Need to vs. *have/got to*: Four socio-pragmatic corpus studies

Lelia Glass

To Appear in Selected Papers from NWAV 43 (U. Penn. Working Papers in Linguistics)

1 Introduction

A growing chorus of researchers has been urging the fields of sociolinguistics and semantics/pragmatics to engage more with one another. Since at least the 1970s, the father of sociolinguistics has avoided the term "sociolinguistics" because "it implies that there can be a successful linguistic theory or practice which is not social" (Labov 1972: xv). Eckert 2011 agrees, calling on semantics/pragmatics in particular: "It is time to integrate the study of variation with the study of meaning in language more generally." As she points out, both sociolinguistics and semantics/pragmatics deal with meaning. Both also rely on the idea of a competition between forms, where meaning is conveyed not only by the forms themselves, but also by the speaker's choice among them.

At the same time, it is getting much easier to mine large corpora and perform experiments quickly. These advances allow researchers to ask new questions, expanding the types of variation that can be studied to better include semantic and pragmatic phenomena (Eckert 2011, de Marneffe and Potts 2014). Now that linguists are more eager and equipped to address such questions, recent work has started to unify the study of meaning. I am particularly inspired by the idea of deriving a linguistic variant's social meaning from the way its semantics differs subtly from that of its competitors (Torres Cacoullos 2001, Acton and Potts 2014). These studies use corpora as their main evidence, again highlighting that usage is key for understanding how subtle semantic differences play out socially.

When the notion of a sociolinguistic "variable" is loosened so that competing variants are allowed to have slightly different meanings, these different meanings might relate to each other in various ways. Each of these semantic relationships opens up a different dimension along which semantically distinct variants can compete, and thus a different way that a variant's social meaning can arise. For example, it is well known that speakers may choose a weaker or more ambiguous form for social reasons. When a speaker uses *some* instead of *all*, she might want to implicate that she does not believe *all* would be true. But there might be a more subtle, social reason: perhaps she wants to remain truthful while sparing the hearer's feelings. So she might say that *some people hated your poem* when all of them did (Bonnefon *et al* 2000). When a speaker asks her date if he wants to come upstairs for coffee (an offer for sex), she chooses the more indirect form so that she could *plausibly deny* her intention in case he refuses (Pinker *et al* 2008). These social reasons for choosing a weak or ambiguous form are ripe to be explored in the growing socio-pragmatic literature.

In this paper, I investigate a pragmatic competition between more and less ambiguous variants where the choice between them is influenced by interactional factors. I consider the strong root necessity modal *need to* compared to *have/got to*. It has been observed that *need to* describes an obligation that originates with some party's "internal compulsion" (Smith 2003, Leech 2003, Nokkonen 2006) or priorities (Rubinstein 2012), whereas *have/got to* are more ambiguous about the source of the obligation. Researchers have also suggested that this subtle semantic difference gives *need to* a unique social meaning (Nokkonen 2006, Smith 2003, Leech 2003), especially in the second person (*you need to*), which is the focus of this paper. However, the exact nature of this social meaning remains elusive: it has been described as "polite" (Smith 2003) and "democratic" (Leech 2003) since it appeals to the addressee's needs as opposed to the speaker's authority; but it has also been called "hierarchical" (Nokkonen 2006), "infantilizing" (Yagoda 2006) and "underhanded" (Yagoda 2006).

I attempt to sort out the unique socio-pragmatic meaning of *need to* by examining how it competes with its more ambiguous alternatives, *have to* and *got to*. Because *need to* ties the obligation to someone's internal needs or priorities (usually the hearer's in the second person), the speaker who uses *you need to* unambiguously acts as if she is familiar with the hearer's priorities and licensed to advise him on what would serve these priorities. In contrast, with *you have/got to*, she does not

necessarily claim to know his priorities or know what would be good for him, because she could simply be reporting a more general, external obligation.

Since *you need to* reveals that the speaker (thinks she) knows what is good for the hearer, it can be risky. If the hearer does not feel that the speaker has a legitimate claim to tell him what is good for him in the context, it can be perceived as presumptuous. But if the speaker does have a legitimate claim to use *you need to*, it can also allow her to appear considerate of the hearer's priorities.

As a result, I predict that *you need to* will be more appropriate, and thus more commonly used, by people who have some claim to know what is good for the hearer. This might include the hearer's close friends; people in authority over him, especially people who play a mentoring role in his life; and people with recognized expertise about what he should do in a given situation. I also predict that *you need to* will be used less often by people without a claim to know what is good for the hearer, including strangers or acquaintances who have less legitimate concern with his well-being; subordinates to authority figures; and people with no expertise about what he should do in a given situation. Finally, I predict that when *you need to* is used by someone without license to do so, it may be perceived as presumptuous. I find evidence consistent with these predictions in a series of corpus studies.

The sociolinguistic variable has traditionally been defined as "two ways of saying the same thing" (Labov 1972: 271). To adapt the concept to semantic or pragmatic features, researchers argued that the notion of the variable could be defined more loosely "on the basis of common discourse function", (Dines 1980; Cheshire 2005), "functional comparability" (Lavandera 1978) or "rough semantic equivalence" (Weiner and Labov 1983).

When variants of a variable are allowed to differ in meaning as long as they are sufficiently similar, it becomes possible for subtle *semantic* differences between variants to give rise to differences in social meaning. This strategy has been especially effective in the domain of function words (closed-class items such as determiners and negation) and function phrases (tag questions, fillers, and hedges), perhaps because classes of such words clearly serve a common function. While function words and phrases may seem utilitarian, a long tradition of research has revealed that they are rich in social meaning (Lakoff 1974, Stubbe and Holmes 1995, Dixon and Foster 1997, Moore and Podesva 2009, Potts 2011, a.o.). Within this line of work, I am particularly inspired by cases where a particular function word's unique social meanings is derived from the way its semantics differs subtly from that of its competitors.

For example, Torres Cacoullos 2001 investigates two different progressives in Mexican Spanish, the *estar* progressive (where *estar* historically means "to be located at") and the *andar* progressive (where *andar* historically means "to go around"). She finds that these two forms convey different social meanings which derive from their different denotational meaning. Using corpus evidence, she argues that the *andar* "go around" passives index rural or popular urban identity because it is more associated with physical, outdoor activity, whereas the *estar* passive is associated with more educated people and their characteristically indoor interests. Thus, these two terms have subtly different semantic meanings (*andar* suggests physical movement, *estar* does not) which give rise to their different social meanings and different distribution.

Acton and Potts 2014 consider demonstratives (*this, that, these, those*) as compared to other determiners. They argue that *that* (as in *that liberal media that twists everything you say*) strongly presupposes that the hearer shares a referent for the item in question. If the hearer agrees that there is a liberal media that twists everything you say, he might feel a strong sense of shared perspective with the speaker; but if he does not agree, he might feel manipulated by this false presupposition. They use former U.S. vice-presidential candidate Sarah Palin as a case study, showing that Palin uses demonstratives more than other politicians and arguing that this variable contributes to her polarizing persona.

As with the Spanish progressives and the determiners, the current study also uses corpus evidence; investigates a class of function words; and attempts to derive one variant's unique social meaning from the way its semantics differs slightly from that of its competitors. But unlike the earlier studies, it focuses on a case where the variants differ in ambiguity, illuminating a new way that social meaning can be grounded in semantic meaning.

2 *Need to*

Modals are expressions for discussing obligation, inference, ability, permission, and much more. Semantically, modals come in different flavors: they can discuss states of knowledge and uncertainty (epistemic modals), or ability, permission, and obligation (root modals). Some modals can only express certain modal flavors: for example, in American English, *might* can only express epistemic possibility, not permission; *she might be here* can only mean it is possible in view of our knowledge that she will go to the store, not that she is permitted to do so. Other modals are extremely flexible: *She has to be here* could express an epistemic conclusion or a deontic obligation.

Within a given modal flavor, such as *deontic*, the exact nature of the inference or obligation is sensitive to context. There is almost always some uncertainty about exactly what contextually given body of rules is invoked by a particular use of a modal. For example, if an usher tells a theatergoer *You have to sit down*, it is unclear whether this obligation stems from the theatre's rules or social customs more generally. The context usually serves to narrow down the choices rather than to provide a precise answer. Throughout the paper, I will call this contextually-narrowed source of the obligation the *obligation source*, echoing the term *ordering source* used in certain semantic formalisms (Kratzer 1981 and related work), but without committing to any particular analysis.

Modals also come in varying degrees of strength: *must, have to, got to* and *need to* are stronger than *ought* and *should*, which are stronger than *can, may* and *might*.

When two modals have the same strength and the same obligation source, they entail each other. We can use a conditional clause to fix the obligation source (to "what is necessary to win") and thus pin down a context where it is a contradiction to assert one strong root necessity modal while denying another. For any combination of modals in 1, the sentence is contradictory.

(1) #You **have to/must/need to/have got to** practice if you want to win, but it's not the case that you **have to/must/need to/have got to** practice if you want to win.

We see that all the strong root necessity modals – including *need to* – mutually entail each other when their obligation sources are fixed. Thus, the strong root necessity modals form a cohesive class of roughly functionally equivalent items.

While *must* can be a root necessity modal just like the others, I will exclude it from the rest of my analysis because corpus hits for *must* are far more likely to be confounded by its relatively high percentage of epistemic uses (Tagliamonte and D'Arcy 2007). Deontic *must* has also been claimed to indicate that the speaker herself is imposing the obligation (Coates:1990: 56). Here, I exclude *must* and focus only on *need to* in comparison to *have/got to*.

With so many necessity modals (*must, have to, got to, need to*) able to express quite similar meanings, it is not surprising that these modals have competed for dominance in English. Recently, many of the older, "true" modals such as *must* and *may* have been losing out to "semi-modals" such as *have to, got to* and *need to* (Krug 2000, Leech 2003). *Need to*, in particular, has increased by 123.2% in American English and 249% in British English between the 1960s and 1990s (Leech 2003), and, according to Google NGrams, has continued to gain ground since. To explain why *need to* has increased so dramatically, many authors (Nokkonen 2006, Leech 2003, Smith 2003 a.o.) have claimed that *need to* has a special semantic and social meaning. However, the literature does not agree on what constitutes this unique meaning.

On the one hand, *need to* is sometimes claimed to be more "polite" than its competitors in the second person because it appeals to the hearer's needs rather than the speaker's own authority (Müller 2008, Smith 2003, Nokkonen 2006, Leech 2003). For example, Nokkonen 2006: 1 writes: "...the obligation expressed by need to often is more polite, impelling that the action demanded is in the addressee's own interest."

At the same time, Nokkonen has also noticed that in corpora, *need to* tends to be used by those with authority, such as trainers to trainees in business training sessions in the BNC (Nokkonen 2012) and teachers to teenagers in the COLT corpus (Nokkonen 2006). She concludes that "*need to* is useful when addressing subordinates politely" (Nokkonen 2012) and even that it is more frequently used among teenagers because "the world of teenagers is perhaps more openly hierarchical" (Nokkonen 2006: 46). Since "polite" phrasing usually seems egalitarian or even deferential, it is surprising that

need to is described as both "polite" and associated with authority.

We have seen that *need to* is conflictingly described not only as "polite," but also as "authoritative," "hierarchical" and "infantilizing." It may seem that only one of these characterizations can be true, since politeness is usually characterized by deference or friendliness (Brown and Levinson 1987), not reminders of authority. In my analysis, I try to sort out these reactions by proposing that the effect of *you need to* is not monolithic or consistent across contexts. Instead, it depends on how the speaker and hearer relate to each other.

3 Proposed social meaning for *need to*

Many researchers have noticed that *need to* is unique among the strong root necessity modals, in that it is thought to evoke "internal compulsion" whereas its competitors tend to evoke more "objective," "external" (Nokkonen 2006, Coates 1990) criteria. For example, 2a might simply mean that, given that people generally admire perseverance (the obligation source), the hearer ought to admire some person. In contrast, 2b seems to suggest that it is in the hearer's interest to admire her.

(2) a. You { **have to/gotta** } admire her for persevering.
b. You **need to** admire her for persevering.

To use the terminology introduced above, *need to* is more restricted in its obligation source than *have to* and *got to*. *Have/got to* can pick out any contextually supplied obligation source (including an epistemic one), but *need to* seems to require an obligation source that is related to some party's "internal compulsion," priorities, or interests (Rubinstein 2012). In contrast, *have to* and *got to* allow any contextually plausible obligation sources.

I take this semantic difference between *need to* and *have/got to* to be basic and primary. From that semantic difference derives a difference in ambiguity, which in turn gives rise the social consequences of choosing one or another. Finally, these different predicted social consequences give rise to predictions about the interactional contexts where each variant is predicted to be most appropriate and thus most prevalent. The only assumption needed to get the analysis off the ground is the difference in available obligation sources for *need* versus its competitors, which is well supported by linguistic evidence such as 2a–2b and previous literature.

In view of this semantic difference, 3a admits of more interpretations than 3b: the obligation could be tied to the hearer's priorities, the rules imposed by the speaker's authority, or simply the rules of graduate school. Or all of these options might be left open in the context. In contrast, 3b is less ambiguous. Here the obligation must be tied to the hearer's priorities in some way. It could be that the hearer's priorities include fulfilling the external rules of graduate school (the "pseudo-deontic" reading of Rubinstein 2012), but this must be subsumed under his priorities.

(3) a. You { **have to/'ve gotta** } finish your paper.
b. You **need to** finish your paper.

This difference in ambiguity influences the contexts in which a speaker will choose one form or another. If the obligation does not relate in any way to the hearer's priorities, only his external obligations, *have to* or *got to* are more truth-conditionally appropriate. The speaker who utters 3a does not necessarily know what is good for the hearer in view of his priorities; she only claims familiarity with the rules of school.

But if the speaker believes it *is* (or should be) a priority for the hearer to finish his paper, the choice is more subtle. On the one hand, *need to* describes the situation less ambiguously than *have/got to*, so according to Gricean principles (Grice 1989 (1967)), it should be favored. But on the other hand, in using *you need to*, the speaker presupposes that she knows the hearer's priorities and is licensed to tell him what would be good for him in view of these priorities. This presupposition could potentially come across as presumptuous if the hearer does not feel that the speaker is licensed to tell him what is good for him. To avoid this socially risky presupposition, the speaker might choose *you have/got to* instead of *you need to* even if she does believe that the obligation stems from the

hearer's priorities. In doing this, she leaves open an interpretation where she is simply referencing some unrelated body of rules.

If *you need to* is socially riskier than *you have/got to*, why do speakers use it at all? I suggest that when a speaker is indeed licensed to speak to the hearer's priorities, *need to* is a more compelling and potentially more considerate alternative. Given the semantics I've proposed, I agree with the earlier literature on *need to* that this form frames the obligation as for the hearer's own good. Explicitly relating the obligation to the hearer's priorities might inspire him to fulfill the obligation more than relating it to some potentially arbitrary external rules. Moreover, the speaker might come across as more considerate if she explicitly attends to the hearer's priorities.

In view of its semantics, when a speaker uses *you need to*, she presupposes that she is licensed to speak to the hearer's priorities and advise him about how to further them. Such a discourse move might be appropriate from someone who has a close relationship with the hearer and thus is familiar with his priorities; a mentor who is thus is licensed to give advice; an authority figure who can act as if she knows more than the hearer about what's good for him in a particular domain; or a person with knowledge on the particular domain she is advising about. In all of these relationships, the speaker is generally licensed to tell the hearer what would be good for him. For the same reasons, *you need to* would be *inappropriate* from people the hearer does not know well, unless they are experts on the relevant domain; people the hearer does not accept as mentor figures in his life; subordinates; or people with little knowledge about the relevant domain, because such speakers would not be in a position to tell the hearer what is good for him.

This analysis explains why *you need to* is described in such conflicting ways throughout the literature. On this account, *you need to* is not monolithically more or less "polite" than *you have to*, but will come across as considerate or presumptuous depending on how the speaker and hearer relate to one another. It may be perceived as "hierarchical" and "infantilizing" because people in power feel more licensed to tell others what to do in general, and because people in power may feel more licensed to speak to hearers' priorities using *need*. These different faces of *you need to* arise out of a unified socio-pragmatic story, depending on how the hearer feels about the speaker.

4 Testing the predictions

In order to test interactional predictions like the ones I am making here, we need specific corpora that provide information about how speakers and hearers relate to each other. I have tried to find four such corpora that allow me to test different aspects of the theory I outlined above: the Providence section of CHILDES (as compared to spoken CoCA); the Michigan Corpus of Academic Spoken English; the Stanford Politeness Corpus; and the U.S. television show *The Office*.

These four corpora allow me to test different facets of my analysis. The child-directed speech study focuses on adults who have an intimate relationship with their children, play a mentoring role in their life, and often believe they know what is good for their children; so it covers many of the cases where I predict *you need to* will be most appropriate and thus most common. The MiCASE study examines academic advisors at a university, allowing me to examine speakers who have an official concern with their hearer's well-being and some expert knowledge about what is good for them, but who do not have any sort of intimate relationship with them. The Stanford Politeness Corpus study – in which I compare highly-rated users of a question-answering site to low-rated users – focuses exclusively on the speaker's expertise about the subject matter. Finally, the *Office* study – examining the power-hungry, obnoxious character Dwight in comparison to others – illuminates what happens when *you need to* misfires from the addressee's perspective.

In all these studies, I focus not just on the absolute rate of *you need to*, but on the ratio of *you need to* to its competitors. Thus I do not measure the total *number* of strong-modal obligations issued by certain types of speakers/writers, but the rate at which these obligations are issued using *you need to*. This methodology allows us to ignore different people's propensity to issue second-person strong-modal assertions and focus on effect created by different modals in these expressions. Moreover, since my prediction focuses on *you need to* in declarative rather than interrogative form, I exclude questions from all of my corpus searches.

In each of these studies, I find the number of times *you need to, you have to* and *you've got to* are uttered in each corpus or sub-corpus. I did a traditional χ^2 test to see if the rate differs in different corpora or sub-corpora as predicted by the different social dynamics of each corpus, or if the rate is the same in each one. All four corpus studies yield statistically significant results this way. But the null hypothesis of this test is that the ratio of *you need to* to *you have/got to* is the same between the two corpora. This assumption may be dangerous when comparing two different (sub-)corpora. Perhaps in one (sub-)corpus, people simply talk about external obligations a lot, whereas in the comparison corpus, they talk about people's priorities. To focus on this social dimension of *need to*, I need a way of making sure that the speaker's use of *you need to* is not just an artifact of her greater use of *need* in general.

Therefore, I performed a random sampling test on the larger (sub)corpus in each pair (suggested to me by Simon Todd), a test that makes no assumptions about how the data are distributed. I began by identifying, in each corpus, all of the trigrams beginning with *you*; all of the trigrams ending with *need to*; and all of the trigrams ending with *have/got to*. Then I found the overlap of *you* and *need* trigrams, and of the *you* and *have/got* trigrams in the smaller of the corpora (which is just the same as the counts for *you need to* and *you have/got to* in that corpus, except that questions are not excluded). I randomly sampled the larger corpus 1000 times, always choosing the same number of *you, need* and *have/got* trigrams as were in the smaller corpus. Each time, I calculated the overlap between the *you* and *need* trigrams, and the *you* and *have/got* trigrams. I used these 1000 values to calculate 95% confidence intervals. Then I assessed whether the values found in the smaller corpus fall inside or outside this interval. I used the same method to calculate the *ratio* of *you need to* to *you have/got to* in the smaller corpus and in random samples of the larger corpus. Essentially these tests tells us: if the bigger corpus used *you, need to* and *have/got to* the same number of times as the smaller corpus, would the rate of *you need to* still be different across the two corpora? All four corpus studies yielded significant results using this methodology as well.

In these corpus studies, I do not distinguish between personal and generic uses of *you* (*Do you want to get coffee?* vs. *In general, you should get at least 30 minutes of exercise a day*). I recognize that this distinction matters for the social meaning of *you need to*: the personal use is probably more face-threatening since it attempts to shape the addressee's behavior rather than stating a general rule. However, since so many uses of *you* are difficult to code, I do not attempt to distinguish personal vs. generic *you* in my analysis. I believe these two types of *you* should be evenly distributed across the different types of speakers I compare in my corpus study, so should not distort the results.

4.1 Child-directed speech

Based on my analysis, I predict that parents may use *you need to* quite frequently, because they may think they know better than their children what's good for them, and they occupy a nurturing role in their children's lives. Generally speaking, parents have at least some power over their children. They usually perceive children as intimates, thus low in social distance from them. They may also try to persuade the child to obey by invoking his or her needs instead of just their own wishes. Thus, I predicted that this child-directed speech would use a higher ratio of *you need to* to *you have/got to* than adult-directed speech. I tested this hypothesis by comparing adults in the Providence section of the CHILDES corpus to speakers in the spoken section of CoCA (Davies 2008).

The CHILDES (Child Language Data Exchange System) corpus is a multilingual, multimedia database of phonetically and orthographically transcribed parent-child interactions, used mainly to study language acquisition. The Providence Corpus (Demuth *et al* 2006) is a subset of the CHILDES database containing 360 hours of transcripts and audio/video recordings of "spontaneous speech interactions" between American mothers and toddlers from 2000-2004.

In the Providence data, parents chat and play with their children to teach them language, motor skills and world knowledge. Since the parents see it as their responsibility to cultivate their children's minds, they seem to play a mentoring role in the children's lives, such that they are certainly licensed to tell their children what is good for them. They also know more than their children, and so are in a position to give advice and instruction.

The Corpus of Contemporary American English (Davies 2008) is a 450-million word collection

of American English from a variety of genres from 1990 to 2012. For this study, I focused on the Spoken genre of the corpus, 85 million words from nearly 150 television and radio programs. The Providence section of CHILDES was recorded in the early 2000s, right in the middle of the spoken CoCA timespan, 1990-2012, controlling for diachronic changes in the use of *need to*.

In the Spoken CoCA data, radio and television reporters may see it as their responsibility to provide the audience with knowledge, but this knowledge is frequently about third parties such as political figures. Although the media certainly tries to make itself relevant to the audience's interests, it plays more of an informational role in the audience's life than a mentoring one.

To test whether speakers use *you need to* differently in these two corpora, I counted the number of hits for *you* + modals for adults in Providence, and for everyone in Spoken CoCA, excluding questions. As predicted, caregivers uses *you need to* significantly more ($p < 0.001$ in χ^2):

	Total words	*you have/got to*	per mill	*you need to*	per mill
Caregivers	1,696,469	179	106	219	129
Spoken CoCA	85,000,000	24,885	293	4,460	53

Using the random sampling method described above, I identified a 95% confidence interval for how often *you need to* and *you have/got to* would appear in Spoken CoCA if *you, need* and *have/got* were used an equal number of times in each corpus. The results remain statistically significant:

	you have/got to	you need to	need : have/got
Caregivers - actual count	251	168	1.49
Spoken CoCA - 95% conf.	89 to 126	119 to 160	0.60 to 0.95

4.2 MiCASE

In my next corpus study, I compared the ratio of *you need to* to *you have/got to* in academic advising sessions to peer study groups in the Michigan Corpus of Academic Spoken English (MiCASE, Römer 2002). Academic advisors are charged with ensuring that students find a suitable major, finish their coursework, and prepare to enter the workforce. Thus, they are professionally concerned with their advisees' interests and supposed to be knowledgeable about this domain. They also take time to ask the student about his/her priorities, so the advisors are qualified to speak to these priorities. Their role at the university is to help students, so it is not presumptuous of them to do so. I therefore predicted that academic advisors should use a high rate of *you need to* compared to *you have/got to*.

In contrast, students in a study group are more egalitarian. Study-mates generally do not occupy a nurturing role in one another's lives. While some students may know more than others about the material, they might try not to emphasize this fact in order to maintain a collaborative atmosphere. Certainly no student is institutionally in charge of instructing the others in the way that an advisor is. They tend to ask each other questions about the material, but no student is in charge, and the conversation often focuses on fulfilling the course's requirements rather than furthering any particular student's goals. According to my analysis, study-mates should therefore use a lower rate of *you need to* compared to *you have/got to*, in comparison to academic advisors.

The Michigan Corpus of Academic Spoken English is a collection of almost two million words of transcribed speech from various interactions around the University of Michigan in the late 1990s and early 2000s. I created a subset of the corpus containing only academic advising sessions and the study group sessions that did not appear to be led by a teaching assistant. The table below summarizes the number of occurrences of *you* plus the modals in each corpus (excluding questions). Advisors use *you need to* much more frequently than study-mates ($p < 0.001$ in χ^2):

	Total words	*you have/got to*	per mill	*you need to*	per mill
Advisors	25,033	19	759	27	1,079
Studymates	86,520	38	439	5	58

To control for differences between the corpora, I used the random sampling technique described above to identify 95% confidence intervals for the raw counts of *you need to*, *you have/got to*, and their ratio. Since this technique does not remove questions, the numbers differ slightly from above. The results remain significant.

	you have/got to	you need to	need : have/got
Advisors - actual count	18	26	1.44
Studymates - 95% conf.	14 to 21	19 to 25	0.22 to 0.73

4.3 Stack Exchange experts and amateurs

In my next corpus study, I used the Stack Exchange section of the Stanford Politeness Corpus (Danescu-Niculescu-Mizil *et al* 2013) to compare expert users to less knowledgeable users. Stack Exchange is an online forum where users can ask and answer topics on a variety of questions (programming, cooking, chess, etc.) When someone posts a question, other uses can up-vote the question if they think it is a good question, or they can answer it. Answers can also be up-voted if they are helpful, or down-vote if they are confusing or incorrect. The corpus records every user's net number of up-votes. One can earn up-votes by asking *or* answering questions; but good answers are almost always given more up-votes than good questions, so I believe the users with the most up-votes have mainly earned them from their answers. Thus, a highly-upvoted user is knowledgeable about the subject matter, writes clearly enough to be perceived as helpful, and has to avoid angering people to be rewarded with up-votes. A low-upvoted user may also be knowledgeable about the topic, but if s/he is, s/he does not dedicate as much time to writing answers. Or s/he may be on the forum to seek advice as a beginner, and may not have the expertise to post helpful answers.

I use up-votes as a proxy for a user's expertise on the topic and reputation within the community. Since highly up-voted users are experts on the topic, they may feel that they know enough to advise others about what is good for them in the relevant domain, so I predicted them to use a higher rate of *you need to* compared to the amateur users. However, unlike the previous two studies, they do not have a lasting relationship with the people whose questions they are answering, so they do not occupy a mentoring role in those people's lives to the extent that parents and academic advisors do. They also do not have much power over the questioners, since users generally do not interact outside of this forum. This corpus therefore allows us to zero in on only one factor: how much a speaker thinks she knows about the topic, which qualifies her to advise others about what is good for them.

The Stanford Politeness Corpus is a large collection of requests made by users of Wikipedia and Stack Exchange. For the Stack Exchange section of the corpus, there is also a file listing every user and his or her net number of up-votes.

I divided the Stack Exchange users into quartiles based on their reputation. The top quartile of users had between 764 and 341,413 up-votes; the bottom quartile had between 1 and 42 up-votes. Then I counted the number of hits for *you need/have/got to* (excluding questions) for users in the top quartile and the bottom quartile, predicting that those in the top quartile would use a higher rate of *you need to*. The table below reports the counts (raw and per million) for users in both of these quartiles. The top-quartile users do indeed use *you need to* at a greater rate ($p < 0.01$ in χ^2):

	Total words	*you have/got to*	per mill	*you need to*	per mill
Top Quartile	5,092,491	672	132	1,731	340
Bottom Quartile	713,248	32	45	38	53

Again, this result might be confounded if the top-quartile users simply use *need to* more often with any grammatical subject. I controlled for this by using the random sampling method to find 95% confidence intervals for the raw counts of *you need to*, *you have/got to*, and their ratio. This technique does not exclude questions, so the numbers differ slightly from above:

	you have/got to	you need to	need : have/got
Bottom - actual count	50	52	1.04
Top - 95% conf.	41 to 67	127 to 172	2.09 to 3.80

The top-quartile users do use *you need to* significantly more often – and at a greater ratio to *you have/got to* – than bottom-quartile users. Again, the results remain significant when we control for the baseline rates of *you, need,* and *have/got to* across the two corpora.

4.4 *The Office*

So far, all of these corpus studies have predicted *you need to* to be most common where it is most appropriate. In my final corpus study, I focus on a speaker who uses *you need to* where it is *not* appropriate, because he misjudges whether he can tell others what's good for them.

This speaker is a character from the American television show *The Office*, a hyper-realistic "mockumentary" about everyday life at an uninspiring office. This character, Dwight Schrute, desperately wants to be in charge. He cares deeply about the office, even watering the boss's plants without his knowledge. The actor who plays him describes Dwight on Dwight's Wikipedia page as "someone who does not hate the system, but has a deep and abiding love for it." I expected Dwight to use a greater rate of *you need to* because he wants to shape everyone's behavior to conform to his cherished ideal for the office. Dwight is also socially insensitive, so I suspected that he would be less sensitive to the interactional risk of using *you need to*.

I created a corpus of all ten seasons of *The Office* (including deleted scenes). Then I counted all hits for *you need/have/got to* uttered by Dwight, and uttered by everyone else on the show. Consistent with my hypothesis, Dwight uses *you need to* significantly more ($p < 0.01$ in χ^2):

	Total words	*you have/got to*	per mill	*you need to*	per mill
Dwight	99,893	32	320	26	260
All Others	653,162	267	409	87	133

I controlled for the baseline rate of *need* across the corpora by using the random-sampling technique. Again, this technique does not exclude questions, so the numbers differ slightly from above:

	you have/got to	you need to	need : have/got
Dwight - actual count	37	22	0.84
Others - 95% conf.	7 to 20	20 to 36	0.24 to 0.81

The results remain significant.

5 Conclusion

I have argued that *you need to* has a slightly different semantics than *you have/got to*, and that this subtlety gives it a unique social meaning that depends on whether the speaker is licensed to tell the hearer what's good for him. I have shown four corpus studies that are consistent with predictions stemming from this analysis. More broadly, I hope this analysis has further shown (following the work cited here) that speakers use function words to communicate not just propositional content, but also their attitudes about themselves and each other; and that the propositional content can ground the affective content in a systematic way.

Acknowledgements Thanks to Chris Potts, Penny Eckert and Dan Lassiter for guidance. I am also indebted to the reviewers and audience at NWAV 43, and to Simon Todd for advice on statistics.

References

Acton, Eric, and Christopher Potts. 2014. That straight talk: Sarah Palin and the Sociolinguistics of Demonstratives. *Journal of Sociolinguistics* 18: 3–31.

Bonnefon, Jean-Francois, Aidan Feeney, and Gaëlle Villejoubert. 2009. When some is actually all: Scalar inferences in face-threatening contexts. *Cognition* 112: 249–258.

Brown, Penelope and Stephen Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Coates, Jennifer. 1990. Modal Meaning: The Semantic-Pragmatic Interface. *Journal of Semantics* 7: 53–63.

Danescu-Niculescu-Mizil, Cristian, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of the 2013 Annual Conference of the Association for Computational Linguistics*, 250–259.

Davies, M. 2008. The Corpus of Contemporary American English. 450 million words, 1990-present. Brigham Young University.

de Marneffe, Marie-Catherine and Christopher Potts. 2014. Developing linguistic theories using annotated corpora. In *The Handbook of Linguistic Annotation*, ed. N. Ide and J. Pustejovsky. Berlin: Springer.

Demuth, Katherine, Jennifer Culbertson, and Jennifer Alter. 2006. Word-minimality, epenthesis, and coda licensing in the acquisition of English (and corresponding corpus, the Providence Corpus). *Language and Speech* 49: 137–174.

Dines, Elizabeth. 1980. Variation in discourse - "and stuff like that". *Language in Society* 9: 13-31.

Dixon, John A., and Don H. Foster. 1997. Gender and hedging: From sex differences to situated practice. *Journal of Psycholinguistic Research* 26: 87–107.

Eckert, Penelope. 2011. The future of variation studies. Paper presented in *All-star panel on the past, present and future of variation studies New Ways of Analyzing Variation 40*.

Eckert, Penelope. 2012. Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology* 41: 87100.

Grice, Herbert Paul. 1989 (1967). *Logic and Conversation*. In: *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Kratzer, Angelika. 1981. The notional category of modality. In: *Words, Worlds, and Context*, ed. Eikmeyer, H.-J., and Rieser, H, 38–74. de Gruyter.

Krug, Manfred. 2000. *Emerging English Modals: A corpus-based study of grammaticalization*. Walter de Gruyter.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Lakoff, Robin. 1974. Remarks on this and that. In: *Proceedings of the Chicago Linguistic Society 10*.

Lavendera, Beatriz. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7(2): 171–182.

Leech, Geoffrey. 2003. Modality on the move: the English modal auxiliaries 1961–1992. In *Modality in contemporary English*, ed. R. Facchinetti, M. Krug, Palmer, F., 311-360. Berlin: Mouton de Gruyter.

Moore, Emma, and Rob Podesva. 2009. Style, indexicality and the social meaning of tag questions. *Language and Society* 38: 447–485.

Müller, Friederike. 2008. From degrammaticalisation to regrammaticalisation? Current changes in the use of NEED. *Arbeiten aus Anglistik und Amerikanistik*: 7194.

Nokkonen, Soili. 2006. The semantic variation of NEED TO in four recent British English corpora. *International Journal of Corpus Linguistics* 11: 29–71.

Nokkonen, Soili. 2012. Need to and the domain of Business in spoken British English. In *English Corpus Linguistics: Looking back, Moving forward Papers from the 30th International Conference on English Language Research on Computerized Corpora*, ed. S. Hoffman, P. Rayson and G. Leech, 131–147.

Pinker, Steven, Martin A. Nowak, and James J. Lee. 2008. The logic of indirect speech. In *Proceedings of the National Academy of Sciences* 105(3), 833–838.

Potts, Christopher. 2011. On the Negativity of Negation. In *Proceedings of Semantics and Linguistic Theory 20*, ed. N. Li and D. Lutz, 636–659.

Römer, Ute. 2002. Michigan Corpus of Academic Spoken English (MiCASE). 1.8 million words, transcribed speech from University of Michigan Ann Arbor.

Rubinstein, Aynat. 2012. Roots of Modality. Doctoral dissertation, University of Massachusetts, Amherst.

Smith, Nicholas. 2003. Changes in the modals and semi-modals of strong obligation and epistemic necessity in recent British English. In *Modality in Contemporary English*, ed. R. Facchinetti, M. Krug and F. Palmer, 241–266. Berlin: Mouton de Gruyter.

Stubbe, Maria, and Janet Holmes. 1995. You know, eh, and other exasperating expressions: An analysis of social and stylistic variation in the use of pragmatic devises in a sample of New Zealand English. *Language and Communication* 15: 63–88.

Tagliamonte, Sali, and Alexandra D'Arcy. 2007. The modals of obligation/necessity in Canadian perspective. *English World-Wide* 28: 47–87.

Torres Cacoullos, Rena. 2001. From lexical to grammatical to social meaning. *Language in Society* 30: 443–478.

Yagoda, Ben. 2006. You need to read this: How need to vanquished have to, must, and should. *Slate Magazine*.

Department of Linguistics
Margaret Jacks Hall
Stanford University
Stanford, CA 94305
lelia@stanford.edu