ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

THE SPECIAL PHONOLOGICAL CHARACTERISTICS OF MONOSYLLABIC FUNCTION WORDS IN ENGLISH

Stefanie Shattuck-Hufnagel[1] *and Nanette Veilleux* [2]

[1]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Mass. USA and
[2]Department of Mathematics and Computer Science, Simmons College, Boston, Mass. USA

ABSTRACT

Monosyllabic Function Words of American English, such as articles (e.g. *the*, *a*), pronouns (*him*, *them*) and conjunctions (*and*, *or*), are notorious for their pronunciation variability in continuous speech. This study explores one potential correlate of this variation: the phonological character of Function Word forms. The Brown Corpus of 1 million words of written text, with each word token labeled for part of speech, provides a quasi-comprehensive listing and categorization of the Function Words of English, making it possible to compare the phonological characteristics of Function Words with those of a substantial sample of Content Words. Results show that Function Words are more likely to begin with a vowel than Content Words are; in addition, when an onset consonant is specified, it is less likely to be a stop consonant for a Function Word than for a Content Word. A set of words whose categorization is uncertain, like *quite, such all, many* etc, show intermediate values on these two phonological dimensions. The differences are consistent with the hypothesis that Function Words are phonologically weaker than Content Words, perhaps contributing to their susceptibility to severe phonetic modification in continuous speech.

1. INTRODUCTION

Monosyllabic Function Words of American English often undergo severe phonetic modification from their citation forms when they occur in continuous speech. These modifications can involve severe reduction, such as loss of the initial consonant, the nucleus vowel or even both, as in e.g. *send'em*, *he'd, Larry'll, up 'n' down*, etc. Such severe reduction processes have not been reported for Content Words like Nouns, Verbs and Adjectives, and both the reduction phenomena and the set of words which undergo them are of interest to a variety of disciplines, including linguistics, cognitive science and speech engineering.

However, the evidence from these different approaches showing that Function Words (FWds) are different from Content Words (CWds) is not always based on the same definition of the set of FWds, or on the same aspects of their behavior. Linguists focus on characteristics such as the special behavior of words like auxiliary verbs in sentence structure (e.g. *do, have* and *is* can appear at the beginning of a question sentence, as in *Do you know what time it is?*, while other verbs cannot, at least in modern English: **Know you what time it is?*), the less-referential quality of grammatical morphemes like *the* and *but* compared with lexical morphemes like *dog* and *crunch*, and the process of cliticization, by which a FWd joins prosodically with a preceding or following word, creating a new sound structure such as *He'd've* from *He would have.*

Cognitive scientists have studied the role of FWds in sentence processing, hypothesizing a special lexicon and lookup process for these words, and emphasizing the closed nature of FWd classes: speakers can productively add new nouns and verbs to the language, but seldom develop a new conjunction. In addition, a distinct role for FWds in the speech production planning process has been proposed, as evidenced by errors such as *He told her* for *She told him*, where number and gender information is misordered but case information is not, suggesting a post-error stage of phonological spellout (Garrett 1980, Shattuck-Hufnagel 1983.)

For automatic speech processing, interest in FWds has focussed on their wide range of pronunciation variation, and its association with their very high rates of occurrence and predictability. Phonologists too have described strong and weak forms for monosyllabic FWds, but from the earliest days of linguistic analysis of these words, it has been noted that pronunciations are in fact graded, i.e. a range of reductions is possible: compare *he had, he'ed, he'd* (Sweet 1883, Selkirk 1972). If FWd modifications are graded, then the question arises of what factors govern the distribution of variants with different degrees of reduction. Factors that have been shown to be correlated with FWd reduction include their specific part of speech, frequency of occurrence, predictability, prosodic and segmental context and adjacent disfluency in the utterance (Jurafsky et. al., to appear).

In this paper we focus on a different factor which may interact with the susceptibility of monosyllabic FWds to phonetic modification: their phonological structure. L. Nakatani noted in a series of lectures in the 1980's (Nakatani, p.c.) that FWds are likely to begin with a vowel (e.g*. I, am, in, are, and, a, it*), and for those that do have an onset consonant, segments such as h *(e.g. he, him, her, here),* w *(we, were, was, what),* y *(you, yet)* and *dh (the, them, that, thus, this, then, they)* are disproportionately represented in the onsets. In this study, we test two hypotheses about the phonological structure of FWds in a comprehensive sample of English text: first, that their structure is different from that of CWds, and second, that the difference corresponds to greater phonological ‘weakness’ for FWds that might be related to their propensity to occur in reduced form in continuous speech. Results reported here are a first step in a larger investigation of the range of phonetic variation found in FWds in spoken utterances, and the factors which govern this variation; later studies will examine differences between monosyllabic and polysyllabic FWds in their modification behavior.

2. METHOD

To compare the phonological shapes of FWds and CWds quantitatively requires a definition of the words which belong in each class, to enable development of a FWd list that is as comprehensive as possible. To reach this goal, we analysed the Brown Corpus of nearly a million words of English text.

2.1 The Text Corpus

The texts. The Brown Corpus of American English (Francis and Kucera, 1979) is a compilation of just over one million words of text, drawn from a variety of printed genres. The corpus was collected to provide as much coverage of American English as possible and has been widely analysed for patterns of word usage and frequency. It is available with Part-of-Speech tags.

Part of speech labels. Each token word of the Brown Corpus has been identified with a Part-of-Speech (PoS) label that reflects its role in the sentence where it occurs. The PoS labels, originally assigned automatically and later hand-corrected over a span of years, consist of 87 categories. Two types are irrelevant for the purposes of this study (6 punctuators, like period, comma and dash, and 4 markers for discourse type, like Foreign Words and Headlines). The remaining 77 categories included 45 base forms and 32 inflectional variants. The 45 base form labels were grouped as described in the next section.

2.2. Defining the Function Word Set

Content Words: The traditional categories of Nouns, Verbs, and Adjectives, as well as Numerals, and some Adverbs. Adverbs included in the CWd class were those formed productively with the suffix *–ly* added to a root morpheme, such as *warmly* and *exuberantly*. Other adverbs, whose status as CWds was less clear, were moved to a third set, Intermediate Words (IWds); these included non-*ly* adverbs such as *also*, and adverbs which happen to end in *–ly* but for which the status of the root morpheme is unclear, at least for speakers of modern English (e.g. *only*).

Function Words: The remainder of the 45 base forms were categorized as FWds, with the following exceptions:

- Words like *according*, which were labeled as Prepositions, were removed from the FWd set and categorized as IWds, since they include a recognizable root morpheme and their status as Prepositions is unclear.
- Four other word types were categorized as IWds because of their uncertain FWd status: Exclamations (e.g. *uh, yeah*), Post-determiners (*many*), Quantifiers (*all*) and Qualifiers (*quite, so*).

Intermediate Words: The label types removed from the CWd and FWd sets, then, included some Adverbs, some Prepositions, Exclamations, Post-determiners, Quantifiers and Qualifiers (Table 1). Results will be reported separately.

Table 1: Word classification based on Brown Corpus PoS Tags

Part of Speech Tag	**CW**	**FW**	**IW**
• nouns, • verbs, • adjectives, • numerals	X		
• adverbs, • qualifiers	CWd+ly		all others
• conjunctions, • be • do • have verbs • determiners, • existential there • modal aux, • pronouns, • particles, • inf marker		X	
• prepositions		all others	CWd+ing
• quantifiers, • post-determiners, • interjections			X

Removal of foreign, rare and archaic words. Words which occurred only once in the text sample were also removed, because many of them were archaic forms like *hast*, which modern speakers and listeners are unlikely to encounter. Foreign words including Latin phrases (*pro forma*) were removed as noted above, since this study focuses on the phonological characteristics of English words.

The result of these manipulations was an adjusted lexicon of 23,237 words, and an adjusted text sample of 981,220 words. Of entries in the lexicon, 22,587 or 97% are CWds, 241 or 1% are FWds, and 409 or 1.7% are IWds. Of word tokens in the text sample, 463,963 or 47% are CWds, 468,036 or 48% are FWds, and 49,221 or 5% are IWds. Thus, the ambiguous-as-to-category IWds make up only a small proportion of word tokens, and the 241 FWds account for almost half of the words in the text.

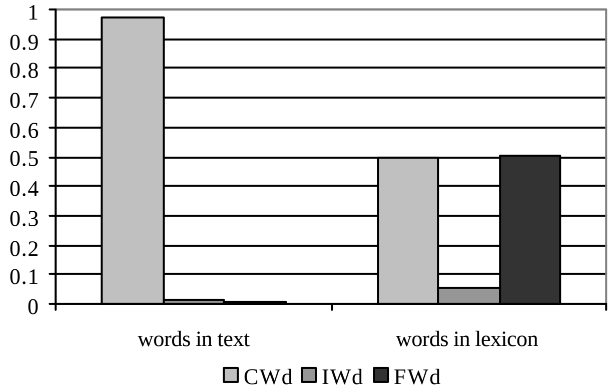

Figure 1: Frequency of CWds and FWds in the Brown Corpus text and lexicon.

2.3. Types of Analyses

Two different aspects of phonological specification are of interest for comparing CWds vs. FWds: syllable structure, and the distribution of phonemic segments. We tested Nakatani's suggestion that FWds are more likely to begin with a VC syllable, and that FWds have a different distribution of onset consonants. Specifically we asked whether CWds are more likely than FWds to begin with a stop consonant /bdgptk/. Both of these asymmetries would be consistent with the hypothesis that FWds are phonologically weaker than CWds, in the sense that a) vowel-onset syllables can be viewed as weaker than canonical CV or CVC syllables, and b) non-stop-consonant onsets are weaker than stop-consonants. Stops represent the most extreme articulation (full oral and nasal closure with cessation of air flow), marked by abrupt acoustic changes in the signal at both closure and release and low or zero amplitude between these two gestures (Stevens 1998). Stops also correspond to the extreme non-vocalic end of the sonority hierarchy (Clements 1990). Thus, stop consonants can be viewed as the strongest class of phonemic segments, perhaps less likely to undergo phonetic modification in context, or at least more likely to retain their consonantal character despite reduction or weakening. We tested these hypotheses with two types of comparison: distribution in the word forms in the lexicon, and distribution in the words in the text corpus. The distinction is an important one, since the number of FWds is small in comparison to the number of CWds, but the frequency of occurrence of many FWds is so high that this set makes up over half of the words we hear and speak every day (Veilleux and Shattuck-Hufnagel 1998).

3. RESULTS

Quantitative evaluation of Nakatani's informal analyses of phonological differences between CWds and FWds showed that these two classes of word differ in the two dimensions examined in this study: the likelihood of an initial vowel, and the likelihood that an initial consonant, if specified, will be a stop. Results for IWds showed intermediate values between those for CWds and FWds on these two dimensions.

3.1. VOWEL-INITIAL SYLLABLE STRUCTURE IN FUNCTION WORDS VS CONTENT WORDS

In the corpus examined here, FWds are more likely than CWds to begin with a vowel both in the lexicon and in the text sample. For example, within the lexicon, 35% of the FWds (as defined in this study) begin with a vowel vs. 19% of the CWds. Of the word tokens in the text corpus, 41% of the FWds are vowel-initial, compared with 17% of the CWds. Thus, FWds begin with vowels about twice as often as CWds (Figures 2 and 3).

3.2. STOPS AS ONSET CONSONANTS IN FUNCTION WORDS VS CONTENT WORDS

In the corpus examined here, FWds are less likely than CWds to begin with a stop consonant. For example, 16% of the FWds in the lexicon begin with a stop, while 34% of CWds are stop-initial. Similarly, 13% of the FWds in the text begin with a stop, compared with 32% of the CWds. Thus for both analyses, the proportion of words beginning with the strongest type of consonant, a stop, is more than twice as high for CWds as for FWds (Figures 2 and 3).

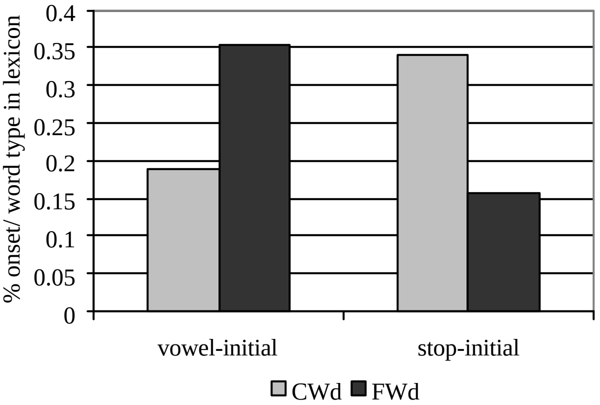

Figure 2: The left pair of bars shows the proportion of CWds and of FWds in the lexicon which begin with a vowel. The right pair of bars shows the proportion of CWds and of FWds that begin with a stop consonant. FWd tokens in the lexicon are more likely to begin with a vowel and less likely to begin with a stop consonant than CWds.

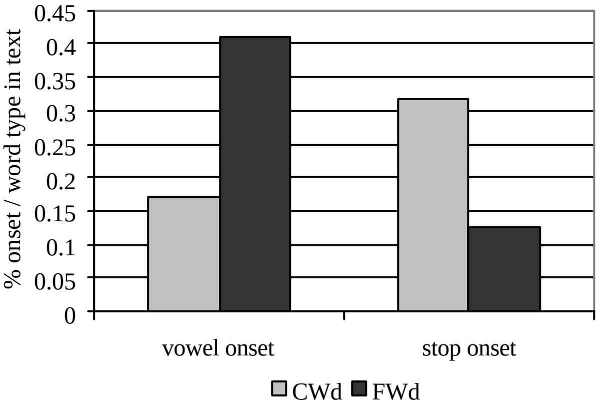

Figure 3: The distribution of vowel-initial and stop- consonant-initial structures among CWds vs. FWds in the text corpus; results are similar to those for the distribution in the lexicon, shown in Figure 2.

3.3 INTERMEDIATE WORDS

In general, values for the IWds were intermediate between FWd and CWd values: 32% are vowel-initial in the lexicon (between 35% for FWds and 19% for CWds), and 36% are vowel-initial in the corpus (between 41% and 17%). Similarly, 21% of IWds are stop-initial in the lexicon (between 16% and 34%), and 10% in the text (lower than both 13% and 32%). These results suggest that many of the IWds resemble FWds on these phonological dimensions.

4. DISCUSSION

The special susceptibility of monosyllabic FWds to phonetic variation in American English has often been noted, but few studies have specifically addressed the question of whether these phonetic modification patterns are limited to all and only the monosyllabic FWds. To answer this question requires a number of steps: developing a list of the FWds, carrying out an acoustic survey of their pronunciation variants (preferably in spontaneous speech but at least in continuous speech), and conducting a systematic investigation of the factors that have been hypothesized to govern this variation. As initial steps toward addressing these issues, we have identified the FWd tokens in a PoS-labelled sample of English texts, and determined that, in this quasi-comprehensive word sample, a FWd is more likely than a CWd to begin with a vowel. Moreover, if there is an onset consonant, it is less likely to be a stop for FWds than for CWds. Similar results were found for the distribution in the lexicon and in the text. The small proportion of words considered to be ambiguous between FWd and CWd classification show intermediate values for these two comparisons; since they make up only 5% of the word tokens in the corpus, they do not affect the conclusions that can be drawn from these results.

These findings are consistent with the hypothesis that FWds as a set are phonologically different from CWds, and that the nature of this difference favors weaker syllable structures and segments for FWds. That is, the vowel-initial syllables that are more common among FWds are not only less canonical than consonant-initial syllables; they may also be more easily restructured into a new constituent with a preceding word, as in *up 'n the tree* or *black 'n' white*. Thus, the more common vowel-onset syllable structure among FWds may be related to their proclivity for restructuring and reduction. Similarly, stop consonants, which are among the strongest consonants on several dimensions, may be more resistant to phonetic weakening than other consonant types. If so, the phonological characteristic of more common stop-consonant onsets among CWds may be related to their reported lesser likelihood of severe phonetic modification.

The potential relation between phonological 'weakness' and susceptibility to modification in context remains to be explored. It would be supported if evidence were found that the phonological characteristics of words in a particular FWd class reflected the types of restructuring and modification that those words are observed to undergo. For example, FWds which are more likely to cliticize leftward might be weak at their left edge, and those likely to cliticize rightward might be weaker at their right edge. Or, FWds which are likely to cliticize might be weaker than those which rarely if ever cliticize. One suggestive piece of evidence along these lines comes from a comparison of subject vs. object pronouns in English. The subject pronouns are more likely to contain a tense vowel (*I*, *you*, *he*, *she*, *we*, *you*, *they* vs. *it*), whereas the object pronouns are slightly more likely to contain a lax vowel (*him*, *her*, *it*, *us*, *them* vs. *me*, *you*, *you*). Since subject pronouns are less likely to cliticize, the predominance of the stronger tense vowels in this FWd set is consistent with the claim that phonological weakness and phonetic modifiability may be related.

5. REFERENCES

1. Clements, G. N., "The role of sonority cycle in core syllabification". In Papers in Laboratory Phonology, ed. Kingston, J. and Beckman, M. Cambridge: Cambridge University Press, 1990.

2. Francis, W. Nelson, and Kucera, Henry, "Frequency Analysis of English Usage: Lexicon and Usage", Boston: Houghton Mifflin Company,1982.

3. Garrett, M.F., The limits of accommodation: Arguments for independent processing levels in sentence production. In V.A. Fromkin (ed.), Errors in Linguistic Performance. NY: Academic Press, 1980

4. Jurafsky, Daniel, Alan Bell, Michelle Gregory, and William D. Raymond, "Probabilistic Relations between Words: Evidence from Reduction in Lexical Production." To appear in Bybee, Joan and Paul Hopper (eds.). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

5. Selkirk, E.O., Phonology and Syntax. Cambridge: MIT Press, 1984

6. Shattuck-Hufnagel, S., Sublexical units and suprasegmental structure in speech production planning. In P.F. MacNeilage (ed.), The production of speech, NY: Springer, 1983

7. Stevens, K., Acoustic Phonetics. Cambridge: MIT Press, 1998.

8. Sweet, H., A Primer of Spoken English. Oxford: The Clarendon Press, 1890

9. Veilleux, N. and Shattuck-Hufnagel, S., "Phonetic modification of the syllable /tu/ in two spontaneous American English dialogues". Proceedings of the International Conference on Spoken Language Processing, Sydney, 1998.

This work is supported by Grant # BCS-98-20126 from the National Science Foundation, USA.