Disentangling frequency effects and grammaticalization

Nicole Hildebrand-Edgar
University of Victoria
nchil@uvic.ca

This paper presents a case study of I DON'T KNOW to investigate the nature of the phonetic reduction using data from the Victorian English Archive (D'Arcy 2011-2014). This phrase has a high usage frequency and is commonly reduced in speech, two concomitant processes in grammaticalization. Further, I DON'T KNOW has use beyond its referential function of "lack of knowledge": it serves various pragmatic functions as a discourse marker. The relationship between phonetic form and semantic function is investigated using quantitative variationist analysis. Similar patterning to that previously reported for other varieties of English is found, suggesting a universal pathway of grammaticalization. Moreover, the frequency of phonetically reduced pragmatic tokens increases in apparent time, suggesting ongoing change of the discourse marker. This change is argued to constitute ongoing phonetic reduction of an already grammaticalized form. The role of frequency effects in driving ongoing change is restricted to one form while the semantic functions remain stable.
Keywords: Grammaticalization; frequency effects; variationist sociolinguistics; language variation and change

1 Introduction

In the course of everyday speech, common phrases are often subject to phonetic erosion, or the reduction of speech sounds. This includes such commonly heard forms as *whatcha doing* (in lieu of *what are you doing*), *hafta* (for *have to*), and many others. A particularly salient example is *I dunno,* a phonetically reduced form of I DON'T KNOW[1] (first noted in Kaisse, 1985). I DON'T KNOW may be reduced to such a degree that it surfaces as a prosodic grunt, only identifiable by its rise-fall intonation. The fact that *I dunno* appears orthographically in text and Internet correspondence is further testament to its prevalence in the minds of speakers. The alternation between the full and reduced variants of I DON'T KNOW does not appear to be conditioned by any linguistic factor; rather, Scheibman (2000) suggests the variation between forms is subject to speaker choice. On the assumption that seemingly free variation between variants is rarely, if ever, random (Labov, 1969), a Variationist Sociolinguistic approach is

[1] The surface form is represented by italics, and the underlying form by capital letters.

adopted to determine the relationship between the full and reduced variants of I DON'T KNOW and the internal and external factors that influence this alternation. In particular, this analysis aims to determine whether grammaticalization is implicated in the variation.

Quantitative studies have shown that phonological reduction is linked to token frequency within the context of grammaticalization (e.g. Thomson & Mulac, 1991), and that the frequency of a collocation drives the propensity for contraction or reduction (Lorenz, 2013). Bybee (2006) outlines the process of reduction in the context of grammaticalization: widening contexts of use and increased frequency lead to entrenchment of collocations into single processing units (or "chunks"); subsequently, they are accessed and produced with less effort, and are thus subject to phonetic reduction. In a corresponding process, the loss of semantic content and prosodic weight may cause loss of stress, promoting reduction. This proposed two-part process is what Bybee, Perkins, & Pagliuca (1994) term the *Parallel Reduction Hypothesis*. The reduction that ensues affects both articulatory gestures and temporal durations. Consequently, an erstwhile multi-morphemic phrase (such as *go + ing to*) comes to be fused and compressed (*gonna)* (Bybee et al. p. 6). Lorenz argues that this "parallel" process is more accurately described as cyclical: as forms progress through stages of grammaticalization, desemanticization (loss of semantic content) leads to phonetic reduction, but phonetic reduction does not lead to further desemanticization. He proposes that the Parallel Reduction Hypothesis be rephrased as a cycle of reanalysis leading to reduction.

Thus, to show that a form is commonly reduced does not, in itself, imply that the form has grammaticalized. As argued in Hopper (1991), while reduction processes such as condensation (shortening of forms) and coalescence (collapsing of adjacent forms) typically accompany grammaticalization, they are neither necessary nor sufficient diagnostics. Rather, they are typical of forms that are advanced in terms of grammaticalization. Forms that are not grammaticalized may also be reduced. The challenge for identifying instances of grammaticalization in synchronic studies is disentangling general frequency effects and the frequency-driven changes associated with grammaticalization: semantic fading (or bleaching), phonological reduction, positional fixing (or syntactic rigidity), and erasure of word boundaries (Bybee et al., 1994; Hopper & Traugott, 2003).

Several of these changes are demonstrated in conversational use of I DON'T KNOW. In addition to its referential use in expressing insufficient knowledge, studies have shown that I DON'T KNOW is deployed for pragmatic functions, including turn-management, hedging, politeness, and face-saving (Baumgarten & House, 2010; Beach & Metzger, 1997; Bybee & Scheibman, 1999; Pichler, 2009; Diani, 2004; Weatherall, 2011). Following is an example of I DON'T KNOW as a politeness device provided in Grant (2010, p.2286) from the CANCODE corpus:

[At a travel agent's]
S1: Did you want to take out insurance?
S2: Erm I'd like to ask about it but I don't know if I want to do that today.
S1: Okay. (CANCODE)

Here, I DON'T KNOW serves to soften S2's refusal of insurance, thereby protecting the face of S1. Both Scheibman (2000) and Pichler (2009) report strong correlations between the full form and the referential function—expressing lack of speaker knowledge—and between the reduced form and pragmatic functions. This form-function split is attributed to grammaticalization of the construction.

The question that arises is whether the findings reported in Scheibman (2000) and Pichler (2009) for I DON'T KNOW are due to a universal path of grammaticalization based on its semantic function, or whether this variation is conditioned by community-specific factors. To address this question, this project investigates the functional and social conditioning of I DON'T KNOW in a corpus of speech data from Victoria, BC. If the findings reported in Scheibman (2000) and Pichler (2009) are indeed indicative of a universal path of grammaticalization, similar results are predicted in this different variety of English. This exploratory analysis will contribute to the literature concerning the grammaticalization of constructions, and will have implications for analyzing the complex relationship between frequency, phonetic reduction and grammaticalization.

2 Literature Review

2.1 Frequency and reduction in grammaticalization

Though Meillet's (1912) original conceptualization of grammaticalization applied to single word-forms, recent work has shown that constructions may also be grammaticalized (e.g. Bybee, 2006; Torres Cacoullos & Walker 2009a, Lorenz 2013). Thompson & Mulac (1991) make a case for grammaticalization that extends beyond individual lexemes in an analysis of *that*-deletion and epistemic parentheticals in English. They find that the most frequent subject-verb combinations without *that* occur most frequently as epistemic parentheticals, the verbs encoding subjective meanings associated with belief and mode of knowing. Their findings indicate that grammaticalization is reliant on discourse frequencies and recurrent patterns.

The notion that grammar arises from one's experience with patterns of language is foundational to usage-based models of language. From this framework, Bybee (2006) investigates the role of frequency in grammaticalization of constructions. She provides evidence that frequency is an important factor in grammaticalization, as it promotes both the autonomization of new constructions (that is, cognitive independence from their source forms), as well as phonetic reduction of these constructions. When constructions are encountered with increasing levels of frequency, they may become conventionalized (as idioms or prefabrications). With higher frequency, new constructions with their own categories may be established. Extremely high frequency may then lead to grammaticalization of these new constructions and changes in constituency. Bybee

states that certain changes associated with grammaticalization are, in part, conditioned by frequency: autonomy, semantic bleaching (or semantic change), and reanalysis (loss of morphosyntactic boundaries) (pp. 720-721). As constructions are encountered more frequently, they are produced more fluently, and this phonetic reduction accumulates in the cognitive representation. This reduction process is recurrent, as already reduced variants of high-frequency phrases are more often selected for production, and subsequently undergo further reduction. Bybee argues that these frequent constructions are single processing units (or, what she calls chunks), making them susceptible to further reduction and grammaticalization. Therefore, this process of reduction only occurs on the grammaticalized form (p. 724).

An important caveat is that new constructions may arise without grammaticalization: certain general constructions may develop new pragmatic meaning without being completely disassociated from their source meanings (such as *How do you do?,* which is still associated with the source question *What are you doing?*) (Bybee, 2006, p. 723). Lorenz (2013) addresses this issue in regards to the reduced forms *gonna, gotta,* and *wanna.* He asks whether they are simply typical ways of pronouncing *going to, got to,* or *want to*, or whether they have independent meanings from their source forms and distinct cognitive representations. He argues that the contracted forms *gonna, gotta,* and *wanna* are emancipated (autonomous) from their source forms as a result of becoming entrenched in memory through frequent usage. In this process of emancipation, a full form becomes phonetically reduced, and as this reduced form is frequently used, it becomes a conventional expression encoding particular meanings (the process of divergence). As the reduced form becomes its own lexical item, speakers stop interpreting it as the full form, and the initial motivation for reduction is lost.

The phrase I DON'T KNOW is both commonly reduced and extremely frequent in discourse. Investigating both the BNC and COCA corpora, Baumgarten & House (2010) find that I DON'T KNOW is the most frequently occurring negative collocation. Previous analyses also show that I DON'T KNOW is a highly frequent collocate across varieties of English (Kaisse, 1985, Scheibman 2000). This high frequency of usage, in addition to the existence of the reduced for *I dunno,* suggests that I DON'T KNOW may very well be a grammaticalized construction. However, while frequency and reduction are processes that occur within grammaticalization, there must be evidence that an erstwhile lexical (content) form has changed in such a way as to assume characteristics of a grammatical (functional) form in order to validate this claim (Hopper & Traugott, 2003). The following section provides a review of the pragmatic functions that are associated with I DON'T KNOW.

2.2 Functional analyses of I don't know

In everyday conversation, I DON'T KNOW is deployed for a much broader range of functions than to simply claim lack of knowledge. Grant (2010), analyzing usage patterns of *I don't know* and *I dunno* in text corpus data of British and New Zealand speech, finds that both *I don't know* and *I dunno* can be used as epistemic devices and expressions of stance (p. 2290). *I dunno* is used especially as a politeness device to soften disagreement. Grant reports that the full form *I don't know* has a greater range of usage, while the reduced form *I dunno* is predominantly a hedging device. While Grant's findings are useful for acknowledging the various functions that *I don't know* and its reduced variant can serve, the form-function regularities she outlines are questionable: as she acknowledges, the data is only written, and there is no way of confirming the criteria by which the transcribers differentiated the two forms.

Weatherall (2011) examines the functional distribution of I DON'T KNOW in British, New Zealand, and American speech. Her analysis focuses specifically on instances of I DON'T KNOW that have scope over the following proposition (as opposed to those that are responses or follow an assessment). She finds that these prepositioned tokens fall into two broad categories: those used in first assessments (signalling exaggeration or non-seriousness), and those used in approximations. In both cases, I DON'T KNOW indexes lack of speaker commitment to what follows. Weatherall argues that these prepositioned epistemic hedges function to disclaim knowledge authority (especially in the first assessment cases), which indicates that source meaning (lack of knowledge) persists. Similarly, Diani (2004) finds that I DON'T KNOW can function to avoid explicit disagreement, avoid commitment, minimize face-threatening acts, and mark uncertainty—all of which retain the central meaning of lack of knowledge. However, neither Weatherall (2011) nor Diani (2004) makes a distinction between full- and reduced-forms of *I don't know*.

The variation between full and reduced forms of I DON'T KNOW is addressed in Scheibman (2000). The linguistic conditioning that drives the reduction of the negative auxiliary *don't* is explored by analyzing its use in everyday conversation. Conversational data from American speakers reveals that reduced DON'T occurs in limited but highly frequent collocations, predominantly I DON'T KNOW. By comparing the semantic and interactive contexts of the full and reduced forms of I DON'T KNOW, Scheibman finds a form-function regularity: both full and reduced forms may express the referential meaning of insufficient knowledge, but only reduced *I dunno* is used in pragmatic (textual or organizational) or subjective (face saving or politeness) functions. These functional correlations are therefore inconsistent with those reported in Grant's (2010) text corpus analysis, where the more full form could serve either referentially or pragmatically, and the reduced form only pragmatically. Scheibman contends that grammaticalization is not implicated for reduced *don't* itself; rather, in the spirit of Thompson & Mulac's (1991) proposal, the conventionalized expressions in which it most frequently occurs are

grammaticalized (e.g. *I don't know, I don't think*). These conventionalized expressions are processed as single units or "chunks", and tend to have pragmatic functions. The full forms continue to exist, though with different functions (layering), and the new forms come to take on subjective and textual meanings (semantic bleaching). Scheibman's results must be taken with caution, however, as they are based on very few tokens (N=36).

Pichler (2009) also addresses the phonetic variants of *don't.* The differential distribution of discourse markers I DON'T KNOW and I DON'T THINK is investigated in everyday speech in Berwick-upon-Tweed. The study employs qualitative methods of conversation analysis in determining functional and social meanings, and quantitative methods of Labovian sociolinguistics in analyzing linguistic conditioning. Pichler identifies three non-localized phonological variants of I DON'T KNOW: a full form *I don't know* (with a marked boundary between the *n* of DON'T and the *n* of KNOW and a full vowel *o* in DON'T), an intermediate form *I donno* (with no marked boundary and a full vowel), and a reduced form *I dunno* (with no marked boundary and a reduced vowel). The reduced form is found to be the most frequent variant across social groups (aside from older males), and has the greatest potential to occur in all pragmatic functions. The full form is found to correlate strongly with referential functions. In addition, a localized variant, *I divn't knaa,* is identified, and found to be socially conditioned. The functional conditioning of non-localized variants is claimed to be a result of grammaticalization, as the distribution of forms exhibits various indices of grammaticalization from Hopper (1991): the full form dominates in referential contexts while an intermediate form is used across functions (layering), the reduced form is very rarely intervened by adverbial modification (decategorialization), and the source meaning of "lack of knowledge" is maintained in the reduced epistemicity meaning of the grammaticalized forms (persistence). As in Scheibman (2000), it is argued that I DON'T KNOW is a formulaic, single processing unit—a fact which has enabled its grammaticalization. Pichler (2009) further suggests that the reduced variable may still be increasing in positional mobility and discourse functions.

2.3 Summary of literature review

Frequency and phonetic reduction are inherent in grammaticalization, although they are not, in themselves, sufficient for identifying forms that have grammaticalized. I DON'T KNOW has been found to be a highly frequent collocation across varieties of English, and its propensity for reduction has been noted in multiple studies. It has also been observed in functional analyses that I DON'T KNOW is used in everyday conversation to encode a variety of pragmatic meanings in addition to its referential meaning. Several authors, notably Scheibman (2000) and Pichler (2009), have found a form-meaning relationship for variants of I DON'T KNOW, and attribute this relationship to grammaticalization of the construction. If these findings imply a universal path

of grammaticalization of I DON'T KNOW, similar results should emerge from analyses of its distribution in other English speaking communities.

3 Methodology and Data

3.1 Theoretical assumptions

To further explore the form-function regularities of full and reduced variants, and to test whether the findings in Scheibman (2000) and Pichler (2009) are due to a universal path of grammaticalization of I DON'T KNOW, I likewise examine its use in natural spoken conversation. As stated in the introduction, a Variationist Sociolinguistic approach is adopted for this analysis. In this framework, the fact that individual speakers will exhibit variable behaviour is recognized; thus, inherent variability in everyday language is taken into account. Further, generally accepted indices of grammaticalization—layering, persistence, semantic bleaching, syntactic generalization, and phonetic erosion—make predictions that can be tested using a variationist approach (Walker, 2010, p. 106). A multivariate analysis is employed using GoldVarbX (Sankoff, Tagliamonte, & Smith, 2012) to tease apart the complex interaction of social and linguistic factors that influence speaker choice. The resultant form-function patterns will be examined in order to determine if grammaticalization is implicated in the variation.

This analysis further assumes the concept of a cline of grammaticalization (Hopper & Traugott, 2003). That is, grammaticalization does not involve abrupt shifts from one category to another, but consists of a series of small transitions that emerge synchronically as a continuum between a fuller, less grammatical form, to a reduced, more grammatical form (p. 6). This assumption becomes important when interpreting results that emerge from the data.

3.2 Data and Coding

Data was extracted from the Synchronic Corpus of Victoria English (SCVE), housed at the University of Victoria Sociolinguistics Research Lab (SLRL). The corpus consists of sociolinguistic interviews with 162 speakers from Victoria, BC, born between 1913 and 1996. A total of 24 speakers were selected based on the factors of age and gender (Table 1). In total, this smaller set of interviews comprises 21 hours of speech and over 275 000 words. Three age groups were defined: younger (18-25), middle (30-49), and older (63-85) to enable an apparent-time analysis of the distribution of variants (Tagliamonte & D'Arcy, 2007), that is, an analysis of different generations at one point in time. As this analysis does not address localized variants or prestige forms, socioeconomic status was not included as an independent variable; all speakers have mid- or upper mid-range SES scores.

	Male	Female	N Speakers	N Tokens
17-25	4	4	8	111
30-49	4	4	8	91
63-85	4	4	8	83
TOTAL	12	12	24	285

Table 1: speaker sample and tokens extracted from SCVE

All instances of the negative periphrastic DO in collocation with the verb *know* and the first-person pronominal subject *I* (or a zero-subject that is coreferential with *I*) were extracted from a 30-minute segment of each interview. On the assumption that speech is more monitored at the beginning of an interview, the 30-minute segment after first 10 minutes of the interview was used for analysis. To control for vast divergences in rates of usage that may confound the results, this 30-minute window was decreased or increased so that speakers had no more than 20 tokens or no fewer than 5 tokens. Where the form or function could not be unambiguously determined, such as in utterances that were cut off or obscured by other sounds, tokens were excluded. Twelve tokens were also excluded because they included adverbial modification (e.g. *I really don't know, I don't even know*). These forms will be addressed in the discussion. Following these methods, a total of 185 tokens were retained and coded for social and linguistic factors.

Each token of *I don't know* was coded auditorily for phonetic form. The full form (*I don't know*) has a distinct morpheme boundary (normally a glottal stop) between the nasals of *don't* and *know*, and the full vowel [o] in *don't*. The intermediate form (*I donno*) has no distinct morpheme boundary between *don't* and *know*, but still has a full vowel in *don't.* The reduced form (*I dunno*) has no distinct morpheme boundary between *don't* and *know*, and the vowel in *don't* is reduced to [ə]. These three forms are similar to those identified in Pichler (2009). The present analysis also includes a category for further reduced forms (*I d'no*), which have no morpheme boundary between *don't* and *know*, a reduced vowel, and some further reduced aspect, such as a lenited [d] in *don't* (e.g. [əno]), no vowel at all in *don't* (e.g. [dno]), a complete fusion of *don't* and *know* (e.g. [ɾo] with a flap), or a complete lack of phrase-medial consonants (observed as a 'prosodic grunt' with a rise-fall intonation that identifies it as I DON'T KNOW).

Following Pichler (2009), syntactic configuration was coded by determining whether tokens have an overt complement. Bound tokens are either preceded or followed by an overt complement, as in (1) and (2) respectively:

(1) WB/79/f well <u>what became of him</u> **I don't know** but I suppose he'd have been relocated

(2) JS/23/m **I don't know** <u>about weddings and stuff</u>

Unbound tokens have no overt complement and are grammatically independent, as in (3):

(3) KA/18/m my most embarrassing moment **I don't know** . I don't think that's a very good question for me

Semantic function was determined following various observations in the literature regarding the pragmatic functions of I DON'T KNOW (e.g. Baumgarten & House, 2010; Beach & Metzger, 1997; Bybee & Scheibman, 1999; Diani, 2004; Weatherall, 2011), and noting indications from the prosody, conversational context, and occurrence of other discourse markers (Pichler, 2009). Tokens that indicated a lack of knowledge were coded as Referential, as in (4) and (5):

(4) CA/21/f so there was actually like a T-A at the school who would take me outside . on my bike and show me how to ride a bike **I don't know** why I got this weird special treatment

(5) JF/84/m so I've been there eleven years . and I applied for Quadra why I did **I don't know**

I DON'T KNOW is also used to maintain rapport and mitigate face threats. Tokens that functioned as markers of reduced epistemicity, politeness devices, or hedges in communicating lack of commitment to a following or preceding utterance were coded as Interpersonal, as in (6) and (7):

(6) INT: would you . put birds in there ? are you interested in doing that ?
BD/30/m uh . no not really it seems **I don't know** it seems weird to . [INT:laughs] keep them in a cage for your viewing . when they could just fly around . so <yeah> . [clicks tongue] yeah

(7) INT how did your pajama pants turn out ?
BL/31/f they were great but . um . pajama pants **I don't know** . they're not that special [laughs]

I DON'T KNOW may be deployed to structure dialogue. It is available to mark topic boundaries, initiate or prevent turn exchange, and link aborted and recast statements (repair). These tokens were coded as Textual, as in (8) and (9):

(8) BL/31/f she's good she was always around . <yeah> because she um . yeah she stayed at home she had a few jobs occasionally but <right> mostly she was at home <mm-hm> .um yeah had a good childhood . <yeah> um . we . you know **. I don't know**

(9) DK/63/m we'd jump down the laundry shoots to land on the mattresses
INT that's /awesome/
DK/63/m /**I don't know** / and then they had those dumb {unclear} we used to crawl up that but uh .

Finally, as Pichler (2009) notes, discourse markers are polypragmatic devices. In the case of I DON'T KNOW, many occurrences serve both interpersonal and textual functions. Rather than subjectively choosing between Interpersonal and Textual, these tokens were coded as Polypragmatic, as in (10) and (11):

(10) *Repair (textual) and hedge (interpersonal)*
CA/21/f she'd talk about how . **I dunno** she'd she made a lot of . World-War-Two jokes [laughs] . i-- with the understanding that it was a terrible thing but it you know . {unclear} . you could make light of it

(11) *Turn-yield (textual) and disclaimer (interpersonal)*
JF/84/m um I worked at Macaroni-Grill on Davie that was the {unclear} Mansion ? <oh okay> **I don't know** yeah

In coding tokens in such a manner the effects of age, gender, syntactic form, and semantic function can be quantified and statistically analyzed. This enables a statistical analysis of the internal and external factors that affect variant choice.

4 Results

4.1 Overall frequency

To ascertain the frequency of usage of I DON'T KNOW, the entirety of the 18 interviews selected from the corpus were analyzed using AntConc concordance software (Anthony, 2011). Consistent with results reported in previous corpus analyses (E.G. Baumgarten & House, 2010; Grant, 2010), I DON'T KNOW is the most frequent 3-word phrase in these materials, occurring a total of 707 times. The next most frequent 3-word phrase, *a lot of*, occurs 447 times.

Following Thompson and Mulac (1991), type frequency of I DON'T KNOW was compared to the token frequency of the negative periphrastic DO

construction. Negative periphrastic DO occurs a total of 1545 times, meaning that nearly half of its occurrences are in the construction I DON'T KNOW (707/1545 = 46%). As shown in Table 2, the first person singular pronoun *I* is by far the most frequent subject collocating with DON'T (1153/1545= 75%). Table 3 shows that when the following word is taken into account, *know* is the most frequent collocate (748/1545=48%).

	N	%
I + don't	**1153**	**75**
you + don't	107	7
they + don't	59	4
we + don't	36	2
OTHER	190	12
TOTAL	1545	

Table 2: relative frequencies of subjects collocating with negative periphrastic DO

	N	%
don't know	**748**	**48**
don't think	124	8
don't have	63	4
don't like	49	3
don't want	46	3
don't remember	45	3
OTHER	470	30
TOTAL	1545	

Table 3: relative frequency of verbs following negative periphrastic DO

Considering the findings regarding frequency and reduction from Thompson & Mulac (1991) cited in §2, the high token frequency of I DON'T KNOW is conceivably a major cause of its phonetic reduction. Whether or not this frequency has led to phonetic reduction in the context of the emergence of grammar will be determined by analyzing the social and linguistic conditioning of the variants. If a connection is established between phonetic form and function, this will indicate that the phonetic reduction is not simply an effect of frequency, and will support the grammaticalization hypothesis.

As outlined in §3, 268 tokens were coded for 3 linguistic factors: phonetic form, syntactic form, and semantic function. Cross tabulation revealed that the factor groups Syntactic form and Semantic function strongly interact: nearly categorically, syntactically bound tokens (those with an overt complement) were used in referential contexts (88/93=95%), and syntactically unbound in other pragmatic functions (163/174=94%). Bound tokens were therefore excluded from the analysis, leaving a total of 174 unbound tokens. This

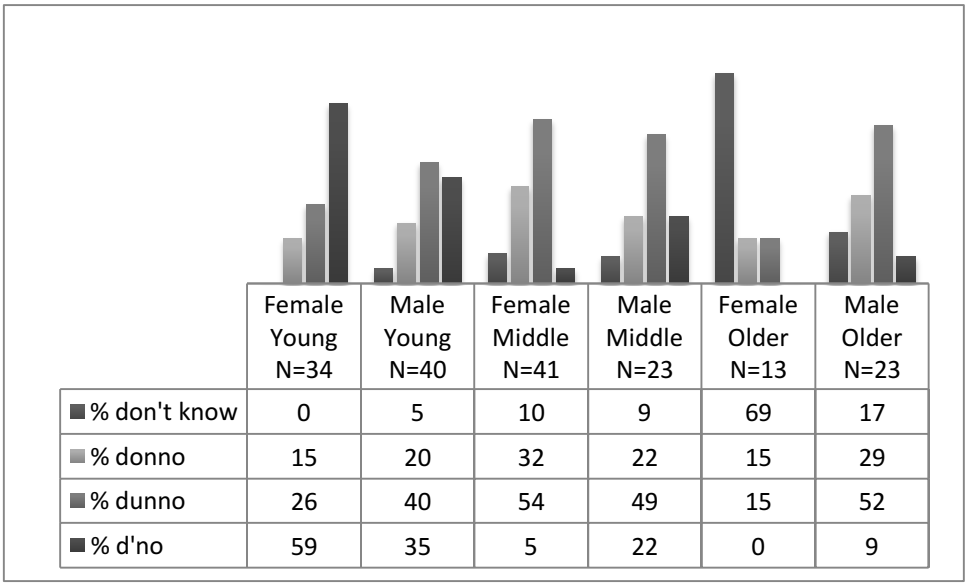
resulted in excluding over half of *don't know* and *dunno* tokens, as well as every instance where *dunno* and *d'no* were used referentially. This will be further discussed in §5. Table 4 shows the overall distribution of variants after bound tokens were removed.

	don't know	***donno***	***dunno***	***d'no***	**TOTAL**
%	12	22	41	25	
N	21	38	72	43	174

Table 4: overall distribution of variants

4.2 Social factors contributing to distribution of variants

Figure 1 shows the distribution of variants with respect to age and gender. The figure shows the percentage of each of the variants in each of the 6 social groups: young female, young male, middle female, middle male, old female, old male. While Pichler (2009) finds that the reduced form *I dunno* is the most common variant across social cohorts (aside from old male), the present data reveal a different pattern: the full form *I don't know* is the most common variant in the old female group, the reduced form *I dunno* is favoured by both middle aged groups and older males, and the further reduced form *I d'no* is the most favoured in the young cohorts. The intermediate form *donno* is not conditioned by age, but does correlate with male speakers. This may contribute to the difference between the old male and old female cohort; cross tabulating the results did not reveal any other factor that may affect this difference, although with such small numbers (N=13 for old female and N=23 for old male), ideolectal effects could easily obscure the results.



	Female Young N=34	Male Young N=40	Female Middle N=41	Male Middle N=23	Female Older N=13	Male Older N=23
% don't know	0	5	10	9	69	17
% donno	15	20	32	22	15	29
% dunno	26	40	54	49	15	52
% d'no	59	35	5	22	0	9

Figure 1: social distribution of variants of I DON'T KNOW.

This age effect shown in figure 1 is an unexpected result given previous analyses of I DON'T KNOW. The incrementally increasing frequency of the very reduced form and the decreasing frequency of the full form in apparent time, as observed in figure 2, suggests that this reduction is not an age-graded effect, but ongoing generational change (Tagliamonte & D'Arcy, 2007). However, the only decisive method for resolving this would a study in real time. Determining whether this change has to do with phonetic reduction only, or whether this is an instance of the emergence of grammar, requires an analysis of the functional distribution of forms.

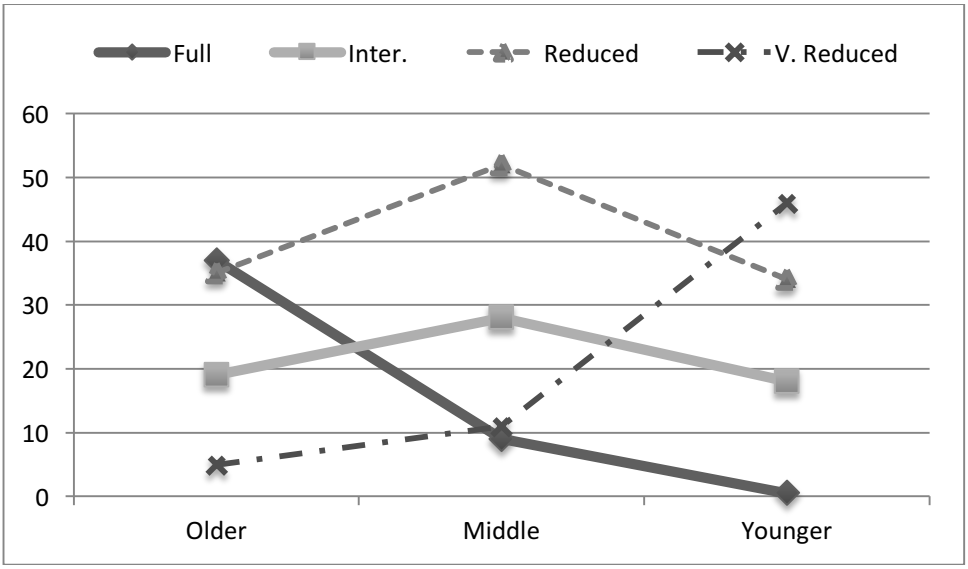
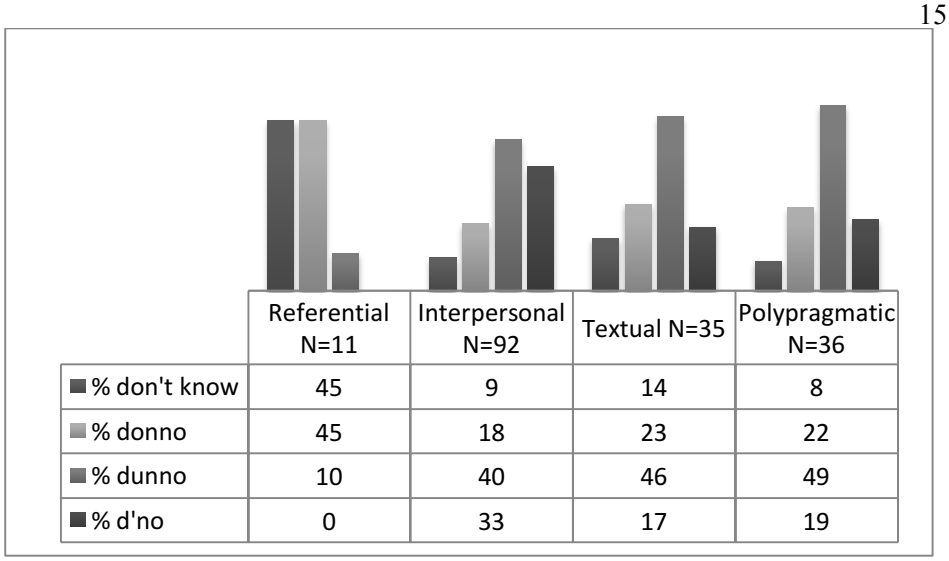


Figure 2: age distribution of variants of I DON'T KNOW

4.3 Linguistic factors contributing to the distribution of variants

Figure 3 gives a breakdown of the functional distribution of the variants of I DON'T KNOW. It tracks the percentages of each of the four categories (referential, interpersonal, textual, and interpersonal-textual) across the four variants. The patterning here is similar to that reported in Scheibman (2000) and Pichler (2009): the full variant correlates with referential uses, though this result must be treated cautiously due to the low number of referential tokens. Recall that, in removing the bound tokens, the majority of tokens functioning referentially, including all instances of *I dunno* and *I d'no* that function referentially, were also removed. The remaining 8 tokens that function referentially are *I don't know* or *I donno*. The reduced variants *I dunno* and *I d'no* correlate with pragmatic uses. These results are suggestive of functional conditioning of variants of I DON'T KNOW as a result of the variable's grammaticalization. The patterning in figure 3 also shows that the pragmatic functions—interpersonal, textual, and polypragmatic—are not particularly differentiated, as they all pattern in the same way. For this reason, they will be collapsed into a singular pragmatic category for the following analysis.



	Referential N=11	Interpersonal N=92	Textual N=35	Polypragmatic N=36
% don't know	45	9	14	8
% donno	45	18	23	22
% dunno	10	40	46	49
% d'no	0	33	17	19

Figure 3: functional distribution of Variants of I DON'T KNOW.

4.4 Multivariate analysis of the distribution of variants

The results in §4.2 suggest that the distribution of variants seems to be shifting in apparent time: older speakers are more likely to use the full form, middle speakers the reduced form, and younger speakers the very reduced form. In §3.3, the results indicate that reduced forms, and primarily *dunno*, correlate with pragmatic functions. To test whether the apparent ongoing change in phonetic reduction is an effect of differential uses of semantic function across age groups, a multivariate analysis is performed using GoldVarbX. Referential function was included only for the analyses of the full and intermediate forms (Tables 5 and 6). As none of the 8 referential tokens were reduced or very reduced variants, they were excluded from the multivariate analysis of these tokens (Table 7 and 8). [2] During the initial analysis, it was noted that one speaker categorically produced the very reduced form *I d'no.* She contributed nearly half (18/43=42%) of all *I d'no* tokens. Further, this speaker was in the young cohort. As the total number of tokens is only 174, such a speaker effect can have a huge difference on the distributional results. The data from this speaker was excluded, and the remaining 156 tokens were rerun through GoldVarb.

[2] Investiation of these 8 referntial tokens revealed no speaker effects: the 8 tokens came from 7 different speakers, spread across age groups (2 young speakers, 2 middle, and 4 older) and speaker sex (4 female speakers and and 4 male).

		don't know	
Input		.081	
Overall %		14	
Total N		156	
	FW	**%**	**N**
Sex			
Female	**.65**	17	70
Male	**.37**	9	86
Range	***28***		
Age			
Older	**.86**	36	36
Middle	**.43**	9	64
Younger	**.30**	3	56
Range	***56***		
Semantic Function			
Referential	**.86**	46	11
Pragmatic	**.47**	10	145
Range	***39***		

Table 5: multivariate analyses of the contribution of internal and external predictors (significant and non significant) to the probability of full form.
(Log likelihood= -46.86, p=0.044)

		donno	
Input		.237	
Overall %		24	
Total N		156	
	FW	**%**	**N**
Sex			
Female	[.55]	29	70
Male	[.46]	21	86
Age			
Middle	[.54]	28	64
Younger	[.51]	23	56
Older	[.42]	19	36
Semantic Function			
Referential	[.74]	46	11
Pragmatic	[.48]	23	145

Table 6: multivariate analyses of the contribution of internal and external predictors to the probability of intermediate form.

		dunno	
Input		.045	
Overall %		46	
Total N		156	
	FW	**%**	**N**
Sex			
Female	[.50]	47	70
Male	[.50]	45	86
Age			
Middle	[.56]	52	64
Younger	[.48]	45	56
Older	[.44]	39	36
Semantic Function			
Pragmatic	**.54**	49	145
Referential	**.11**	9	11
Range	***43***		

Table 7: multivariate analyses of the contribution of internal and external predictors (significant and non significant) to the probability of reduced form (Log likelihood = -114.98, p= 0.015)

		d'no	
Input		.12	
Overall %		16	
Total N		156	
	FW	**%**	**N**
Sex			
Male	**.67**	24	86
Female	**.30**	5	70
Range	***37***		
Age			
Younger	**.67**	29	56
Middle	**.50**	11	64
Older	**.24**	6	36
Range	***33***		

Table 8: multivariate analyses of the contribution of internal and external predictors (significant and non significant) to the probability of very reduced form (Log likelihood = -78.25, p= 0.000).

The results from table 5 show that for the full form, variant choice is significantly favoured for referential uses, and among speakers who are female and in the older category. In table 6, no factors reach significance for the intermediate form, though there is a robust effect of referential function. Table 7 shows that for the reduced form, semantic function is the only factor selected as significant, and the variable is favoured for pragmatic uses. There is also a direction of effect that privileges the middle age cohort. Finally, table 8 shows that age has a very strong effect for the very reduced form, with younger speakers being the most likely to use it. This variant is also more favoured among male speakers, though this did not reach significance. The very reduced form was categorically pragmatic, so the semantic factor group is not included in the analysis.

In Pichler (2009), the differential distribution of the full and reduced form is found to be significant across referential uses and pragmatic uses *combined.* Because there were no referential tokens that surfaced as reduced or very reduced variants in the present data, a similar comparison could not be made. An second analysis was performed which included syntactic function as a factor group instead of semantic function, but this did not result in a better fit for the data. Further, it resulted in a similar problem: removing all referential tokens from the analysis also removed the majority of bound tokens.

5 Discussion

In analyzing these variants, it must be noted that the four categories identified are not well-defined, immutable groups: they represent different points on a continuum, from full articulation to nearly complete erosion (e.g. a consonantless prosodic grunt). This is a key characteristic of changes that occur in the context of grammaticalization: small shifts occur on a cline from a more full, less grammatical form to less full, more grammatical form (Hopper & Traugott 2003). Admittedly, gradual change is difficult to differentiate from abrupt change in an apparent-time analysis (Walker, 2010). Further, as addressed in Lorenz (2013), evidence is needed to show that reduced forms are not simply easier ways of pronouncing lexical items, but that they have independent meanings and a distinct cognitive representation.

5.1 Grammaticalization

Unlike Scheibman (2000) and Pichler (2009), the present results do not yield a phonetic form-function split. However, a split was identified in syntactic context and function, which complicated comparison of the distribution of phonetic variants. As stated in §3.1, bound tokens are nearly categorically used for a referential function (94%), and unbound tokens for pragmatic functions (93%). Figure 4 shows the distribution of variants in these bound tokens. The majority of the excluded bound tokens are the full form *I don't know*.

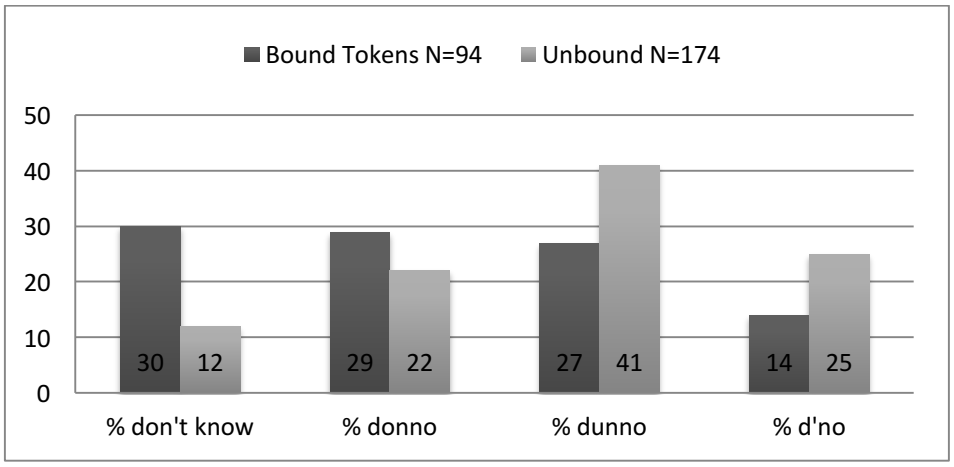


Figure 4: distribution of variants across bound tokens (N=94) compared to unbound tokens (N=174). The distribution of variants across bound and unbound functions is highly significant (χ_2=22.80, df=3, p<0.01)

This split indicates two different kinds of I DON'T KNOW: one form encodes the referential meaning of "lack of knowledge" and resides predominantly in matrix-complement (bound) constructions[3], and another that encodes pragmatic meaning and occurs mostly in unbound constructions. This suggests that at one point, the frequent construction I DON'T KNOW spread to wider syntactic contexts, and its use in these context came to be associated with semantically faded, pragmatic meanings. These observations, along with the phonetic erosion of I DON'T KNOW and its high frequency (described in §3.1) are indicative of grammaticalization. I would argue that the generational change observed in the patterning of phonetic variants is not indicative of an emerging new form, but rather the ongoing phonetic reduction of an already grammaticalized form.

Furthermore, several of Hopper's (1991) indices of grammaticalization are observed in the distribution of I DON'T KNOW. Decategorialization is indicated by the fact that the grammaticalized form is rarely used in the matrix-complement construction. Divergence is also observed, as the original lexical form is still comprised of autonomous elements: it may be adverbially modified (e.g. *I don't* ***even*** *know*, *I* ***really*** *don't know)*, while the grammaticalized form rarely has an intervening adverb (Pichler, 2009). Of the 12 adverbially modified tokens that were removed from the analysis, 11 had referential meaning. Persistence is shown in the pragmatic meanings of reduced epistemicity and lack of commitment; nuances of the referential meaning of lack of knowledge remain. Finally, layering is observed in the coexistence of the two forms.

[3] When I DON'T KNOW is in second position as a referential response to a question (as opposed to a polite response, Diani, 2004), there may be no overt surface complement, but a complement is implied, e.g. "I don't know (the answer to the question)"

The finding that phonetic reduction of the grammaticalized form appears to be advancing in apparent time reflects Bybee et al.'s (1994) Parallel Reduction Hypothesis described in the introduction. I DON'T KNOW, as a result of high frequency and widened syntactic context, has become entrenched as a single processing unit autonomous of its source form. Thus, it is accessed and produced faster, and morphological boundaries lose significance and are subject to erosion. The loss of semantic content has enabled I DON'T KNOW to function as a pragmatic marker, which leads to loss of prosodic weight and stress, further catalyzing reduction.

5.2 Speaker effects

Because of the small size of my sample, I had to consider how speaker effects complicated my analysis. While some speakers took over an hour to produce 5 tokens, one speaker in particular produced 20 in less than 15 minutes. Not only were tokens highly frequent in her speech, they were nearly categorically very reduced—almost half of the very reduced tokens came from this speaker. This speaker is a female, and is also the youngest in the sample. In initial analyses, when this speaker was included, sex was never selected as significant for any variant. After she was removed, not only did sex emerge as a significant factor for the full form, the direction of effect for the very reduced form switched from female to male.

Though excluded from statistical analyses, qualitative consideration of data from this speaker does give weight to the parallel reduction hypothesis: the only referential tokens in her data were adverbially modified, which indicates that grammaticalization is at a very advanced stage. Adverbial modification is now required to encode referential meaning, and the grammaticalized form is becoming increasingly reduced. Whether this speaker's patterns indicate ongoing change, or linguistic marketplace effects in the heterosexual talk market of high school (Eckert, 2011), or the speaker's own idiolect remains an open question, but an interesting avenue for further research.

5.3 Moving forward

This project is an exploratory investigation of variation that potentially indicates grammaticalization. As such, it has brought about many questions and directions for future research. The results reported here are not entirely consistent with those reported by Scheibman (2000) or Pichler (2009). While there are similar correlations observed between the full-form and referential uses and the reduced form and pragmatic uses, the affect of age was not reported in either study. This indicates that something different is happening to I DON'T KNOW in Victoria English—is this a community effect, or is this a stage on the universal pathway of grammaticalization of I DON'T KNOW that was undetected by Scheibman or Pichler? To answer this question requires more data from a wider range of speech communities. It was further reasoned that this increased reduction was a

generational change, not an age-graded effect, and that the grammaticalization of I DON'T KNOW has already taken place. These hypotheses can only be verified by a study in real-time.

The relatively low number of tokens created problems for analysis, especially as no reliability test was performed for the coding procedures. Observed patterns are easily obscured by small differences in raw numbers. Additionally, the small number of speakers and tokens makes it difficult to ensure that patterns observed are not due to ideolectal features. Further, to ensure that the data are representative of individual patterns of usage, Labov (1966, p. 181) advocates for 10-20 instances per speaker, while others call for even more (e.g. Guy, 1993). Several of the speakers in the sample had fewer than 10 tokens. Future analyses will benefit from analyzing a greater number of tokens from a greater number of speakers.

Reduction was defined primarily in terms articulatory gestures for this project. However, Bybee et al. (1994) indicate that phonetic reduction occurs in articulatory gestures as well as duration, though measuring duration was beyond the scope of this project. Further, the unit of focus in coding for reduction was DON'T, following Scheibman (2000) and Pichler (2009). Reduction was most salient for the consonant [d] in DON'T, the morpheme boundary [n?n] between DON'T and KNOW, and the vowel in DON'T. Impressionistically, the initial vowel [aj] ("I") and the final vowel [ow] in KNOW also vary in their phonetic form, though this was not included in coding. The pattern of reduction for these vowels may have interesting implications for the concept of phonetic reduction: does a hierarchy exist for which segments are reduced first? Analyzing variants based on the production of all segments and for phonetic duration would enrich this analysis, potentially yielding interesting results for patterns of distribution.

6 Conclusion

The interaction of frequency and phonetic reduction is a well-documented phenomenon in language change (Bybee, 2006). While these two phenomena are inevitable components of grammaticalization, they are not, in themselves, necessary or sufficient in identifying cases of grammaticalization. This study sought to untangle these interwoven processes by performing a variationist analysis of the distribution of phonetic variants of the highly frequent collocation I DON'T KNOW. Results similar to those reported in previous studies were uncovered concerning the form-function regularities of I DON'T KNOW, indicating a universal path of grammaticalization of this construction. Previously unreported results were also found: the increased frequency of reduced and very reduced variants in apparent time indicates ongoing change in the reduction of I DON'T KNOW. Whether this increasing reduction is an age-graded effect or a generational change, and whether it is a community-specific or universal tendency, calls for further examination of this form across time and speech communities.

Acknowledgements

A huge thank-you to Alex D'Arcy for her generous expertise and guidance through the many iterations of this project, and to my partner John Edgar for being my ever-patient statistics consultant.

References

Anthony, L. (2011). AntConc (Version 3.2.2) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved November 27, 2014, from http://www.laurenceanthony.net/

Beach, W.A., & Metzger, T. (1997). Claiming insufficient knowledge. *Human Communication Research, 23*(4), 562-588.

Baumgarten, N., & House, J. (2010). I think and I don't know in English as lingua franca and native English discourse. *Journal of Pragmatics*, *42*(5), 1184-1200.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language, 82*(4), 711-733.

Bybee, J., Perkins, R., & Pagliuca, W. (1994). *The evolution of grammar: Tense, aspect, and modality in the languages of the world.* Chicago, IL: University of Chicago Press.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of *Don't* in English. *Linguistics*, *37*(4), 575–596.

D'Arcy, A. (2011-2014). Victoria English: its development and current state. Standard Research Grant no. 410-2011-0219. Social Sciences and Humanities Research Council of Canada.

Diani, G. (2004). The discourse functions of *I don't know*. In K. Aijmer & A. Stenström (Eds.): *Discourse patterns in spoken and written corpora*. Amsterdam: John Benjamins. 157-171.

Eckert, P. (2011). Language and power in the preadolescent heterosexual market. *American Speech, 86*(1), 85-97.

Grant, L. (2010). A corpus comparison of the use of I don't know by British and New Zealand speakers. *Journal of Pragmatics*, *42*(8), 2282-2296.

Guy, G. (1993). The quantitative analysis of linguistic variation. In D. Preston (Ed.), *American dialect research*. Philadelphia, PA: John Benjamins. 223-249.

Hopper, P. (1991). On some principles of grammaticization. In E.C. Traugott & B. Heine (Eds.) *Approaches to grammaticalization: Focus on theoretical and methodological issues*, vol. 1. Philadelphia, PA: John Benjamins. 17-35.

Hopper, P. and Traugott, E. (2003). *Grammaticalization.* 2nd edition. Cambridge: Cambridge University Press.

Kaisse, E. (1985). Connected speech: The interaction of syntax and phonology. San Diego, CA: Academic Press.

Labov, W. (1966). *The social stratification of English in New York City.* Washington, DC: Centre for Applied Linguistics.

Labov, W. (1969). Contraction, deletion, and inherent variability of English copula. *Language, 45*(4), 715-762.

Lorenz, D. (2013). Contractions of English semi-modals: The emancipating effect of frequency. NIHN Studies. Freiburg: Rombach.

Meillet, A. (1912). L'evolution des forms grammaticales. *Linguistique général et linguistique historique.* Paris: Champion. 130-148.

Pichler, H. (2009). The functional and social reality of discourse variants in a northern English dialect: I DON'T KNOW and I DON'T THINK compared. *Intercultural Pragmatics*, *6*(4), 561–596.

Sankoff, D., Tagliamonte, S., & Smith, E. (2012). *Goldvarb Lion: A variable rule application for Macintosh*. Department of Linguistics, University of Toronto.

Scheibman, J. (2000). *I dunno*: A usage-based account of the phonological reduction of *don't* in American English conversation. *Journal of Pragmatics, 32,* 105-124.

Tagliamonte, S. & D'Arcy, A. (2007). Frequency and variation in the community grammar: Tracking a new change through the generations. *Language Variation and Change, 19,* 199-217.

Thompson, S., & Mulac, A. (1991). A quantitative perspective on the grammaticalization of epistemic parentheticals in English. In E. C. Traugott and B. Heine (Eds.), *Approaches to Grammaticalization*, vol. 2, 313-330. Amsterdam: John Benjamins.

Torres Cacoullos, R., & Walker, J. (2009). The Present of the English Future: Grammatical Variation and Collocations in Discourse. *Language, 85*(2), 321-354.

Walker, J. (2010). *Variation in linguistic systems.* New York, NY: Routledge.

Weatherall, A. (2011). I don't know as a Prepositioned Epistemic Hedge. *Research on Language & Social Interaction*, *44*(4), 317-337.