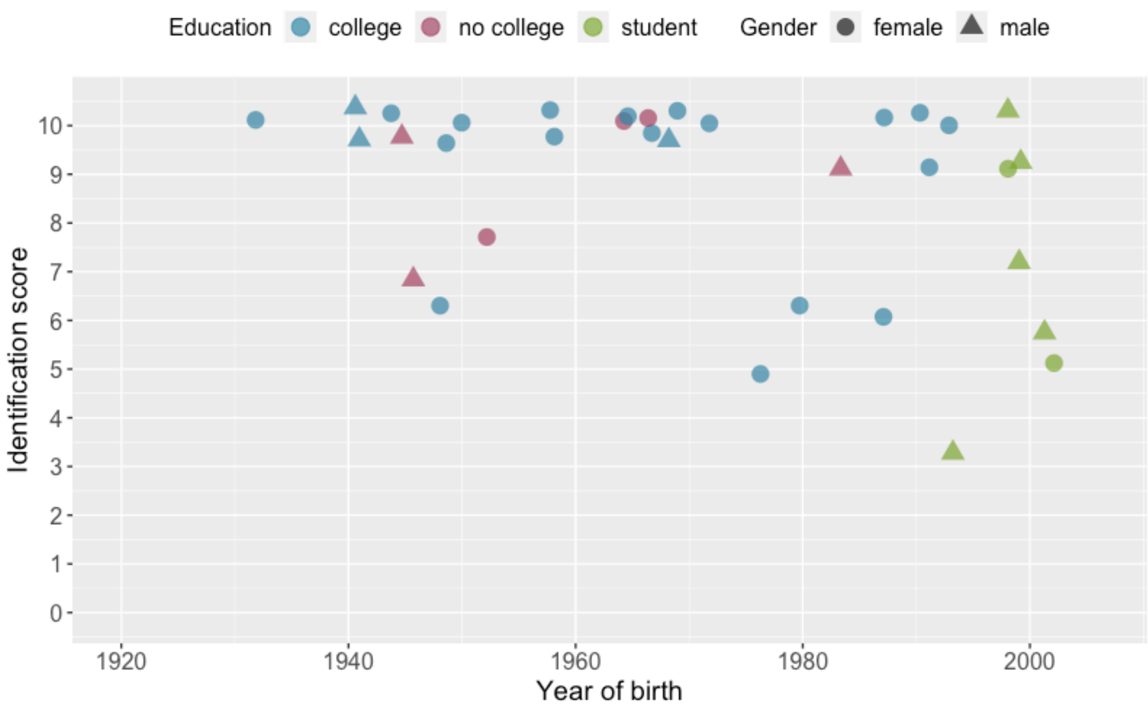


Figure 169: Commutation test scores for the participants' own minimal pairs.

A plausible influential factor on the self-commutation test scores is the participants' degree of merger in production. This, however, only seems to be the case for a few of these speakers. The participant who scored lowest on the commutation test, with only three words identified correctly, is Daniel, born in 1993. He produced his LOT-THOUGHT minimal pairs with a Euclidean distance of only 64 Hz, and with almost 68% overlap. Thus, he appears to be merged to a relatively high degree in minimal pair production. This is also true for the youngest speaker in the sample, Grace, who produced her minimal pairs with 85 Hz distance and 78% overlap and was only able to identify half of her words correctly. Both of these speakers' degree of overlap of LOT and THOUGHT and the means of the two vowels in minimal pair production are illustrated in Figures 170 and 171. Owing to this relatively high degree of merger for these two speakers in minimal pairs, it is not surprising that identifying them correctly might pose a problem. Therefore, these particular cases are not a strong indicator of a fading identification ability.

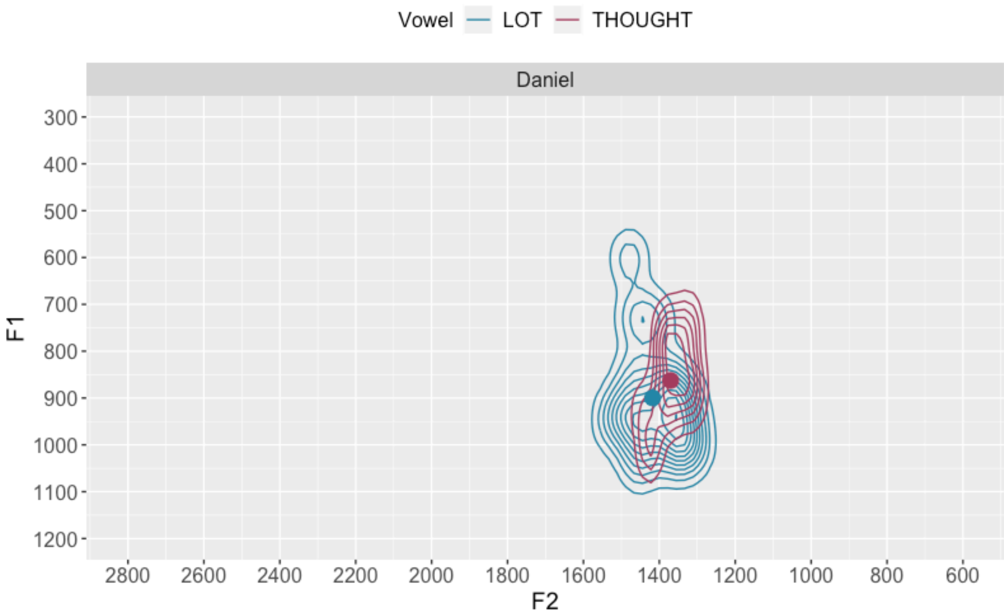


Figure 170: LOT-THOUGHT distance and overlap in Daniel's minimal pairs.

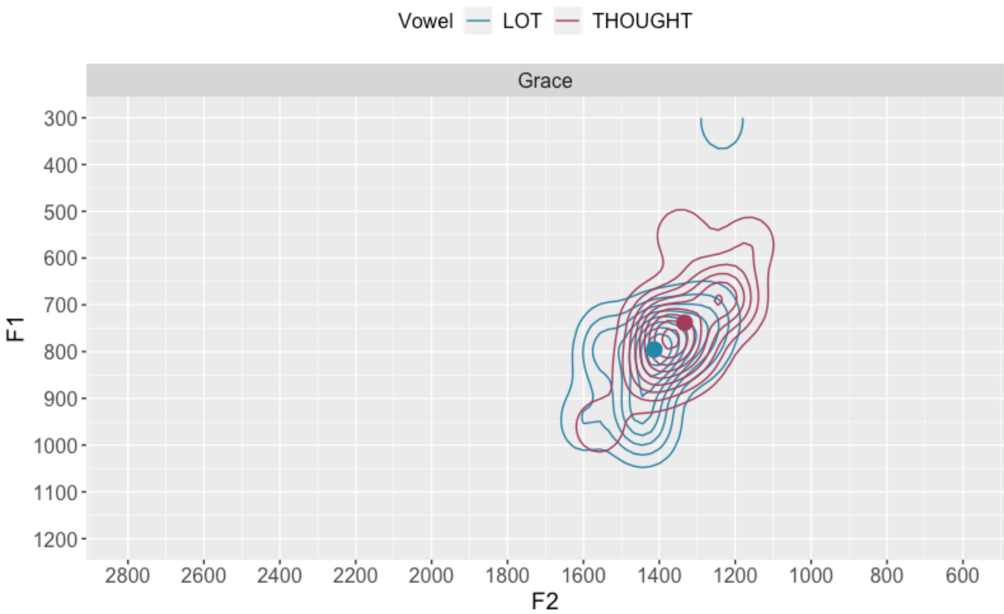


Figure 171: LOT-THOUGHT distance and overlap in Grace's minimal pairs.

Other younger speakers make stronger cases for a diminishing identification ability. Mark and Lindsey, for example, born in 2001 and 1987 respectively, failed to identify four of their 10 stimuli correctly. In production, however, neither of the two are merged to a great extent, with 128 to 130 Hz distance, and 49 to 52% overlap of their minimal pair LOT and THOUGHT, as shown in Figures 172 and 173. Nevertheless, although notably less merged than Grace (Figure 171), they did not score significantly higher in their identification of *cot* and *caught*.

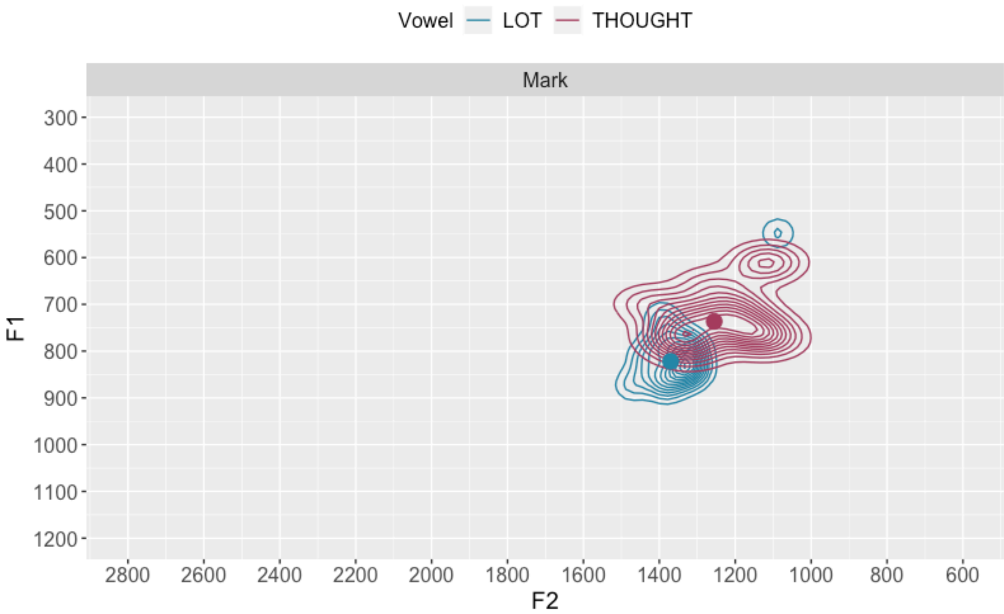


Figure 172: LOT-THOUGHT distance and overlap in Mark's minimal pairs.

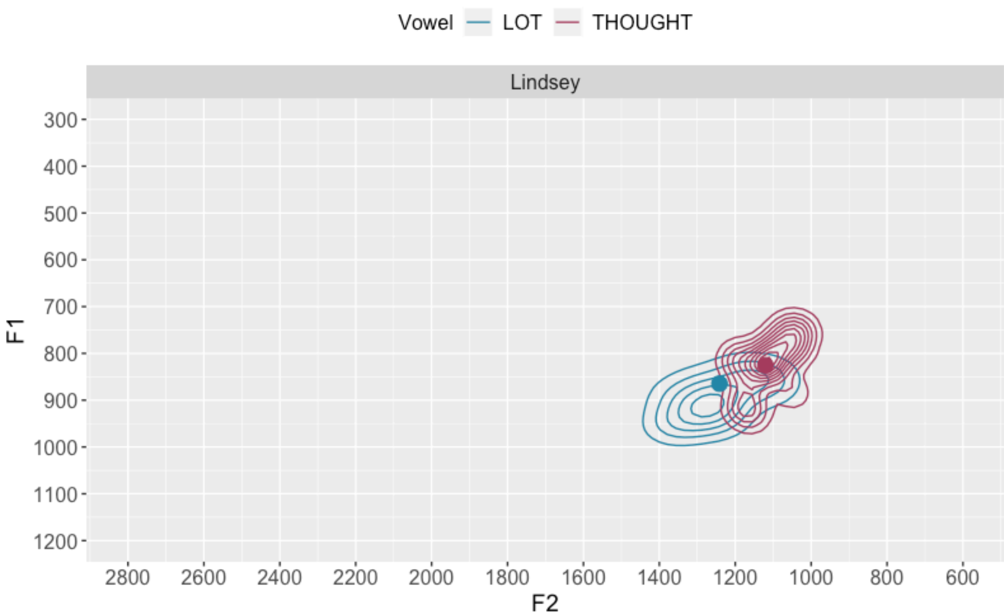


Figure 173: LOT-THOUGHT distance and overlap in Lindsey's minimal pairs.

Some speakers who scored relatively low in the self-commutation test do not appear to be merged at all in production, as can be seen in Figures 174 and 175. One of these speakers is Rachel, born in 1976. Despite a distance of 196 Hz between her minimal pairs, and only 25% overlap, she was only able to identify half of her *cot-caught* stimuli correctly. Another example is Scott, born in 1946. In his minimal pairs, the distance between LOT and THOUGHT is 380 Hz, and the amount of overlap is 9%. Thus, he is clearly distinct in his production. However, he misidentified three of his own *cot* and *caught* stimuli.

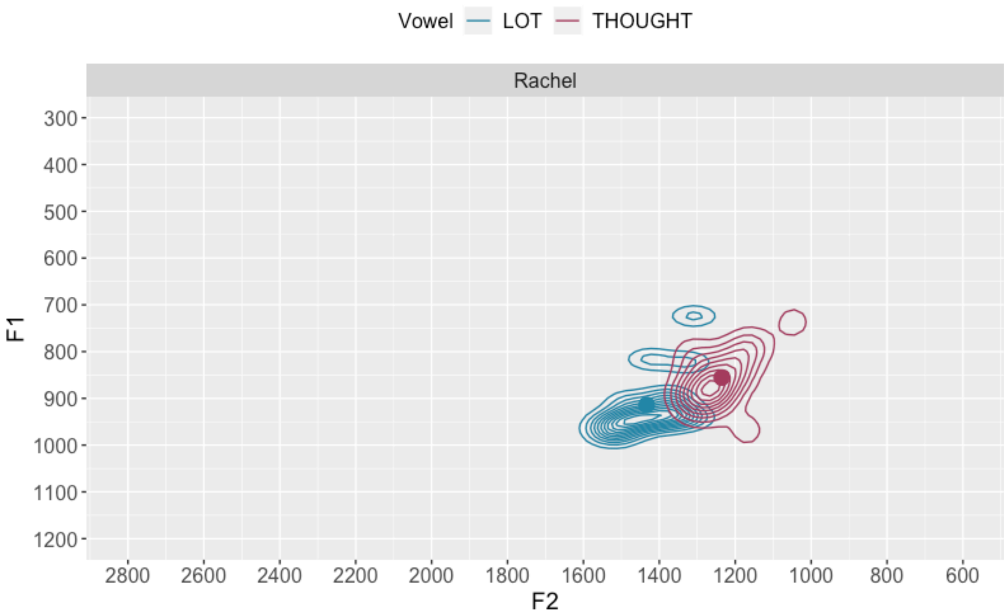


Figure 174: LOT-THOUGHT distance and overlap in Rachel's minimal pairs.

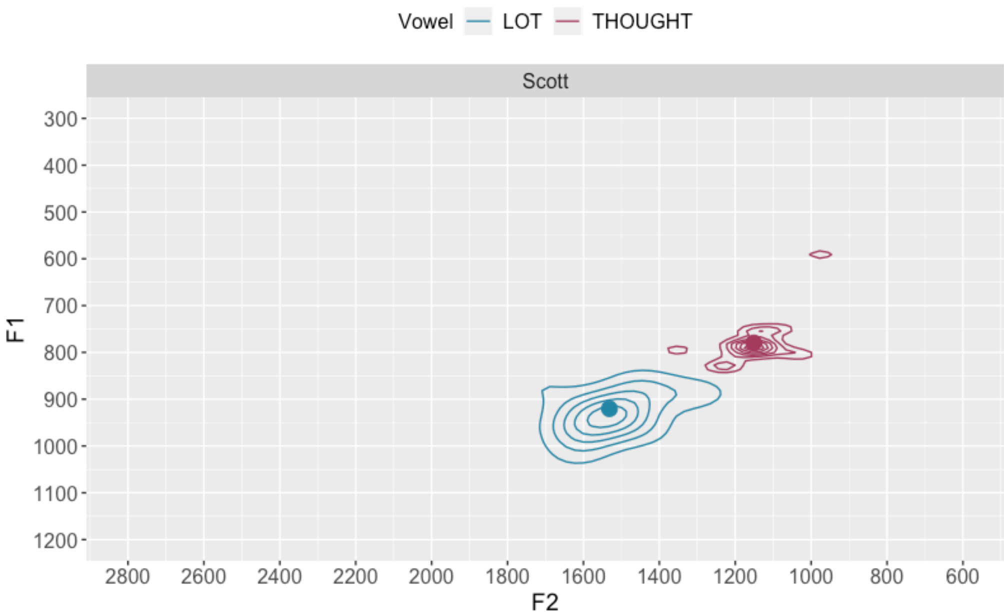


Figure 175: LOT-THOUGHT distance and overlap in Scott's minimal pairs.

6.3.3 Summary: Merger in Perception

Overall, the results from the self-judgment test and the commutation test suggest the same general pattern: Students and (female) college educated speakers are more merged in perception than most other speakers. All but one student have transitional minimal pair judgments, and some of them have at least some trouble correctly identifying even clearly distinct minimal pairs. College educated females who are transitional in perception are an exception rather than a general rule, and it seems that it is mostly younger speakers in this group who are showing signs of an advancing merger in perception. For the most part, this pattern matches the observations of merger in production, however, as was shown above, even some speakers who do not seem to be merged in minimal pair production are not entirely able to reliably identify LOT and THOUGHT correctly, which may suggest that merger in perception precedes merger in production for these speakers. There is, however, no evidence for this on a larger scale.

6.4 The Social Evaluation of Shifted and Merged LOT & THOUGHT

In the previous chapters, it was found that the advancing merger in production is mainly due to the retraction and raising of LOT, which constitutes a contra-movement to NCS-LOT fronting. In this chapter, I will explore whether potential social awareness and evaluation of fronted and/or merged production may be an underlying factor behind this development, taking into account anecdotal evidence as well as the matched guise ratings for fronted and unfronted LOT, as well as for the COT-CAUGHT merger.[78] Like TRAP, fronted LOT has been reported to have attracted at least some overt social commentary, and results from matched guise experiments show that fronted LOT is rated less favorably than unfronted LOT, especially among younger participants (Savage et al., 2016). Similar results have been reported for subjective reactions to merged and distinct production of LOT and THOUGHT (Di Paolo, 1992).

In the present study, I found that at least some participants seem to have a sense of awareness of differences in the realizations of the low back vowels. In reaction to the merger guises, two of the participants offered the following opinions:

[78] Because the matched guise experiment did not include guises for THOUGHT, the social perception of this vowel cannot be assessed here.

> *To me it would sound more like a Boston thing, like I /kat/ the ball or in New York City, I /kat/ it … you know /pak θə ka/, /kat/ the ball, I /kat/ the ball."*
>
> *(Monica, born in 1965)*
>
> *He sounded more like Boston kind of, ya /tʃak/ did you hear the /pap/ and the /tʃak/ … ya you hear the a, like there's an a in there or something."*
>
> *(Sarah, born in 1969)*

In explaining which feature she relies on when identifying Canadian English, Chloe, born in 1998, points out:

> *It's mainly with like the o's, like the way they pronounce their o's."*

Similarly, Ruth, who was found to rely on DRESS in her description of Canadian speech (Chapter 4.3.1), also includes LOT and/or THOUGHT in her accounts of Canadian speech, and Patrick imitated the speech he finds characteristic of Syracuse, Buffalo and Chicago with a notably fronted LOT in the word *sausage* (Chapter 3.3). Thus, it seems that variation in the realizations of the low back vowels has reached the level of conscious awareness for at least some of the participants, and thus may have become a marker in the community, which may be subject to social evaluation in a matched guise experiment.

The five categories tested in the matched guise experiment in the present study include friendliness, age, education, localness, and Canadian-ness. Noteworthy differences could be observed in the ratings for the perceived level of education, both for fronted and unfronted LOT as well as for merged and distinct production of LOT and THOUGHT. Furthermore, merged and distinct production is rated differently in terms of perceived localness, and fronted and unfronted LOT in terms of perceived friendliness. The rating patterns for each of these categories, and the social factors that influence these ratings, will be presented in the following subchapters.

6.4.1 Perceived Level of Education

The plots below illustrate the ratings for the perceived level of education of merged and distinct production of LOT and THOUGHT (Figure 176 below) and of fronted and unfronted LOT (Figure 177 below). They show that listeners appear to perceive unfronted LOT as more educated sounding than fronted, LOT. Likewise, merged production, which involves unfronted LOT, is perceived as more educated sounding than distinct production, although this is only the case for younger speakers. Older speakers, on the other hand, rate distinct production as the more educated way of speaking, so long as LOT is not fronted. This *age* pattern is reversed in the ratings of the LOT guises, where the youngest speakers do not appear to distinguish between fronted and unfronted, while older speakers do.

The regression model in Table 104 below[79] supports the *age* differentiation in the ratings of merged and distinct guises. While ratings for the distinct guise do not change over time, ratings for the merged guises are increasing significantly by an estimated 0.3 units per 10 years on a 6-point scale. This appears to have led to a reversal of the perception of merged and distinct LOT and THOUGHT in terms of how educated the speakers are perceived. Younger listeners rate merged guises significantly higher than distinct guises, i.e. perceive merged production as more educated, while older listeners rate distinct production more favorably. A regression model that considers raters born before 1975 separately[80] estimates the rating difference between the merged and unmerged guises at 0.6 units on a 6-point scale, with higher ratings for the distinct guises (p= 0.006).

The regression model in Table 105 below[81] supports the observation that unfronted LOT is perceived as more educated sounding than fronted LOT. However, this appears to apply only to female listeners, who rate unfronted guises almost 1 unit higher than fronted guises (p= $2x10^{-5}$). Males, on the other hand, have a significantly different rating pattern, and the fronted variant of LOT is rated 0.3 units higher than the unfronted one. A regression model including only male listeners does not find this to be a significant difference (-0.296, p= 0.336). Thus, unfronted LOT is perceived as more educated sounding by female listeners, while males make no distinction between the two variants.

[79] An interaction between *guise* and *age* does not reach the level of significance, and therefore is not included in the regression model.
[80] The factors of distance between and overlap of LOT and THOUGHT in the listeners' minimal pair production had to be removed from the regression when testing subsamples owing to scaling issues.
[81] Again, an interaction between *guise* and *age* does not reach the level of significance, and therefore is not included in the regression model.

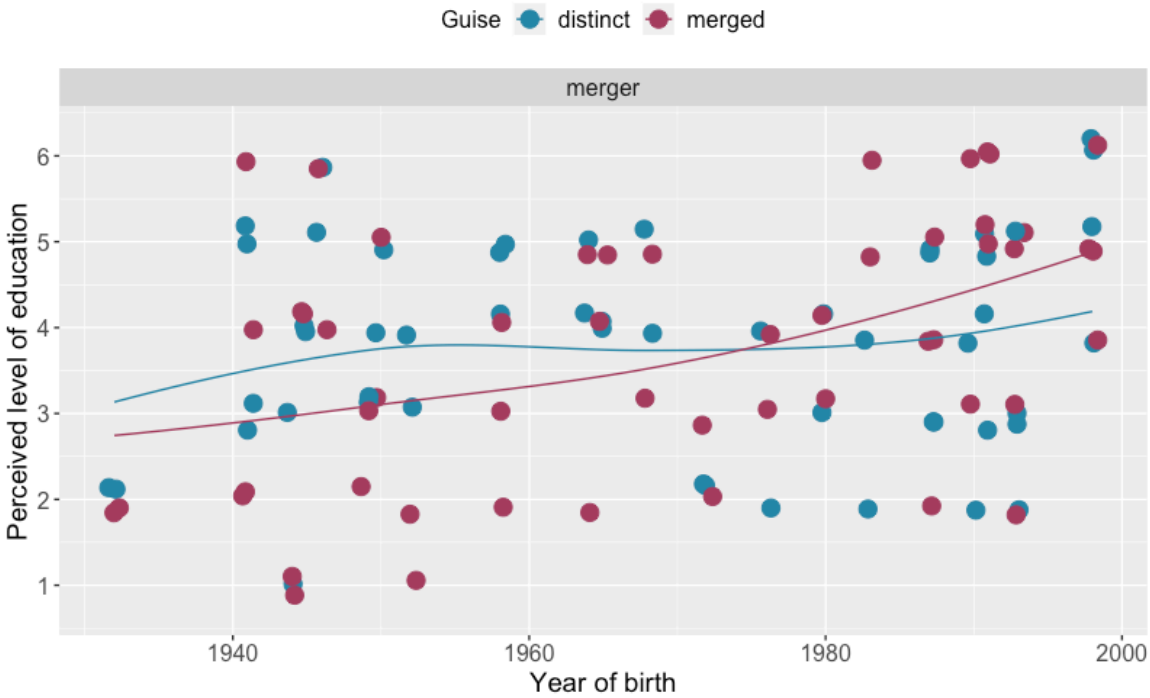


Figure 176: Merger matched guise ratings for the perceived level of education.

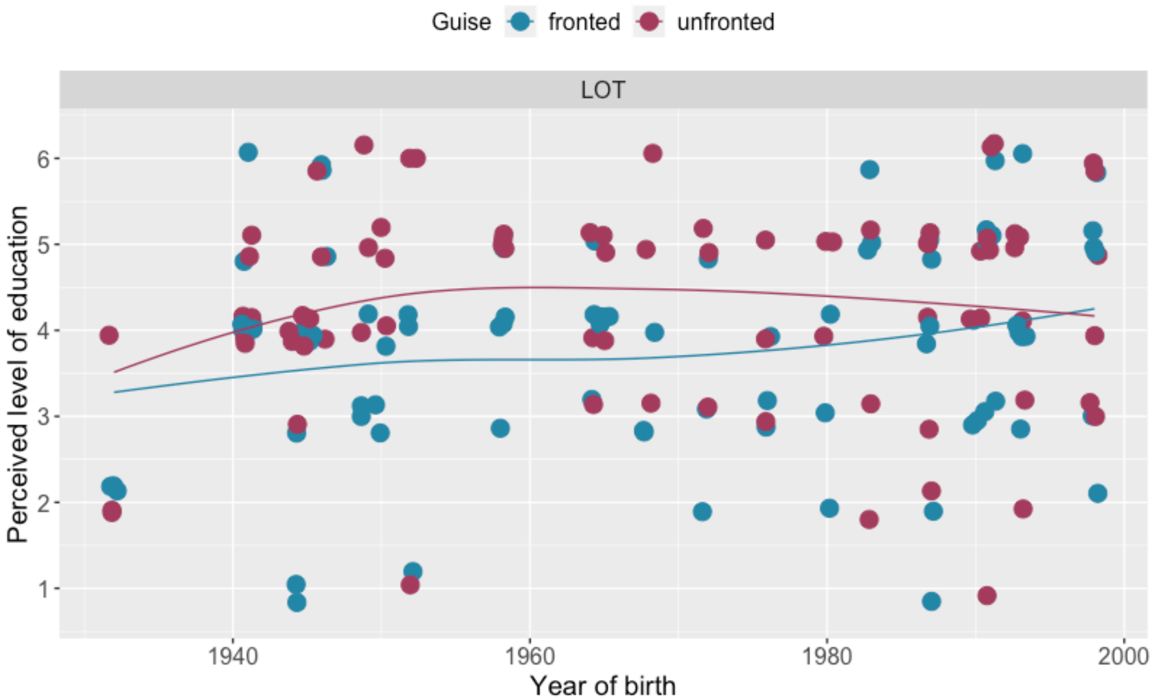


Figure 177: LOT matched guise ratings for the perceived level of education.

Predictor	merger	
	Coefficient	*p*
(Intercept)	3.914	
Age	-0.005	
Gender (Male)	0.769	0.131
Education *(No college)* *(Student)*	0.36 0.3	0.781
Minimal pair distance	-0.001	0.648
Minimal pair overlap	-0.303	0.831
Guise (merged)	1.105	
Voice *(R)*	0.31	0.1
*Age*merged*	-0.025	0.008

Table 104: Merger matched guise ratings for the perceived level of education. Reference levels: female, college educated, distinct, voice J. Random effects: *listener*. *n*= 106

Predictor	LOT	
	Coefficient	*p*
(Intercept)	1.881	
Age	-0.004	0.69
Gender (Male)	1.1	
Education *(No college)* *(Student)*	0.414 -0.585	0.448
Spontaneous LOT F1	0.009	0.353
Spontaneous LOT F2	-0.003	0.512
Shifting F1[82]	-0.002	0.652
Shifting F2	-0.001	0.7
NCS score *(1)* *(2)* *(3)*	-0.653 -0.042 -0.655	0.677
Guise (unshifted)	0.9	
Voice *(R)* *(T)*	-0.084 -0.104	0.864
*Male*unshifted*	-1.2	0.001

Table 105: LOT matched guise ratings for the perceived level of education. Reference levels: female, college educated, NCS score 0, shifted, voice J. Random effects: *listener*. *n*= 159

Thus, it seems that, while the perception of distinct speakers has not changed over time in the community, merged speakers are increasingly perceived as more educated sounding, to the extent that younger listeners rate merged production more favorably than distinct production. How far forward LOT is produced in distinct production also seems to be of relevance, as fronted LOT is perceived as less educated sounding than unfronted LOT, though this only seems to be the case for women. This pattern is in agreement with the results presented above for the production of LOT, as younger speakers, especially those with a college education, are retracting and raising LOT and are moving toward more merged production, especially in more careful speech. However, the regression models presented in Tables 104 and 105 above provide no evidence of a correlation between speech production and ratings. If the retracting of LOT and subsequent merger with THOUGHT is in fact a result of social evaluation, participants with the highest degree of merger would be expected to be ones who rate retracted or merged production more favorably than fronted production. That this does not appear to be the

[82] Shifting here refers to style shifting from spontaneous speech to wordlist style.

case according to the statistical results in Tables 104 and 105 above warrants a closer look at the social distribution of the ratings for both the merger guises and the separate LOT guises.

Although the regression model in Table 104 above provides no evidence for significant effects of *gender* or *education* on the merger ratings, Figure 178 illustrates quite clearly that, when divided by these two factors, there is only one group of listeners that follows the overall rating patterns for merger guises presented above: Males without a college degree tend to rate merged production more favorably than distinct production, while all other groups appear to perceive distinct production as more educated sounding, or do not distinguish between the two variants in their ratings. One potential exception might be younger college educated females. As Figure 178 shows, this group of listeners also seems to slightly favor merged guises over distinct guises.

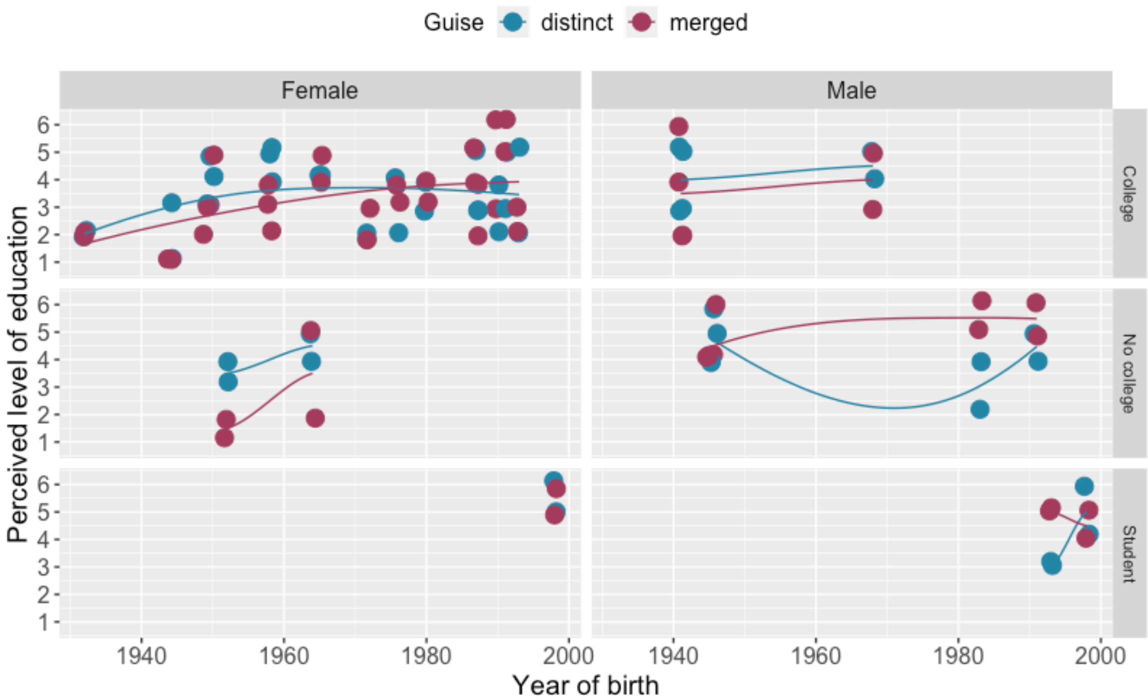


Figure 178: Merger matched guise ratings for the perceived level of education across the listeners' *gender* and *level of education*.

The regression models presented in Tables 106 and 107 below support the observation that males without a college degree and younger college educated females rate merged guises as more educated sounding than distinct guises, though the rating differences are estimated to be three times as high for males without a college degree than for younger college educated females. For neither of the two groups, however, are the rating differences predicted to be statistically significant. Nevertheless, these patterns explain why the regression model in Table 104 above finds no significant correlation between the degree of merger among the raters and their ratings for merged and distinct production. As was described in Chapter 6.2, speakers who show the highest degree of merger, and the most progress toward more merged production, are students and younger college educated females. In their ratings of merged and distinct production, however, these two groups of listeners make only minimal distinctions between the two guises. On the other hand, males without a college degree differentiate between the two guises more clearly, rating merged guises higher than distinct guises. In their own production, however, they do not participate in the merger or in the progress toward more merged production.

Predictor	Coefficient	*p*
(Intercept)	-11.98	
Age	-0.03	1
Minimal pair distance	0.047	1
Minimal pair overlap	9.548	1
Guise (merged)	0.75	0.183
Voice (R)	0.25	0.644

Table 106: Merger matched guise ratings for the perceived level of education among male listeners without a college degree. Reference levels: distinct, voice J. Random effects: *listener*.

Predictor	Coefficient	*p*
(Intercept)	1.599	
Age	-0.021	0.857
Minimal pair distance	0.007	0.403
Minimal pair overlap	3.595	0.398
Guise (merged)	0.25	0.613
Voice (R)	0.25	0.613

Table 107: Merger matched guise ratings for the perceived level of education among female college educated listeners born after 1980. Reference levels: distinct, voice J. Random effects: *listener*.

The ratings for the perceived level of education of fronted and unfronted LOT, on the other hand, resemble the production patterns of each demographic group more closely. As can be seen in Figure 179 below, the majority of raters in each group rates unfronted LOT as the more educated sounding variant, which is in agreement with significant retraction of LOT in production. The only exception to this are males without a college degree, who seem to rate fronted LOT higher than unfronted LOT, which mirrors their production patterns, as they do not seem to participate in LOT retraction in any of the three speech styles. Thus, the social perception of fronted and unfronted LOT does appear to be closely correlated to production patterns, regardless of the lack of statistical significance.

Figure 179: LOT matched guise ratings for the perceived level of education across the listeners' *gender* and *level of education*.

6.4.2 Perceived Localness of Merged and Distinct LOT and THOUGHT

The ratings for the perceived level of localness of the merger guises, illustrated in Figure 180 below, suggest that the difference in how local merged and distinct production is perceived to sound depends on the raters' *age*. Listeners born before 1970 do not appear to distinguish between merged and distinct LOT and THOUGHT to a great extent in terms of how local they sound. Younger participants, on the other hand, appear to differentiate the guises slightly more, and increasingly perceive merged production as less, and distinct production as more local sounding. The regression model presented in Table 108[83] below does not detect an overall significant difference in the ratings for merged and distinct guises. The same model for a subset of speakers born after 1960 predicts a nearly significant difference of 0.7 units on a 6-point scale (p= 0.054). Thus, it appears that younger listeners perceive distinct production of LOT and THOUGHT to be the more local sounding way of speaking.

[83] Again, an interaction between *guise* and *age* does not reach the level of significance, and therefore is not included in the regression model.

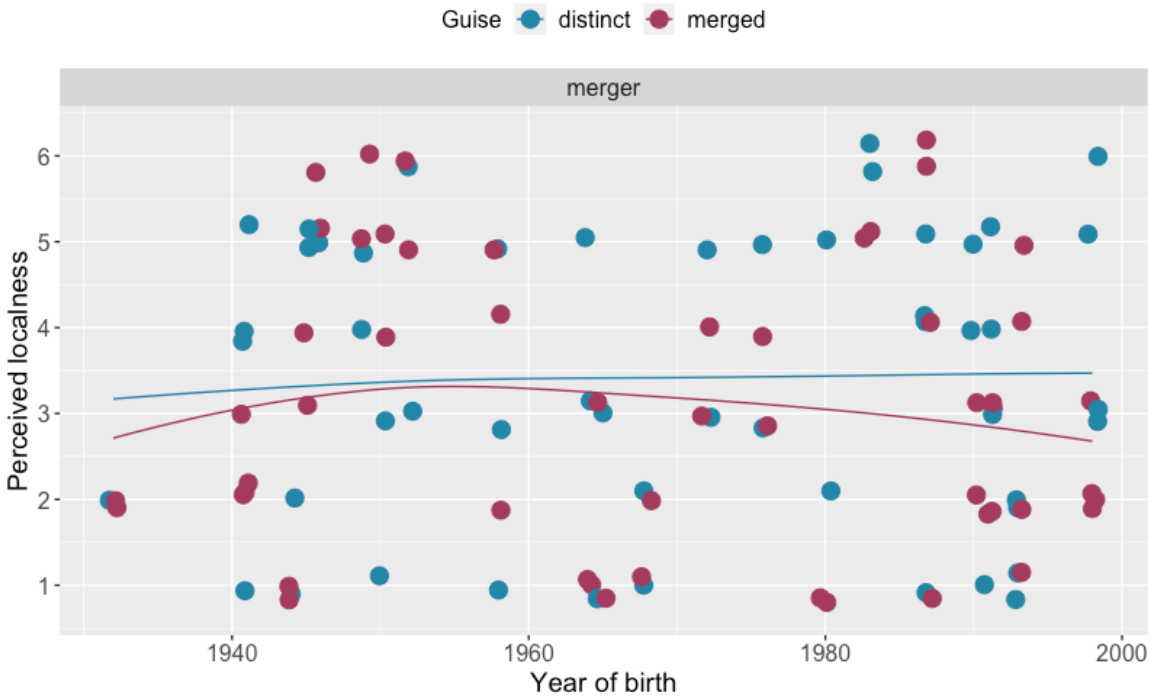


Figure 180: Merger matched guise ratings for perceived localness.

Predictor	merger	
	Coefficient	*p*
(Intercept)	3.358	
Age	-0.004	0.81
Gender (Male)	0.14	0.823
Education *(No college)* *(Student)*	 1.42 0.578	0.123
Minimal pair distance	0.0003	0.915
Minimal pair overlap	-0.875	0.626
Guise (merged)	-0.351	0.169
Voice *(R)*	 -0.143	 0.575

Table 108: Merger matched guise ratings for perceived localness. Reference levels: female, college educated, distinct, voice J. Random effects: *listener*. *n*= 104

In addition to the listeners' *age*, their *level of education* appears to impact their rating patterns for perceived localness of merged and distinct LOT and THOUGHT as well. The regression model presented in Table 108 above suggest that listeners without a college degree as well as students appear to rate both guises higher than college educated listeners. While this in itself is not a particularly relevant or statistically significant find, it does warrant a closer look at the distribution of ratings across the three educational groups.

The ratings for each educational group are visualized in Figure 181. The plot shows that college educated raters as a group do not seem to differentiate between merged and distinct LOT and THOUGHT in terms of how local they sound. Raters without a college degree, on the other hand, appear to differentiate to a much greater extent between the merged and distinct guises, with the exception of older raters in this group. Students appear to be mixed in their perception. While the older one of the three, born in 1993, seems to perceive the merged guises to sound more local, the opposite pattern can be observed for the two younger students, both born in 1998.

Figure 181: Merger matched guise ratings for perceived localness across the listeners' *level of education*.

Regression models that test the effect of *guise* separately for raters with and without a college degree support these observations. For college educated listeners, the estimated rating differences is an insignificant 0.123 (p= 0.677), while the model for listeners without a college degree estimates that distinct guises are rated 1.1 units higher, i.e. more local sounding, than merged guises (p= 0.027). Thus, it appears that the *age* patterns in the ratings for perceived localness of merged and distinct LOT and THOUGHT presented above are mainly accounted for by participants without a college degree as well as by students.

6.4.3 Perceived Friendliness of Fronted and Unfronted LOT

Figure 182 suggests that for the majority of participants, unfronted LOT is the more friendly sounding variant. Although the rating differences appear rather small, they are more or less consistent across different ages, with the exception of the youngest participants.

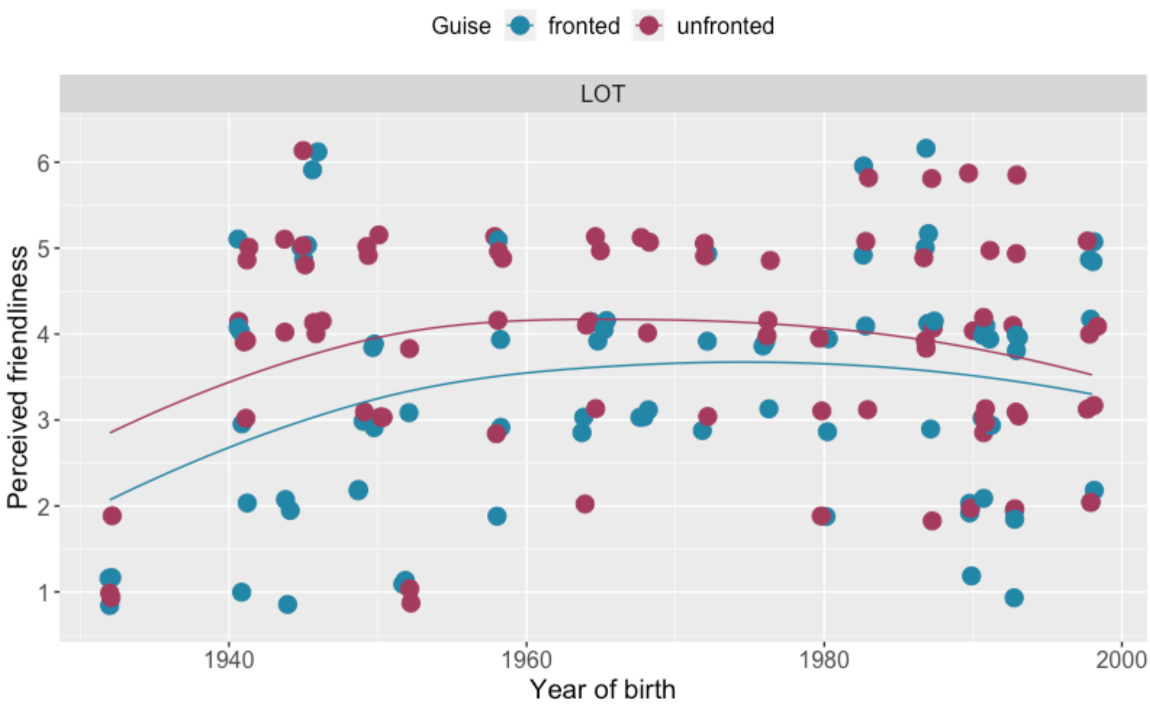


Figure 182: LOT matched guise ratings for perceived friendliness.

The regression model presented in Table 109[84] below estimates that rating differences between fronted and unfronted LOT differ across the three voices (i.e. the people who read the carrier phrases: J, R and T). On average, participants rated unfronted guises 0.7 units higher than the fronted guises for voice J, and 1.1 units higher for voice R. Only for the guises of voice T are there no notable rating differences. Regression models that test for the effect of *guise* separately for each voice confirm the statistical significance of these

[84] Again, an interaction between *guise* and *age* does not reach the level of significance, and therefore is not included in the regression model.

results for voice J (*p*= 0.009) and R (*p*= 0.001), and the lack thereof for voice T (*p*= 0.273). The regression model in Table 109 suggests that ratings differ notably between males and females, raters of different educational backgrounds, and raters with different NCS scores. Although none of these factors are estimated to have a significant effect on the overall ratings, the relatively high estimated coefficients warrant a closer look at the distribution of the ratings for these different groups of listeners.

Predictor	LOT	
	Coefficient	*p*
(Intercept)	0.55	
Age	-0.008	0.557
Gender (Male)	0.727	0.206
Education (No college) (Student)	0.073 -0.801	0.661
Spontaneous LOT *F1*	0.003	0.831
Spontaneous LOT *F2*	0.001	0.923
Shifting F1	-0.003	0.573
Shifting F2	0.002	0.557
NCS score (1) (2) (3)	-0.681 -0.466 -0.862	0.85
Guise (unshifted)	0.704	
Voice (R) (T)	-0.058 0.667	
Voice*unshifted (R) (T)	0.391 -0.992	0.003

Table 109: LOT matched guise ratings for perceived friendliness. Reference levels: female, college educated, NCS score 0, shifted, voice J. Random effects: *listener*. *n*= 157

Listeners of both genders make a similar distinction between both LOT guises. This distinction is somewhat stronger among older raters than it is among younger raters; in fact, only for older female raters are the rating differences between fronted and unfronted guises statistically significant (0.7, *p*= 0.002). For older males, they do not quite reach the level of statistical significance (0.5, *p*= 0.08). Younger raters, on the other hand, do not seem to differentiate to a great extent between fronted and unfronted LOT in their friendliness ratings of the guises regardless of *gender*. On average, male and female listeners born after 1980 rate unfronted guises 0.3 units higher than fronted guises *(p*= 0.317). Thus, the overall pattern of unfronted LOT being the favored variant of LOT in terms of perceived friendliness holds for both genders regardless of *age*, but is statistically significant only for older females.

The raters' *level of education* also appears to influence the ratings of LOT in terms of its perceived friendliness, as shown in Figure 183. It appears that only college educated raters differentiate between fronted and unfronted LOT in their ratings, favoring the unfronted guises of LOT over the fronted guises in its perceived friendliness. A regression model that considers only college educated raters estimates a difference of 0.8 on a 6-point scale between the ratings of the two variants *(p*= 0.0001). Raters without a college degree, on the other hand, do not seem to distinguish between fronted and unfronted LOT at all in terms of perceived friendliness (0.2, *p*= 0.39). The same observation applies to student raters (0.3, *p*= 0.487).

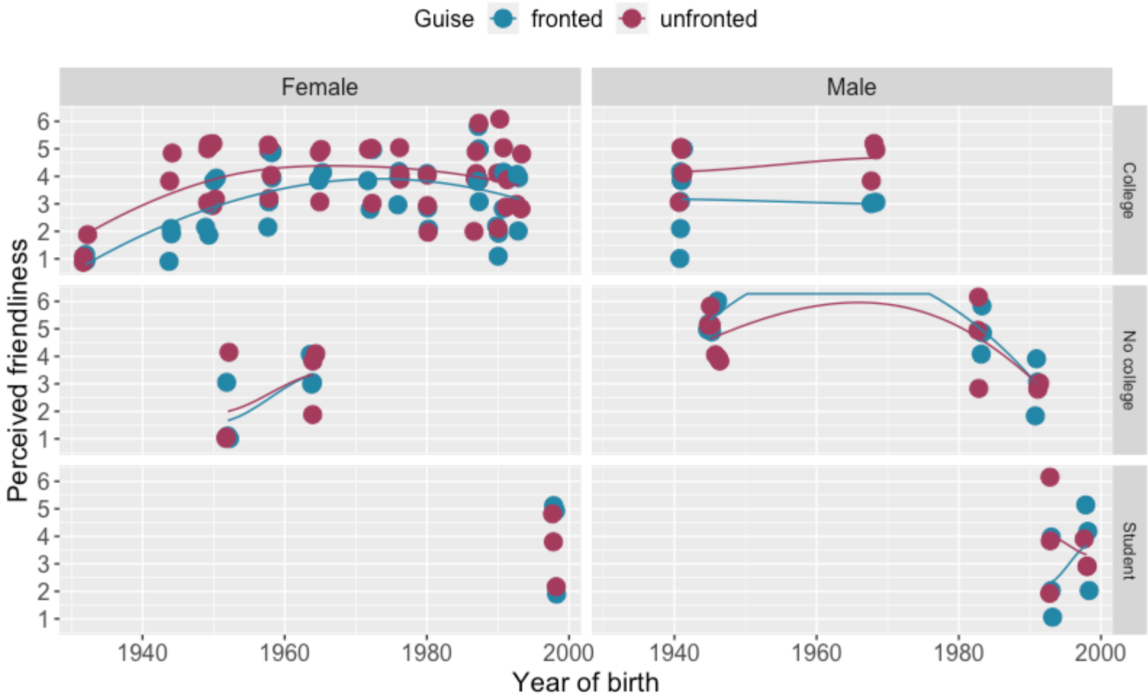


Figure 183: LOT matched guise ratings for perceived friendliness across the listeners' *gender* and *level of education*.

6.4.4 Summary: Social Evaluation

The analysis of the matched guise data has shown that fronted, unfronted and merged realizations of LOT may be subject to a certain amount of social evaluation. The results show that retracted LOT, whether merged with or distinct from THOUGHT, is perceived as more educated sounding than fronted LOT. Additionally, merged production is perceived as less local sounding in comparison to distinct production, and fronted realization of LOT is perceived as sounding less friendly than the unfronted variant.

However, the social distribution of these rating patterns does not necessarily match production patterns. Merged production is rated as more educated and less local sounding primarily by raters without a college degree. These are, however, the same participants who show little to no progress toward the merger in their own production. Participants who do show this progress, i.e. college educated speakers, do not appear to differentiate to a great extent in their ratings of distinct and merged LOT and THOUGHT.

The ratings for fronted and unfronted LOT match production patterns more closely. Unfronted LOT is rated as more educated and more friendly sounding primarily by college educated and female raters, i.e. the same participants who themselves tend to produce more retracted variants of LOT. However, the youngest speakers, who have the most retracted realizations of LOT, do not seem to follow these rating patterns, and instead make no distinctions between the two guises in their ratings.

6.5 Discussion: LOT, THOUGHT and the Merger

The analyses of LOT and THOUGHT, and the distance between and overlap of these two vowel classes have shown that, while the community as a whole is still predominantly distinct, there is clear evidence for a merger in progress, both in production and perception. The bulk of this progress appears to stem from LOT retraction, which not only implies progress toward the merger, but also a reversal of LOT's NCS trajectory. This development is notably different from what has previously been reported for Ogdensburg, as Dinkin (2009) found relative stability of LOT in the community; however, it is in agreement with observations in other Inland North communities, both in and outside of New York State. To a certain degree, the merger also appears to be advanced by the lowering and fronting of THOUGHT, which corroborates Dinkin's (2009) observation of ongoing THOUGHT lowering in the community. However, changes in THOUGHT are noticeable primarily among students, i.e. the youngest participants, and only in more careful speech.

6.5.1 Near Merging Toward a Near-Merger?

The question that remains is which benchmark should be used in the classification of speakers into "merged", "distinct" or, potentially, near-merged. In the present sample, 39 out of 48 speakers have 50% or more overlap of LOT and THOUGHT in spontaneous speech, meaning that 81% of the speakers in this sample are almost fully merged in spontaneous speech by Boberg's (2001) standard (see Chapter 2.6.3). At the same time, only 20 of these 39 speakers have a Cartesian distance between 100 and 200 Hz. Thus, among those with more than 50% overlap is a good number of speakers for which mean LOT and THOUGHT are securely distinct, which raises the question of whether they should be considered as nearly merged speakers or not. There are certainly speakers that can be described as almost fully merged in spontaneous speech, as was shown in Chapter 6.2; however, the majority of speakers would be best described as transitional in this speech style. In the two more careful styles, potential nearly complete mergers appear to be less ambiguous. In wordlist style, four out of six speakers with more than 50% overlap also have less than 200 Hz distances between LOT and THOUGHT; in minimal pairs, it is seven out of eight. While these cases present more robust instances of almost complete mergers, they do seem to be the exception rather than the rule in the community. However, combined with the observation that the merger is advancing faster in apparent time in these two more careful styles than in spontaneous speech, a community-wide progress toward a near-merger seems likely.

Speech style	Distance		Overlap	
	≤ 100 Hz	≤ 200 Hz	≥ 80%	≥ 50%
Spontaneous speech	0	20	3	39
Wordlist style	1	3	0	6
Minimal pairs	3	10	0	8

Table 110: Status of the COT-CAUGHT merger in production in Ogdensburg.

The *age* distribution among the speakers listed in Table 110 above supports this prediction. Of the 20 speakers with over 50% overlap and less than 200 Hz distance, 11 were born before 1980, and nine after 1980. While this indicates an advantage among older speakers, these 11 speakers actually only represent 39% of speakers in this age group, while the nine younger speakers represent 45% of their age group, so that more merged production in this speech style is slightly more common among younger speakers. Likewise, in wordlist style, the four speakers with more than 50% overlap and less than 200 Hz distance are split evenly between speakers born before and after 1980. This makes 7% for older, and 10% for younger speakers. In minimal pairs, this difference is significantly more prominent; 100% of the speakers with more than 50% overlap and less than 200 Hz difference were born after 1980, constituting 25% of the speakers in this age group. Thus, it appears that younger speakers do lead the merger in production, as is common for near-mergers. However, a rapid increase in merger that generally characterizes near-mergers can only be observed in minimal pairs.

The perception data, on the other hand, provide no evidence for a near-merger. As a whole, the community is still predominantly distinct in perception, as summarized in Table 111. The majority of participants judge themselves as distinct in minimal pair production and were able to correctly identify minimal pairs of LOT and THOUGHT. Only one participant judged her production to be the merged, though this judgment is not particularly reliable. Likewise, there is one speaker who falls into the "merged in perception" category based on an identification score of less than 50%. However, it should be noted that the low score is a result of mistaking LOT for THOUGHT and vice versa 100% of the time in the distinct line-up. In the self-commutation test, this participant scored 9/10, despite 60% overlap in her own production, indicating that she clearly is able to correctly categorize LOT and THOUGHT.

Identification scores			*Self-judgment*		
≤ 50% ≤ 10/20	≤ 75% ≤ 15/20	100% 20/20	Merged	Transitional	Distinct
1	5	19	1	13	28

Table 111: Status of the COT-CAUGHT merger in perception in Ogdensburg.

According to the characteristics of a near-merger, especially older participants should be more merged in perception than they are in production. This is, however, not the case in the present sample. Of the 15 speakers who scored lower than 20/20 on the identification task, only five were born before 1980, and only one of these five (born in 1976), scored lower than 15/20. On the other hand, 10 participants born after 1980 scored lower than 20/20, four of them 15/20 or lower, and one lower than 10/20. Likewise, only two of the 14 speakers who do not judge their own minimal-pair production as distinct were born before 1980. Thus, it seems that younger participants not only lead the merger in production, but also in perception, which is not a common pattern for a near-merger.

Furthermore, based on these findings, there is little evidence that merger in perception is preceding merger in production, as has commonly been found with advancing COT-CAUGHT (near-) merger. Instead, it appears that both processes are progressing simultaneously, as those speakers who are the most merged in production tend to be the same speakers who have lost some of their perceptual ability, with only a few exceptions. This is in line with reports of other NCS communities in New York, where progress toward the merger, if any, was observed in production rather than perception, while for Inland North communities outside of New York that show any signs of progress toward the merger, this progress is evident in perception, but not production (*ANAE*, p. 63).

Overall, the notable amount of inter and intra-speaker variation in the production and perception of the merger suggests that the community is progressing toward a near-merger. Younger speakers are clearly in advance of older speakers, both in perception and production. While, in production, the lead of younger over older speakers is substantial only in minimal pair production, this does fit the description of near-mergers being more advanced in minimal pairs than casual speech. While this is not yet the case in the community, it does appear to be developing in this direction. The distinction between LOT and THOUGHT that does still exist in the community is mainly evident on the F2 dimension, another common observation for near-mergers. The only near-merger feature that *cannot* be observed in Ogdensburg is the expected advantage of merger in perception among older speakers. Thus, this study provides evidence that near-mergers may be a common step in the process of the merging of two vowels in a community.

Why perception and production of the merger are progressing differently in New York than in the rest of the Inland North, where perception does seem to be in advance of production, is not clear. The most plausible explanation at this point appears to be its proximity to areas that were among the first to adopt the merger in virtually all directions: Canada to the north and to the west, Pennsylvania to the south, and Northern New England to the east. All of these regions have long been described as merging LOT and THOUGHT in low back position, so that it is possible that frequent and long-standing exposure to LOT in retracted, rather than fronted, position has triggered the retraction of LOT in the New York portion of the Inland North sooner than in other states, alongside weakening perceptive ability. Dinkin (2010) came to the same conclusion, though, like in the present study, there is no strong evidence for this to be the case. His argument was built around the fact that, at that time, LOT retraction seemed to have been unique to the New York section of the Inland North; however, recent research has shown that the same processes were ongoing in other parts of the Inland North during that time.

While there certainly are Inland North communities outside of New York that may have been exposed to retracted LOT and the merger owing to their proximity to Canada, diffusion of phonological features across the US-Canadian border has been found to be much less effective than diffusion within national borders. Thus, border communities such as Detroit may not have been affected by merged production as much as speakers in New York. Additionally, Detroit is located relatively far from the nearest merged urban center, Toronto (about 230 miles). The distance from Buffalo to Toronto, on the other hand, is less than 100 miles. Likewise, the closest urban centers to Ogdensburg are Ottawa and Kingston, both Canadian, both merged. Although the role of Canada and urban centers in general in the diffusion of the merger is questionable, these differences provide some explanations for why Ogdensburg and other Inland North communities in New York may be more advanced in production than other communities in the Inland North, and thus show no difference in the advancement of the merger in production and perception. While the merger may not have diffused to these communities from Canada, increased exposure may at least have contributed to a weakened barrier between both phonemes, and thus helped to lay the groundwork that would enable the merger to occur sooner in production than in other communities without similar exposure.

There are also Inland North communities which, like Ogdensburg, are in closer proximity to merged regions other than Canada. The majority of them are located in the western end of the Inland North, where the merger has been found to be progressing as a result of diffusion from the neighboring State of Minnesota. This progress seems to have affected perception earlier than production for half of the speakers in this part of the Inland North (*ANAE*, p. 63), i.e. the opposite of what has been observed in New York. It is possible that this difference is due to the history of the merger in Minnesota, where it seems to be a much more recent development than in New England, Pennsylvania and Canada. Although parts of Minnesota were reported as merged in 1966, *ANAE* found speakers in this state to be merely transitional. Thus, because the merger in Minnesota is a relatively new development, it may simply not have been established enough to influence the production of speakers in neighboring NCS communities as early as the long-established merger in New England and Pennsylvania did. It was, however, able to weaken their perception.

Thus, while it is possible that greater exposure to the merger in areas where it has been well-established for a significant amount of time affects the relative progress of the merger in perception and production in transitional communities, the evidence for this is not particularly strong, and further research would be required to answer this question definitively.

Experiment design might offer an alternative explanation for the limited degree of merger in perception. As outlined in Chapter 2.4.4, the participants were asked for the self-judgment of minimal pair production after they completed the reading tasks, which ended with the *cot-caught* line-up used for the self-commutation test. Thus, it is possible that the participants' judgments refer predominantly to this particular minimal pair, which, arguably, tends to be marked as distinct more frequently than other minimal pairs of LOT and THOUGHT (Gordon, 2006; Johnson, 2007). Additionally, Gordon (2006) suggests that participants may be reluctant to identify pairs as the same when explicitly asked. Thus, the participants' self-judgment might be understating the degree of merger in perception in this study. However, this applies to only one measure of merger in perception; identification scores based on the commutation test should not be affected by this.

6.5.2 The Diffusion Issue

One question that remains to be discussed is how the merger has reached Ogdensburg. The most likely source of the merger in Ogdensburg appears to be the neighboring North Country, from where the merger seems to be spreading to Ogdensburg through contagious diffusion (see Chapter 1.3.2.). If this is the case, we would expect the "new" phoneme in Ogdensburg to be similar in quality to that of the North Country and Northwestern Vermont, the most likely source for the merger in the North Country. A comparison of the spectral qualities of LOT and THOUGHT in the merger in Burlington, Vermont (Boberg, 2001), the North Country (Dinkin, 2009) and the present data suggests that this is indeed the case. In Burlington, speakers appear to merge LOT and THOUGHT in low back position, with an F1 range between 650 and 900 Hz, and an F2 range between 1050 and 1550 Hz for both vowels. Likewise, the merger in the North Country appears to occur between 750 and 850 Hz for F1, and between 1200 and 1400 Hz for F2. The ranges that have been identified for four nearly merged speakers in the present sample fall securely within these ranges. On the flip-side of this, they are also not notably different from the realization of the merger that has been identified for Canada (*ANAE*; Boberg, 2001) or Pennsylvania (Labov, 2010).

However, if the merger is in fact spreading to Ogdensburg through diffusion, we would expect a different *age* pattern than the one that has become evident in the analysis. In Ogdensburg, like most other communities studied so far, progress toward the merger is led by the youngest speakers in the sample. This pattern, according to Labov (2007) is common for transmission and incrementation; however, as discussed in Chapter 1.3, a change that is adopted through diffusion would be expected to be more advanced among adults rather than adolescents. The same discrepancy was noted for the development of the unraised TRAP system in Chapter 3.4.2 and will be discussed further in Chapter 8.3.1.

6.5.3 The Merger as the Incoming Norm

In its social distribution, the merger in Ogdensburg follows the same patterns that have been identified in previous research of advancing COT-CAUGHT merger. The change appears to be led by college educated, and potentially female speakers. This lead is particularly prominent in perception, as younger college educated females are the only participants in the sample, who, aside from students, describe their own minimal pair production as transitional. This social distribution of the merger in the community supports the assumption that the merger is still in its early stages in the community and suggests that it is adopted as the incoming norm.

Stylistic differentiation in the adoption of the merger lend further support to the assumption that the merger is perceived as the new incoming norm. Progress toward merged production appears to be most vigorous in minimal pair production both in terms of pace and participating social groups. Furthermore, retracted LOT has become the favored variant in more careful speech, a reversal that seems to have occurred around 1960. While, in the LOT production data in this study, 1960 was not found to be a significant point in time, it has been pointed out in previous studies as the cut-off year for LOT retraction (Dinkin, 2009, 2011). While the validity of conclusions drawn from style-shifting patterns has been questioned based on the argument that merged production in minimal pairs might merely be the result of formal testing, the progress toward merged production in minimal pairs in Ogdensburg is validated by the same, albeit less robust, developments in the two other speech styles. Thus, while the majority of speakers in

Ogdensburg are *less* merged in minimal pairs than they are in spontaneous speech, the fact the transition toward the merger in minimal pairs is the most robust compared to the other two speech styles may be an indication of the merger being perceived as the incoming norm.

This tentative conclusion is supported, at least partially, by the matched guise data. The ratings for fronted, unfronted and merged LOT suggest that NCS-fronted LOT is perceived as less educated, i.e. less standard sounding than unfronted and merged LOT, as well as less friendly than unfronted LOT. Thus, it appears that the retraction of LOT and the resulting progress toward the merger may be triggered by a more favorable perception of retracted and/or merged LOT. If LOT retraction, much like the low TRAP system, is in fact diffusing to Ogdensburg as the new standard from the North Country, there should be evidence of a positive attitude toward the North Country dialect area among Ogdensburgers, which will be explored further in Chapter 8.3.1.

Interestingly, the evaluation data does not necessarily correlate with the speakers' production. For fronted and unfronted LOT, the rating patterns generally match the patterns observed in the production of LOT among the participants: Those who rate backed LOT more favorably are also those who produce it further back in the vowel space. The ratings for the merger guises, on the other hand, do not show this correlation. Here, it is the group who shows the least amount of merger in production that rates merged guises more favorably, while those who are more merged in production do not differentiate between merged and distinct in their ratings. This mismatch is most likely the result of methodological choices and the limits of social evaluation of phonological variation. In Chapter 1.5, I introduced Labov's hypothesis that listeners do not react to or evaluate phonological distinction, or the lack thereof; but instead respond to perceived differences in phonetic realization. Because the merger guises in the present study were created by replacing instances of THOUGHT with a nearly merged LOT stimulus (see Chapter 2.3.2.1), it is possible, if not expected, that participants reacted to a perceived fronted, lowered and unrounded variant of THOUGHT, rather than to the fact that LOT and THOUGHT sounded the same in the merged guises. Thus, it is feasible that the ratings for the merger guises reflect the social perception of shifted, i.e. lowered, fronted and unrounded THOUGHT, rather than that of merged vs. distinct production. In this case, the rating patterns correlate with the speakers' production more closely. Participants who differentiate their merger ratings, i.e. react to lowered and fronted THOUGHT, are the same participants who, in their own production, have a somewhat fronted THOUGHT. Participants without fronted THOUGHT, on the other hand, do not appear to distinguish between shifted and unshifted THOUGHT.

There is, however, at least one piece of evidence that suggests that speakers are in fact aware of a distinction vs. a lack of distinction between LOT and THOUGHT. To the question whether the minimal pairs she was asked to read and listen to sound different, Monica (born in 1965) responded:

> “*Yes, they're always different here. Not always everywhere, though.*”

In Chapter 6.4 above, Monica was also quoted as associating lowered and fronted THOUGHT in *caught* at least partially correctly with Boston and, for some reason, with New York City, where the merger is certainly not a traditional feature. Thus, while Monica is the only one of the participants in Ogdensburg to comment specifically on the lack of a LOT-THOUGHT distinction in other dialect areas, this indicates that awareness of underlying phonological variation is not necessarily completely outside the realm of possibility. This example provides the kind of justification for future experimental studies testing people's perceptual abilities, including social responses to mergers, which, according to Labov (2001, p. 343), has been missing from the literature.

If the participants in this study did in fact react to the presence and lack of distinction in the merger guises rather than to a lowered and fronted realization of THOUGHT, the question of why rating patterns do not match production patterns remains unanswered. One possible explanation is that, while being able to perceive both phonological and phonetic change, listeners might react differently to the two concepts. Thus, while LOT retraction (entailing less distinction to THOUGHT) may be evaluated in a certain way, the presence and absence of distinction between LOT and THOUGHT may be subject to a completely different evaluation. In other words, the LOT guises and merger guises are asking for the evaluation of two different concepts, and the results are therefore not comparable. A second potential answer to this question might be that speakers who favor merged production in their ratings, i.e. male speakers without a college degree, simply do not wish to participate in this particular speech pattern in spite of perceiving it as sounding more educated, i.e. more standard – or potentially because of it.

Because of the asymmetry in the ratings and production of the merger, the ratings for LOT may be a more reliable indication of the social perception of the different realizations of this vowel at this point. Since retracted LOT is rated more favorably than fronted LOT in terms of the perceived level of education and perceived friendliness, it could be assumed that the same applies to merged production, so long as speakers do not actually react to merged or distinct production. This conclusion is similar to that of previous research on the social perception of LOT and the COT-CAUGHT merger, which has found that merged speakers are perceived as sounding more standard, being more successful and having a more favorable personality.

Whether participants react to the phonetic realization of one member of a merger in merger guises rather than to the identical realization of both members, and whether the social perception of *retracted* LOT can be equated with the social perception of *merged* LOT remains a question that will have to be dealt with in future research. To test whether the ratings for the merger guises were actually reactions to lowered and fronted THOUGHT, a separate analysis of the social evaluation of variants of THOUGHT are necessary, which, owing to a lack of data in both the present study as well as previous research cannot be accomplished here. While *raised* THOUGHT in other parts of the country, e.g. New York City, is known to be highly stigmatized far beyond city limits (e.g. Becker, 2014), so far, there is no evidence in the literature suggesting that *lowered* and *fronted* THOUGHT has attracted any kind of social meaning, and because this study did not include separate guises for THOUGHT, this cannot be tested further at this point. The field would benefit from more advanced methodologies in the elicitation of subjective reactions toward mergers, which allow for the possibility to discern whether the participants' responses refer to the altered phonetic realization of sound A or sound B, or to the fact that sound A and sound B are identical.

Chapter 7: What *Else* Is Going On?

7.1 The Elsewhere Shift – An Introduction

Labov (1991) labelled North American dialect regions outside the North and the South the "Third Dialect", which he described as having a fairly stable vowel system. However, in some regions included in this Third Dialect, most notably California and Canada, research has reported a vowel shift that has come to be known as the California Shift and/or Canadian Shift. Both shifts are assumed to be triggered by the COT-CAUGHT merger, whereby the retraction of LOT toward THOUGHT creates a vacuum in the low front corner of the vowel space, causing TRAP to retract and DRESS and KIT to lower and/or retract. It has, however, also been argued that, instead of a drag chain triggered by the COT-CAUGHT merger, the shift is operating as a pull chain, initiated by the lowering of KIT, independent of the merger (Kennedy & Grama, 2012). Other changes commonly observed in communities undergoing this shift are the fronting of GOOSE, FOOT, STRUT and GOAT (e.g. Clarke et al., 1995; Eckert, 2011; Fought, 1999; Kennedy & Grama, 2012). GOOSE fronting, in combination with TRAP retraction, has been found to lead to a reversal of the relative position of these two vowels on the front-back dimension (Boberg, 2011).

Both the California and the Canadian Shift are marked by the features outlined above; however, there are some subtle differences, which is why some scholars insist on differentiating between the two (e.g. Kennedy & Grama, 2012). For example, while a low nasal or continuous TRAP system, i.e. raised pre-nasal TRAP and low pre-oral TRAP, is commonly found in California (e.g. Podesva, D'Onofrio, Van Hofwegen, & Kim, 2015), TRAP in Canada has been described as lowered and retracted regardless of phonological environment. In California, the centrally merged LOT/THOUGHT vowel has been found to be retracting and raising (D'Onofrio et al., 2016), leaving pre-oral TRAP as the lowest vowel in the system, which is similar to the Canadian pattern (Boberg, 2011). In Canada, however, there is evidence of increasing rounding of the low back merged LOT/THOUGHT vowel (Woods, 1993).

Although the Third Dialect was initially assumed to have little influence on other varieties American English (Labov, 1991), the vowel shift that characterizes it has reportedly been spreading across the US, to the extent that it is now commonly referred to as the Elsewhere Shift. This shift has, for example, been reported in Vermont (Boberg, 2001), Ohio (Durian, 2012), Southern Illinois (Bigham, 2009), Kansas (e.g. Kohn & Stithem, 2015), South Carolina (Baranowski, 2013) and Alaska (Bowie et al., 2012). Outside of North America, Hickey (2018) reports that similar developments have been observed in Ireland (Hickey, 2016), Scotland (Holmes–Elliot & Smith, 2015), South Africa (Chevalier, 2016) and Australia (Cox and Palethorpe, 2008, 2012).

In recent years, the Elsewhere Shift and the co-occurring fronting of the back vowels appear to have been affecting Inland North communities as well. Here, communities have been found to be developing a low nasal or continuous TRAP system (e.g. Morgan et al., 2017; Wagner et al., 2016), backing LOT (Nesbitt & Mason, 2016; Morgan et al., 2017), backing DRESS (Nesbitt & Mason, 2016; Morgan et al., 2017; Wagner et al., 2016), fronting STRUT (e.g. Wagner et al., 2016), fronting GOOSE (*ANAE*; Gordon, 2001; Labov, 1994; McCarthy, 2007; Morgan et al., 2017; Nesbitt & Mason, 2016; Wagner et al., 2016) and fronting GOAT (Nesbitt & Mason, 2016; McCarthy, 2007). KIT retraction has only been observed in parts of Michigan (Morgan et al., 2017), not including Lansing (Wagner et al., 2016).

In New York, some of these developments toward the Elsewhere system have been observed as well. LOT is retracting all over New York State, and in Rochester and several smaller Inland North Fringe communities, there is evidence for an allophonic split into pre-nasal and pre-oral TRAP (i.e. a low continuous or nasal TRAP system) (Dinkin, 2009, 2011, 2013; Driscoll & Lape, 2015; King, 2017). The COT-CAUGHT merger, however, has been reported as being absent in those communities (Dinkin, 2009, 2010, 2019; King, 2017), although Dinkin has found evidence of an incipient merger. In Syracuse, Driscoll and Lape (2015) also report STRUT fronting.

Most research into the Elsewhere Shift and the associated changes in the back vowels have found similar social distributions. Most changes have been found to be in progress, led by younger speakers, and primarily by females (e.g. *ANAE*; Boberg, 2011; Clarke et al., 1995; D'Onofrio et al., 2016; Kennedy & Grama, 2012) and college educated speakers (*ANAE*). Whether this holds true for Inland North communities remains to be investigated. So far, a female lead in the Inland North has only been reported for GOOSE fronting (Clopper et al., 2005; Labov, 1991).

In the chapters above, it has been found that the vowels involved in the NCS are undergoing significant restructuring in Ogdensburg: LOT is retracting toward a merger with THOUGHT, TRAP is developing into a low nasal or continuous system, DRESS is backing and lowering and STRUT appears to be fronting. In combination, these changes indicate that the reconfiguration of the vowel system might be orienting toward the Elsewhere system, the only exception being KIT, which shows little to no variation in the community, and has, if anything, been *fronted* in recent years, countering both NCS and Elsewhere trajectories. Three missing pieces that have not yet been discussed are GOOSE, FOOT and GOAT. While Dinkin (2009) did report substantially centralized GOOSE in Ogdensburg, the status of FOOT and GOAT is entirely unknown. All three of these variables will be analyzed in the following subchapters. Because neither of the three vowels are involved in the NCS or in the COT-CAUGHT merger and were therefore not originally intended to be target vowels in this study, they were only recorded in spontaneous speech in the 2016 recordings. Thus, *speech style* becomes redundant and will not be considered in the analyses. Likewise, neither GOOSE, FOOT nor GOAT were included in the matched guise experiment, their potential social evaluation will not be analyzed here.

7.2 Results: GOOSE

Figure 184 below suggests a good amount of inter-speaker variation in the realization of GOOSE, especially in terms of frontness. In terms of height, the majority of speakers realize GOOSE at a similar height, with F1 means between 400 and 500 Hz; F2, however, varies by over 600 Hz. Some older speakers have F2 means of about 1500 Hz or less, while others, especially younger ones, have fronted GOOSE as far as 2000 Hz or more. The majority of speakers, regardless of *age*, realize GOOSE front of center. As the name implies, frontness is the more relevant dimension in GOOSE fronting; thus, the analysis below will focus on changes in F2, both in apparent and real time.

Figure 185 below suggests that GOOSE is fronting quite rapidly in apparent and real time. The oldest speakers in 2008 produced GOOSE with an F2 of less than 1100 Hz, while the younger speakers in this sample average around 1500 Hz. This average is the same with which the oldest speakers in the 2016 sample realize GOOSE, indicating notable real-time fronting of the vowel. Younger speakers in the 2016 sample appear to have advanced this change even further, producing GOOSE with F2 means around 1900 Hz. While all speakers seem to be participating in GOOSE fronting, college educated speakers do appear to lead this change with a slight advantage over speakers without a college degree.

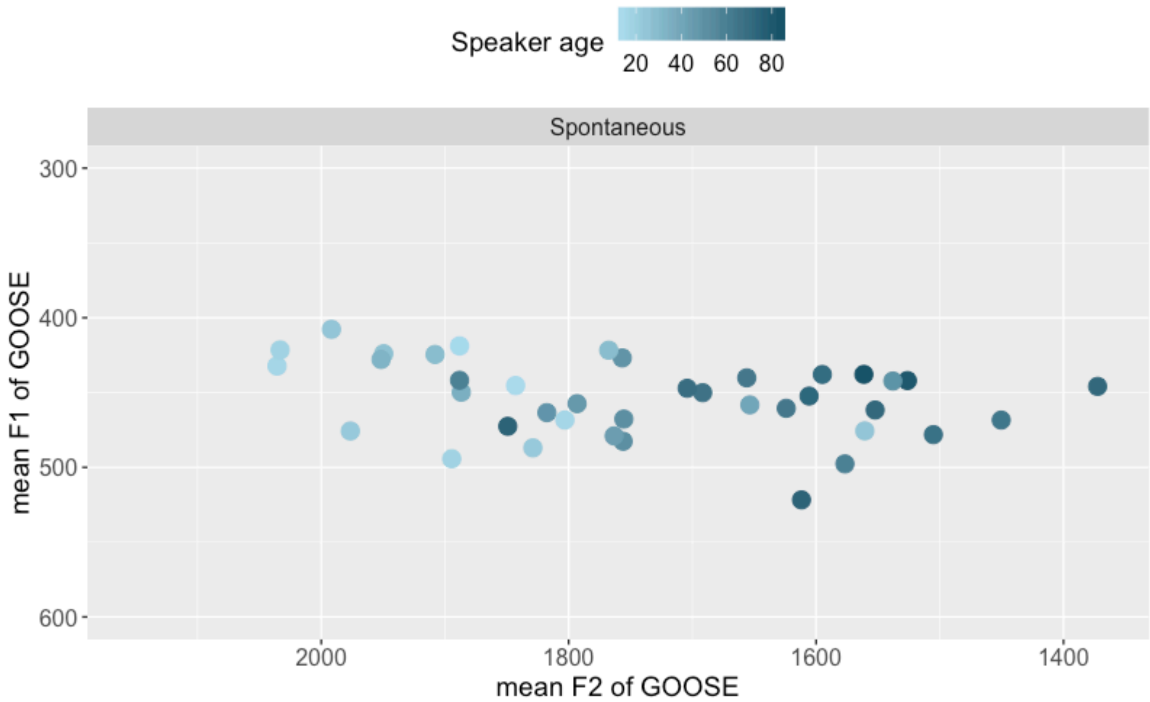


Figure 184: GOOSE F1 and F2 means in 2016 by *age*. Lighter shades represent younger speakers, darker shades older speakers.

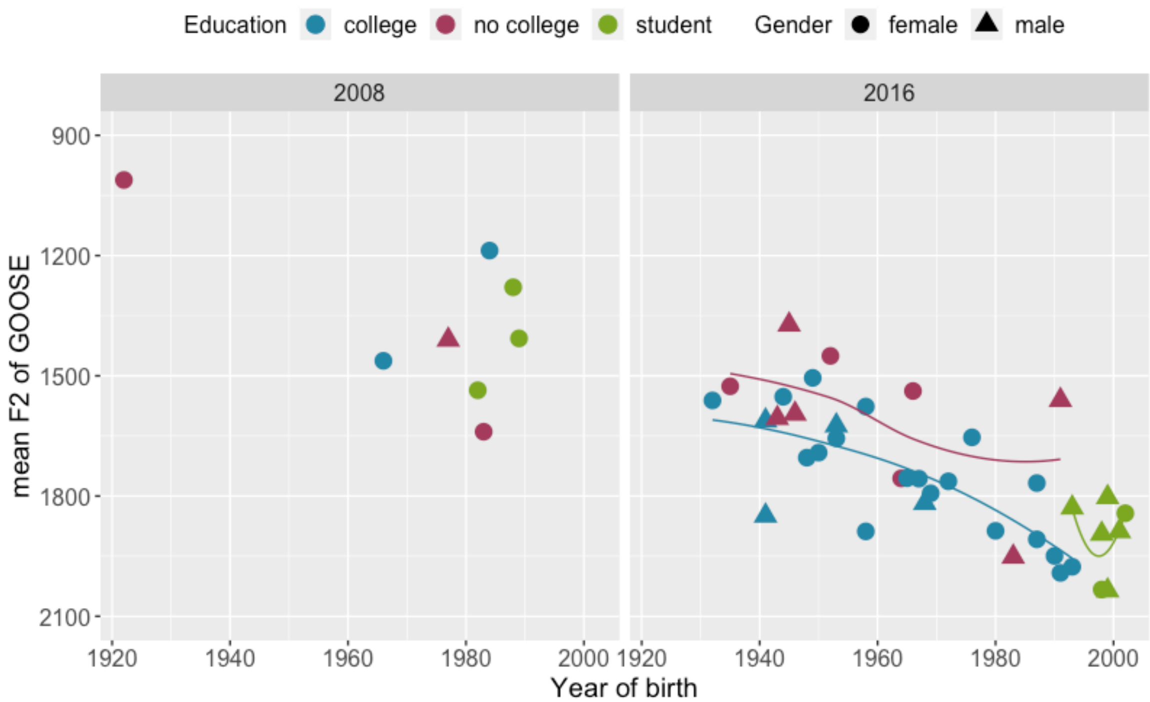


Figure 185: GOOSE F2 means in 2008 and 2016 by *education* and *gender*.

The regression model in Table 112 corroborates the observations presented above. In the combined sample, F2 of GOOSE is increasing significantly in apparent time, by an estimated 32 Hz per 10 years. In the 80 years tracked in the data, this amounts to nearly 250 Hz. The bulk of this trend seems to stem from the 2016 sample (3.5 Hz, p= 0.0002), as the effect of *age* in 2008 is not found to be statistically significant (1.6 Hz, p= 0.57). On the other hand, the lack of a significant interaction between *age* and *sample year* suggests that the rate of fronting does not differ to a great extent between the two samples. Apparent-time fronting is supported by significant real-time differences, with 2016 speakers producing GOOSE an estimated 160 Hz fronter in the vowel space than 2008 speakers. Otherwise, the frontness of GOOSE does not appear to be influenced by social factors to a significant extent; the difference between speakers of different *educational backgrounds* is not significant.

Predictor	Coefficient	*p*
(Intercept)	1783.816 Hz	
Age	-3.235 Hz	0.0003
Gender (Male)	1.428 Hz	0.963
Education *(No college)* *(Student)*	 -39.942 Hz 12.016 Hz	0.441
Sample year *(2008)*	 -161.597 Hz	 0.001
Environment		$5x10^{-7}$

Table 112: GOOSE F2 in 2008 and 2016. Reference levels: females, college educated, 2016, /p/. Random effects: *speaker, word*. n= 3168[85]

7.3 Results: FOOT

Figure 186 below suggests that there is some inter-speaker variation in the frontness of FOOT. While the majority of speakers in the 2016 sample appear to realize FOOT just back of center, some speakers have F2 means of less than 1300 Hz, and others have fronted FOOT somewhat, to an F2 greater than 1600 Hz. However, no apparent *age* pattern in this distribution emerges from Figure 186 below, i.e. there is no evidence that FOOT is fronting in apparent time. In fact, the speaker with the backest FOOT F2 mean appears to be one of the youngest in the sample.

---

[85] Though not discussed here, the linguistic constraints for GOOSE fronting in the community seem, as might be expected, quite strong for the speakers in this sample (see Appendix H).

The regression model in Table 113 supports the absence of a significant *age* correlation. In fact, none of the social factors tested are found to be significant, although males are estimated to realize a slightly fronter FOOT than females. In terms of height, the majority of speakers realize FOOT at a similar height, with F1 means between 500 and 600 Hz. Whether or not FOOT has undergone changes in real time cannot be assessed here, as I did not recode FOOT in the 2008 data to match the 2016 data.

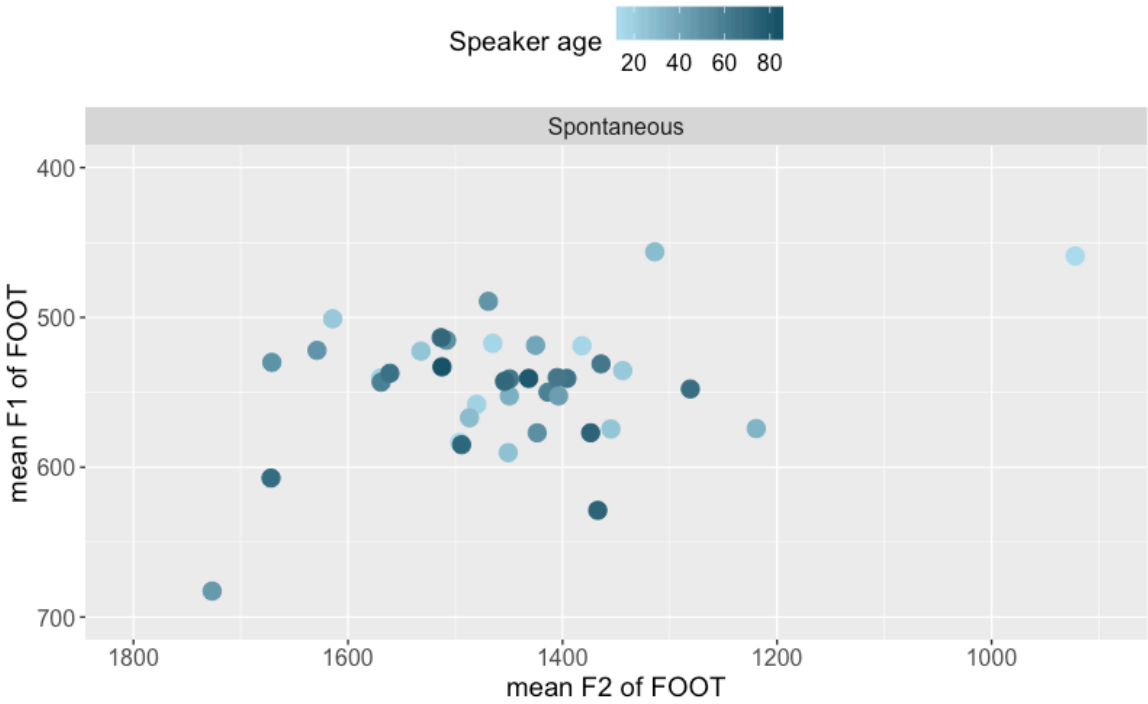


Figure 186: FOOT F1 and F2 means in 2016 by *age*. Lighter shades represent younger speakers, darker shades older speakers.

Predictor	Coefficient	*p*
(Intercept)	1261.064 Hz	
Age	0.126 Hz	0.874
Gender (Male)	83.252 Hz	0.007
Education *(No college)* *(Student)*	 7.141 Hz 47.388 Hz	0.649
Environment		0.0002

Table 113: FOOT F2 in 2016. Reference levels: females, college educated, /p/. Random effects: *speaker, word*. *n*= 772

7.4 Results: GOAT

For GOAT, measurements from multiple measurement points (20%, 35%, and 50%) will be considered in the analysis.[86] Because multiple-point measurements are not available for the 2008 data, the analysis of GOAT will be based on the 2016 sample only.

Figure 187 suggests that, much like GOOSE, GOAT is subject to some variation in the community, particularly on the front-back dimension. In terms of height, the majority of speakers realize GOAT at a similar height, with F1 means between 600 and 640 Hz at 20% vowel duration, i.e. the onset of the vowel. F2, however, varies considerably, ranging from 1100 Hz to nearly 1400 Hz. This variation in F2 will be explored further below.

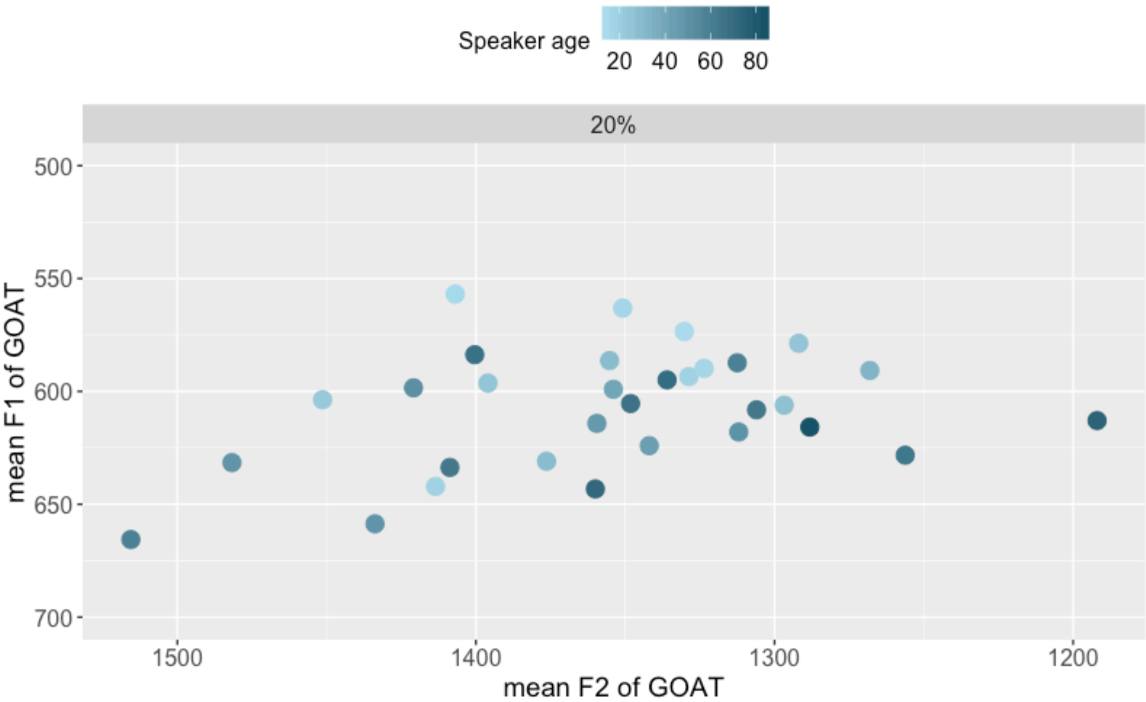


Figure 187: GOAT F1 and F2 means at 20% vowel duration in 2016 by *age*.

Figure 188 below suggests that GOAT fronting is conditioned by the speakers' *level of education* and *age*. It seems as though the onset of GOAT was fronted by college educated speakers born before 1970, but that it has remained relatively stable for this group since then. For speakers without a college degree, on the other hand, there is no indication of

[86] Because diphthongs were not originally part of this study, formant measurements at the 20%, 50%, 65% and 80% marks were not renormalized after data cleaning. For reasons of consistency, un-renormalized measurements will be used for all measurement points in this analysis of GOAT.

GOAT fronting, and as a result, they appear to produce a somewhat backer GOAT than speakers with a college degree. Students appear to have followed the trend of college educated speakers in the community.

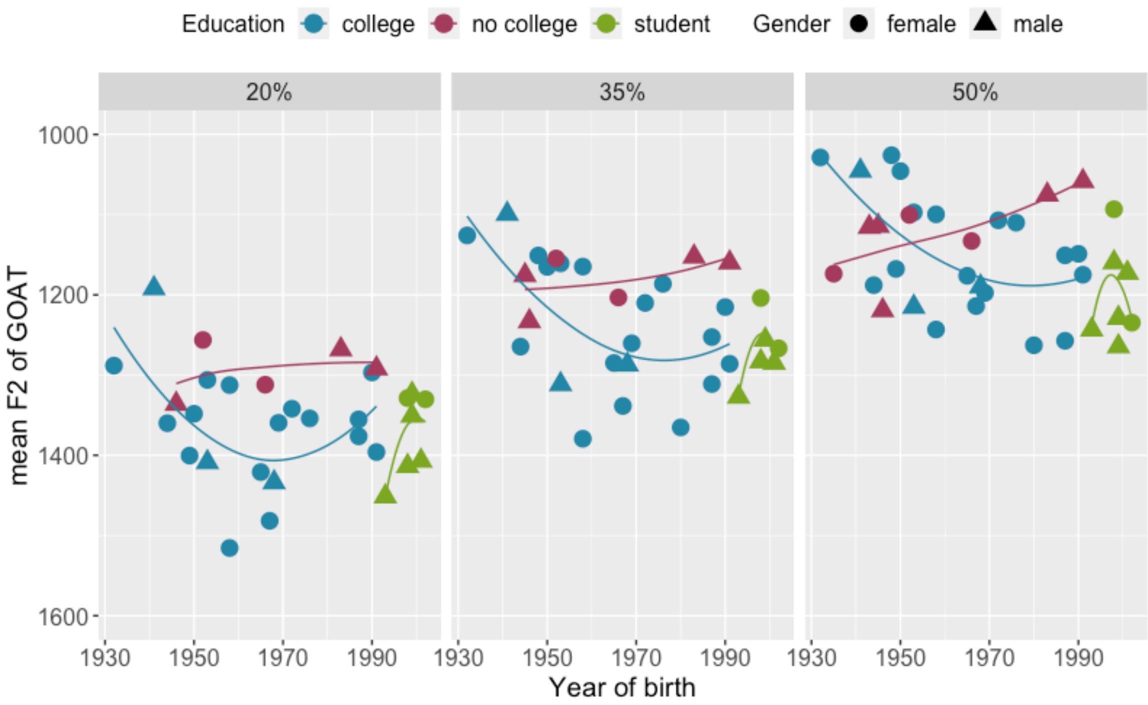


Figure 188: GOAT F2 means at different measurement points in 2016 by *education* and *gender*.

The regression models in Table 114 below support these observations. At all three measurement points, i.e. 20%, 35%, and 50%, college educated speakers are found to produce a notably fronter GOAT than speakers without a college degree. Additionally, the significant interactions between *age* and *education* at the 35% and 50% marks confirm that apparent-time developments differ between these two groups. In models that consider only college educated speakers, the estimated changes amount to 17 Hz per 10 years at the 20% mark (p= 0.045), 22 Hz per 10 years at the 35% mark (p= 0.009), and 23 Hz at the 50% mark (p= 0.006).[87] For speakers without a college degree, on the other hand, no notable changes over time are predicted at 20% (0.36 Hz, p= 0.618) or 35% (0.64 Hz,

[87] In models that consider only college educated speakers born before 1970, these numbers increase to 25 Hz per 10 years at the 20% mark (p= 0.122), 29 Hz per 10 years at the 35% mark (p= 0.094), and 28 Hz at the 50% mark (p= 0.086), but lose their statistical significance.

p= 0.319). At 50%, GOAT is predicted to be *backing* at a significant level for this group (1.73 Hz, *p*= 0.027).

Predictor	20%		35%		50%	
	coefficient	*p*	coefficient	*p*	coefficient	*p*
(Intercept)	1352.149 Hz		1191.432 Hz		1171.075 Hz	
Age	-1.731 Hz		-2.122 Hz		-2.113 Hz	
Gender (Male)	-7.737 Hz	0.765	5.89 Hz	0.816	27.452 Hz	0.269
Education (No college)	-160.874 Hz		-205.355 Hz		-254.736 Hz	
*Age*No college*	1.862 Hz	0.168	2.747 Hz	0.041	3.903 Hz	0.004
Environment		0.073		0.019		0.006

Table 114: GOAT F2 in 2016. Reference levels: females, college educated, 2016, /p/. Random effects: *speaker, word*. *n*= 6965

7.5 Summary: GOOSE, FOOT and GOAT

The results regarding the frontness of GOOSE, FOOT and GOAT suggest that two of the three vowels, GOOSE and GOAT, have undergone notable changes in the community.

GOOSE shows significant ongoing apparent and real-time fronting. In both data sets, GOOSE is fronting rapidly in apparent time, and these developments are corroborated by real-time evidence, as 2016 speakers have notably greater F2 means than 2008 speakers, i.e. produce a notably fronter GOOSE, as can be seen in Figure 189.

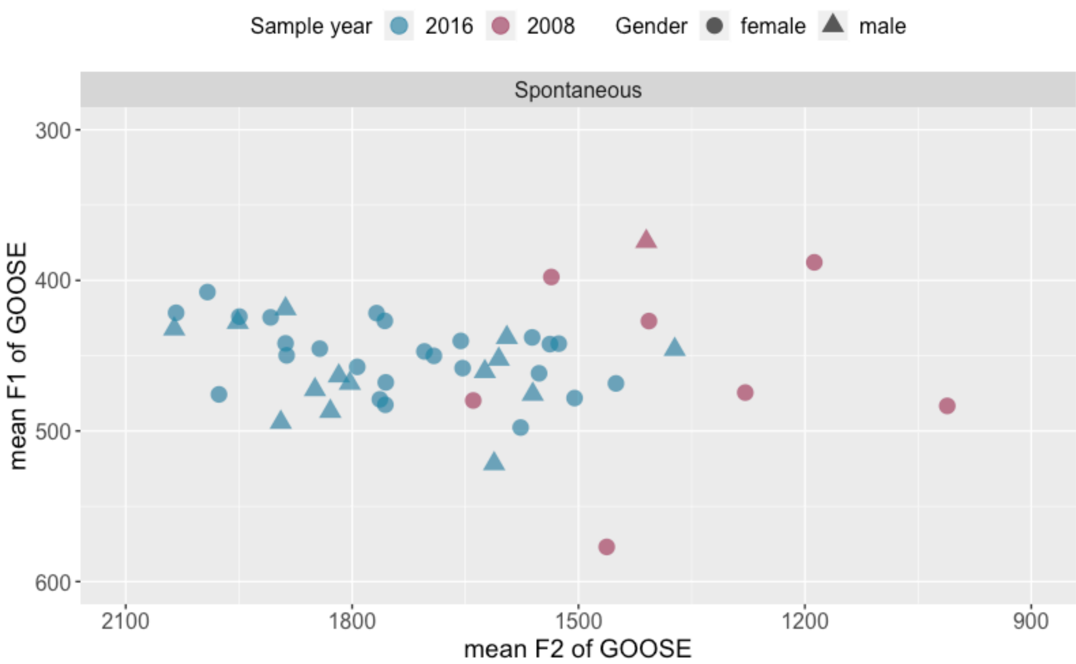


Figure 189: GOOSE F1 and F2 means in 2008 and 2016 by *gender*.

GOAT has been fronted by some speakers in the sample, though this change appears to have been conditioned by their *level of education*. For speakers without a college degree the onset of GOAT remains relatively far back and does not appear to be fronting in apparent time. For college educated speakers, the data indicates an apparent time fronting of the GOAT onset that seems to have amounted to about 100 Hz. However, it appears that this change occurred for speakers born before 1970, and that GOAT has maintained its somewhat fronted position since then for this group of speakers and continues to be realized with similar qualities by the students in the sample.

7.6 Discussion: The Elsewhere Shift in Ogdensburg

The results presented above for the fronting of GOOSE and GOAT are largely in agreement with reports of both developments in the literature. Both vowels have fronted at a statistically significant level in apparent time. However, this appears to have been a much faster process for GOOSE, while GOAT was fronting at a slower rate, a find that has been reported for other communities underdoing GOOSE and GOAT fronting as well. Furthermore, while GOOSE is realized front of center by the majority of speakers, the GOAT onset remains back of center, which appears to be common for North American English as well (*ANAE*).

Contrary to other communities, the fronting of GOAT appears to have come to a halt in Ogdensburg. This is likely due to GOAT having fronted as far as it possibly can. *ANAE* suggests that the F2 limit for GOAT is 1400 Hz in triangular vowel systems, and that few speakers exceed an F2 of 1550 Hz for GOAT. Although these absolute numbers are based on a different normalization system than the measurements in the present study and can therefore not form the base for direct comparison, in relative terms, they suggest that GOAT generally does not move front of center, which seems to be the reason for discontinued fronting of GOAT in Ogdensburg.

GOOSE, on the other hand, continues to front in apparent time. In combination with TRAP retraction (see Chapter 3.2) this appears to have led to a reversal of the relative F2 positions of these two vowels, which has been reported for other communities as well. Figure 190 below shows that the F2 distance between GOOSE and TRAP has been decreasing continuously in Ogdensburg, to a point at which the youngest speakers in the combined 2008 and 2016 sample realize GOOSE in a fronter position than TRAP.

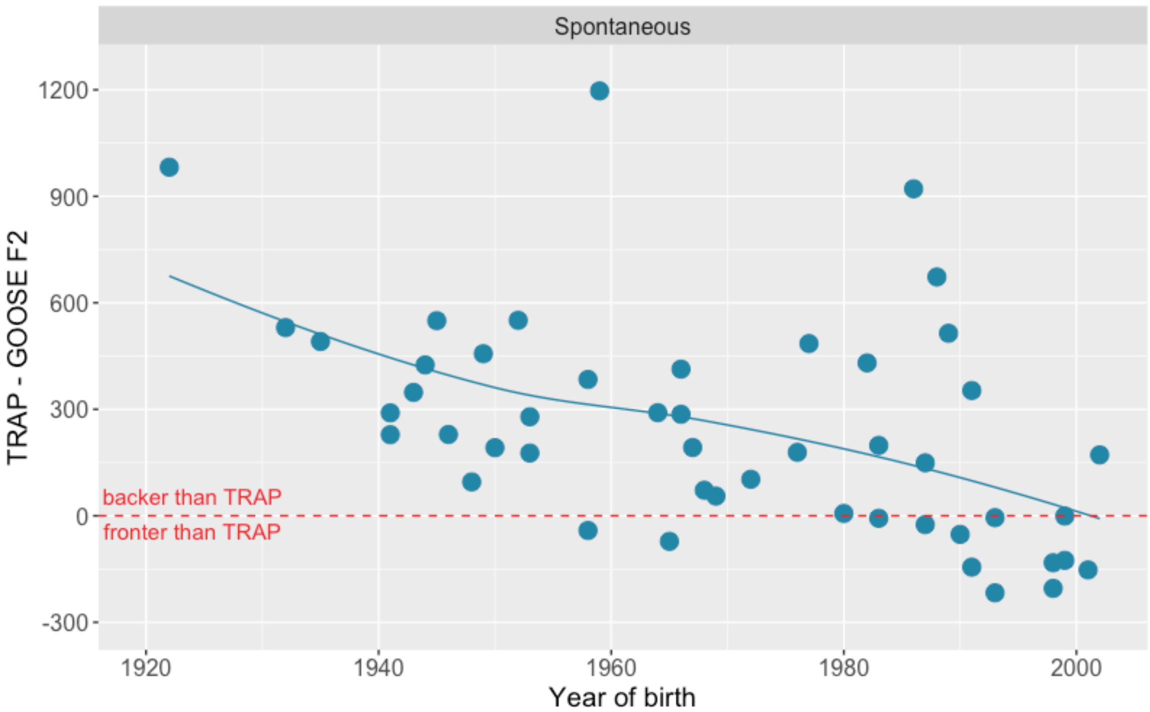


Figure 190: F2 distance between GOOSE and TRAP in spontaneous speech. A negative value indicates that mean GOOSE is fronter than mean TRAP.

Contrary to other reports of GOOSE and GOAT fronting, the social stratification of these two changes is relatively limited. Neither of the two developments show the female lead that has been reported in virtually all other communities, and only the fronting of GOAT appears to have been led by college educated speakers, which *is* in line with earlier reports.

These developments in GOOSE and GOAT, and the lack thereof in FOOT, help to answer the question of whether or not Ogdensburg is orienting toward the Elsewhere system. GOOSE fronting alone, of course, is not a particularly strong indicator of the development of the Elsewhere system, as it has been found to be common to numerous North American dialects, including the Inland North at the height of the NCS (*ANAE*). Regarding GOAT, on the other hand, the Inland North has been described as being resistant to fronting (*ANAE*), which no longer seems to be the case as the data in the present study suggest. In combination with the restructuring of TRAP, continued lowering and centralization of DRESS, progress toward the COT-CAUGHT merger and fronting of STRUT, the possibility of the development of an Elsewhere-like pattern cannot be ruled out. In fact, the vowel systems of some of the most advanced speakers do resemble a typical Elsewhere system. Figures 191 and 192 below show the vowel systems of Allison, a recent college graduate born in 1993, and Jason, a recent high school graduate born in 1998, in spontaneous speech.

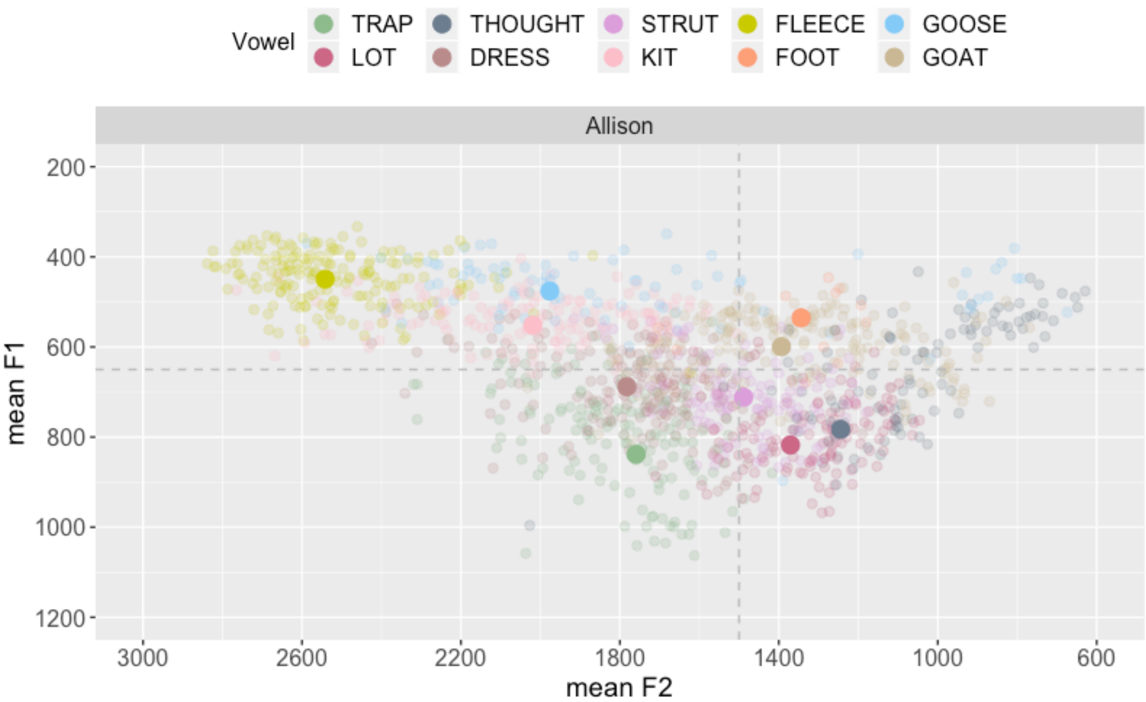


Figure 191: The vowel system of Allison, born in 1993, in spontaneous speech.

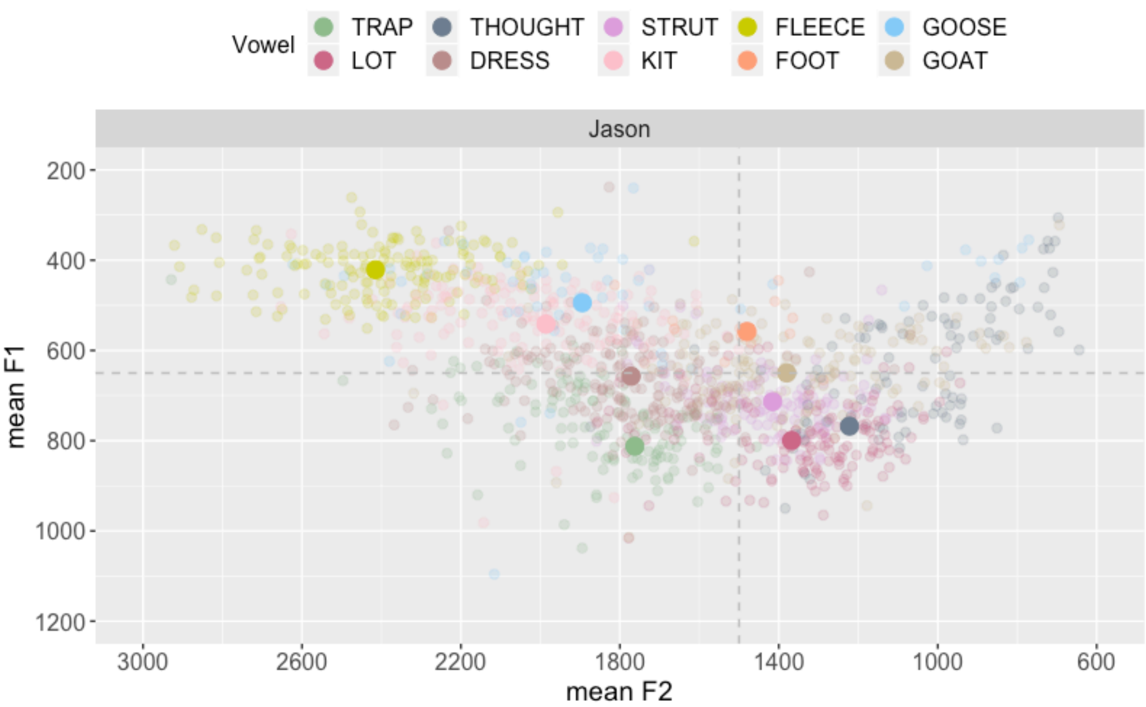


Figure 192: The vowel system of Jason, born in 1998, in spontaneous speech.

For both speakers, LOT is somewhat raised, creating a good amount of overlap with THOUGHT, leaving some of their TRAP tokens, and TRAP means, to occupy the lowest positions in their vowel space, in a position that is backer than DRESS. GOOSE and GOAT are notably fronted; GOOSE is approaching KIT in a position notably fronter than TRAP, and GOAT is approaching the center of the vowel space. Additionally, Jason appears to have fronted FOOT to a certain degree, a change that, for the most part, appears to be absent from the community. In the plots below, the small translucent dots represent the speakers' individual tokens for each vowel, the bigger and solid dots represent their means.

For both of these speakers, it should be noted that the F1 differences between mean TRAP and mean LOT are very small, i.e. TRAP is only slightly lower than LOT. However, compared to the community as a whole, this appears to be quite an advanced configuration. As Figure 193 shows, LOT and THOUGHT are clearly distinct in the community, with LOT occupying low central position; TRAP is still in a position higher than LOT and fronter than DRESS. GOOSE is fronted, but not to the same extent as for the speakers above, and GOAT does not appear to be fronted at all. Thus, the two younger speakers presented above indicate a clear development toward an Elsewhere-like pattern, and given the continuing changes in the community, it seems likely that this will be the predominant pattern in Ogdensburg in the near future. The two exceptions to the Elsewhere pattern, of course, are KIT and FOOT, which as of yet show no signs of centralization in Ogdensburg.

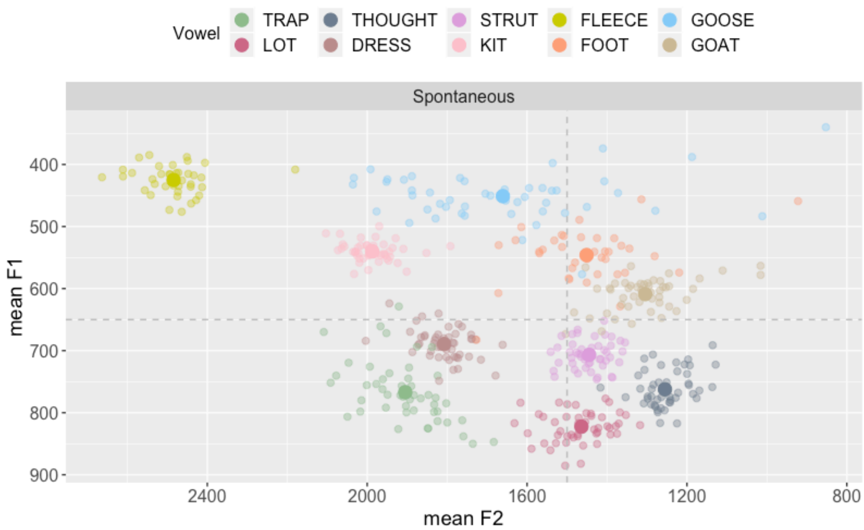


Figure 193: Summary of speaker and community means. The small translucent dots represent the speaker means for each vowel in spontaneous speech, the bigger and solid dots represent community means in spontaneous speech.

Chapter 8: Discussion & Conclusion

8.1 Research Questions - A Revisit

The findings presented in this study contribute to the growing body of research on the recession of the NCS and the spread of the Elsewhere Shift, including the COT-CAUGHT merger. The current study is the first systematic and complete analysis of an NCS community adopting the Elsewhere system, analyzing developments in all variables involved in both chain shifts, along with co-occurring changes in the back vowels. It is also the first study to focus in depth on these processes in a rural area. In doing so, the study addressed a number of issues which, so far, have remained unanswered in research regarding the recession of the NCS in an attempt to improve the understanding of the underlying processes that shape these changes. The purpose of this study was threefold:

- It was intended to investigate the status of the NCS alongside potential progress toward the COT-CAUGHT merger, and GOOSE, FOOT and GOAT fronting, in order to determine the likelihood of an incipient Elsewhere system in Ogdensburg.
- A second concern was potential social evaluation of the variables involved in these changes, and how potential positive or negative social perception might affect the treatment of the phonemes in production.
- A third purpose was to address the question whether rural communities may be treating the NCS (and the withdrawal from it) and the adoption of the Elsewhere system (including the COT-CAUGHT merger and changes in the back vowels) differently than urban communities.

The findings regarding each of these issues will be discussed in the following subchapters, alongside the questions they raise and their theoretical and methodological implications.

8.2 The NCS Going Elsewhere

The results presented in the previous chapters have shown that Ogdensburg seems to only have been affected by the NCS temporarily. Though often lacking statistical significance, the trends in the data suggest that, in 2008, the NCS was in full force in

Ogdensburg: It appears that TRAP was raising, KIT, DRESS and STRUT were backing, and THOUGHT was lowering in apparent time. The only vowel that does not appear to have moved along its NCS trajectory is LOT, although it seems to have been realized in a relatively fronted position all along. By 2016, there are virtually no traces of categorical TRAP raising, or STRUT and KIT backing in the community; THOUGHT and DRESS continue to progress along their respective trajectories; and LOT is moving in the opposite direction. As a result of these changes in recent years, the NCS appears to be virtually absent from the community now. Figure 194 illustrates that, while in 2008 the majority of speakers in Ogdensburg participated in the majority of NCS features, by 2016, it seems to have disappeared from the community.

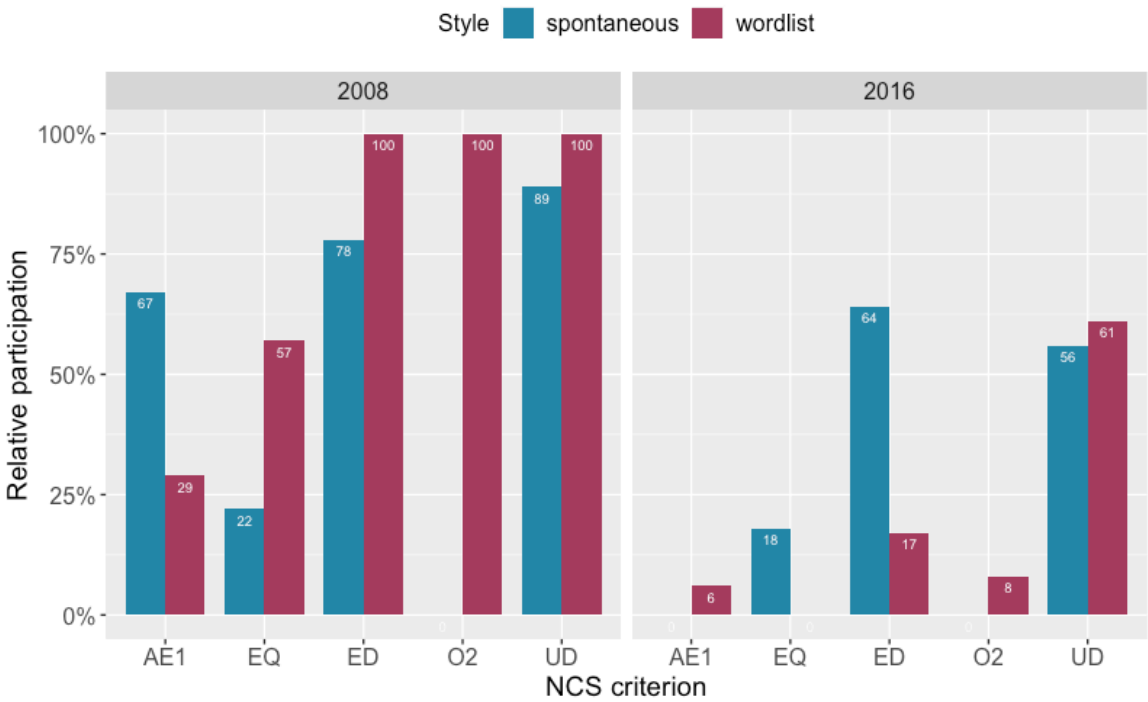


Figure 194: Relative participation in NCS criteria in 2008 and 2016 by *speech style*.

Especially the two most prominent features, categorical TRAP raising (AE1) and LOT fronting (O2), can no longer be heard in Ogdensburg. In fact, in my interviews, I noticed only one participant who had notably fronted LOT in the word *top*, which I initially falsely transcribed as *tap*, and only realized it was *top* because *tap* did not make sense in the context of shopping for clothes. However, this seems to have been an isolated occurrence. The two criteria that appear to still be common to the community are ED and UD, i.e. small F2 distances between DRESS and LOT, and LOT being fronter than STRUT. Neither of these are surprising, as DRESS is backing alongside LOT, and STRUT has always been realized far back in the vowel space. However, participation in these two criteria is receding both in apparent and real time as well.

Figure 195 below provides a different view of the NCS that is no more. The plot illustrates that, while in 2008 the majority of speakers fulfilled three or more of the five NCS criteria, in 2016, the numbers have reduced drastically to zero to two, generally maintaining ED and UD. Furthermore, while a slight shift away from NCS features can be observed among 2008 speakers in wordlist style, this shift is much more evident among 2016 speakers, where most speakers further reduce their participation in the NCS to zero to one feature, most commonly maintaining UD. Seven of the 2016 speakers shift in the opposite direction, however, and have higher NCS scores in wordlist style than in spontaneous speech. For most of these speakers, this is the result of the adoption of either UD or ED, which is not particularly surprising as pointed out above. Some of them adopt EQ instead, i.e. they reverse the relative height and frontness of TRAP and DRESS, which, given the ongoing lowering and retraction of DRESS in combination with a still relatively high TRAP is not surprising either. Only one of these speakers who adopt EQ in careful speech also adopts AE1, i.e. raises TRAP to an F1 of less than 700 Hz, so that her EQ is likely a result of raised TRAP rather than lowered and retracted DRESS. This particular speaker is Breanna, born in 1964, who is a notable outlier in the 2016 data, with an NCS score of two in spontaneous speech (UD, ED) and a score of four (UD, ED, EQ, AE1) in wordlist style, which makes her the speaker with the highest NCS scores in the entire 2016 data set.

Interestingly, as Figure 195 illustrates, there is no apparent difference in NCS scores between age groups in the 2016 sample, suggesting that the entire speech community dropped the NCS nearly simultaneously. This is in agreement with Thiel and Dinkin's (under review) suggestion that the abandonment of raised TRAP in Ogdensburg was the result of communal change, and the real-time patterns observed for DRESS, STRUT, KIT and THOUGHT support this conclusion. Not only are the apparent-time trends *toward* NCS from 2008 no longer present in the 2016 data, but TRAP was *lower*, STRUT, DRESS and KIT were *fronter*, and THOUGHT was *higher* than they had been in 2008. This suggests that, rather than generational change in which each younger cohort had less NCS than their predecessors, Ogdensburg lost the entire NCS through communal change. For TRAP and STRUT, this communal change appears to have been heralded by apparent-time trends (though they are not always statistically significant), while for KIT, no such trend exists, and for DRESS and THOUGHT, the opposite is the case, as they continue to move along NCS trajectories in 2016, though, presumably, no longer as part of the NCS.

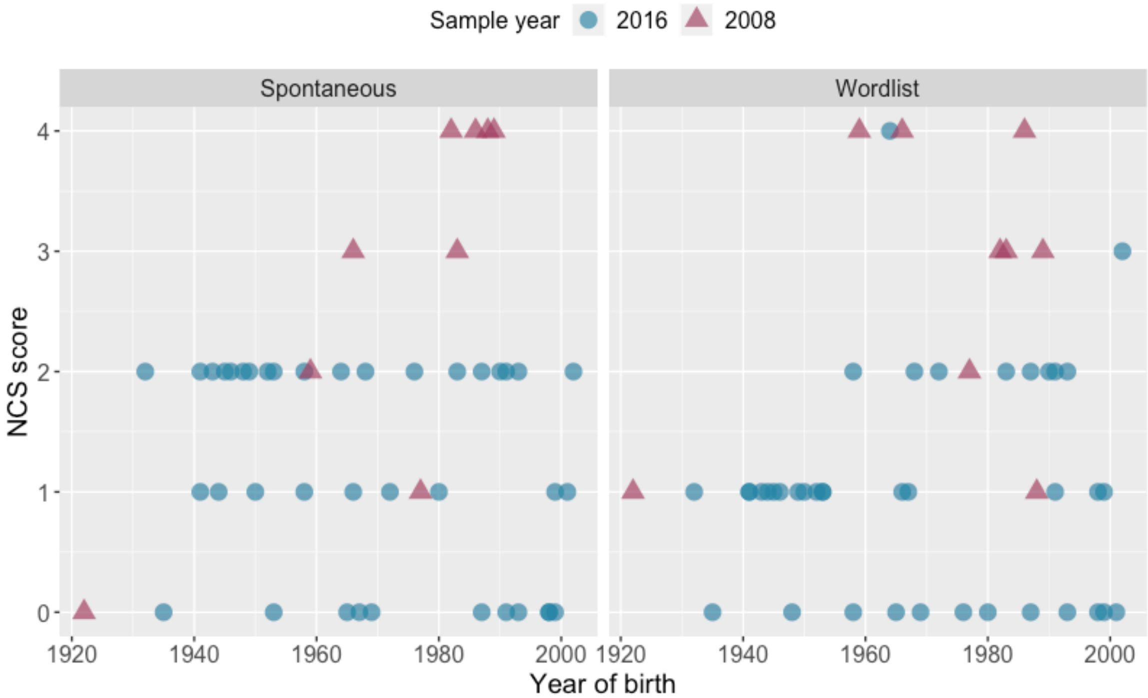


Figure 195: NCS scores in 2008 and 2016 across *age* by *speech styles*.

Based on the example of TRAP, Thiel and Dinkin (under review) suggest that communal change, i.e. the abrupt abandonment of the NCS, resulted from increasingly negative evaluation of raised TRAP in the community. Thus, they propose that communal change may be part of the transition of a linguistic indicator into a marker, in the sense that sufficient negative evaluation of a linguistic feature leads to the collective retreat from this feature. The patterns observed for DRESS, STRUT, KIT and THOUGHT in the present study, however, do not support this hypothesis. While DRESS and THOUGHT might have reached the level of conscious awareness in the community, the comments made by the participants do in no way indicate that shifted variants are perceived negatively, nor do the rating patterns, or lack thereof, in the matched guise experiment suggest stigmatization of any kind, assuming that the merger guises are a reflection of the social perception of lowered and fronted THOUGHT. For DRESS, STRUT and KIT, style-shifting patterns might indicate that the unshifted variants are favored over their shifted counterparts, as speakers who notably retract these variables in 2008 style-shift toward fronter, i.e. less NCS-shifted, variants for all three vowels in more careful speech. As the same style-shifting pattern was observed for TRAP, this may suggest that all of these changes operate under very similar social pressures as TRAP. Further research will be necessary to verify whether the changes observed in Ogdensburg are progressing as changes from above or below, and to determine the conditions for communal change.

The apparent-time shifting of TRAP, DRESS, STRUT, KIT and THOUGHT in NCS direction in the 2008 data strongly suggests that the NCS came to Ogdensburg around 1980 - quite late relative to its development in Western New York. This raises the question of *how* the NCS reached the community. Under the assumption that the NCS diffused to small communities like Ogdensburg, it would be expected that contact between Ogdensburg and the Inland North intensified, thereby facilitating diffusion of the NCS, around the point in time when NCS adoption becomes evident. This, however, does not seem to be the case. As described in Chapter 1.8, Ogdensburg lost its industrial base in the second half of the 20th century, at a time *before* the adoption of the NCS becomes apparent in the data. Because their industry was their main link to the Inland North, it seems more likely that, along with the industry, contact with major NCS communities was on the *decline* during this time, making it *less* likely for the NCS to diffuse. In other words, diffusion of the NCS would have been more likely in the first half of the 20th century, when trade facilitated intensive contact with the Inland North. There are 10 speakers in the 2016 data born in the first half of the 20th century who do have an NCS score of two, which is the highest anyone has scored in 2016, which might suggest that the NCS did come to Ogdensburg during this time. However, for all 10 of these speakers, the two criteria met are ED and UD, i.e. LOT fronter than STRUT and a relatively small DRESS-LOT F2 distance, both of which, as explained earlier, are not particularly surprising and not necessarily an indication of incipient NCS in Ogdensburg during this time. In fact, the absence of TRAP raising (i.e. AE1) among these speakers suggests that ED and UD were long-standing features in the community rather than early signs of a diffusing NCS, as TRAP raising is generally considered one of the earliest and most prominent NCS features, and would therefore likely have been one of the first to diffuse. A full investigation of how the NCS developed in Ogdensburg would require more resources than this project allows, but might be a fruitful starting point for future research, especially in this part of the Inland North, as the triggers and mechanisms here might differ from those identified in other parts of the Inland North (e.g. Durian & Cameron, 2018), as indicated by the seemingly long-standing retracted STRUT, which differs from ongoing STRUT retraction in e.g. Chicago.

8.2.1 Methodological Concerns

The differences in vowel realizations between the two data sets is so striking, and the real-time separation between them so short, that alternate possible explanations for these differences must be considered before accepting the conclusion that the NCS simply disappeared from the community. One possible explanation might be the methodological differences that do exist between the 2008 and 2016 studies, summarized in Table 115 (see Chapter 2.4.1 for detail).

	2008	2016
Data collection	3 days Spontaneous encounters Short socioling. interviews 15–30 mins Spontaneous speech + wordlist + other elicitation tasks 2 phone interviews Spontaneous speech + other elicitation tasks	3 months Scheduled interviews Full-length socioling. interviews 1–2 hours Spontaneous speech + wordlist + other elicitation tasks
Interviewer	Native English speaker (non-NCS), male	Non-native English speaker, female
Corpus	9 speakers 7 f + 2 m Born: 1922–1989, mean: 1972 White 3 college, 3 no college, 3 students	39 speakers 25 f + 14 m Born: 1932–2002, mean: 1969 White (except one) 23 college educated, 9 no college, 7 students
Formant extraction	Hand measured Lobanov (re)normalized	Automated vowel extraction (FAVE) Lobanov normalized
Mean calculation	*ANAE* methodology	*ANAE* methodology

Table 115: Comparison of methods for the 2008 and 2016 data sets.

Firstly, it is possible that the difference between the methods of formant measurement led to the differences observed between the two samples. However, this seems improbable, as research has shown that Dinkin's hand measurements and FAVE's results for the same data do not differ significantly from each other (Severance et al., 2015). A second issue that may have led to differences in the data sets are the circumstances under which the two data sets were collected. The somewhat more formal nature of the interviews in 2016 may have elicited a more careful speech style in 2016, thus creating the illusion of less NCS and a higher degree of merger in the community than there actually is. At the same time, the longer duration of the interviews in 2016 may have allowed the speakers more time to become accustomed to the conversation and the interviewer, and relax into a more casual speech style (Chambers, 2008; L. Milroy & Gordon, 2003). Indeed, the participants' frequent use of "the Burg" to refer to Ogdensburg as well as numerous examples of colloquialisms used by both male and female speakers of all ages indicate that the language they used was quite casual and uncensored:

> "*Wow I'm speaking like an idiot." (Ben, 1999)*
>
> *I was like ugh, this sucks." (Chloe, 1998)*
>
> *Holy crap it's gonna be awful." (Jason, 1998)*
>
> *[He] is like a big ... local history buff." (Daniel, 1993)*
>
> *You're just frigging bullshitting." (Anthony, 1991)*
>
> *Being a complete asshole to students." (Megan, 1990)*
>
> *Like [he] had a joint and I was like huh that's fucking weird." (Ryan, 1983)*
>
> *What the hell is going on." (Sarah, 1969)*
>
> *Holy cow." (Patrick, 1968)*
>
> *I'm like yeah dude." (Amber, 1967)*
>
> *[They] didn't give a rat's rear." (Richard, 1941)*

Thirdly, the fact that I am a non-native English speaker may have prompted speakers in 2016 to attempt a more standard or careful pronunciation than Dinkin's participants did in 2008, out of concern that I might not understand more regionally marked or casual variants. However, this too seems unlikely for two reasons: a) they often used idiomatic expressions, exemplified in some of the quotes above and by phrases such as *"shoot off an email" (Summer, 1987), "a hole in the wall" (Ashley, 1966), and "once in a blue moon" (Anthony, 1991)*, which they certainly would have avoided too if they were trying to accommodate perceived second-language deficiencies; and b) my language skills were rarely commented on, and if so, participants usually assumed I was a native or bilingual speaker of English. On the other hand, this impression may have prompted them to refrain from more salient and marked features for a different reason: to match a perceived high standard, academic variety of English (Corbett, 2017), though again this does not seem to have been the case as the examples above show. Furthermore, it seems that at least some of the phonetic changes observed in Ogdensburg are proceeding below the level of conscious awareness, and for these variables, this level of control over the usage of certain variants is questionable.

However, even if for these reasons speech in 2016 was produced in a more careful style than in 2008, Thiel and Dinkin (under review) found that this hardly seems likely to account alone for the differences that exist between the two data sets based on the example of TRAP. As summed up in Table 116, the mean difference between spontaneous and wordlist TRAP F1 for the youngest speakers in the 2016 data, born between 1980 and 2002, is about -35 Hz. The difference between their spontaneous F1 mean and that of their peers in the 2008 data is more than three times as high, about -118 Hz. Thus, if the difference between speakers interviewed in 2008 and 2016 is a result of two different spontaneous speech styles affected by the identity of the interviewer, the difference between these two styles is at least three times as wide as intra-speaker variation in 2016. While this cannot be ruled out as a possibility, it seems improbable that the difference between two spontaneous-speech interview styles with different interviewers would be so much greater than the difference between spontaneous speech and wordlist style with a single interviewer, even taking into account potential accommodation toward a non-native interviewer. Therefore, it is likely that at least some of the differences between the 2008 and 2016 samples are the result of real-time change. Further research into interview methods, including potential interviewer effects of non-native speakers, would be needed to clarify how much of an effect these methodological differences really had. Additionally, further real-time evidence from earlier recordings of local speakers would help solve this problem. Unfortunately, no such data was available for the present study.

Sample year	2008	2016	
Age range	1980–1990	1980–2002	
Spontaneous	665 Hz	783 Hz	Δ -118 Hz
Wordlist	730 Hz	818 Hz	Δ -88 Hz
	Δ -65 Hz	Δ -35 Hz	

Table 116: F1 TRAP means in young age cohorts in spontaneous speech and wordlist style in 2008 and 2016.

8.2.2 Theoretical Implications

If real-time differences observed in this study are in fact real, they have serious implications for the reliance on apparent-time reasoning in the study of sound change in progress, as the study has shown that inferences made from apparent-time trends do not always hold true in real time: Not only do younger speakers generationally move away from the NCS-shifted variants of their elders; the very same age cohort whose vowels were most shifted in 2008, those born in the 1980s, have them no more shifted than anyone else in 2016. As pointed out earlier, this pattern suggests that the NCS is being lost in Ogdensburg not via generational change, but via communal change. At least for some of the NCA features, this is likely the result of increased negative evaluation, which, by 2016, had become sufficiently strong enough for speakers who share this evaluation to collectively retreat from these features. This also included adults, who, in the apparent-time paradigm, are assumed to *not* participate in linguistic change, or at least do so only sporadically and at a lower rate than younger speakers (Labov, 2007). Under this scenario, the gradual generational change in the social evaluation of e.g. raised TRAP in combination with a retreat from this feature led by college educated speakers seems to have foreshadowed an eventual communal change away from this NCS feature, once its negative evaluation was sufficiently strong (Thiel and Dinkin, under review). The present study therefore sheds light on the rapidity with which regional dialect features can be reversed once they become stigmatized.

The case of LOT, however, does not support this hypothesis. Much like TRAP, LOT appears to have risen to the level of conscious awareness, and its retracted variant seems to be favored over the fronted one. However, other than raised TRAP, fronted LOT does not seem to have been suddenly abandoned by the entire community. Instead, LOT has been retracting gradually in apparent and real time. Why the variants of these two variables are treated differently in production despite (seemingly) similar social perceptions is not immediately apparent. It is possible, however, that their social evaluations are not as similar as they may seem. This idea will be pursued further in Chapter 8.3.2 below.

Regardless of the differences in the developments of TRAP and LOT over time in Ogdensburg, the findings in this study stress the importance of implementing real-time comparison into sociolinguistic studies, especially if there is reason to suspect increasing negative social evaluation around the variable of interest. Doing so, this study presented an unexpected opportunity to catch a community just at the point of retreating from the

NCS. In the absence of the 2008 data, it would not have been clear that Ogdensburg had ever been an NCS community.

8.3 Orientation Toward a New Standard

Potential social evaluation of the variables of interest was assessed on the basis of various indices, including apparent and real-time change, style-shifting patterns, advantages among female and college educated speakers, and matched guise ratings.

The shift between speech styles observed in the analyses indicates that some of the variables of interest in the study are stylistically sensitive. Particularly TRAP and LOT appear to be subject to a significant amount of intra-speaker variation. In combination with TRAP lowering and LOT retraction as well as matched guise judgments, the data suggest that both variables have progressed from a sociolinguistic indicator to a marker (see Chapter 1.5): As both features become a marker in Ogdensburg, both of the predicted indices of social stigma, i.e. a steep slope of style shifting and negative evaluations, emerge in apparent time in the community. The overt comments about raised realizations of TRAP and fronted realizations of LOT further strengthen the argument that both variants have risen to the level of social awareness in the community, to the point that both changes were reversed.

It is possible that the same argument can be applied to THOUGHT. While the style-shifting slope is not as steep for THOUGHT as it is for TRAP and LOT, and it is unclear whether the reactions to the merger guises can be taken as evidence for reactions to lowered and fronted variants of THOUGHT, some of the participants' comments about the matched guises do suggest a certain level of awareness of lowered and fronted THOUGHT. Additionally, the timing of the reversal of style shifting the height of THOUGHT between 1960 and 1970 coincides with that of the height of TRAP (1960) and the frontness of LOT (1950-1960), suggesting that all three changes are interconnected, striving toward the same perceived standard.

While some style effects can be observed for STRUT and DRESS as well, they do not necessarily imply orientation toward a new standard. This seems to be particularly true for STRUT, which was not overtly commented on by any of the participants, and for which style shifting is not in agreement with apparent-time fronting (though this change is led by college educated speakers). For DRESS, the implications of the observed style patterns are less clear. While they are ambiguous, they do to a certain degree corroborate the apparent and real-time changes, which do also seem to be led by college educated speakers. They also match the style-shifting patterns observed in Lansing, where variation in DRESS seems to have reached the level of social awareness, and where DRESS appears to continue on its downward trajectory owing to positive social perception of the lowered variant.

Overall, it appears that the ongoing changes in TRAP and LOT along trajectories that reverse the NCS are proceeding above the level of consciousness, while those in DRESS, THOUGHT, and STRUT seem more likely to progress as changes from below. Although there is some evidence that DRESS and THOUGHT may have reached the level of social awareness as well, whether or not their variants have become social markers in Ogdensburg is a question that will have to be answered in future research. An indicator of orientation toward the standard might be potential differences in the rate of shifting (i.e. backing/lowering of DRESS and fronting/lowering of THOUGHT) depending on word frequency. Changes that orient toward a new standard have been found to be most advanced in lower frequency words, while otherwise sound change is more likely to be more advanced in more frequently used words (Dinkin, Forrest, & Dodsworth, 2017). Thus, if DRESS and THOUGHT are progressing as changes from above, these changes would be expected to be more notable in low-frequency words. The opposite case would suggest change from below instead. While an assessment of the effect of lexical frequency on the observed changes is outside the scope of the present study, it seems to be an interesting hypothesis to be tested in future research.

While the ongoing phonetic changes in Ogdensburg appear to be robust, they do not seem to affect all speakers in the community to the same extent. All target vowels, with the exception of KIT, show more or less pronounced differences between speakers of different educational backgrounds and/or between male and female speakers. Most changes appear to be led by college educated speakers, and these differences are particularly prominent for TRAP, LOT and STRUT. Here, speakers without a college degree not only lag behind speakers with a college background, they either show no participation in the ongoing changes at all or, in fact, change in the opposite direction. For TRAP and LOT, this is in agreement with previous research and suggests that whatever is left of the NCS might be developing into a sociolect, as speakers with more access to social prestige are retreating from newly stigmatized variants and are adopting new prestige features instead. For STRUT, on the other hand, social perception of fronted and backed variants has not yet been studied.

However, more often than not, these social differences that emerge very clearly in the visualized data lack statistical significance, for reasons that are not immediately clear. One explanation might be that the differences are particularly prominent among younger speakers, who are not only divided by their *educational backgrounds*, but also by *gender*. Younger college educated speakers in the sample happen to be female, while the two younger speakers without a college degree are male. Thus, a potential effect of *education* may have been statistically assigned to *gender* rather than *education*, which was indeed the case in some of the tested models. However, in many cases, neither of the two factors were found to be of statistical significance. A second explanation might be that there are, overall, too few data points for the differences to be statistically significant. This does, however, not explain why *some of* the differences between speakers with and without a college degree are found to be significant, while others are not. The estimated effect sizes do not appear to be a relevant element here, as there are cases of small effect sizes that do reach the level of statistical significance, while larger effect sizes do not. Thus, for the most part, inferential statistics suggest that the differences between these two social groups in Ogdensburg are not representative of the community as a whole, which contradicts research in other NCS communities, where a significant lead *away from* NCS variants was found for middle class and college educated speakers. It seems unlikely, therefore, that the same social distribution of phonetic changes observed in Ogdensburg is simply due to chance. The fact that very similar patterns were apparent in the plots of five of the six NCS vowels analyzed in the study, as well as in the plots for both GOOSE and GOAT, further strengthens the legitimacy of the effect of *education.* This suggests that this effect should be taken seriously, at least for TRAP and LOT, which have been shown to be socially marked, regardless of the lack of statistical significance.

Indeed, this highlights the responsibility of the researcher to not only ensure the use of a combination of descriptive and inferential statistics, but also to ascribe more weight to one over the other based on reasoned judgment. Sole reliance on inferential statistics may lead to irresponsible oversights of pivotal effects in the analysis of sociophonetic variables. Had I relied predominantly on inferential statistics to determine how the changes in the vowels are conditioned by social factors, and in doing so ignored strong evidence from the visualization of the data, I would have run the risk of making erroneous conclusions about phonetic changes in English in Ogdensburg. Future research would benefit from greater insight into both the application of mixed effects multiple linear regression models as well as analysis of variance (ANOVA) in sociophonetic research, and how they should be incorporated and balanced, alongside descriptive statistics.

If the effect of *education* is in fact authentic, it may mean that the results presented in the previous chapters overstate the rate of phonetic changes in Ogdensburg overall, since (female) speakers with a college education are overrepresented in the sample relative to their actual proportion in Ogdensburg (see Chaper 1.7).

8.3.1 Regional Reorientation and the Significance of 1960

If the rejection of raised TRAP and fronted LOT in Ogdensburg is in fact a result of both variants having become socially marked, this raises the question of what caused the increasing negative perception of raised TRAP and fronted LOT. A definitive answer to this question lies outside the scope of this study; however, the timing of the changes observed in both variables suggests that they are connected to Ogdensburg's economic decline, which started in the 1960s – the same decade in which we observe unraised TRAP and retracted LOT (and potentially lowered THOUGHT) becoming the favored variants in production and perception.

Early studies on the NCS found that raised TRAP and fronted LOT were perceived as urban, and middle class, i.e. standard, and that these variants were generally targeted in careful speech. Thus, before the industries in the Inland North collapsed, Inland North speech patterns may have held a certain kind of prestige toward which Ogdensburgers used to strive, potentially as an expression of affiliation with the Inland North, to which the city used to be tightly connected through industry and trade. This would explain the momentary participation in NCS features that can be observed in data collected in 2008. Today, Ogdensburgers still appear to associate raised TRAP and fronted LOT with urbanity, particularly with Inland North cities in New York such as Rochester, Syracuse and Buffalo, but also Chicago, as their comments about these features have shown. However, the prestige these variants may once have carried seems to have disappeared along with the industries that gave these urban centers their status. Consequently, Ogdensburgers had to orient toward a new source of standard language. The most likely choice for a new source of pride may have been the (dialectological) North Country, bordering Ogdensburg to the east, which is characterized by a low nasal TRAP system as well as retracted and nearly-merged LOT – both of which are features currently developing in Ogdensburg, likely because they have become the new perceived standards toward which Ogdensburgers now strive.

The likelihood of the occurrence of this kind of regional reorientation, and the potential linguistic implications it may bring with it, is to a great extent dependent on the absence or presence of physical, psychological and political barriers. Britain (2014), for example, found that the presence of negative attitudes and a lack of infrastructure prevented a shift in orientation across dialect boundaries in the English Fens. Before large-scale drainage, the Fens had been physically separated from the surrounding areas, as it was largely covered in marshland. After the area was drained and more accessible, negative stereotyping of the Fens and its inhabitants, and limited roads and public transportation to and through this area, created new barriers between the Fens and the neighboring areas, preventing reorientation from taking place. For Ogdensburg and the North Country, the opposite seems to have been the case: Not only has the North Country always been easily accessible to Ogdensburgers, it also appears to be perceived in a positive light.

Ogdensburg has a long-standing regional affiliation with the area that encompasses the dialectological North Country. As mentioned in Chapter 1.8, the term North Country not only refers to a particular dialect area, it also describes the northernmost region of New York State, including Ogdensburg. It is commonly used by Ogdensburgers (and others) to express regional affiliation with nearby communities, especially the surrounding towns, including Canton and Potsdam – both of which belong to the same county as Ogdensburg, and both of which are part of the dialectological North Country. Thus, there are no political boundaries separating Ogdensburg from the neighboring dialect which may have hindered the weakening of the dialect boundary. Additionally, Ogdensburg is relatively well connected to the dialectological North Country, as several well-travelled state and county routes link it to Canton and Potsdam, as illustrated in Figure 196 below. Thus, face-to-face interaction with speakers from Canton and Potsdam has been the norm rather than the exception for Ogdensburgers.

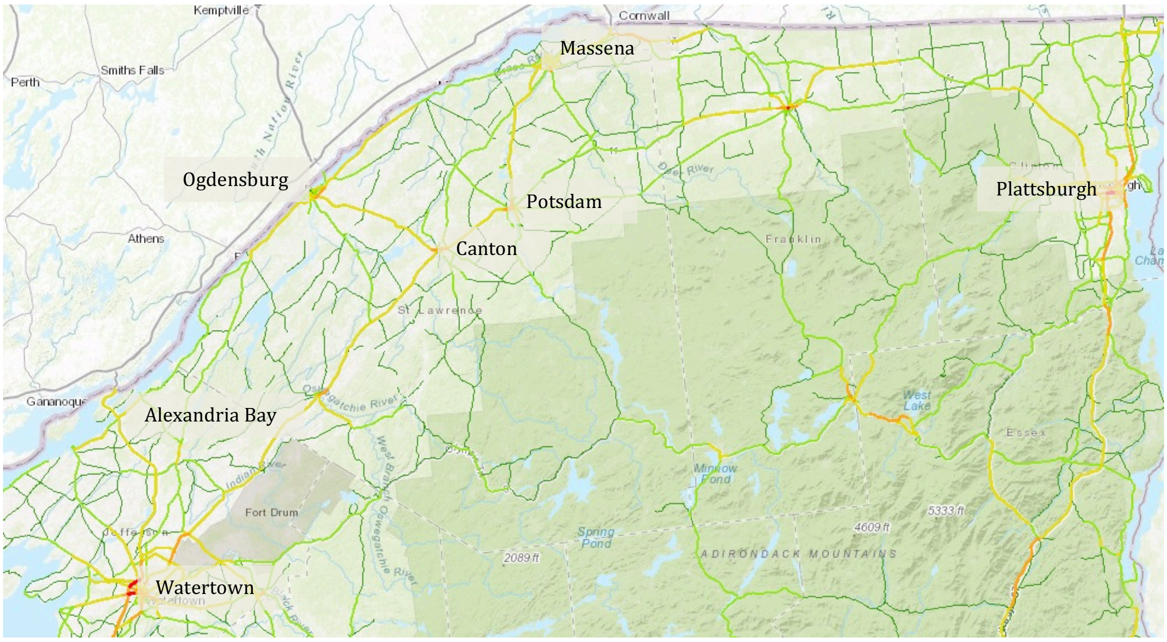


Figure 196: Average daily traffic flow in 2015 in the geographic North Country (NYSDOT). Traffic increases from green to yellow to orange to red.

Canton and Potsdam are two of the more prominent communities in the North Country. Potsdam, in terms of population, is the biggest town in this part of New York State, and Canton, while notably smaller, is the St. Lawrence County seat. Both of these towns have been college towns since the 19th century, each being home to two universities. Additionally, both Canton and Potsdam are in much better economic shape than Ogdensburg, with median household incomes that were 38% and 42% higher than that of Ogdensburg in 2016 (US Census Bureau, n.d.). Thus, Canton and Potsdam have long been characterized by a greater professional class, a bigger tax base and a better socioeconomic standing than Ogdensburg, and Ogdensburgers are fully aware of these socioeconomic differences between the towns, as is evidenced by the two following quotes from two of the participants in the present study:

> *I characterize [Ogdensburg] as a blue collar, hard head town, unlike Canton and Potsdam, which have colleges, you know. We used to have one here, but it's gone."*
>
> *(Richard, born in 1941)*

[Canton and Potsdam] had a, well they have a, a bigger tax base there and they, you know, they have the professional class."
(Tracy, born in 1944)

It is possible that these socioeconomic differences between Ogdensburg on the one hand, and Canton and Potsdam on the other hand, may, until recently, have contributed to the *maintenance* of the dialect boundary between these towns. As long as the industry in the Inland North and Ogdensburg was running at full speed, the Inland North may have carried just as much, though likely a different kind of prestige as the North Country. Rather than being sophisticated, as the North Country, the Inland North was industrious and hard-working. It is likely that Ogdensburg, which was long characterized in the same way, associated more closely with this kind of prestige than with that carried by Canton and Potsdam, leading them to orient toward the Inland North, socially and linguistically, rather than the dialectological North Country. In other words, despite their close affiliation with the area that comprises the North Country dialect, their connection to the Inland North had a stronger impact on their linguistic behavior during this time.

However, when Ogdensburg lost its industrial base, they may have become disconnected from the Inland North, and when the Inland North as a whole became less industrious and less productive and lost its status as a source of prestige and pride, Ogdensburg's eastern neighbors, Canton and Potsdam, may have been the only places to orient toward. At this point, Canton and Potsdam not only surpassed Ogdensburg in terms of socioeconomic status, they had also managed to maintain their old-town charm in the form of historic buildings and lively downtown areas, which Ogdensburgers seem to value, as they themselves have lost much of both. A potential shift in regional orientation may have been further facilitated simply by the *need* for these two near-by towns. After Urban Renewal, Ogdensburg lost most of its local stores, so that they had to rely more heavily on Canton and Potsdam than they did before. Both Canton and Potsdam have also been able to preserve their movie theaters, which Ogdensburgers have been visiting frequently since their own closed down. Thus, while it is not likely that this situation has led to an *increase in contact* between Ogdensburg on the one hand and Canton and Potsdam on the other hand, it seems plausible that the *nature* and *salience* of this contact and of the relationship between Ogdensburg, Canton and Potsdam has changed. Although it would have been possible for this new dependency to have let to feelings of resentment among

Ogdensburgers, the opposite appears to be the case, and the following quotes evidence the positive attitudes Ogdensburgers seem to have toward Canton and Potsdam:

In terms of restaurant choices, shopping opportunities, and recreational and cultural activities, participants find Canton and Potsdam to be superior to Ogdensburg:

> “*We're fortunate that we do have, uhm, Canton and Potsdam, there's a couple of options.”*
>
> *(Aubrey, born in 1980)*

> *There were a couple record stores here in Canton and Potsdam, and before the internet, uhm, you could, you know, come over here and usually get stuff that you couldn't get in Ogdensburg.”*
>
> *(Charlotte, born in 1958)*

> *We used to have a movie theater but now we have to go to Canton.”*
>
> *(Chloe, born in 1998)*

> *I've got Clarkson and SUNY Potsdam for lectures or art shows or festivals. Potsdam seems to have a lot of outdoor festivals ... I think more, more than Canton and more than Ogdensburg.”*
>
> *(Amanda, born in 1958)*

Many participants also shared their sentiments about Ogdensburg in pre-Urban Renewal times, comparing it to present-day Canton and Potsdam:

> “*Potsdam is beautiful, now we could have been like Potsdam, uhu, Canton also.”*
>
> *(Tracy, born in 1944)*

Like here in Canton, when you're here in Canton you'll notice it still has all the old about it. Downtown is all the brick. Ogdensburg was loaded, like if you're taking pictures, you take a picture of downtown Canton, that's what Ogdensburg used to have … like Ford and State and all through there was just beautiful, with all these old brick buildings. And they've torn them down. Biggest mistake they probably ever could have made."

(Ashley, born in 1966)

Take a look at the, some of the villages around like Potsdam, Canton, uh, they still retain their old buildings, their old structures. And it's working for them so, you know. They still have the box stores and the shopping areas and stuff like that, but they're still able to, uh, hold things together in downtown areas … They're beautiful, yes, and one thing that Canton and Potsdam have done is they said to people that come into the village: 'If you're gonna build here, you're gonna have to build so that it matches the rest of our, our village', uh, like the Clarkson Inn and stuff like that, uhm, they had them build it so it matched, you know, which is good … Potsdam and Canton were, were smarter, and they, they saw probably what happened to Ogdensburg and they didn't allow it to happen to Canton and Potsdam."

(Eddie, born in 1943)

Similar positive attitudes have been expressed toward the North Country towns of Lake Placid and Plattsburgh, though both are too far away for everyday contact, and it is unlikely that they would have an immediate effect on the speech of Ogdensburgers, at least in comparison to Canton and Potsdam. Overall, it seems that Ogdensburgers perceive the dialectological North Country in a positive light, which increases the likelihood that they have reoriented toward this dialect area as a new source of pride. This, in turn, may have led them to perceive North Country speech patterns as more favorable, leading them to reject the speech patterns associated with the *urban* Inland North, and to adopt those of the *professional* North Country instead.

Evidence for the possibility of this kind of a regional reorientation with linguistic consequences can be found in British dialectology, in the study of Middlesbrough in Northeastern England (Llamas, 2007). Middlesbrough is located on a regional and dialect border between Yorkshire to the south and Tyneside to the north. Politically, it used to be part of North Yorkshire; however, owing to constant repoliticizing and redrawing of administrative boundaries, it has been put in closer connection with Newcastle to the northeast. Llamas found that this led to a shift in orientation among speakers in Middlesbrough away from Yorkshire and toward Tyneside, which also affected their speech patterns, as they adopted dialect features commonly found in the Northeast and rejected those that are characteristic of Yorkshire.

The flip side of the argument that the North Country to the east may be the new source of pride is Alexandria Bay, located to the west of Ogdensburg in the Inland North. Alex Bay, like Ogdensburg, is located on the St. Lawrence River. Other than Ogdensburg, however, Alex Bay has been very successful in using their waterfronts to capitalize on tourism and leisure travel. Thus, Ogdensburgers look up to the town and its economic development, and it might be possible that, as a result, the Inland North speech patterns of Alex Bay might still be perceived as prestigious. However, while some of my participants do report occasional travel to Alex Bay, it is too far away for everyday interactions, and thus is less likely to have as much as an impact on the speech of Ogdensburgers as Canton and Potsdam. Indeed, Figure 196 (p. 375) shows that traffic flow from Ogdensburg to Canton (and, potentially Potsdam through Canton[88]) is notably higher than from Ogdensburg to Alex Bay. Likewise, the most direct route to Watertown, another Inland North Fringe community, is less frequently travelled than those between Ogdensburg, Canton and Potsdam.

While the scenario outlined above makes social and historical sense and appears to be a likely underlying cause for the sound changes that can be observed in Ogdensburg, verification of the hypothesis lies outside the scope of the current project. Further, preferably anthropological research will be necessary to shed light onto how Ogdensburg dealt with the socioeconomic changes they encountered in the second half of the 20th century. Under the assumption that the hypothesized explanation holds true, the scenario raises several questions, which would also need to be dealt with in future studies.

The first of these questions concerns the timing of the industrial decline and that of the community's linguistic reaction to it. As described above, Ogdensburg's economy started to collapse in the 1960s, the same decade during which locals become more likely to produce lower TRAP, retracted LOT and lowered THOUGHT in more careful speech than

[88] Although there are alternative routes, going through Canton is in fact the most direct, shortest and, depending on traffic, quickest way to get to Potsdam from Ogdensburg.

they do in spontaneous speech. In other words, the community appears to have reacted *immediately* to these socioeconomic changes. Other research, however, argues that the first generation to react linguistically to social and economic changes is the generation that first experiences these changes as young adults (e.g. Nesbitt, 2018). In the case of Ogdensburg, this would refer to speakers born in the 1940s and 1950s, i.e. somewhat earlier than what has been observed in the data. It is certainly the case that speakers born before the 60s started to react to changes in Ogdensburg, as *progress* toward the style reversal for TRAP, LOT and THOUGHT appears to have started earlier than 1960. However, the significance of these early processes cannot be determined here, owing to limited data for this age group. Under the assumption that industrial decline in the community was a gradual process, Ogdensburg's economic situation would have been worse in the 1970s and 80s, when speakers born around 1960, i.e. those who first *reverse* style-shifting patterns, linguistically came of age. This, then, may have further promoted the shift away from urban variants such as raised TRAP and fronted LOT, especially in careful speech, leading to the reversal that can be observed in the style-shifting data. In any case, it surely is not a coincidence that the timing of linguistic and social changes coincides so precisely.

The apparent time progress toward the NCS in the 2008 data might constitute a counter-argument to the hypothesis that the linguistic changes that can be observed in Ogdensburg result from a shift in regional in the community. If the community did in fact reorient toward the North Country and away from the Inland North as a reaction to industrial decline in the second half of the 20th century, why did speakers born *after* 1980 show progress *toward* the NCS instead of moving away from it? The case of TRAP may suggest that the more favorable evaluation of its lowered variant did not become strong enough as to affect spontaneous speech of speakers who acquired TRAP raising as children until sometime between the two data collections in 2008 and 2016 (Thiel & Dinkin, under review). Speakers of the same age interviewed in 2016, i.e. those in their 30s, have already reoriented regionally and linguistically, even in spontaneous speech, while for those interviewed in 2008, the pressure to lower TRAP was only strong enough to affect more careful speech at that time. For LOT, however, this argument does not hold true. While there is no indication of apparent-time LOT fronting in 2008, speakers in this data set shift to a significantly fronter LOT in wordlist style compared to spontaneous speech. Thus, the assumption that, until at least 2008, social pressure to retreat from the NCS was just strong enough to affect careful speech does not hold here, as the opposite appears to be the case, at least in wordlist style. It is, of course, possible that the social evaluation around LOT started to form somewhat later than it did for TRAP, and thus had not yet affected speakers in 2008 at all. This does, however, seem unlikely, given that the reversal of style shifting the frontness of LOT in 2016 seems to slightly pre-date that of TRAP, so that LOT appears to have become socially marked around the same time as, if not earlier than TRAP. However, given that no such effect can be observed in the shift from spontaneous speech to minimal pairs, it is also possible that the style shift from spontaneous speech to wordlist style in 2008 is merely due to hyper-articulation, though this would arguably put LOT in a more retracted rather than fronted position.

The third concern regarding the hypothesis that regional reorientation is the underlying reason for changes in progress in Ogdensburg is the age pattern observed in these changes. If lowered TRAP, retracted LOT and lowered/fronted THOUGHT (i.e. the COT-CAUGHT merger) are in fact spreading to Ogdensburg as the new standard through contagious diffusion from the North Country, a different *age* pattern than the one that has become evident in the analysis would be expected. In Ogdensburg, like most other communities studied so far, progress toward lowered TRAP and the merger is gradual, being advanced further with each successive generation. This pattern, Labov (2007) argues, is common for changes progressing through transmission and incrementation; however, as discussed in Chapter 1.3, a change that is adopted through diffusion would be expected to be more advanced among adults rather than adolescents.

The most plausible explanation for the gradual development of these changes is that they are not spreading by means of diffusion alone, but instead are advanced through a combination of different mechanisms, including initial diffusion and subsequent transmission and incrementation. Under this scenario, adults may have been the first to adopt a somewhat lowered TRAP and a weakened distinction between LOT and THOUGHT through contact with North Country speakers. As the results in the previous chapters have shown, this contact was not incentive enough for the immediate adoption of a complete merger; however, they suggest that it may have weakened the contrast between LOT and THOUGHT among adult speakers, which is in line with Dinkin's (2009) suggestion that the diffusion of a weakened distinction is more likely than the diffusion of a merger as such. Once secure distinction among adults had destabilized, this may have been passed on to their children through transmission, and these children have further reduced the distinction, eventually leading to a full merger in the community. This would explain the exceptional behavior of the students in their realizations of the low back vowels. The same logic might apply to TRAP, though the real-time differences between the 2008 and 2016 data suggest that for TRAP, contact with the North Country may have been enough to cause the entire community to abandon raised TRAP simultaneously and at the same rate.

It is also possible that these changes are not *spreading* to Ogdensburg at all, but that they are independent developments instead, which would explain the gradual progress of each of them. Because Ogdensburg is surrounded by unraised TRAP and the merger in Canada and the North Country, this is rather unlikely. However, mergers *can* develop as independent features even with neighboring dialects that are already merged, if the two involved dialects have similar vowel systems (Johnson, 2007). This is certainly not the case for Ogdensburg and Canada, as Ogdensburg has been characterized by the NCS, and Canada by the Canadian shift. Ogdensburg and the North Country may be more similar in this respect. Although the NCS as a whole is not found in the North Country, some of its features have been identified there, most notably the reversal of the relative positions of LOT and STRUT and LOT and DRESS[89] (Dinkin, 2009), the same two features that characterize the speech of older speakers in the community and can still be heard frequently in Ogdensburg. And, since Ogdensburg has abandoned categorical TRAP raising, the vowel systems of its residents are now even more similar to that of the North Country. These similarities between the North Country and Ogdensburg are likely due to similar settlement patterns, a factor that Johnson (2007) identified as vital in independent developments of mergers in communities adjacent to merged communities. However, there is no clear evidence for an independent development of the merger in Ogdensburg, and it would require further research to determine whether or not this is the case.

Another factor that might be promoting the merger in Ogdensburg is in-migration of merged speakers, facilitating contact between distinct and merged speakers, especially children, thereby advancing the merger in an otherwise distinct community. Table 117 below lists the eight speakers who, in relative terms, are the most merged in production (though still distinct) across all three speech styles. As the Table shows, five of these eight speakers have at least one parent from a potentially transitional or merged dialect area. This may suggest that speakers who grew up with at least one parent merged or

[89] Because the North Country was found to be transitional in their adoption of the merger, these reversals suggest a relatively back DRESS and STRUT rather than fronted LOT.

transitional in production might themselves adopt a more merged, but still distinct, production when growing up around distinct speakers. This, in turn, may result in more merged production among their peers with parents who are distinct. On the other hand, there are speakers in the 2016 sample who are certainly distinct in production and perception despite a parent from a transitional or merged dialect area, so that there is no clear advantage of this group of speakers over speakers with two distinct parents. In-migration to Ogdensburg is, overall, also very rare (see Chapter 1.8), and it seems unlikely that merged speakers would constitute a part of the community that is substantial enough to advance the merger. Further research that includes speakers of fully merged parents would help to clarify the role that in-migration might play in the progress toward the merger in Ogdensburg.

Speaker	Born and raised	Minimal pair judgment	Commutation test score	Mother's origin	Father's origin
Jason	Ogdensburg	transitional	20/20	North Country, NY (transitional)	Kansas (merged)
Ben	Ogdensburg	transitional	19/20	Ogdensburg, NY	Ogdensburg, NY
Daniel	Ogdensburg	transitional	11/20	Tupper Lake, NY (likely transitional)	Ogdensburg, NY
Mark	Ogdensburg	transitional	16/20	Lisbon, NY (likely transitional)	Ogdensburg, NY
Allison	Ogdensburg	distinct	20/20	Syracuse, NY (distinct)	Nashua, NH (merged)
Grace	Ogdensburg	merged	15/20	Ogdensburg, NY	Ogdensburg, NY
Rachel	Ogdensburg	transitional	14/20	Ogdensburg, NY	Watertown, NY (distinct)
Tracy	Ogdensburg	distinct	20/20	Ogdensburg, NY	Canada (merged)

Table 117: Origins of parents of the most merged speakers in 2016. Those highlighted in blue are (likely) transitional or merged in their production of LOT and THOUGHT, those highlighted in wine-red are likely distinct.

8.3.2 Theoretical and Methodological Implications

As was discussed at several points throughout this study and in this chapter specifically, social evaluation of the features of interest appear to have played a major role in the changes that can be observed in their production. For both TRAP and LOT, it appears that the lowered and retracted variants respectively have become the favored way of speaking in Ogdensburg. However, as was pointed out above, the observation that Ogdensburg seems to have reacted differently to seemingly similarly evaluated variants of two variables raises the question of *how similar* their social evaluations really are.

It is possible that the different patterns of change over time observed for TRAP and LOT result from different configurations of the social meanings attached to their variants.

Campbell-Kibler (e.g. 2011) suggests that variants of the same variable can carry meanings that are independent of each other, i.e. one variant being perceived as educated *can* but does not necessarily *have to* imply that the other variant is perceived as uneducated. Yet, this is the conclusion drawn for both TRAP and LOT in this study, based on the comparison of the respective evaluations of their contrasting variants. While this conclusion is largely supported by apparent and real-time change and style-shifting patterns in the directions of the favored variants, both including advantages among female and/or college educated speakers, there remains the possibility that NCS-like variants of TRAP and LOT are not actually evaluated *negatively* by the participants in the present study, or that one of them is, while the other is not.

Under this scenario, the social meanings of TRAP's variants might be linked, while those of LOT are indexed independently. In other words, raised TRAP may be tied to its unraised counterpart in their indexical field, meaning that its social evaluation is the binary opposite of that of the unraised variant. Thus, because unraised TRAP is perceived as more educated sounding, raised TRAP may be associated with the opposite, i.e. uneducated speech, and the community abandoned it rather suddenly because of this negative evaluation. Fronted and retracted LOT, on the other hand, may be evaluated independently of each other, so that retracted LOT being evaluated as sounding more educated does not imply that fronted LOT is perceived as uneducated. Owing to this lack of stigmatization, there was no need for fronted LOT to be rejected on a communal level. Whether or not this is the case cannot be assessed from the matched guise data collected for the present study, as this assessment requires a comparison of the contrasting variants against a neutral alternative (Campbell-Kibler, 2011). If this *is* the reason for the different apparent and real-time patterns in the production of TRAP and LOT, it suggests that, in a *phonetic* variable's indexical field, variants can occupy independent positions, which, so far, has only been found to be the case with the morphological variable *-ing* (Campbell-Kibler, 2011). However, the participants' comments about raised TRAP and fronted LOT do not indicate any negative perception of either of these variants. While participants seem to be aware of them and correctly associate them with Inland North cities, none of their remarks implied any negative evaluation of these features. Further testing of the social perception of TRAP and LOT specifically, and phonetic and phonological variables in general in future research would be beneficial in the exploration of this issue.

A second question that arose in the analysis of the social evaluation data is the possibility of participants reacting to underlying phonological variation rather than to the phonetic differences that constitute this variation. As was discussed in Chapter 6.5.3, the rating patterns for the merger guises did not correlate with the speakers' production, in the sense that it was the *least* merged speakers who rated the merged guises more favorably, while speakers who were more merged did not differentiate between merged and distinct guises. Because merged guises were created with an almost fully merged LOT stimulus in THOUGHT words (see Chapter 2.3.2.1), the most plausible explanation for this asymmetry seems to be that participants responded to the perceived differences in the phonetic realization of THOUGHT rather than to the lack of contrast between LOT and THOUGHT, as per Labov's proposition that observable, surface-level elements of language, such as phonetics, can carry social meaning, while their underlying, more abstract, phonological processes generally do not. However, a comment by one of the participants in the study, who pointed out that *cot* and *caught* would not be pronounced differently everywhere, suggests that awareness of these processes cannot be ruled out completely, at least for mergers. As this particular participant was able to correctly index merged *cot* and *caught* with a geographic location (Boston), she would arguably be able to index it with social characteristics such as education or friendliness as well.

If speakers are in fact aware of phonological distinction, this also has implications for the study of vowel chain shifts, which are similarly abstract in their operation below surface structures. Much like mergers, chain shifts have been argued to not carry social affect, as people do not pay attention to the *relation* of sounds. However, if listeners are aware of structural facts such as distinctions, it is possible that the same applies to interlinked shifts. However, vowel chain shifts might operate on a phonological level that is even more abstract than that of mergers. While the concept of "same vs. distinct" in mergers seems relatively simple and easy to grasp, that of "shifting in order to maintain distinction" in chain shifts seems more likely to be outside of people's perceptual abilities.

8.3.3 Dialectological Implications

The third and last purpose of the present study was to examine similarities and differences in the treatment of NCS features and of changes associated with the Elsewhere Shift in rural and urban communities. The interpretation of the findings suggests that

Ogdensburg treats the variables involved in these changes in ways that are very similar to those identified in more urban areas. The community appears to be moving toward a low nasal or low continuous TRAP system, toward the merger of LOT and THOUGHT, and DRESS continues to lower and retract. Additionally, GOOSE, GOAT and STRUT have moved or are still moving toward fronter positions. In combination, this suggests that Ogdensburg is orienting toward the supra-regional Elsewhere Shift. While this appears to be a community-wide trend, there is evidence that this change is led primarily by college educated speakers. Speakers without a college education continue to make use of some NCS variants (or at least lag behind in the changes away from the NCS), front GOOSE at a slightly lower rate than college educated speakers, and did not participate in the fronting of GOAT. Since similar findings have been reported for urban centers of the Inland North, including Chicago, Buffalo, and Lansing, it seems that the reversal of the NCS affects Ogdensburg, a rather rural community, in much the same ways as urban centers, which supports Britain's (2009, 2012) argument that language variation and change is not fundamentally different in urban and rural areas.

Given that the Elsewhere pattern is common in Canada and is becoming more common in US dialects of English, this raises the question of how different varieties of US English and Canadian English really are at this point. According to *ANAE*, the defining feature of Canadian English is the Canadian Shift, i.e. the Elsewhere Shift. However, as this study and previous research has shown, this shift, along with the associated changes in the back vowels, is spreading quite rapidly across US dialects. While there are subtle differences in the shift across different regions (e.g. nasal/continuous TRAP system in the US vs. non-nasal TRAP system in Canada), the overall outcomes appear to be very similar, as the examples of the most advanced speakers in the present study have shown. Likewise, the second feature that characterizes Canadian English, Canadian Raising, i.e. the centralization of the onset of PRICE and MOUTH before voiceless consonants, is not confined to Canada. Instead, *ANAE* shows that centralization of PRICE is very commonly found in the Inland North as well. Additionally, *ANAE* (p. 130) argues that Canadian Raising in Canada is not consistent enough to define Canada as a dialect region.

Generally, the national border between the US and Canada has been found to serve as a sharp phonological boundary (e.g. Boberg, 2000). The phonetic changes observed in Ogdensburg in the present study support this finding, as the ongoing changes that seem to be reaching Ogdensburg through diffusion are more likely adopted from the eastern neighbors, the dialectological North Country, rather than from Canada to the north, despite frequent cross-border communication. This is evidenced by the development of a nasal or continuous TRAP system in Ogdensburg, which is absent in Canadian English.

If the observation that rural communities are retreating from the NCS to the same extent as urban centers holds true for other parts of the Inland North, the findings of the present study might be an indication of decreasing dialect diversity in the English of North America. Not only is the Inland North becoming *less distinct* from surrounding dialects through the loss of its regionally marked variants, it seems to be becoming *more similar* through the adoption of a supra-local system, which is found in the US with increasing frequency. As the same pattern is characteristic of Canadian English as well, Canadian and regional US varieties appear to be becoming increasingly similar as well, despite limited diffusion of phonetic features across the national border. In 2006, *ANAE*, the most recent and comprehensive overview of speech patterns in the US and Canada, found regional dialect diversity to be *increasing* over time. Apparent-time developments toward vowel shifts like the NCS in the Inland North formed the main premises to support this conclusion. However, if the NCS is in fact disappearing on a regional scale, and Inland North communities are adopting the Elsewhere pattern instead, as the present study in combination with previous research suggests, this may no longer be true, and large-scale dialect leveling may be on the rise.

Appendices

Appendix A: Wordlist

Wordlist items, listed in the order in which they were presented to the participants:

1	tab	19	cut	37	feather
2	sketch	20	born	38	hockey
3	golf	21	pass	39	keg
4	pot	22	kept	40	barn
5	coffee	23	fought	41	cab
6	bus	24	top	42	bag
7	hut	25	bash	43	box
8	hush	26	best	44	cough
9	badge	27	toss	45	jaw
10	bed	28	hug	46	pep
11	bubble	29	cause	47	back
12	pop	30	bet	48	bought
13	path	31	dog	49	beg
14	cop	32	revolve	50	hutch
15	peck	33	chop	51	gawk
16	hub	34	bat	52	but
17	bad	35	huddle	53	pause
18	pod	36	sorry	54	deposit

Appendix B: Commutation Test

The line-up on top represents the order in which *cot* and *caught* were represented to the participants in writing. The second line-up represents the order in which the pre-recorded *cot* and *caught* were presented to them auditorily.

Caught	Cot	Cot	Caught	Cot	Caught	Caught	Cot	Cot	Caught
Cot	Cot	Caught	Cot	Caught	Caught	Cot	Caught	Caught	Cot

Appendix C: Matched Guise Order and Evaluation Sheet

"Linguistic and Cultural Practices in a Border Town" – Listening tasks – Anja Thiel

Part 1: Phrases
Please listen to the following recordings and try to rate each speaker on the scales below:

Rec #	Sentences	Comments
1	That's exactly what happened. Her dad's so sad about it. She has a passion for babbling.	
2	He did better in the second test. They confessed to the theft. They need good cheddar for their guests.	
3	The doctor was in shock about it. Those scotch bottles weren't in stock. He didn't bother stopping for them.	
4	I had my jacket in my bag. Her dad's so sad about it. That's exactly what happened.	
5	My buddy took the bus a couple times. She made double fudge brownies for me. They cuddled with the puppy.	
6	They were assessing the chef. They confessed to the theft. They need good cheddar for their guests.	
7	His jaw popped out of place. We got a good shot of the chalk cliffs.	
8	The doctor was in shock about it. Those scotch bottles weren't in stock. He didn't bother stopping for them.	
9	My buddy took the bus a couple times. She made double fudge brownies for me. They cuddled with the puppy.	
10	He did better in the second test. They confessed to the theft. They need good cheddar for their guests.	
11	I had my jacket in my bag. Her dad's so sad about it. She has a passion for babbling.	
12	My buddy took the bus a couple times. She made double fudge brownies for me. They cuddled with the puppy.	
13	The doctor was in shock about it. Those scotch bottles weren't in stock. He didn't bother stopping for them.	
14	His jaw popped out of place. We got a good shot of the chalk cliffs.	
15	I had my jacket in my bag. Her dad's so sad about it. That's exactly what happened.	
16	They were assessing the chef. They confessed to the theft. They need good cheddar for their guests.	
17	The doctor was in shock about it. Those scotch bottles weren't in stock. We didn't bother stopping for them.	
18	That's exactly what happened. Her dad's so sad about it. She has a passion for babbling.	
19	His jaw popped out of place. We got a good shot of the chalk cliffs.	

20	The doctor was in shock about it. Those scotch bottles weren’t in stock. He didn’t bother stopping for them.	
21	They were assessing the chef. They confessed to the theft. They need good cheddar for their guests.	
22	My buddy took the bus a couple times. She made double fudge brownies for me. They cuddled with the puppy.	
23	I had my jacket in my bag. Her dad’s so sad about it. She has a passion for babbling.	
24	The doctor was in shock about it. Those scotch bottles weren’t in stock. He didn’t bother stopping for them.	
25	My buddy took the bus a couple times. She made double fudge brownies for me. They cuddled with the puppy.	
26	His jaw popped out of place. We got a good shot of the chalk cliffs.	
27	They were assessing the chef. They confessed to the theft. They need good cheddar for their guests.	
28	My buddy took the bus a couple times. She made double fudge brownies for me. They cuddled with the puppy.	

This speaker sounds…	**Strongly agree**	**Agree**	**Weakly agree**	**Weakly disagree**	**Disagree**	**Strongly Disagree**
friendly:						
educated:						
old:						
local to Ogdensburg:						
Canadian:						

Part 2: Word pairs

1) Please indicate which of the following words you hear in your own recording:

1	☐ cot	☐ caught	☐ I can't tell
2	☐ cot	☐ caught	☐ I can't tell
3	☐ cot	☐ caught	☐ I can't tell
4	☐ cot	☐ caught	☐ I can't tell
5	☐ cot	☐ caught	☐ I can't tell
6	☐ cot	☐ caught	☐ I can't tell
7	☐ cot	☐ caught	☐ I can't tell
8	☐ cot	☐ caught	☐ I can't tell
9	☐ cot	☐ caught	☐ I can't tell
10	☐ cot	☐ caught	☐ I can't tell

2) Please indicate which of the following words you hear in the second recording:

1	☐ cot	☐ caught	☐ I can't tell
2	☐ cot	☐ caught	☐ I can't tell
3	☐ cot	☐ caught	☐ I can't tell
4	☐ cot	☐ caught	☐ I can't tell
5	☐ cot	☐ caught	☐ I can't tell
6	☐ cot	☐ caught	☐ I can't tell
7	☐ cot	☐ caught	☐ I can't tell
8	☐ cot	☐ caught	☐ I can't tell
9	☐ cot	☐ caught	☐ I can't tell
10	☐ cot	☐ caught	☐ I can't tell

Appendix D: Index of Sampled Speakers

The following table lists all 48 speakers included in this study, sorted by year of birth (youngest to oldest) and sample year (2016, 2008).

Pseudonym	YOB	Gender	Education	Sample Year
Grace	2002	f	student	2016
Mark	2001	m	student	2016
Ben	1999	m	student	2016
Will	1999	m	student	2016
Chloe	1998	f	student	2016
Jason	1998	m	student	2016
Allison	1993	f	college	2016
Daniel	1993	m	student	2016
Anthony	1991	m	no college	2016
Sophie	1991	f	college	2016
Megan	1990	f	college	2016
Lindsey	1987	f	college	2016
Summer	1987	f	college	2016
Ryan	1983	m	no college	2016
Aubrey	1980	f	college	2016
Rachel	1976	f	college	2016
Melissa	1972	f	college	2016
Sarah	1969	f	college	2016
Patrick	1968	m	college	2016
Amber	1967	f	college	2016
Ashley	1966	f	no college	2016
Monica	1965	f	college	2016
Breanna	1964	f	no college	2016
Amanda	1958	f	college	2016
Charlotte	1958	f	college	2016
Henry	1953	m	college	2016
Kelly	1953	f	college	2016
Stephanie	1952	f	no college	2016
Bethany	1950	f	college	2016
Helen	1949	f	college	2016
Ruth	1948	f	college	2016
Scott	1946	m	no college	2016
Brian	1945	m	no college	2016
Tracy	1944	f	college	2016
Eddie	1943	m	no college	2016
Gary	1941	m	college	2016
Richard	1941	m	college	2016
Donna	1935	f	no college	2016
Nicole	1932	f	college	2016
Shelley	1989	f	student	2008
JessicaJ	1988	f	student	2008

JessM	1986	f	college	2008
StacyB	1983	f	no college	2008
NoreenH	1982	f	student	2008
MikeP	1977	m	no college	2008
Jackie	1966	f	college	2008
Dan	1959	m	college	2008
Wanda R	1922	f	no college	2008

Appendix E: Consent Form

UNIVERSITÄT BERN

Consent for participation in a research interview

"Linguistic and Cultural Practices in a Border Town"

I agree to participate in a research project by Anja Thiel M.A. from the University of Bern, Department of English in Bern, Switzerland. I understand that the project is designed to gather information about linguistic and cultural practices in Ogdensburg, NY. I will be one of approximately 50 people being interviewed for this research. The purpose of this document is to specify the terms of my participation in the project through being interviewed.

1. I have been given sufficient information about this research project. The purpose of my participation as an interviewee in this project has been explained to me and is clear.

2. My participation as an interviewee in this project is voluntary. There is no explicit or implicit coercion whatsoever to participate. I understand that I will not be paid for my participation. I also understand that I may withdraw and discontinue participation at any time without penalty.

3. Participation involves being interviewed by the researcher, Anja Thiel M.A., from the University of Bern. The interview will last approximately 60 minutes. I allow the researcher to take written notes during the interview. I also allow audio recording of the interview. If I don't want to be recorded, I will not be able to participate in the study.

4. I have the right not to answer any of the questions. If I feel uncomfortable in any way during the interview session, I have the right to withdraw from the interview.

5. I have been given the explicit guarantee that, if I wish so, the researcher will not identify me by name or function in any reports using information obtained from this interview, and that my confidentiality as a participant in this study will remain secure. In all cases subsequent uses of the data will protect the anonymity of individuals and institutions. The data obtained from this interview may be used by the researcher for research and teaching purposes only.

6. I have read and understood the statements on this form. I have had all my questions answered to my satisfaction, and I voluntarily agree to these terms.

7. I have been given a copy of this consent form co-signed by the researcher.

________________	________________________	________________________
Date	Signature Participant	Signature Researcher

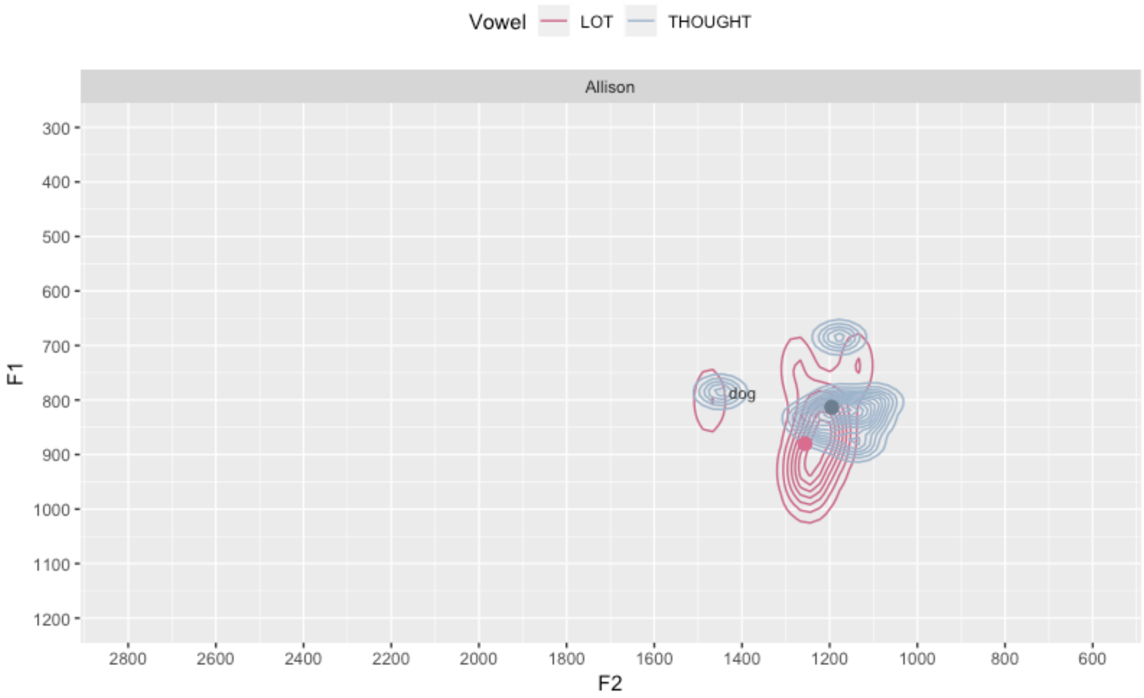
For further information, please contact:
Prof. David Britain
Chair of Modern English Linguistics, Department of English, University of Bern
Phone: +41 31 631 83 81
Email: britain@ens.unibe.ch

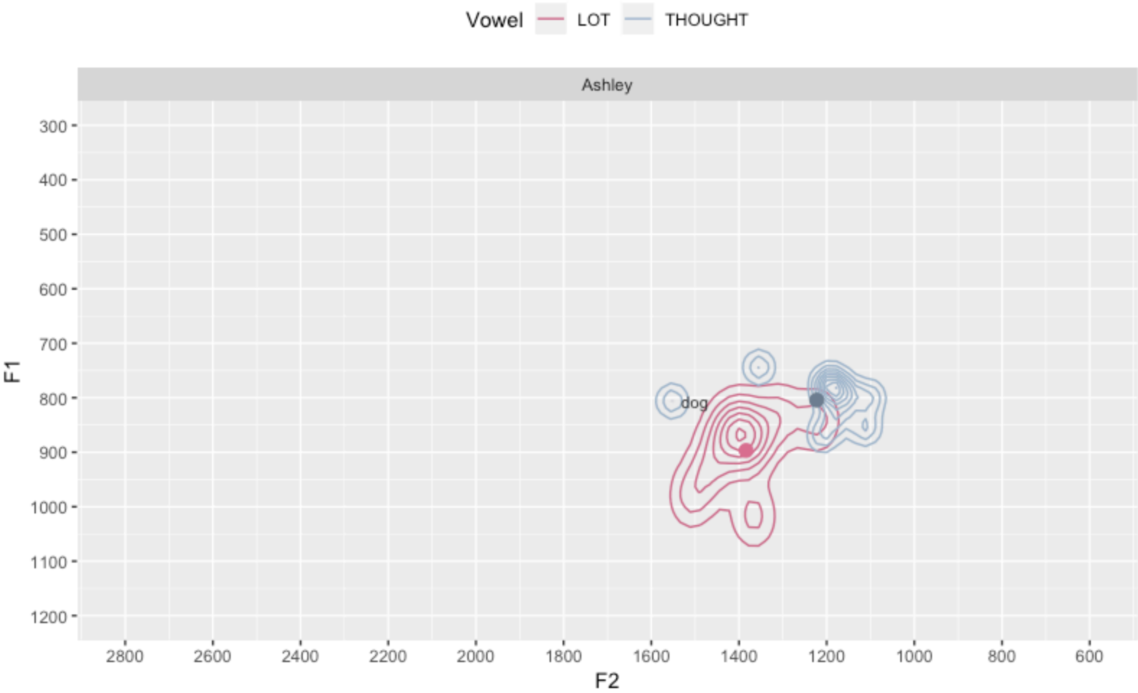
Appendix F: Stop Words

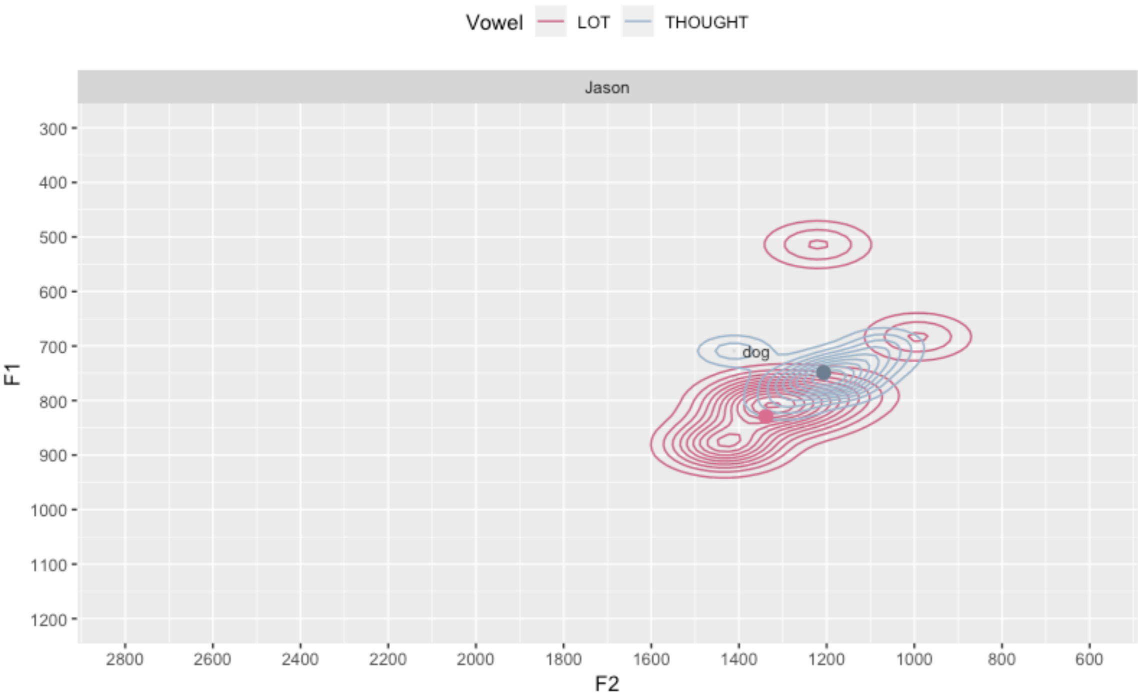
The following words were removed from the data sets and excluded from analyses:

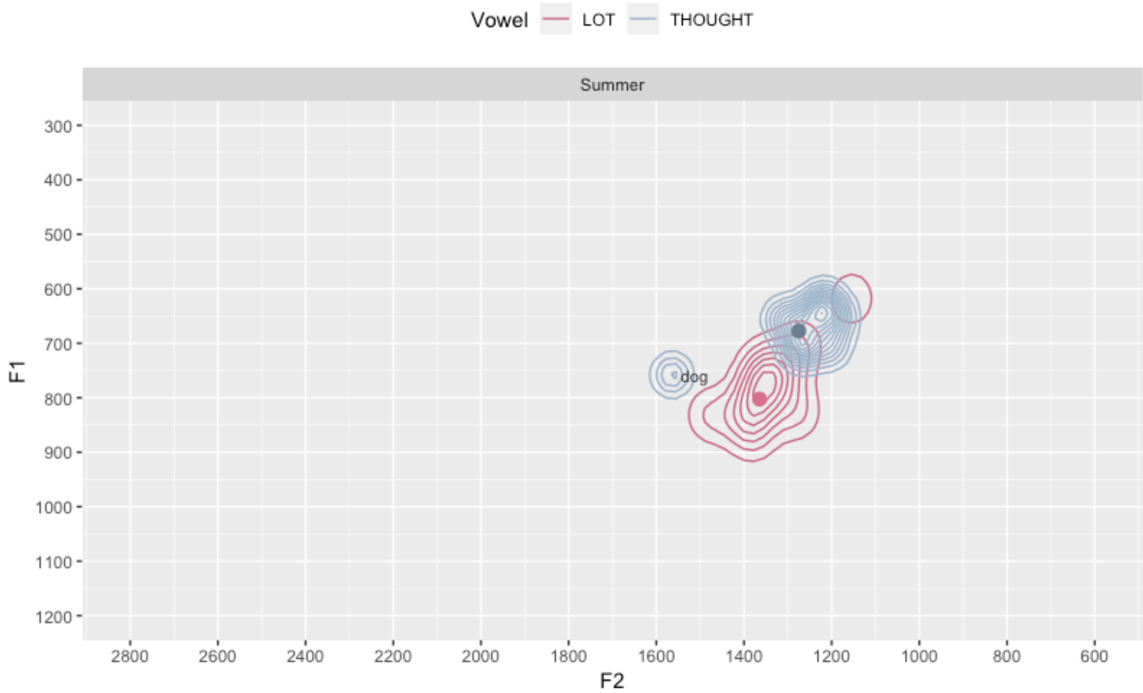
Prepositions	of, off, there, then, up, on, in, at, by, out, to, for, than, with, why, who, when
Pronouns (incl. all contracted forms)	where, what, I, he, his, him, it, my, me, mine, she, her, hers, they, them, we, us, our, you, your, all
Determiners	a, an, the, that, this, no
Interjections	ah, aha, ha, haha, aw, oh, uh, huh, uhm, hm, ya, yab, yeah, yay, okay, kay, nope, cuz, nah, woah, wow, bah, so
Conjunctions	And, but, or, if, as
Enumerators	one, ones, one's
Adverbs	yes, not,
Auxiliaries (incl. contractions, past tense, negations)	can, do, have, be, will, ain't
removed manually	one-letter words, unfinished words

Appendix G: Distribution of *dog* in Wordlist Style




Vowel
LOT
THOUGHT
Ashley
F1
300
400
500
600
700
800
900
1000
1100
1200
dog
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
F2


Vowel
LOT
THOUGHT
Jason
F1
300
400
500
600
700
800
900
1000
1100
1200
dog
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
F2


Vowel
LOT
THOUGHT
Summer
300
400
500
600
700
800
900
1000
1100
1200
F1
dog
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
F2

Appendix H: Phonological Environments

TRAP

Spontaneous speech

predictor	coefficient
s	-16.87
z	-48.45
t	-34.76
d	-58.99
b	-44.56
θ	-11.11
ð	-54.57
l	-44.87
f	-4.21
v	-47.07
tʃ	-32.13
dʒ	-74.46
ʃ	-49.51
ʒ	-67.54
k	-17.19
g	-34.43

Effects of following segment on F1 of TRAP in the regression model in Table 9. Reference level: [p]; p≈ 10^{-6}

predictor	coefficient
s	-17.18
z	47.59
t	3.51
d	79.05
b	13.91
θ	36.85
ð	-230.31
l	-5.47
f	-16.58
v	-46.07
tʃ	40.95
dʒ	33.56
ʃ	71.73
ʒ	35.04
k	-3.79
g	35.32

Effects of following segment on F2 of TRAP in the regression model in Table 10. Reference level: [p]; p≈ 0.016

predictor	coefficient
m	4.15
ŋ	91.27

Effects of following segment on F1 of TRAMP in the regression model in Table 11. Reference level: [n]; p≈ 10^{-13}

predictor	coefficient
m	-3.39
ŋ	-245.34

Effects of following segment on F2 of TRAMP in the regression model in Table 12. Reference level: [n]; p≈ $9x10^{-13}$

predictor	coefficient
s	-13.87
z	-58.69
n	-79.63
t	-35.77
d	-51.66
m	-74.88
b	-34.95
θ	-4.61
ð	-57.01
l	-38.32
f	-4.52
v	-40.92
tʃ	-30.87
dʒ	-70.67
ʃ	-48.11
ʒ	-60.01
k	-17.79
g	-25.55

Effects of following segment on F1 of TRAP and TRAMP in the regression model in Table 13. Reference level: [p]; p≈ $4x10^{-14}$

predictor	coefficient
s	-27.44
z	38.78
n	209.52
t	-8.29
d	59.76
m	203.67
b	1.29
θ	24.76
ð	-230.2
l	-20.94
f	-26.19
v	-61.47
tʃ	29.37
dʒ	33.88
ʃ	53.19
ʒ	25.43
k	-3.63
g	25.91

Effects of following segment on F2 of TRAP and TRAMP in the regression model in Table 14. Reference level: [p]; p≈ $3x10^{-10}$

predictor	coefficient
s	-16.87
z	-48.45
t	-34.76
d	-58.99
b	-44.56
θ	-11.11
ð	-54.57
l	-44.87
f	-4.21
v	-47.07
tʃ	-32.13
dʒ	-74.46
ʃ	-49.51
ʒ	-67.54
k	-17.19
g	-34.43

Effects of following segment on F1 of TRAP in the regression model in Table 15. Reference level: [p]; p≈ $2x10^{-6}$

predictor	coefficient
m	8.73
ŋ	72.92

Effects of following segment on F1 of TRAMP in the regression model in Table 16. Reference level: [n]; p≈ 10^{-10}

predictor	coefficient
s	-16.34
z	19.1
t	7.82
d	74.25
b	9.75
θ	14.64
ð	-111
l	-9.45
f	-19.1
v	-48.76
tʃ	46.74
dʒ	41.38
ʃ	66.25
ʒ	37.18
k	4.3
g	37.1

Effects of following segment on F2 of TRAP in the regression model in Table 17. Reference level: [p]; p≈ 0.035

predictor	coefficient
m	-9.5
ŋ	-187.76

Effects of following segment on F2 of TRAMP in the regression model in Table 18. Reference level: [n]; p≈ $2x10^{-8}$

Wordlist style

predictor	coefficient
t	-3.35
d	-52.06
b	-41.11
θ	-11.72
dʒ	-38.03
ʃ	-32.51
k	-47.48
g	-25.61

Effects of following segment on F1 of TRAP in the regression model in Table 21. Reference level: [s]; p≈ 0.039

predictor	coefficient
t	82.51
d	16.14
b	83.48
θ	34.27
dʒ	116.54
ʃ	111.21
k	-75.83
g	52.65

Effects of following segment on F2 of TRAP in the regression model in Table 22. Reference level: [s]; p≈ $2x10^{-5}$

predictor	coefficient
t	-10.87
d	-65.67
b	-38.78
θ	-8.19
l	-37.12
dʒ	-42.84
ʃ	-34.77
k	-34.72
g	-40.11

Effects of following segment on F1 of TRAP in the regression model in Table 23. Reference level: [s]; p≈ 0.452

predictor	coefficient
t	78.44
d	44.35
b	78.01
θ	30.23
l	-228.52
dʒ	101.74
ʃ	92.14
k	-204.56
g	54.97

Effects of following segment on F2 of TRAP in the regression model in Table 24. Reference level: [s]; p≈ 0.647

Style shifting

predictor	coefficient
s	-13.97
z	-55.9
t	-35.37
d	-53.91
b	-35.19
θ	-5.12
ð	-58.12
l	-40.6
f	-2.59
v	-43.85
tʃ	-30.55
dʒ	-65.33
ʃ	-44.13
ʒ	-61.11
k	-17.51
g	-35.26

Effects of following segment on F1 of TRAP in the regression model in Table 25. Reference level: [p]; p≈ 0.039

predictor	coefficient
s	-17.74
z	35.24
t	8.42
d	71.83
b	11.21
θ	24.32
ð	-220
l	-6.83
f	-22.41
v	-51.01
tʃ	35.13
dʒ	55.14
ʃ	70.16
ʒ	27.24
k	4.49
g	32.79

Effects of following segment on F2 of TRAP in the regression model in Table 26. Reference level: [p]; p≈$2x10^{-5}$

DRESS

Spontaneous speech

predictor	coefficient
s	-8.36
z	-84.14
n	-50.75
t	-40.9
d	-45.98
m	-27.38
b	-62.5
θ	-5.52
ð	-35.74
-	-121.03
l	-25.44
f	-11.47
v	-40.86
tʃ	-73.08
dʒ	-85.39
ʃ	-15.44
ʒ	-50.02
r	-157.06
ŋ	102.87
k	-10.75
g	-48.94

Effects of following segment on F1 of DRESS in the regression model in Table 30. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
s	-7.02
z	28.14
n	136.42
t	119.17
d	110.34
m	51.16
b	6.54
θ	153.81
ð	-5.21
-	497.15
l	-17.51
f	74.65
v	74.12
tʃ	180.31
dʒ	160.81
ʃ	-20.62
ʒ	43.39
r	311.32
ŋ	90.92
k	74.87
g	253.62

Effects of following segment on F2 of DRESS in the regression model in Table 31. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
s	-11.31
z	-50.33
n	-51.42
t	-42.35
d	-47.92
m	-30.58
b	-65.26
θ	-6.38
ð	-36.8
-	-121.88
- *r*	-146.98
l	-25.57
f	-13.69
v	-41.71
tʃ	-64.07
dʒ	-87.74
ʃ	-17.1
ʒ	-50.76
r	-158.2
ŋ	102.55
k	-12.99
g	-49.67

Effects of following segment on F1 of DRESS in the regression model in Table 32. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
s	-3.4
z	43.47
n	140.82
t	124.44
d	117.13
m	62.79
b	11.99
θ	116.02
ð	33.42
-	500.52
- *r*	262.29
l	-21.41
f	69.54
v	77.46
tʃ	173.62
dʒ	161.23
ʃ	-17.12
ʒ	47.41
r	317.62
ŋ	94.01
k	78.88
g	247.27

Effects of following segment on F2 of DRESS in the regression model in Table 33. Reference level: [p]; p≈ $2x10^{-16}$

Wordlist style

predictor	coefficient
s	-66.44
t	-50.57
d	-94.3
ð	-47.76
tʃ	-68.02
k	13.77
g	-95.58

Effects of following segment on F1 of DRESS in the regression model in Table 34. Reference level: [p]; p≈ 0.064

predictor	coefficient
s	1.576
t	160.4
d	203.65
ð	-53.35
tʃ	239.54
k	85.53
g	262.47

Effects of following segment on F2 of DRESS in the regression model in Table 35. Reference level: [p]; p≈ 0.041

predictor	coefficient
s	-34.06
n	-17.99
t	-51.8
d	-86.42
ð	-45.89
- r	-207.3
tʃ	-78.52
k	-13.53
g	-91.93

Effects of following segment on F1 of DRESS in the regression model in Table 36. Reference level: [p]; p≈ 0.299

predictor	coefficient
s	112.89
n	238.65
t	124.96
d	-39.52
ð	-79.39
- r	795.9
tʃ	228.95
k	148.65
g	242.42

Effects of following segment on F2 of DRESS in the regression model in Table 37. Reference level: [p]; p≈ 0.073

Style shifting

predictor	coefficient
s	-10.31
z	-86.3
n	-52.35
t	-42.92
d	-48.13
m	-29.17
b	-64.24
θ	-6.93
ð	-31.59
-	-122.71
l	-27.3
f	-13.46
v	-42.62
tʃ	-57.81
dʒ	-87.19
ʃ	-17.15
ʒ	-51.2
r	-158.82
ŋ	99.84
k	-11.16
g	-54.6

Effects of following segment on F1 of DRESS in the regression model in Table 38. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
s	-4.16
z	31.67
n	139.63
t	124.83
d	114.47
m	53.88
b	9.62
θ	154.95
ð	-4.62
-	501.12
l	-14.61
f	77.02
v	77.15
tʃ	189.1658
dʒ	162.79
ʃ	-16.87
ʒ	43.55
r	314.33
ŋ	94.25
k	78.92
g	273.63

Effects of following segment on F2 of DRESS in the regression model in Table 39. Reference level: [p]; p≈ $2x10^{-16}$

STRUT

Spontaneous speech

predictor	coefficient
s	-26.35
z	-45.55
n	-28.3
t	-1.14
d	-72.15
m	-36.28
b	-45.99
θ	-7.88
ð	-33.76
l	-54.28
f	-18.02
v	-39.05
tʃ	-49.06
dʒ	-13.63
ʃ	4.17
ŋ	-35.1
k	-6.49
g	-43.19

Effects of following segment on F1 of STRUT in the regression model in Table 40. Reference level: [p]; p≈ 0.002

predictor	coefficient
s	-14.87
z	63.16
n	2.48
t	26.35
d	80.66
m	-9.71
b	-37.56
θ	38.69
ð	-9.91
l	-150.59
f	-39.29
v	-66.99
tʃ	153.25
dʒ	163.69
ʃ	87.59
ŋ	-29.87
k	33.61
g	-40.66

Effects of following segment on F2 of STRUT in the regression model in Table 41. Reference level: [p]; p≈ $2x10^{-6}$

predictor	coefficient
s	-26.76
z	-45.3
n	-31.11
t	-1.05
d	-69.7
m	-35.91
b	-48.46
θ	-40.74
ð	-35.05
l	-52.39
f	-18.86
v	-40.01
tʃ	-41.28
dʒ	6.57
ʃ	-53.98
ʒ	-30.62
ŋ	-35.92
k	-21.16
g	-41.51

Effects of following segment on F1 of STRUT in the regression model in Table 42. Reference level: [p]; p≈ 0.025

predictor	coefficient
s	-7.17
z	60
n	12.57
t	28.45
d	92.29
m	-1.37
b	-27.88
θ	48.55
ð	15.15
l	-142.59
f	-34.48
v	-43.95
tʃ	140.35
dʒ	152.57
ʃ	79.1
ʒ	-40.67
ŋ	-33.61
k	63.84
g	-24.3

Effects of following segment on F2 of STRUT in the regression model in Table 43. Reference level: [p]; p≈ 10^{-7}

Wordlist style

predictor	coefficient
s	-49.18
t	-11.19
d	-58.31
b	-67.98
tʃ	-18.62
ʃ	-23.97
g	-24.45

Effects of following segment on F1 of STRUT in the regression model in Table 44. Reference level: [p]; p≈ 0.494

predictor	coefficient
s	-113.44
t	21.17
d	-71.47
b	-169.79
tʃ	1.52
ʃ	-27.92
g	-97.27

Effects of following segment on F2 of STRUT in the regression model in Table 45. Reference level: [p]; p≈ 0.528

predictor	coefficient
s	-49.19
t	-11.2
d	-58.34
b	-67.99
tʃ	-18.67
ʃ	-23.99
k	-50.93
g	-24.45

Effects of following segment on F1 of STRUT in the regression model in Table 46. Reference level: [p]; p≈ 0.541

predictor	coefficient
s	-112.55
t	22.08
d	-70.89
b	-168.9
tʃ	2.43
ʃ	-26.97
k	-133.39
g	-96.23

Effects of following segment on F2 of STRUT in the regression model in Table 47. Reference level: [p]; p≈ 0.576

Style shifting

predictor	coefficient
s	-27.03
z	-45.56
n	-28.26
t	-2.11
d	-69.08
m	-36.37
b	-46.58
θ	-8.1
ð	-34.23
l	-54.87
f	-18.21
v	-38.89
tʃ	-38.18
dʒ	-13.16
ʃ	-3.38
ŋ	-35.2
k	-6.28
g	-39.15

Effects of following segment on F1 of STRUT in the regression model in Table 48. Reference level: [p]; p≈ 0.002

predictor	coefficient
s	-8.15
z	60.64
n	6.09
t	21.38
d	79.02
m	-8
b	-34.98
θ	45.94
ð	-2.52
l	-144.73
f	-34.85
v	-48.48
tʃ	131.3
dʒ	180.6
ʃ	90.83
ŋ	-24.66
k	36.35
g	-31.59

Effects of following segment on F2 of STRUT in the regression model in Table 49. Reference level: [p]; p≈ $9x10^{-7}$

KIT

predictor	coefficient
s	-32.36
z	-67.30
n	-30.03
t	-23.31
d	-37.37
m	-18.51
b	-39.41
θ	-35.72
ð	-22.71
-	-157.60
l	-23.58
f	-31.27
v	-35.28
tʃ	-30.88
dʒ	-59.48
ʃ	-56.18
ʒ	-92.54
r	-92.77
ŋ	-36.10
k	-13.07
g	-58.84

Effects of following segment on F1 of KIT in the regression model in Table 50. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
s	-27.74
z	-13.77
n	159.43
t	39.27
d	118.27
m	38.23
b	-5.84
θ	49.31
ð	-28.67
-	61.17
l	-25.43
f	55.68
v	22.71
tʃ	121.36
dʒ	96.90
ʃ	52.37
ʒ	113.53
r	247.19
ŋ	207.11
k	128.6
g	219.53

Effects of following segment on F2 of KIT in the regression model in Table 51. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
s	-32.59
z	-70.16
n	-27.71
t	-24.43
d	-37.65
m	-16.41
b	-39.25
θ	-35.98
ð	-22.54
-	-156.79
l	-23.32
f	-32.12
v	-37.10
tʃ	-30.74
dʒ	-59.76
ʃ	-51.04
ʒ	-92.73
r	-92.78
ŋ	-36.88
k	-14.67
g	-59.9

Effects of following segment on F1 of KIT in the regression model in Table 52. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
s	-7.79
z	-39.06
n	173.97
t	43.12
d	122.10
m	33.04
b	-3.33
θ	51.99
ð	-26.63
-	65.99
l	-34.11
f	56.70
v	9.04
tʃ	129.53
dʒ	94.92
ʃ	32.50
ʒ	116.65
r	250.33
ŋ	194.20
k	131.23
g	222.92

Effects of following segment on F2 of KIT in the regression model in Table 53. Reference level: [p]; p≈ $2x10^{-16}$

LOT

Spontaneous speech

predictor	coefficient
s	-26.48
z	-17.68
n	-46.32
t	-28.91
d	-26.76
m	-13.97
b	-3.45
θ	-18.7
ð	-11.38
-	-44.21
- *l*	-13.93
l	-54.79
f	-63.11
v	33.59
tʃ	-34.01
dʒ	-31.28
ʃ	-57.23
ʒ	7.76
r	-106.24
ŋ	-8.28
k	1.61
g	-50.9

Effects of following segment on F1 of LOT included in the overall regression model in Table 64. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
s	-87.27
z	-61.48
n	14.58
t	15.01
d	9.59
m	16.58
b	-8.57
θ	-40.81
ð	-43.27
-	-130.87
- *l*	159.97
l	-92.86
f	-160.43
v	-0.26
tʃ	9.02
dʒ	8.02
ʃ	-141.87
ʒ	34.53
r	-34.46
ŋ	-89.16
k	-1.92
g	-75.24

Effects of following segment on F2 of LOT included in the overall regression model in Table 65. Reference level: [p]; p≈ $5x10^{-5}$

predictor	coefficient
s	-19.05
z	2.61
n	-38.44
t	-24.20
d	-19.43
m	-7.88
b	6.27
θ	-11.50
ð	-4.48
-	-37.53
- *h*	-6.76
- *r*	-101.91
l	-54.15
f	-56.94
v	38.91
tʃ	-13.56
dʒ	-24.72
ʃ	-50.45
ʒ	15.02
r	-99.77
ŋ	-1.68
k	5.25
g	-43.73

Effects of following segment on F1 of LOT in the regression model in Table 66. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
s	-85.69
z	-69.29
n	12.62
t	4.51
d	-1.30
m	18.21
b	-17.00
θ	-46.74
ð	-74.04
-	-138.89
- *h*	152.01
- *r*	-181.64
l	-84.65
f	-159.29
v	-25.68
tʃ	-18.03
dʒ	-6.78
ʃ	-139.24
ʒ	27.59
r	-43.05
ŋ	-132.69
k	-11
g	-77.93

Effects of following segment on F2 of LOT in the regression model in Table 67. Reference level: [p]; p≈ $2x10^{-5}$

Wordlist style

predictor	coefficient
z	-5.54
t	9.63
d	-2.72
l	-140.35
r	-158.46
k	-3.95

Effects of following segment on F1 of LOT included in the overall regression model in Table 68. Reference level: [p]; p≈ 0.016

predictor	coefficient
z	-16.34
t	-44.067
d	-39.66
l	-229.56
r	-80.54
k	-34.68

Effects of following segment on F2 of LOT included in the overall regression model in Table 69. Reference level: [p]; p≈ 0.036

predictor	coefficient
z	-11.53
t	3.78
d	-1.84
m	-180.62
l	-119.94
r	-165.86
k	-41.20

Effects of following segment on F1 of LOT in the regression model in Table 70. Reference level: [p]; p≈ 0.119

predictor	coefficient
z	-28.36
t	-48.11
d	-35.18
m	369.54
l	-172.23
r	-94.69
k	81.92

Effects of following segment on F2 of LOT in the regression model in Table 71. Reference level: [p]; p≈ 0.208

Minimal pairs

predictor	coefficient
t	65.15
d	56.67
l	69.81
r	-54.63
k	95.28

Effects of following segment on F1 of LOT in the regression model in Table 72. Reference level: [n]; p≈ 0.002

predictor	coefficient
t	72.76
d	41.51
l	-32.35
r	-129.04
k	56.12

Effects of following segment on F1 of LOT included in the overall regression model in Table 73. Reference level: [n]; p≈ 0.057

predictor	coefficient
t	61.01
d	51.47
θ	1.96
- *r*	-193.97
l	55.03
r	-57.77
k	88.06

Effects of following segment on F2 of LOT in the regression model in Table 74. Reference level: [n]; p≈ 0.0001

predictor	coefficient
t	28.20
d	15.64
θ	-64.13
- *r*	-195.98
l	-62.44
r	-144.96
k	25.73
g	-3.58

Effects of following segment on F2 of LOT in the regression model in Table 75. Reference level: [n]; p≈ 0.059

Style shifting

predictor	coefficient
s	-25.68
z	-13.45
n	-47.04
t	-26.06
d	-26.14
m	-12.6
b	-3.08
θ	-21.89
ð	-10.37
-	-45.64
- *h*	-15.47
l	-56.84
f	-59.97
v	35.19
tʃ	-34.17
dʒ	-33.0985
ʃ	-56.7791
ʒ	6.29348
r	-106.023
ŋ	-10.9446
k	4.40191
g	-49.0948

Effects of following segment on F1 of LOT included in the overall regression model in Table 76. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
s	-92.66
z	-52.65
n	11.88
t	7.11
d	-3.91
m	21.83
b	-15.46
θ	-54.3
ð	-72.76
-	-130.76
- *h*	150.06
l	-97.82
f	-155.68
v	-17.71
tʃ	-2.28
dʒ	-4.0135
ʃ	-139.826
ʒ	32.3465
r	-40.3143
ŋ	-130.494
k	-7.0415
g	-73.7331

Effects of following segment on F2 of LOT included in the overall regression model in Table 77. Reference level: [p]; p≈ $2x10^{-16}$

THOUGHT

Spontaneous speech

predictor	coefficient
z	-27.59
n	-17.06
t	-29.85
d	-30.91
b	41.88
θ	90.54
-	-5.08
l	-37.85
f	1.97
ʃ	-25.81
r	-238.84
ŋ	-30.75
k	-4.20
g	-15.77

Effects of following segment on F1 of THOUGHT included in the overall regression model in Table 78. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
z	-86.46
n	-54.33
t	-48.13
d	-42.16
b	106.70
θ	52.98
-	2.31
l	-121.32
f	-36.08
ʃ	-171.76
r	-266.38
ŋ	-58.91
k	-33.96
g	-8.00

Effects of following segment on F2 of THOUGHT in the regression model in Table 79. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
z	-21.99
n	-10.55
t	-26.72
d	-24.48
b	44.57
θ	74.61
-	-5.64
l	-39.87
f	1.47
ʃ	-22.70
r	-234.87
ŋ	-26.52
k	-1.81
g	-6.78

Effects of following segment on F1 of THOUGHT in the regression model in Table 80. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
z	-85.78
n	-54.75
t	-48.61
d	-42.9
b	106.4
θ	21.09
-	1.73
l	-118.82
f	-36.39
ʃ	-171.79
r	-266.65
ŋ	-57.75
k	-33.91
g	-13.47

Effects of following segment on F2 of THOUGHT in the regression model in Table 81. Reference level: [p]; p≈ $2x10^{-16}$

Wordlist style

predictor	coefficient
z	-47.24
t	-52.28
-	-47.03
f	-5.67
r	-291.27
k	-56.12
g	-44.84

Effects of following segment on F1 of THOUGHT included in the overall regression model in Table 82. Reference level: [s]; p≈ 0.012

predictor	coefficient
z	-31.75
t	-92.06
-	14.78
f	26.47
r	-329.31
k	0.53
g	53.4

Effects of following segment on F2 of THOUGHT in the regression model in Table 83. Reference level: [s]; p≈ 0.033

predictor	coefficient
z	-29.82
t	-80.81
-	-4.24
l	-80.5
f	34.97
r	-347.11
k	38.83
g	49.58

Effects of following segment on F1 of THOUGHT in the regression model in Table 84. Reference level: [s]; p≈ 0.002

predictor	coefficient
z	-29.82
t	-80.81
-	-4.24
l	-80.50
f	34.97
r	-347.11
k	38.83
g	49.58

Effects of following segment on F2 of THOUGHT in the regression model in Table 85. Reference level: [s]; p≈ 0.137

Minimal pairs

predictor	coefficient
t	41.75
d	47.34
l	46.10
r	-206.17
k	46.04

Effects of following segment on F1 of THOUGHT included in the overall regression model in Table 86. Reference level: [n]; p≈ $9x10^{-5}$

predictor	coefficient
t	101.17
d	96.87
l	52.89
r	-238.55
k	66.9

Effects of following segment on F2 of THOUGHT in the regression model in Table 87. Reference level: [n]; p≈ 0.0005

predictor	coefficient
t	39.82
d	44.43
l	40.28
r	-214.37
k	41.36

Effects of following segment on F1 of THOUGHT in the regression model in Table 88. Reference level: [n]; p≈ $2x10^{-15}$

predictor	coefficient
t	83.68
d	75.6
l	26.46
r	-231.87
k	44.24

Effects of following segment on F2 of THOUGHT in the regression model in Table 89. Reference level: [n]; p≈ 0.0003

Style shifting

predictor	coefficient
z	-31.29
n	-30.63
t	-26.65
d	-9.49
b	48.71
θ	93.25
-	-13.92
l	-36.36
f	3.85
ʃ	-29.62
r	-236.91
ŋ	-29.39
k	-5.45
g	-14.40

Effects of following segment on F1 of THOUGHT included in the overall regression model in Table 90. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
z	-78.98
n	-78
t	-54.57
d	-43.3
b	107.79
θ	59.81
-	-2.65
l	-121.61
f	-32.35
ʃ	-182.41
r	-259.25
ŋ	-58.36
k	-37.6
g	-8.51

Effects of following segment on F2 of THOUGHT included in the overall regression model in Table 91. Reference level: [p]; p≈ $2x10^{-16}$

Duration LOT & THOUGHT

predictor	coefficient
s	-0.02
z	0.01
n	-0.02
t	0.00
d	0.02
m	-0.03
b	0.01
θ	-0.04
ð	0.03
-	0.03
- *h*	0.01
l	-0.02
f	0.00
v	-0.01
tʃ	-0.01
dʒ	-0.03
ʃ	0.02
ʒ	0.06
r	-0.05
ŋ	-0.01
k	0.0002
g	0.02

Effects of following segment on vowel duration in spontaneous speech the regression model in Table 102. Reference level: [p]; p≈ $2x10^{-16}$

predictor	coefficient
s	0.00
z	0.05
t	-0.02
d	0.10
-	0.06
l	-0.01
f	-0.05
r	-0.03
k	0.00
g	0.07

Effects of following segment on vowel duration in wordlist style in the regression model in Table 102. Reference level: [p]; p≈ 0.038

predictor	coefficient
t	-0.05
d	0.05
l	-0.11
r	-0.09
k	-0.05

Effects of following segment on vowel duration in minimal pairs in the regression model in Table 102. Reference level: [p]; p≈ $3x10^{-14}$

GOOSE, FOOT & GOAT

predictor	coefficient
s	143.51
z	192.46
n	101.40
t	-7.37
d	51.51
m	-139.57
b	24.08
θ	-61.82
ð	-423.25
-	-27.80
- *h*	-289.63
l	-337.23
f	14.89
v	-318.27
tʃ	79.86
dʒ	66.98
ʃ	4.22
ʒ	107.03
r	-138.00
k	107.36
g	-594.37

Effects of following segment on F2 of GOOSE included in the overall regression model in Table 112. Reference level: [p]; p≈ 5x10^{-7}

predictor	coefficient
d	371.75
m	-276.93
l	-143.15
ʃ	-210.49
r	10.69
k	-11.06
g	304.58

Effects of following segment on F2 of FOOT included in the overall regression model in Table 113. Reference level: [p]; p≈ 0.0002

predictor	coefficient
s	-50.08
z	-4.77
n	45.99
t	8.80
d	60.68
m	2.72
b	24.85
θ	-44.19
ð	-140.81
-	52.96
- *w*	-11.42
- *h*	-37.66
l	-123.19
f	-121.79
v	67.16
tʃ	-35.99
ʃ	97.51
ʒ	-50.56
k	-76.10
g	60.27

Effects of following segment on F2 of GOAT at 20% in the regression model in Table 114. Reference level: [p]; p≈ 0.073

predictor	coefficient
s	84.13
z	115.93
n	84.34
t	99.21
d	159.44
m	104.46
b	106.97
θ	54.07
ð	-0.86
-	96.34
- *w*	176.31
- *h*	358.22
l	-15.03
f	26.98
v	78.17
tʃ	86.91
ʃ	282.86
ʒ	39.81
k	24.84
g	200.38

Effects of following segment on F2 of GOAT at 35% in the regression model in Table 114. Reference level: [p]; p≈ 0.019

predictor	coefficient
s	101.57
z	108.56
n	14.91
t	65.77
d	63.3
m	99.03
b	26.55
θ	17.10
ð	34.54
-	80.67
- *w*	28.36
- *h*	228.27
l	-102.75
f	-18.24
v	-31.83
tʃ	104.8
ʃ	248.13
ʒ	29.49
k	-10.91
g	162.23

Effects of following segment on F2 of GOAT at 50% in the regression model in Table 114. Reference level: [p]; p≈ 0.006

References

Abbuhl, R., Gass, S., & Mackey, A. (2014). Experimental research design. In R. J. Podesva & D. Sharma (Eds.), *Research Methods in Linguistics* (pp. 116–134). New York: Cambridge University Press.

Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America*, *116*(5), 3099–3107.

Ash, S. (1999). Word list data and the measurement of sound change. *New Ways of Analyzing Variation*. Presented at the NWAV28, Toronto.

Ash, S. (2002). The distribution of a phonemic split in the Mid- Atlantic region: Yet more on short a. *University of Pennsylvania Working Papers in Linguistics*, *8*(3 Selected Papers from NWAV30), 1–15.

Audacity Team. (2012). Audacity® (Version 2.0.5). Retrieved from http://audacityteam.org

Baayen, H. R. (2014). Multivariate statistics. In R. J. Podesva & D. Sharma (Eds.), *Research Methods in Linguistics* (pp. 337–372). New York: Cambridge University Press.

Bailey, C. J. (1970). Variation and language theory. *University of Hawaii Working Papers in Linguistic*, *2*, 161–23.

Bailey, George. (2016). Automatic Detection of Sociolinguistic Variation Using Forced Alignment. *University of Pennsylvania Working Papers in Linguistics*, *22*(2 Selected Papers from NWAV44).

Bailey, Guy, Wikle, T., Tillery, J., & Sand, L. (1993). Some patterns of linguistic diffusion. *Language Variation and Change*, *5*(3), 359–390.

Baranowski, M. A. (2007). *Phonological Variation and Change in the Dialect of Charleston, South Carolina*. Durham: Duke University Press.

Baranowski, M. A. (2013). On the role of social factors in the loss of phonemic distinctions. *English Language and Linguistics*, *17*(02), 271–295.

Barton, M. (1830). *Something new, comprising a new and perfect alphabet*. Retrieved from http://books.google.com/books?id=dhxLdimSOS4C

Becker, K. (2014). (r) we there yet? The change to rhoticity in New York City English. *Language Variation and Change*, *26*(2), 141–168.

Beckford Wassink, A. (2006). A geometric representation of spectral and temporal vowel features: quantification of vowel overlap in three linguistic varieties. *The Journal of the Acoustical Society of America*, *119*(4), 2334–2350.

Benson, E. J., Fox, M. J., & Balkman, J. (2011). The Bag That Scott Bought: The Low Vowels In Northwest Wisconsin. *American Speech*, *86*(3), 271–311.

Bigham, D. S. (2007). Vowel Variation in Southern Illinois. *American Dialect Society Annual Meeting*. Presented at the Anaheim. Anaheim.

Bigham, D. S. (2009). Northern California Vowels In Southern Illinois. *American Dialect Society Annual Meeting*, 17. San Francisco.

Bigham, D. S. (2010). Correlation of the Low-Back Vowel Merger and TRAP-Retraction. *University of Pennsylvania Working Papers in Linguistics*, *15*(2 Selected papers from NWAV37).

Boberg, C. (2000). Geolinguistic diffusion and the U.S.–Canada border. *Language Variation and Change*, *12*(1), 1–24.

Boberg, C. (2001). The Phonological Status of Western New England. *American Speech*, *76*(1), 3–29.

Boberg, C. (2004). Real and apparent time in language change: Late adoption of changes in Montreal English. *American Speech*, *79*(3), 250–269.

Boberg, C. (2005). The Canadian shift in Montreal. *Language Variation and Change*, *17*(2), 133–154.

Boberg, C. (2011). Reshaping the Vowel System: An Index of Phonetic Innovation in Canadian English. *University of Pennsylvania Working Papers in Linguistics*, *17*(2 Selected Papers from NWAV39), 21–29.

Boberg, C., & Strassel, S. M. (1995). Phonological Change in Cincinnati. *University of Pennsylvania Working Papers in Linguistics*, *2*(2 Proceedings of the 19th Annual Penn Linguistics Colloqium).

Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer (Version 6.0.19). Retrieved from http://www.praat.org/

Bowie, D., Bushnell, T., Collins, A., Kudenov, P., Meisner, S., Ray, M., … Kubitskey, K. (2012). A very northern California shift? The vowel system of southcentral Alaska. *New Ways of Analyzing Variation*. Presented at the NWAV41, Bloomington.

Britain, D. (2009). *"Big bright lights" versus "green and pleasant land"? The unhelpful dichotomy of "urban" versus "rural" in dialectology* (E. Al-Wer & C. Holes, Eds.). Leiden: Brill.

Britain, D. (2012). Countering the urbanist agenda in variationist sociolinguistics: dialect contact, demographic change and the rural-urban dichotomy. In *Dialectological and Folk Dialectological Concepts of Space*. Berlin, Boston: Walter de Gruyter.

Britain, D. (2014). Where North Meets South? Contact, Divergence and the Routinisation of the Fenland Dialect Boundary. In D. Watt & Llamas, Carmen (Eds.), *Languages, borders and identity* (pp. 27–43). Edinburgh: Edinburgh University Press.

Burke, A. (2018). "We still love you": Border town urging Canadian visitors to keep coming.

Callary, R. E. (1975). Phonological change and the development of an urban dialect in Illinois. *Language in Society*, *4*(02), 155.

Campbell-Kibler, K. (2006a). *Listener perceptions of sociolinguistic variables: The case of (ING)*. Stanford University.

Campbell-Kibler, K. (2006b). Methods for the Study of the Social Structure of Linguistic Variation. *Berkeley Linguistics Society*, *32*(1), 73–84.

Campbell-Kibler, K. (2006c). *Variation and the listener: The contextual meanings of (ING)*.

Campbell-Kibler, K. (2007). Accent, (ING), and the social logic of listener perceptions. *American Speech*, *82*(1).

Campbell-Kibler, K. (2008). *I'll be the judge of that: Diversity in social perceptions of (ING)* (Vol. 37).

Campbell-Kibler, K. (2010). Sociolinguistics and Perception. *Language and Linguistics Compass*, *4*(6), 377–389.

Campbell-Kibler, K. (2011). The sociolinguistic variant as a carrier of social meaning. *Language Variation and Change*, *22*(3), 423–441.

Chambers, J. K. (1995). *Sociolinguistic Theory*. Wiley.

Chambers, J. K. (2008). *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Wiley.

Chevalier, A. (2016). *Globalisation versus internal development: the reverse short front vowel shift in South African English* (PhD dissertation). University of Cape Town, Cape Town.

Clarke, S., Elms, F., & Youssef, A. (1995). The third dialect of English: Some Canadian evidence. *Language Variation and Change*, *7*(2), 209–228.

Clopper, C. G. (2011). Checking for reliability. In M. Di Paolo & M. Yaeger-Dror (Eds.), *Sociophonetics: A student's guide* (pp. 188–197). London: Routledge.

Clopper, C. G., Hay, J., & Plichta, B. (2011). Experimental speech perception and perceptual dialectology. In M. Paolo & M. Yaeger-Dror (Eds.), *Sociophonetics: A Student's Guide* (pp. 149–162). London: Routledge.

Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *The Journal of the Acoustical Society of America*, *118*(3 Pt 1), 1661–1676.

Corbett, C. (2017). Tracking Phonetic Accommodation: An analysis of short-term speech adjustments during interactions between native and nonnative speakers. *New Ways of Analyzing Variation*. Presented at the NWAV46, Madison.

Coupland, N. (2009). Language attitudes, standardization and language change. In P. Quist, F. Gregersen, M. Maegaard, & N. J. Jørgensen (Eds.), *Language attitudes, standardization and language change: perspectives on themes raised by Tore Kristiansen on the occasion of his 60th birthday* (pp. 27–49). Oslo: Novus forlag.

Cox, F., & Palethorpe, S. (2008). Reversal of short front vowel raising in Australian English. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 342–345. Brisbane: Adelaide: Casual Productions.

Cox, F., & Palethorpe, S. (2012). Standard Australian English: The Sociostylistic Broadness Continuum. In R. Hickey (Ed.), *Standards of English: Codified Varieties Around the World* (pp. 294–317). Cambridge: Cambridge University Press.

DADRAS Architects. (2011). *City of Ogdensburg, New York: Downtown Improvement/"Main Street" Revitalization Strategy*. Retrieved from http://www.ogdensburg.org/DocumentCenter/Home/View/895

Deterding, D., & Hvitfeldt, R. (1994). The Features of Singapore English Pronunciation: Implications for Teachers. *Teaching and Learning*, *15*(1), 98–107.

Di Paolo, M. (1988). Pronunciation and categorization in sound change. In K. Ferrara, B. Brown, K. Walters, & J. Baugh (Eds.), *Linguistic change and contact: Proceedings of the Sixteenth Annual Conference on New Ways of Analyzing Variation in Language* (30th ed., pp. 84–92). Austin: Texas Linguistic Forum.

Di Paolo, M. (1992). Hypercorrection in response to the apparent merger of (ɔ) and (ɑ) in Utah english. *Language & Communication*, *12*(3), 267–292.

Di Paolo, M., & Faber, A. (1990). Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change*, *2*(2), 155–204.

Di Paolo, M., & Yaeger-Dror, M. (2011a). Field methods: gathering data, creating a corpus, and reporting your work. In M. Yaeger-Dror & M. Di Paolo (Eds.), *Sociophonetics: A student's guide* (pp. 7–23). London: Routledge.

Di Paolo, M., & Yaeger-Dror, M. (2011b). *The Vowel Categorization Experiment: a test of phonemic status*. Routledge.

Di Paolo, M., Yaeger-Dror, M., & Beckford Wassink, A. (2011). Analyzing Vowels. In M. Di Paolo & M. Yaeger-Dror (Eds.), *Sociophonetics: A Student's Guide* (pp. 87–106). London: Routledge.

Dinkin, A. J. (2009). *Dialect Boundaries and Phonological Change in Upstate New York* (PhD dissertation). University of Pennsylvania, Philadelphia.

Dinkin, A. J. (2010). Weakening Resistance: Progress Toward the Low Back Merger in New York State. *University of Pennsylvania Working Papers in Linguistics*, *15*(2 Selected Papers from NWAV37).

Dinkin, A. J. (2011). Nasal Short-a Systems vs. the Northern Cities Shift. *University of Pennsylvania Working Papers in Linguistics*, *17*(2 Selected Papers from NWAV39).

Dinkin, A. J. (2012). Toward a unified theory of chain shifting. In T. Nevalainen & E. Closs Traugott (Eds.), *The Oxford Handbook of the History of English*. New York: Oxford University Press.

Dinkin, A. J. (2013). Settlement patterns and the eastern boundary of the Northern Cities Shift. *Journal of Linguistic Geography*, *1*, 4–30.

Dinkin, A. J. (2016). Phonological Transfer as a Forerunner of Merger in Upstate New York. *Journal of English Linguistics*, *44*(2), 162–188.

Dinkin, A. J. (2019). Low Back Merger Encroaching at a Stable Dialect Boundary in Northern New York. *American Dialect Society Annual Meeting*. Presented at the New York. New York.

Dinkin, A. J., Forrest, J., & Dodsworth, R. (2017). Word Frequency in a Contact-Induced Change. *New Ways of Analyzing Variation*. Presented at the NWAV46, Madison.

Dixon, J. A., Mahoney, B., & Cocks, R. (2002). Accents of Guilt?: Effects of Regional Accent, Race, and Crime Type on Attributions of Guilt. *Journal of Language and Social Psychology*, *21*(2), 162–168.

Dobson, E. J. (1957). *English Pronunciation, 1500-1700*. Clarendon Press.

D'Onofrio, A., & Benheim, J. (2018). Contextualizing reversal: Sociohistorical dynamics and the Northern Cities Shift in a Chicago neighborhood. *New Ways of Analyzing Variation*. Presented at the NWAV47, New York.

D'Onofrio, A., Eckert, P., Podesva, R. J., Pratt, T., & Van Hofwegen, J. (2016). The low vowels in California's Central Valley. In V. Fridland, T. Kendall, B. Evans, & A. Beckford Wassink (Eds.), *Speech in the Western States* (pp. 11–32). Durham: Duke University Press.

Drager, K., & Hay, J. (2011). *Mergers in Production and Perception*. Presented at the ISLE 2, Boston.

Drager, K., Hay, J., & Walker, A. (2010). Pronounced Rivalries: Attitudes And Speech Production. *Te Reo*, *53*, 27–53.

Driscoll, A., & Lape, E. (2015). Reversal of the Northern Cities Shift in Syracuse, New York. *University of Pennsylvania Working Papers in Linguistics*, *21*(2 Selected Papers from NWAV43), 41–47.

Durant, S. W., & Peirce, H. B. (1878). *History of St. Lawrence Co., New York*. Philadelphia: L. H. Everts & Co.

Durian, D. (2012). *A New Perspective on Vowel Variation across the 19th and 20th Centuries in Columbus, OH*. The Ohio State University.

Durian, D. (2014). Another Look at the Short-a System of Late 19th and Early 20th Century Chicago in Pederson's PEMC Data, DARE, and LANCS. *New Ways of Analyzing Variation*. Presented at the NWAV43, Chicago.

Durian, D., & Cameron, R. (2018). Another Look at the Development of the Northern Cities Shift in Chicago. *New Ways of Analyzing Variation*. Presented at the NWAV47, New York.

Eckert, P. (1988). Adolescent Social Structure and the Spread of Linguistic Change. *Language in Society*, *17*(2), 183–207.

Eckert, P. (1989). The whole woman: Sex and gender differences in variation. *Language Variation and Change*, *1*(3), 245–267.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453–476.

Eckert, P. (2011). Language and Power in the Preadolescent Heterosexual Market. *American Speech*, *86*(1), 85–97.

Eckert, P., & Labov, W. (2017). Phonetics, phonology and social meaning. *Journal of Sociolinguistics*, *21*(4), 467–496.

EIS. (1973). *Ogdensburg Area Study, Recommended Transportation Plan, Ogdensburg-Oswegatchie: Environmental Impact Statement*. Northwestern University: United States. Federal Highway Administration.

EIS. (1979). *Ogdensburg Harbor: Environmental Impact Statement*. Department of the Army, Buffalo District, Corps of Engineers.

EIS. (1982). *Ogdensburg Harbor Dredging to 27 Feet: Environmental Impact Statement*. Buffalo: US Army Corps of Engineers.

ELAN (Version Version 4.9.4). (2016). Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/

Elazar, D. Judah. (1972). *American Federalism: A View from the States* (2nd ed.). New York: Crowell.

Emerson, O. F. (1891). The Ithaca dialect: a study of present English. *Dialect Notes*, *1*, 85–173.

Evanini, K., Isard, S., & Liberman, M. (2009). Automatic Formant Extraction for Sociolinguistic Analysis of Large Corpora. *Interspeech*, 1655–1658. Brighton.

Fabricius, A. H., Watt, D., & Johnson, D. E. (2009). A comparison of three speaker-intrinsic vowel formant frequency normalization algorithms for sociophonetics. *Language Variation and Change*, *21*(03), 413–435.

Fasold, R. W. (1969). *A sociolinguistic study of the pronunciation of three vowels in Detroit speech*. Washington: Center for Applied Linguistics], Sociolinguistics Program.

Fasold, R. W. (1984). *The Sociolinguistics of Society*. Oxford: Blackwell.

Fought, C. (1999). A majority sound change in a minority community: /u/-fronting in Chicano English. *Journal of Sociolinguistics*, *3*(1).

Fox, M. J. (2014). Free from the Northern Cities Vowel Shift: /ae/-raising in Northwestern Wisconsin. *American Dialect Society Annual Meeting*. Presented at the Minneapolis. Minneapolis.

Fox, M. J. (2016). The Structural Antagonism and Apparent-time Change of the Northern Cities Shift and the Low Back Vowel Merger in Northwestern Wisconsin. *New Ways of Analyzing Variation*. Presented at the NWAV45, Vancouver.

Frazer, T. C. (1979). The speech island of the American bottoms: A problem in social history. *American Speech*, *54*, 185–93.

Friedman, L. (2014). *The St. Louis Corridor: Mixing, Competing, and Retreating Dialects* (PhD dissertation). University of Pennsylvania, Philadelphia.

Garand, P. S. (1927). *The History of the City of Ogdensburg*. Ogdensburg: Rev. Manuel J. Belleville.

Genesee, F., & Holobow, N. E. (1989). Change and Stability in Intergroup Perceptions. *Journal of Language and Social Psychology*, *8*(1), 17–38.

Giles, H., & Coupland, N. (1991). *Language: Contexts and Consequences*. Open University Press.

Giles, H., Smith, P., Browne, C., Whiteman, S., & Williams, J. (1980). Women's speech: The voice of feminism. In S. McConnell-Ginet, R. Borker, & N. Furman (Eds.), *Women and Language in Literature and Society* (pp. 150–156). New York: Praeger.

Gordon, M. (2001). *Small-Town Values, Big-City Vowels: A Study of the Northern Cities Shift in Michigan (Publications of the American Dialect Society 84)*. Durham: Duke University Press.

Gordon, M. (2006). Tracking the low back merger in Missouri. In T. E. Murray & B. L. Simon (Eds.), *Language Variation and Change in the American Midland: A New Look at 'Heartland' English* (pp. 57–68). Amsterdam/Philadelphia: John Benjamins.

Gordon, M. (2013). Investigating Chain Shifts and Mergers. In J. K. Chambers & N. Schilling (Eds.), *The Handbook of Language Variation and Change* (2nd ed., pp. 203–219). Wiley-Blackwell.

Gordon, M., & Strelluf, C. (2016). Working the Early Shift: Older Inland Northern Speech and the Beginnings of the Northern Cities Shift. *Journal of Linguistic Geography*, *4*(1), 31–46.

Hall-Lew, L. (2013). 'Flip-flop' and mergers-in-progress. *English Language and Linguistics*, *17*(02), 359–390.

Hall-Lew, L. (2017). When does a (sound) change stop progressing? *New Ways of Analyzing Variation*. Presented at the NWAV46, Madison.

Harris, J. (1985). *Phonological Variation and Change: Studies in Hiberno-English*. Cambridge: Cambridge University Press.

Havenhill, J. (2018). Audiovisual cue enhancement in the production and perception of the COT-CAUGHT contrast. *New Ways of Analyzing Variation*. Presented at the NWAV47, New York.

Hay, J. (2011). Statistical Analysis. In M. Di Paolo & M. Yaeger-Dror (Eds.), *Sociophonetics: A Student's Guide* (pp. 198–214). London: Routledge.

Hay, J., Drager, K., & Thomas, B. (2013). Using nonsense words to investigate vowel merger. *English Language & Linguistics*, *17*(2), 241–269.

Hay, J., Drager, K., & Warren, P. (2009). Careful Who You Talk to: An Effect of Experimenter Identity on the Production of the NEAR/SQUARE Merger in New Zealand English. *Australian Journal of Linguistics*, *29*(2), 269–285.

Hay, J., Nolan, A., & Drager, K. (2006). From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review*, *23*(3), 351–379.

Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*(4), 458–484.

Herndobler, R. (1993). Sound Change and Gender in a Working-Class Community. In T. C. Frazer, *"Heartland" English: Variation and transition in the American Midwest* (pp. 137–156). The University of Alabama Press.

Herold, R. (1990). *Mechanisms of merger: The implementation and distribution of the low back merger in eastern Pennsylvania* (PhD dissertation). University of Pennsylvania, Philadelphia.

Hickey, R. (2016). English in Ireland: Development and varieties. In R. Hickey (Ed.), *Sociolinguistics in Ireland* (pp. 3–40). Basingstoke: Palgrave Macmillan.

Hickey, R. (2018). 'Yes, that's the best': Short front vowel lowering in English today: Young people across the anglophone world are changing their pronunciation of vowels according to a change which started in North America. *English Today*, *34*(2), 9–16.

Hollett, P. (2006). Investigating St. John's English: Real- and Apparent-time Perspectives. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique*, *51*(2–3), 143–160.

Holmes-Elliott, S., & Smith, J. (2015). DRESS-down: /ɛ/-lowering in apparent time in a rural Scottish community. *Proceedings from the XVIII International Congress of Phonetic Sciences*. Presented at the International Congress ofPhonetic Sciences, Glasgow.

Hough, F. B. (1853). *A History of St. Lawrence and Franklin Counties, New York: From the Earliest Period to the Present Time*. Albany: Little & Company.

Irons, T. L. (2007). On the status of low back vowels in Kentucky English: More evidence of merger. *Language Variation and Change*, *19*(2), 137–180.

Ito, R. (2001). Belief, Attitudes, and Linguistic Accommodation: A Case of Urban Sound Change in Rural Michigan. *University of Pennsylvania Working Papers in Linguistics*, *7*(3 Selected Papers from NWAV29), 129–143.

Ito, R., & Preston, D. R. (1998). Identity, Discourse, and Language Variation. *Journal of Language and Social Psychology*, *17*(4), 465–483.

Jasewicz, E., Fox, R. A., & Salmons, J. (2011). Cross-generational vowel change in American English. *Language Variation and Change*, *23*(1), 45–86.

Johnson, D. E. (2007). *Stability and change along a dialect boundary: The low vowels of southeastern New England* (PhD dissertation). University of Pennsylvania, Philadelphia.

Johnson, D. E. (2014). Descriptive statistics. In R. J. Podesva & D. Sharma (Eds.), *Research Methods in Linguistics* (pp. 288–315). New York: Cambridge University Press.

Katz, J. (2016). *Speaking American: How Y'all, Youse, and You Guys Talk: A Visual Guide*. Boston; New York: Houghton Mifflin Harcourt.

Kennedy, R., & Grama, J. (2012). Chain Shifting and Centralization in California Vowels: An Acoustic Analysis. *American Speech*, *87*(1), 39–56.

Kenrick, W. (1846). *Notes on Ogdensburg, Its Position, Its Rivers and Lakes, and Proposed Rail-road*. Boston: Dutton and Wentworth Printers.

King, S. (2017). African American Identity and Vowel Systems in Rochester, New York. *New Ways of Analyzing Variation*. Presented at the NWAV46, Madison.

Kircher, R. (2015). The Matched-Guise Technique. In Z. Hua (Series Ed.), *Research Methods in Intercultural Communication: A Practical Guide* (pp. 196–211). Hoboken: John Wiley & Sons.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208–1221.

Kohn, M., & Stithem, C. (2015). The Third Vowel Shift in Kansas: A supra-regional shift with regional variation. *New Ways of Analyzing Variation*. Presented at the NWAV44, Toronto.

Koops, C., Gentry, E., & Pantos, A. (2008). The effect of perceived speaker age on the perception of PIN and PEN vowels in Houston, Texas. *University of Pennsylvania Working Papers in Linguistics*, *14*(2 Selected Papers from NWAV 36).

Kurath, H. (1939). *Handbook of the linguistic geography of New England*. Providence: Brown University.

Kurath, H., Bloch, B., & Hansen, M. L. (1939). *Linguistic atlas of New England*. Providence, R.I.: Brown University.

Labov, W. (1963). The Social Motivation of a Sound Change. *Word*, 273–309.

Labov, W. (1966). *The Social Stratification of English in New York City*. Cambridge University Press.

Labov, W. (1972). The Social Motivation of a sound change. In *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. (1984). *Field Methods of the Project on Linguistic Change and Variation* (J. Baugh & J. Sherzer, Eds.). Englewood Cliffs: Prentice-Hall.

Labov, W. (1991). The three dialects of English. In P. Eckert (Ed.), *New Ways of Analyzing Sound Change*. New York: Academic Press.

Labov, W. (1994). *Principles of Linguistic Change: Internal Factors* (Vol. 1). Oxford: Blackwell.

Labov, W. (2001). *Principles of Linguistic Change: Social Factors*. Oxford: Blackwell.

Labov, W. (2006). A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics*, *34*(4), 500–515.

Labov, W. (2007). Transmission and Diffusion. *Language*, *83*(2), 344–387.

Labov, W. (2010). *Principles of Linguistic Change: Cognitive and Cultural Factors* (Vol. 3). Oxford: Blackwell.

Labov, W. (2017). Plotnik (Version 10.3). Retrieved from http://www.ling.upenn.edu/~wlabov/Plotnik

Labov, W., Ash, S., Baranowski, M. A., Nagy, N., Ravindranath, M., & Weldon, T. (2006). Listeners' sensitivity to the frequency of sociolinguistic variables. *University of Pennsylvania Working Papers in Linguistics*, *12*(2 Selected Papers from NWAV34).

Labov, W., Ash, S., & Boberg, C. (2006). *Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: Mouton de Gruyter.

Labov, W., Mark, K., & Miller, C. (1991). Near-mergers and the suspension of phonemic contrast. *Language Variation & Change*, (3), 33–74.

Labov, W., Rosenfelder, I., & Fruehwald, J. (2013). One Hundred Years of Sound Change in Philadelphia: Linear Incrementation, Reversal, and Reanalysis. *Language*, *89*(1), 30–65.

Labov, W., Yaeger, M., & Steiner, R. (1972). *A quantitative study of sound change in progress*. U. S. Regional Survey.

Landon, H. F. (1932). *History of the North Country*. Reprint Services Corp.

Langstrof, C. (2006). *Vowel Change in New Zealand English - Patterns and Implications*.

Lass, R. (1987). *The shape of English: structure and history*. London, Melbourne: J. M. Dent.

Lass, R. (1992). What, If Anything, Was the Great Vowel Shift? In M. Rissanen, O. Ihalainen, T. Nevalainen, & I. Taavitsainen (Eds.), *History of Englishes: New Methods and Interpretations in Historical Linguistics* (Reprint 2011, pp. 144–155). Berlin, Boston: De Gruyter Mouton.

Lee, S. (2018). Patterns of the Mainstream Sound Change in a Liminal Region: Low Back Merger in Washington DC. *Journal of English Linguistics*, *46*(4), 267–292.

Llamas, C. (2007). “A place between places”: Language and identities in a border town. *Language in Society*, *36*(4), 579–604.

Lobanov, B. (1971). Classification of Russian vowels spoken by different listeners. *Journal of the Acoustical Society of America*, *49*, 606–608.

MacKenzie, L., & Turton, D. (2013). Crossing the pond: Extending automatic alignment techniques to British English dialect data. *New Ways of Analyzing Variation*. Presented at the NWAV42, University of Pittsburgh/Carnegie Mellon University.

Maclagan, M., & Hay, J. (2007). Getting fed up with our feet: Contrast maintenance and the New Zealand English “short” front vowel shift. *Language Variation and Change*, *19*(1), 1–25.

Maguire, W., Clark, L., & Watson, K. (2013). Introduction: what are mergers and can they be reversed? *English Language & Linguistics*, *17*(2), 229–239.

Majors, T. (2005). Low back vowel merger in Missouri speech: acoustic description and explanation. *American Speech*, *80*(2), 165–179.

Martinet, A. (1955). *Économie des changements phonétiques: traité de phonologie diachronique*. Bern: A. Franke.

Mason, A. (2018). It’s a TRAP!: The trigger for the Elsewhere Shift in Lansing, Michigan. *New Ways of Analyzing Variation*. Presented at the NWAV47, New York.

McCarthy, C. (2007). Social Correlates of Vowel Shifting in Chicago: A Pilot. *McGill Working Papers in Linguistics*, *21*, 21–34.

McCarthy, C. (2010). The Northern Cities Shift in Real Time: Evidence from Chicago. *University of Pennsylvania Working Papers in Linguistics*, *15*(2 Selected Papers from NWAV37).

Merriam, N. (1907). *The first settlement of Ogdensburg* (Swe-Kat-Si Chapter). New York: Silver, Burdett, & Co.

Milholland, A. (2018). Reversal of the Northern Cities Shift in Buffalo, NY. *New Ways of Analyzing Variation*. Presented at the NWAV47, New York.

Milroy, J. (1992). *Linguistic Variation and Change: On the Historical Sociolinguistic of English*. Oxford, UK; Cambridge, USA: Blackwell Pub.

Milroy, J. (1995). Review of the book Principles of linguistic change. Volume I: Internal factors, by William Labov. *Journal of Linguistics*, *31*(02), 435–439.

Milroy, J., & Harris, J. (1980). When is a merger not a merger?: The MEAT/MATE problem in a present-day English vernacular. *English World-Wide*, *1*(2), 199–210.

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics*, *21*(02), 339.

Milroy, J., Milroy, L., Harris, J., Gunn, B., Pitts, A., & Policansky, L. (1983). *Sociolinguistic Variation and Linguistic Change in Belfast.* Social Science Research Council.

Milroy, L., & Gordon, M. (2003). *Sociolinguistics: Method and Interpretation*. Wiley.

Monroe, B. S. (1896). Pronunciation of English in New York State. *Dialect Notes*, *1*, 445–456.

Morgan, B.-K., DeGuise, K., Acton, E., Benson, D., & Shvetsova, A. (2017). Shifts toward the supra-regional in the Northern Cities region: Evidence from Jewish women in Metro Detroit. *New Ways of Analyzing Variation*. Presented at the NWAV46, Madison.

Murray, T. E. (2002). Language variation and change in the urban midwest: The case of St. Louis, Missouri. *Language Variation and Change*, *14*(3), 347–361.

Neary, T. M. (1977). *Phonetic Feature Systems for Vowels* (PhD dissertation). University of Alberta, Reprinted 1978 by the Indiana University Linguistics Club.

Nesbitt, M. (2016). Rising above the level of awareness: The Northern Cities Shift in Lower Michigan. *Graduate Linguistics Expo at Michigan State*. Presented at the East Lansing. East Lansing.

Nesbitt, M. (2018). Economic Change and the Decline of Raised TRAP in Lansing, MI. *University of Pennsylvania Working Papers in Linguistics*, *24*(2 Selected Papers from NWAV46), 66–76.

Nesbitt, M., & Mason, A. (2016). Evidence of the Elsewhere Shift in the Inland North. *New Ways of Analyzing Variation*. Presented at the NWAV45, Vancouver.

Niedzielski, N. (1999). The Effect of Social Information on the Perception of Sociolinguistic Variables. *Journal of Language and Social Psychology*, *18*(1), 62–85.

Nycz, J., & Hall-Lew, L. (2014). Best practices in measuring vowel merger. *Proceedings of Meetings on Acoustics: The Acoustical Society of America through the American Institute of Physics*, *20*, 1–19.

Ogdensburg, NY Population. (2018). Retrieved April 4, 2019, from World Population Review website: http://worldpopulationreview.com/us-cities/ogdensburg-ny-population/

Ping, T. L. (2005). Does the survey response scale format matter? *IMTA - International Military Testing Association*. Presented at the IMTA - International Military Testing Association, Singapore.

Plichta, B., Preston, D. R., & Rakerd, B. (2007). “It’s Too Hat in Here?” The Perception of NCS a-Fronting. *Linguistica Atlantica*, *27*(0), 92–95.

Podesva, R. J., D’Onofrio, A., Van Hofwegen, J., & Kim, S. K. (2015). Country ideology and the California Vowel Shift. *Language Variation and Change*, *27*(2), 157–186.

Podesva, R. J., & Zsiga, E. (2014). Sound recordings: acoustic and articulatory data. In R. J. Podesva & D. Sharma (Eds.), *Research Methods in Linguistics* (pp. 169–194). New York: Cambridge University Press.

Power, R. L. (1953). *Planting Corn Belt culture: The impress of the upland southerner and Yankee in the Old Northwest*. Indianapolis: Indiana Historical Society.

Preston, D. R. (1996). Where the worst English is spoken. In E. W. Schneider (Ed.), *Focus on the USA* (pp. 297 – 360). Amsterdam: John Benjamins.

Preston, D. R. (2009). Are you really smart (or stupid, or cute, or ugly, or cool)? Or do you just talk that way? In P. Quist, F. Gregersen, M. Maegaard, & Jørgensen, Normann (Eds.), *Language attitudes, standardization and language change*. Oslo: Novus.

Preston, D. R. (2013). Language with an Attitude. In J. K. Chambers & N. Schilling (Eds.), *The handbook of language variation and change* (2nd ed., pp. 157–182). Chichester: Wiley-Blackwell.

R Core Team. (2016). R: A Language and Environment for Statistical Computing (Version 1.0.136). Retrieved from https://www.R-project.org

Rasinger, S. M. (2013). *Quantitative Research in Linguistics: An Introduction*. A&C Black.

Robinson, L. (2017). Decades, millions spent on Ogdensburg’s contaminated shore. *Watertown Daily Times*. Retrieved from http://www.watertowndailytimes.com/article/20170813/NEWS05/170819356

Robinson, L. (2018). 100,000 passengers and counting at Ogdensburg International Airport. *The Journal*. Retrieved from http://www.ogd.com/article/20181016/OGD/181019272

Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). *FAVE (Forced Alignment and Vowel Extraction) Program Suite*. Retrieved from http://fave.ling.upenn.edu

Ryan, E. B., Giles, H., & Hewstone, M. (1987). The measurement of language attitudes. In U. Ammon, N. Dittmar, & K. Mattheier (Eds.), *Sociolinguistics, an International Handbook of the Science of Language and Society* (Vol. 1, pp. 1068–1081). Berlin: Walter de Gruyter.

Sandburg, B. (2010). For a Newsman and His Town, the End of an Era. *The New York Times*. Retrieved from https://www.nytimes.com/2010/10/03/nyregion/03editor.html

Sandburg, B. (2015). Steele's photos captured Ogdensburg's glory. *Watertown Daily Times*. Retrieved from http://www.ogd.com/adv/steeles-photos-captured-ogdensburgs-glory-20150411

Sandburg, B. (2016). History of a hospital: an Ogdensburg landmark. *Watertown Daily Times*. Retrieved from http://www.watertowndailytimes.com/news05/history-of-a-hospital-an-ogdensburg-landmark-20160918

Sankoff, G., & Blondeau, H. (2007). Language Change across the Lifespan: /r/ in Montreal French. *Language*, *83*(3), 560–588. Retrieved from JSTOR.

Savage, M., & Mason, A. (2018). Style and attitude: The social evaluation of the BET vowel. *New Ways of Analyzing Variation*. Presented at the NWAV47, New York.

Savage, M., Mason, A., Nesbitt, M., Pevan, E., & Wagner, S. E. (2016). Ignorant and annoying: Inland Northerners' attitudes towards Northern Cities Shift short o. *American Dialect Society Annual Meeting*.

Severance, N., Evanini, K., & Dinkin, A. J. (2015). Examining the performance of FAVE for automated sociophonetic vowel analyses. *New Ways of Analyzing Variation*. Presented at the NWAV44, Toronto.

Shuy, R. W. (1962). *The northern-midland dialect boundary in Illinois*. University of Alabama Press.

Smith, M. K., & Bailey, G. (1980). Attitude and activity: Contextual constraints on subjective judgments. In H. Giles, W. P. Robinson, & P. M. Smith (Eds.), *Language: social psychological perspectives*. Pergamon Press.

Stanford, J. N., & Kenny, L. A. (2013). Revisiting transmission and diffusion: An agent-based model of vowel chain shifts across large communities. *Language Variation and Change*, *25*(02), 119–153.

Stockwell, R., & Minkova, D. (1988). The English Vowel Shift: Problems of Coherence and Explanation. In D. Kastovsky & G. Bauer (Eds.), *Luick Revisited* (pp. 355–394). Tübingen: Narr.

Stockwell, R., & Minkova, D. (1997). On Drifts and Shifts. *Studia Anglica Posnaniensia: International Review of English Studies*, (31), 283–303.

Styler, W. (2016). *Using Praat for Linguistic Research*. Retrieved from http://savethevowels.org/praat

Tagliamonte, S. A. (2006). *Analysing Sociolinguistic Variation*. Cambridge University Press.

Taylor, M. C. (SSJ, PhD). (2017). Industrious Immigrants in the Early 1900's. In *Under Four Flags* (pp. 306–315). Society of the Sisters of St. Joseph.

Thiel, A., & Dinkin, A. J. (under review). Escaping the TRAP: Losing the Northern Cities Shift in Real Time. *Language Variation & Change*.

Thomas, C. K. (1935-1937). Pronunciation in Upstate New York I - VII. *American Speech*, *10–12*.

Thomas, E. R. (2001). *An Acoustic Analysis of Vowel Variation in New World English* (Pck edition). Durham, NC: Duke University Press Books.

Thomas, E. R. (2002). Sociophonetic Applications of Speech Perception Experiments. *American Speech*, *77*(2), 115–147.

Thomas, E. R. (2004). Rural Southern white accents. In *A Handbook of Varieties of English: A Multimedia Reference Tool.* Berlin, Boston: De Gruyter Mouton.

Trudgill, P. (1974). The Social Differentiation of English in Norwich. In N. Coupland & A. Jaworski (Eds.), *Sociolinguistics: A Reader* (pp. 179–184). London: Macmillan Education UK.

Trudgill, P. (1986). *Dialects in Contact*. Oxford, UK; New York, NY, USA: Wiley-Blackwell.

Trudgill, P. (1988). Norwich Revisited: Recent Linguistic Changes in an English Urban Dialect. *English World-Wide*, *9*(1), 33–49.

Trudgill, P. (1996). Dialect typology: Isolation, social network and phonological structure. In G. R. Guy, C. Feagin, D. Schiffrin, & J. Baugh (Eds.), *Towards a Social Science of Language: Papers in honor of William Labov. Volume 1: Variation and change in language and society*. John Benjamins Publishing.

Trudgill, P., & Foxcroft, T. (1978). On the sociolinguistics of vocalic mergers: Transfer and approximation in East Anglia. In P. Trudgill (Ed.), *Sociolinguistic patterns in British English* (pp. 69–79). Baltimore: University Park Press.

Trudgill, P., Schreier, D., Long, D., & Williams, J. P. (2009). On the Reversibility of Mergers: /w/, /v/ and Evidence from Lesser-Known Englishes. *Folia Linguistica Historica*, *37*(Historica vol. 24,1-2), 23–46.

US Census Bureau. (n.d.). American FactFinder. Retrieved October 1, 2018, from https://factfinder.census.gov

Vita Nuova. (2012). *City of Ogdensburg, NY: Economic and Market Trends Analysis*.

Wagner, S. E., Mason, A., Nesbitt, M., Pevan, E., & Savage, M. (2016). Reversal and re-organization of the Northern Cities Shift in Michigan. *University of Pennsylvania Working Papers in Linguistics*, *22*(2 Selected Papers from NWAV44), 171–179.

Walker, J. A. (2014). Variation analysis. In R. J. Podesva & D. Sharma (Eds.), *Research Methods in Linguistics* (pp. 440–459). New York: Cambridge University Press.

Wells, J. C. (1982). *Accents of English:* Cambridge University Press.

Wolfram, W. (1969). *A sociolinguistic description of Detroit Negro speech*. Center for Applied Linguistics.

Woods, H. B. (1993). A synchronic study of English spoken in Ottawa. In S. Clarke (Ed.), *Focus on Canada* (pp. 151–178). Amsterdam: Benjamins Pub.

Wright, J. (1905). *The English dialect grammar*. Retrieved from http://archive.org/details/englishdialectg00wrig

Yang, C. (2009). *Population Structure and Language Change*. University of Pennsylvania.