Strong Necessity Modals: Four Socio-pragmatic Corpus Studies

Lelia Glass*

1 Introduction

The sociolinguistic variable has traditionally been defined as "two ways of saying the same thing" (Labov 1972:271), such as contrasting pronunciations of a word. To adapt the concept to semantic or pragmatic features, researchers have defined the sociolinguistic variable 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 slightly in meaning, 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). While function words and phrases may seem utilitarian, they are surprisingly rich in social meaning (Lakoff 1974, Dixon and Foster 1997, Moore and Podesva 2009, Potts 2011, a.o.). Moreover, it is sometimes possible to derive a function word's unique social meaning from the way its semantics differs subtly from that of its competitors, creating a shared project for semantics and sociolinguistics.

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" progressives index rural or popular urban identity because this form is more associated with physical, outdoor activity, whereas the *estar* progressive 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. More recently, Acton and Potts (2014) consider demonstratives (*this, that, these, those*) as compared to other determiners. They argue that *that* (as in *<u>that</u> 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.

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: they might be asymmetrically entailing, they might have slightly different presuppositions, one might be more ambiguous than the other, and so on. Each of these semantic relationships opens up a different dimension along which semantically distinct variants can compete, and thus a different way that the choice of that variant can convey social or interpersonal meaning. For example, it is well known that speakers may choose a weaker or more ambiguous form for social reasons, even though this choice violates Grice's conversational maxim of Manner ("avoid ambiguity"). 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.

*Thanks to Chris Potts, Penny Eckert and Dan Lassiter for very helpful comments. I am also indebted to the reviewers and audience at NWAV 43, and to Simon Todd for valuable advice on statistics.

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. For example, (1a) might simply mean that, given that people generally admire perseverance (what I'll call the "obligation source"), the hearer ought to admire some person. In contrast, (1b) seems to suggest that it is in the hearer's interest to admire her.

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

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 the relevant domain. 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 without relevant expertise. 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.

As with the Spanish progressives and the determiners, the current study also uses corpus evidence to derive a function word'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 semantics.

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).

Within a given modal flavor, such as *deontic* (relating to obligations), the exact nature of the 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.

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), unlike other deontic modals. 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). 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 (Nokkonen 2012) and teachers to teenagers (Nokkonen 2006). She concludes that "*need to* is useful when addressing subordinates politely" (Nokkonen 2012) and suggests 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 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" and "hierarchical." 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, but 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. 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, but *need to* seems to require an obligation source that is related to some party's "internal compulsion," priorities, or interests (Rubinstein 2012).

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 to 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 will 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 (1a–1b) and previous literature.

In view of this semantic difference, a sentence like *You have to finish your paper* admits more interpretations than the same sentence with *need*: 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, *You need to finish your paper* 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.

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/got to* are more appropriate. By using *have/got to*, the speaker does not claim to know what is good for the hearer; 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 move, 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 have 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 licensed to give advice, an authority figure, or a person with relevant expertise. 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 without relevant expertise, 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*. It may be perceived as "polite" because it frames the obligation in terms of the hearer's interests rather than in terms of rules imposed from above. These different faces of *you need to* arise out of a unified socio-pragmatic story.

4 Testing the Predictions

In order to test interactional predictions like these, we need specific corpora that provide information about how speakers and hearers relate to each other. I have tried to find four such corpora to test different aspects of the theory outlined above: the Providence section of CHILDES (as compared to the Spoken section of the Corpus of Contemporary American English), the Michigan Corpus of Academic Spoken English (MiCASE), a corpus of posts and comments from an online discussion site called Stack Exchange, and the text of all seasons of the American television show *The Office*.

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 the effect created by different modals in these expressions.

These corpora were chosen because they contain information about how speakers relate to hearers. Thus, I do not make predictions about the rate of *need to* with grammatical subjects other than *you*. Perhaps the speakers who know what's best for the hearer also think they know what's best for miscellaneous third parties, making their rate of *need to* higher all around; or perhaps both the high *you need to* speakers and the low *you need to* speakers feel equally licensed to comment on what is good for third parties, making their rate of *need to* differ only with regard to *you*. I have no reason to prefer either of these hypotheses. However, because the analysis only predicts a difference with *you*, I wanted to be sure that the different rates of *you need to* are not simply an epiphenomenon arising from different rates of *need to* with other subjects. Therefore, a statistical technique was designed to control for the baseline rates of *you*, *have/got to* and *need to* across the two corpora. This technique tests whether the pairs of corpora differ particularly in their rates of *you need to* use, as predicted by this analysis; or in their use of *need to* overall, which is not predicted.

I performed a random sampling test on the larger (sub)corpus in each pair, 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 sampled the larger corpus's *you, need* and *have/got* trigrams 1,000 times, each time choosing the number of those trigrams in the smaller corpus. For example, imaging that the smaller corpus had 3,000 *you* trigrams and the larger had 6,000, I would choose a random 3,000 *you* trigrams from the larger corpus's 6,000 *you* trigrams. If the smaller corpus had 1,500 *need* trigrams and the larger had 3,500, I chose a random 1,500 of the larger corpus's 3,500. Each time I performed this sampling, I calculated the overlap between the *you* and *need* trigrams, and the *you* and *have/got* trigrams. Imagine that 455 of the sampled *you* trigrams were *you need to*, then 461 the next time I sampled, then 490, and so on, 1,000 times.

I used these 1,000 values to calculate 95% confidence intervals. Then I assessed whether the values found in the smaller corpus fall inside or outside this interval. Essentially these tests tell 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? Would the ratio of *you need to* to *you have/got to* be the same?

As a side note, 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 save money*). 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 assume that they should be evenly distributed across the different types of speakers I compare in my corpus studies, 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 to their children, because they may think they know what's good for their children, and they play a nurturing role in their children's lives. They may also try to persuade the child to obey by invoking his/her needs instead of their own wishes. Thus, I predicted that 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 multimedia database of 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 transcribed interactions between mothers and children in New England in the early 2000s.

In the CHILDES-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 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.

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. 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.

I chose the Providence section of CHILDES because it was recorded in the early 2000s, right in the middle of the CoCA timespan, 1990-2012. Since *you need to* has increased diachronically, I wanted to compare two corpora from roughly the same time.

To test whether speakers use *you need to* differently in these two corpora, I counted the number of hits for *you need to, you have to* and *you got to* for adults in CHILDES-Providence, and for everyone in Spoken CoCA, excluding questions. I found the raw counts for each modal with *you* in the two corpora, excluding questions, as well as the scaled count per million words (pmw). I also calculated the percent of the time that modal is used over its alternative. For example, for *you need to*, the percentage represents *you need to / (you need to + you have/got to*), along with 95% confidence intervals. As predicted, caregivers use *you need to* significantly more:

	Total words	*you have/got to* raw • pmw • %	*you need to* raw • pmw • %	95% conf.
Caregivers	1,696,469	179 • 106 • 44.97%	219 • 129 • 55.03%	± 5.11%
Spoken CoCA	85,000,000	24,885 • 293 • 84.80%	4,460 • 53 • 15.20%	± 0.17%

Table 1: Raw and per-million counts for *you*+modal for CHILDES parents vs. CoCA.

Using the random sampling method described above, I sampled Spoken CoCA 1,000 times, each time choosing the same number of *you, need* and *have/got* trigrams as are found in the caregiver corpus (since this technique does not exclude question, the numbers are slightly larger than above). These 1,000 samples give us 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	186	254	1.37
Spoken CoCA - 95% conf.	87 to 125	115 to 160	0.61 to 0.98

Table 2: Random sampling results comparing caregivers to Spoken CoCA.

4.2 MiCASE

In the 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 major, finish their coursework, and prepare for the workforce. Thus, they are professionally concerned with their advisees' interests and supposed to be knowledgeable about this domain. Advisors also take time to ask the student about his/her priorities, so they are qualified to speak to these priorities. I therefore predicted that 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 advisors.

The Michigan Corpus of Academic Spoken English is a freely available collection of around 200 hours (1,848,364 words) of transcribed speech from various interactions around the University of Michigan in the late 1990s and early 2000s. The corpus can be searched by the type of interaction, including classes, colloquia, tours of the university, office hours, and so on.

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 TA. I calculated the number of occurrences (raw and per million) of *you* plus the modals in each corpus (excluding questions), as well as the percentages. We see that the advisors use *you need to* much more frequently than study-mates.

	Total words	*you have/got to* raw • pmw • %	*you need to* raw • pmw • %	95% conf.
Advisors	25,033	19 • 759 • 41.30%	27 • 1,079 • 58.70%	± 15.61%
Studymates	86,520	38 • 439 • 88.37%	5 • 58 • 11.63%	± 10.94%

Table 3: Raw and per-million counts for *you*+modal for advisors vs. study-mates.

To control for differences between the corpora, I used the random sampling technique described above to created 1,000 random samples of the larger "studymate" corpus in which *you, need* and *have/got* are used the same number of times as in the smaller "advisor" corpus. These 1,000 samples give us 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.

	you have/got to	*you need to*	*need : have/got*
Advisors - actual count	20	27	1.35
Studymates - 95% conf.	13 to 21	19 to 25	0.24 to 0.78

Table 4: Random sampling results comparing advisors to study-mates.

The advisors used *you need to* more often, and at a higher rate compared to *you have/got to*, than the studymates, even controlling for the frequencies of *you, need* and *have/got* individually.

4.3 Stack Exchange Experts and Amateurs

In my next corpus study, I compared expert users to less knowledgeable users in a database of online Q&A forums. The database, Stack Exchange, is a set of online forums where users can ask and answer questions on a variety of topics (programming, cooking, chess, etc.) When someone posts a question, other users 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-voted if they are confusing or incorrect. The corpus records every user's "reputation" on the site. Reputation, according to the site, is "a rough measurement of how much the community trusts you; it is earned by convincing your peers that you know what you're talking about." A few of the main ways to earn reputation are if one's question is voted up (+5), one's answer is voted up (+10), one's answer is endorsed by another user (+15), one's suggested edit is accepted (+2). One loses reputation when one's question or answer is voted down (-2), when one down-votes an answer (-1), or when one posts spam (-100).

One can earn up-votes by asking *or* answering questions; but good answers earn more reputation points than good questions according to the reputation formula, so the users with the most up-votes have mainly earned them from their answers, indicating that they are knowledgeable about the topic. A low-reputation user may also be knowledgeable, 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.

Since high-reputation 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. In contrast to the previous two studies, high-reputation users 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.

I created a corpus of writings in 50 randomly chosen communities of the over 250 on the site. The corpus's 27 million posts and comments are dominated by the site's largest community, Stack Overflow, where users ask and answer questions about computer programming, but 49 other communities such as fitness, parenting, and anime are also included. For all of these users, numbering over 900,000, the corpus records each one's comments and posts along with his/her reputation score.

I conducted a binomial logistic regression on these data, asking whether the log of a user's reputation predicts her log odds of using *need to*, and including "user" as a random effect. I found that as predicted, higher-reputation users tend to use *need to* more often ($\beta = 0.028$, standard error = 0.002461, $p < 0.0001$).

Since the random sampling test requires two discrete groups rather than an ordering on speakers, I also pulled out the top and bottom deciles of users based on reputation. Then I counted the number of hits for *you need/have/got to* (excluding questions) for users in the top 10% and the bottom 10%, predicting that those in the top 10% would use a higher rate of *you need to*. I found the counts, both raw and scaled per million words (pmw), for users in both of these percentiles. As predicted by the logistic regression, top-10% users do indeed use *you need to* at a slightly greater rate. Because the numbers are so large, this difference is statistically significant ($p < 0.001$ in χ^2).

	Total words	*you have/got to* raw • pmw • %	*you need to* raw • pmw • %	95% conf.
Top 10%	295,432,959	650,250 • 220 • 31.76%	139,709 • 473 • 68.24%	± 0.21%
Bottom 10%	51,714,132	11,067 • 214 • 34.18%	21,312 • 412 • 65.82%	± 0.53%

Table 5: Raw and per-million counts for *you*+modal for top-decile vs. bottom-decile users.

Again, this result might be confounded if the top-10% 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 10% - actual count	11,657	21,975	1.89
Top 10% - 95% conf.	11,871 to 12,206	24,622 to 25,072	2.03 to 2.10

Table 6: Random sampling results comparing top-decile users to bottom-decile users.

The top-10% users do use *you need to* significantly more often – and at a greater ratio to *you have/got to* – than bottom-10% 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 employee, Dwight Schrute, desperately wants to be in charge. He insists on being referred to as the "Assistant Regional Manager" rather than his official title, "Assistant to the Regional Manager." He cares deeply about the office, even watering the boss's plants without his knowledge. 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, frequently exasperating other employees, 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) using transcripts from officequotes.net. Then I counted all hits for *you need/have/got to* uttered by Dwight, and by everyone else on the show. I found raw and per-million counts of *you* with each modal (excluding questions) for Dwight and other characters, as well as the percentages. Consistent with my hypothesis, Dwight uses *you need to* significantly more.

	Total words	*you have/got to* raw • pmw • %	*you need to* raw • pmw • %	95% conf.
Dwight	99,893	32 • 320 • 55.17%	26 • 260 • 44.83%	± 13.92%
All Others	653,162	267 • 409 • 75.42%	87 • 133 • 24.58%	± 4.72%

Table 7: Raw and per-million counts for *you*+modal for Dwight vs. other *Office* characters.

It is possible that these results might be warped because Dwight might use *need* more with any grammatical subject. I controlled for this possibility by using the random-sampling technique to find 95% confidence intervals for the raw counts of *you need to*, *you have/got to*, and their ratio. Again, this technique does not exclude questions, so the numbers differ slightly from above. The results remain significant.

	you have/got to	you need to	need : have/got
Dwight - actual count	32	27	0.84
Others - 95% conf.	20 to 36	7 to 20	0.24 to 0.81

Table 8: Random sampling results comparing Dwight to other characters on *The Office.*

5 Conclusion

This paper has 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. The paper has described 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.

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.

Davies, Mark. 2008. 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, Georgetown University.

Grice, Herbert Paul. 1989 (1967). *Studies in the Way of Words*. Cambridge: Harvard University Press.

Kratzer, Angelika. 1981. The notional category of modality. In *Words, Worlds, and Context*, ed. H.-J. Eikmeyer, and H. Rieser, 38–74. Berlin: 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*: 345–356.

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

Leech, Geoffrey. 2003. Modality on the move: the English modal auxiliaries 1961–1992. In *Modality in Contemporary English*, ed. R. Facchinetti, M. Krug, and F. Palmer, 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* 33:71–94.

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* 105:833–838.

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.

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