The Buckingham Journal of Language and Linguistics Volume 3 pp 135-162

FUNCTIONAL LOAD: TRANSCRIPTION AND ANALYSIS OF THE 10,000 MOST FREQUENT WORDS IN SPOKEN ENGLISH

Leah Gilner[*] *and Franc Morales*

ABSTRACT

Not all aspects of a language have equal importance for speakers or for learners. From the point of view of language description, functional load is a construct that attempts to establish quantifiable hierarchies of relevance among elements of a linguistic class. This paper makes use of analyses conducted on the 10-million-word spoken subcorpus of the British National Corpus in order to characterize what amounts to approximately 97% of the phonological forms and components heard and produced by fluent speakers in a range of contexts. Our aim is to provide segmental, sequential, and syllabic level rankings of spoken English that can serve as the basis for reference and subsequent work by language educators and researchers.

INTRODUCTION

It has been posited that there are at least two important reasons why pronunciation is not being taught and why learners are left to their own devices when it comes to this crucial component of spoken interaction. First, there is a lack of understanding regarding what aspects of pronunciation have the most value for learners (Breitkreutz et al., 2002; Jenkins, 2000; MacDonald, 2002). And, second, teaching pronunciation is apparently more prone to marginalization than other aspects of language instruction (Fraser, 2002; Setter and Jenkins, 2005). These two observations are interrelated. After all, if the approach adopted by teacher, program curriculum, or material is unsystematic and lacking a rationale concerning sequencing or selection of priorities, it is understandable that teachers and program avoid pronunciation instruction and that little time and resources are dedicated to it. This problematic fact is compounded by the observation made by Derwing and Munro (2005, p. 383) that there is "little published research on pronunciation teaching and very little reliance on the research that does exist".

Regardless of the preparation of teachers and the shortcomings of curricula and materials, the absence of priorities in contemporary education

[*] Bunkyo Gakuin University, Japan

does not imply that these do not exist since it is evident that not all elements of a language have equal bearing in its realization. In some languages, for example, vowels dominate word formation while in others consonants do. Moreover, not all vowels play the same role in word formation, as in English, where four vowels do more work than the remainder of the vowel class together. These observations apply to all elements of a language as well as to its realization. The importance of these observations, for both fluent speakers and learners, has been noted by a number of researchers (Catford, 1987; Kitahara, 2008; Stokes and Surenden, 2005).

Additionally, it has been observed that when speakers use language they do so by exercising selection preferences that give prominence to certain features (George, 1997; Leech et al., 2001; Nation, 2004; Sinclair, 1991). Investigation into the frequency of lexical occurrence in language use reveals that those features that work extensively in, for example, word formation are not necessarily prominent in language use, and vice versa. In English, for instance, the segment /ð/ plays a very small role in word formation, there being but a few words that include this segment. However, inspection of language in use shows that this sound is one of the most frequently heard and produced. The study reported in this paper takes into consideration these two modes of quantifying language (i.e., with and without accounting for frequency of occurrence), regarding them as complementary since each is able to offer information that the other one cannot.

The goal of this study is to contribute to the understanding and assistance of the development of perceptive and productive pronunciation skills. To this end, the 10,000 most frequent words in spoken English (as represented by the British National Corpus) have been identified, transcribed, and analyzed. In this manner, the study focuses on words in isolation rather than connected speech. Segments, clusters, and syllables have been investigated based on their role in word formation as well as their frequency of occurrence in language use. The presentation of results makes extensive use of the construct of functional load (FL) because of its roots in phonetic tradition (see Surendran and Niyogi, 2003 for discussion) and its applicability in pronunciation skill instruction and assessment (Brown, 1991; Catford, 1987; Munro and Derwing, 2006).

FL has been variously defined (Catford, 1987; King, 1967; Hockett, 1955) although within common ground. FL can be formulated as a means of quantifying the relative amount of work elements from a linguistic class do in the language. For instance, if one considers the class of vowels in the context of word formation, a measure of FL reveals that the high-front and reduced vowels are used more often in the lexicon than any other vowels and, thus, do more work. Conversely, FL can be conceptualized as the amount of information lost if elements are eliminated from a linguistic class (Surendran, 2003). Regarding phonemic contrasts, for example, FL reveals that the conflation of the segments /d/ and /z/ would make it impossible to distinguish

(in isolation) a larger amount of words than the conflation of any other two consonants, thereby making this contrast of greater relevance in production and processing.

The usefulness of FL can be appreciated in the findings from two recent studies, Stokes and Surendran (2005) and Munro and Derwing (2006). The first study tested a range of measures in the prediction of the age of emergence of consonants among English-speaking children, finding that the FL measure was the best indicator. The second study investigated the relationship between FL and speech production in ESL adult learners, concluding that "high functional load errors had a greater impact on listeners' perceptions of the accentedness and comprehensibility of L2 speech than did low functional load errors" (Munro and Derwing, 2006, p. 529).

The FL rankings used by Munro and Derwing (2006) came from Brown (1991) which, in turn, are based on raw analyses of language undertaken by Denes (1963). One of the motivations for the present investigation is that Denes' study is one of a kind, thus forcing modern studies (for example, Munro and Derwing, 2006) to use data collected and analyzed some half century ago by a single researcher. A second motivation, and possibly of greater importance, is that inspection of the descriptive study presented in Denes (1963) shows that the size of the language sample used in the analyses was limited to 23,052 tokens (running words) and that the source of the sample was written material from two readers "prepared for teaching English to foreign students" (Denes, 1963, p. 893). The study presented in this paper uses a language sample approximately 400 times larger (9,174,650 running words) and, importantly, the source of the sample is actual spoken language, specifically, spontaneous conversation and task-oriented speech (Leech et al., 2001).

Stokes and Surendran's (2005) raw analyses are of more recent origin (although mostly from written sources) but they are unavailable. This leaves the field without an up-to-date phonetic description of spoken English. Our interest, therefore, centers on the elicitation of a raw description of spoken language that uses a spoken corpus as its sole source and that is based on a sizeable amount of actual language in use.

METHODOLOGY

The data set was comprised of the 10,000 most frequent unlemmatized words from the analyses conducted by Kilgariff (1995) on the 10-million-word spoken subcorpus of the British National Corpus (BNC). The use of unlemmatized forms ensured that the study is faithful to the actual words produced by fluent speakers and, therefore, those words that learners will ultimately be faced with. Note that Kilgariff's word list includes certain words that we have excluded from our data set. Specifically, we have dismissed non-words (i.e. *er*, *mm*, or *ah*), unresolvable contracted forms (i.e. *'s*, *'ll*, or *'ve*),

proper nouns (i.e. *Leicester*, *Banbury*, or *Nottinghamshire*), and lexical phrases (i.e. *a bit*, *of course*, or *as well*). Together, we estimate the dismissed entries reduced the size of the subcorpus from 10,365,623 tokens to 9,399,232 tokens. In this manner, the occurrence of the 10,000 words used in this study amounts to 97.61% (9,174,650 tokens) of the total running words (tokens) in the subcorpus.

Table 1. The vowel system of NAE (shaded areas = +round)

	Front		Central	Back	
High	i		ə ʌ		u
		ɪ		ʊ	
Mid	eɪ				oʊ
		ɛ		ɔ	ɔɪ
Low		æ aɪ		aʊ	ɑ
	Tense	Lax		Lax	Tense

Table 2. The consonant system of NAE

Manner of Articulation		Place of Articulation						
		Bilabial	Labialdental	Interdental	Alveolar	Palatal	Velar	Glottal
Plosive	– voice	p			t		k	
	+ voice	b			d		g	
Fricative	– voice		f	θ	s	ʃ		h
	+ voice		v	ð	z	ʒ		
Affricate	– voice					tʃ		
	+ voice					dʒ		
Nasal	+ voice	m			n		ŋ	
Lateral approximant	+ voice				l			
Approximant	+ voice				r			
Glide	+ voice	w				j		

Transcription procedures followed those described in Gilner and Morales (2008). Each of the 10,000 words was transcribed in broad citation form based on a North American English dialect model (Tables 1 and 2). All transcriptions included syllable boundaries and, if applicable, primary and secondary stress information. The vagaries of syllable boundary identification (Kreidler, 1997; Kreidler, 2004; Ladefoged, 2001) were addressed by consistent application of the Maximum Onset Principle (Anderson, 1982; Pulgram, 1970; Yavaş, 2006), that is, intervocalic consonants were affiliated with syllable-initial positions rather than syllable-final whenever the result was a clustering of consonants in accord with the phonotactic constraints outlined by Kreidler (1997, 2004). Syllables were also consistently transcribed with a vowel nucleus so that syllabic consonants were transcribed as schwa + consonant for the purposes of this study. The transcription procedure was conducted manually and meticulously, each and every word was inspected twice by both authors. Additionally, custom software was developed to facilitate this process and, notably, included a range of background integrity checks aimed at flagging faults and inconsistencies. The size of the task made human error an understandable concern and, therefore, the amount of work invested in securing the accuracy of the transcriptions was substantial.

Once the transcription process was completed, additional custom software was developed to carry out the analyses hereafter presented.

GENERAL CHARACTERISTICS OF THE DATA SET

As mentioned, the 10,000 words (types) in the data set account for 97.61% (9,174,650 tokens) of the total running words (tokens) in the BNC subcorpus. Table 3 shows the amount of types and tokens for unique orthographic and transcribed forms.

Table 3. Number of types and tokens in the data set

	Types	Tokens
Words	10,000	9,174,650
Transcriptions	9,738	9,174,650

The number of transcriptions is smaller than the number of (orthographic) words because of the presence of 504 homophones although, naturally, the number of tokens is equal for both orthographic and transcribed forms. These 504 homophones (226 pairs, 12 triplets, 4 quadruplets) are distributed as follows: 332 monosyllabic words (~65.9% of the 504), 114 disyllabic words

(~22.6%), 48 trisyllabic words (~9.5%), 8 tetrasyllabic words (~1.6%), and 2 pentasyllabic words (~0.4%).

From this point on, results from analyses will be reported using side-by-side tables. The table on the left will reflect the data set as a collection of words without regard to their frequency of occurrence in language use. This will, for example, allow us to determine the amount of work particular segments do in word formation. The quantities reported in table on the right will take into consideration the frequencies with which the words in the data set occur in the language as it is used. This will, for example, allow us to estimate the amount of work particular segments do in language use.

Results are reported by providing raw quantities together with simple descriptive statistics to assist interpretation. We hope that this approach facilitates subsequent application and work by others. Thus, results from analyses are given by providing actual *amounts* as these occur in the data and, for ease of interpretation, the percentage *share* that each element contributes to the whole. Last, we have adopted a measure of FL similar to Catford's (1987), that is, the element with the highest amount is assigned a FL value of 1 while the FL values of other elements are made proportional to this value.

Table 4. Breakdown of data set by number of syllables

Types					Tokens			
Syllable #	Amount	Share	FL		Syllable #	Amount	Share	FL
2	4,100	41.000%	1.00		1	7,281,845	79.369%	1.00
1	2,824	28.240%	0.69		2	1,376,872	15.007%	0.19
3	2,059	20.590%	0.50		3	378,841	4.129%	0.05
4	782	7.820%	0.19		4	111,751	1.218%	0.02
5	209	2.090%	0.05		5	23,581	0.257%	0.00
6	24	0.240%	0.01		6	1,664	0.018%	0.00
7	1	0.010%	0.00		8	49	0.001%	0.00
8	1	0.010%	0.00		7	47	0.001%	0.00
Total	10,000	100.000%			Total	9,174,650	100.000%	

Table 4 presents a breakdown of the words in the data set by number of syllables (ranked according to FL). As just mentioned, analyses provide two views, namely, with and without considering frequency of occurrence. We can already appreciate a difference between the word choices made by fluent speakers *(Types)* and the frequency with which fluent speakers choose to use these words *(Tokens)*. In terms of types, disyllabic words have the highest FL while, in terms of tokens, monosyllabic words have the highest FL.

From the entire lexicon, the 10,000 most frequent words preferred by speakers are largely disyllabics (41.0%), followed by monosyllabics (~28.2%)

and trisyllabics (~20.6%). This distribution contrasts with the use speakers make of these words. Monosyllabics clearly dominate the utterances produced (~79.4%). If frequencies in language use were to be uniform (they are not by any means), each monosyllabic word would be used an average of 2,747 times in the collection of samples that makes up the subcorpus while each disyllabic word would be used an average of 340 times. Equally revealing, trisyllabic words amount to ~20.6% percent of the words in the data set but only to ~4.1% of those occurring in language use.

SEGMENTS

The words that form the data set are made up of 59,793 segments, 21,533 vowels and 38,260 consonants, where the counts refer to occurrence of segments in word formation *(in word types)*. For instance, the consonant /n/ occurs twice in the word *afternoon* and is, therefore, counted twice. When taking into consideration language in use, that is, the frequency of occurrence of the word *afternoon*, the consonant /n/ receives a value of 3,078 (2 x 1,539 where 1,539 is the frequency of the word *afternoon*). In this manner, English segments, as they occur *in word tokens* (i.e. as they occur in the language sample captured by the subcorpus), amount to 29,861,586 instances, 11,747,726 vowels and 18,113,860 consonants.

The FL vowel/consonant ratios are 1:1.78 *in word types* and 1:1.54 *in word tokens*. Collectively, consonants do significantly more work in both word formation and language use. There are, of course, more consonants (n = 24) than vowels (n = 15). If we were to assume that all segments were employed with equal frequency (they are not), each vowel would appear an average of ~1,435 times *in word types* and ~783,182 times *in word tokens* while each consonant would appear an average of ~1,594 times *in word types* and ~754,745 times *in word tokens*.

Naturally, neither individual vowels nor consonants occur with equal frequency in word formation *(in word types)* or in language use *(in word tokens)*. Table 5 provides a summary of results for vowel segments.

In word formation *(in word types)*, the top four vowels account for ~62.7% of all occurrences. In language use *(in word tokens)*, there is a rearrangement of the segments according to FL, particularly noticeable in the values of central vowels. If clustered by vowel type, front vowels amount to half of all occurrences in both types and tokens (~52.9% and ~52.4%, respectively). The two central vowels, however, drop eight percentage points (from ~26.6% *in word types* to ~18.8% *in word tokens*) in favor of back vowels and diphthongs. In other words, back vowels and diphthongs do more work in language use than they do in word formation.

Similar but more uniform trends of distribution can be observed for consonants, possibly due to the larger number of elements in the class. Table

6 shows that the top four consonants account for ~43.7% of all occurrences in word formation *(in word types)*. We also observe a rearrangement of segments

Table 5. Frequency of occurrence of vowels

In word types					In word tokens			
Segment	Amount	Share	FL		Segment	Amount	Share	FL
ə	4,623	21.47%	1.00		ɪ	1,726,282	14.69%	1.00
ɪ	4,523	21.00%	0.98		i	1,624,791	13.83%	0.94
ɛ	2,192	10.18%	0.47		æ	1,152,510	9.81%	0.67
l	2,164	10.05%	0.47		ə	1,122,571	9.56%	0.65
Æ	1,341	6.23%	0.29		ʌ	1,083,276	9.22%	0.63
eɪ	1,177	5.47%	0.25		ɛ	981,996	8.36%	0.57
ʌ	1,100	5.11%	0.24		u	834,943	7.11%	0.48
ɑ	966	4.49%	0.21		aɪ	742,069	6.32%	0.43
aɪ	923	4.29%	0.20		eɪ	670,444	5.71%	0.39
ɔ	741	3.44%	0.16		ɔ	573,513	4.88%	0.33
oʊ	731	3.39%	0.16		oʊ	460,022	3.92%	0.27
U	664	3.08%	0.14		ɑ	383,294	3.26%	0.22
aʊ	202	0.94%	0.04		aʊ	219,047	1.86%	0.13
ɔɪ	101	0.47%	0.02		ʊ	148,300	1.26%	0.09
ʊ	84	0.39%	0.02		ɔɪ	24,668	0.21%	0.01
Total	21,533	100.00%			Total	11,747,726	100.00%	

according to FL in language use, particularly in the case of the voiced interdental fricative. The disparity of values for /ð/ is well known. Very few words (n = 66 or 0.17% of the total) in the language have this segment but these words are extremely frequent in use (n = 1,036,575 or 5.72% of the total).

The data in Table 6 shows that obstruents do more work than sonorants and that voiced consonants do more work than voiceless consonants. This is so regardless of whether we consider their role in word formation or in language use and, in all four cases, FL values coincide at an approximate 3:2 ratio.

From the point of view of place of articulation, alveolars account for ~63.1% of consonants in word formation (*in word types*) and ~56.7% in language use (*in word tokens*). Labials account for ~18.3% and ~19.8%, respectively, while velars/glottals account for ~12.3% and ~10.9%, respectively. Palatals also maintain their presence in both cases and do so at ~5.5%. The significant change takes place in interdentals as already mentioned.

Regarding manner of articulation, plosives account for ~33.7% of consonants in word formation and ~32.5% in language use, fricatives account for ~23.4% and ~26.3% respectively, liquids account for ~20.7% and ~14.9%, nasals for ~17.3% in both cases, affricates for ~2.4% and ~1.5%, and glides for ~2.4% and 7.6%.

SEGMENT CONTRASTS (MINIMAL PAIR ANALYSES)

The relative importance of segments in comprehension and intelligibility is highlighted in those cases where they serve to differentiate words and, in particular, where a single segment is the only phonetic element that

Table 6. Frequency of occurrence of consonants

In word types					In word tokens			
Segment	Amount	Share	FL		Segment	Amount	Share	FL
R	4,931	12.89%	1.00		t	2,371,952	13.09%	1.00
T	4,063	10.62%	0.82		n	2,026,751	11.19%	0.85
N	3,961	10.35%	0.80		r	1,706,548	9.42%	0.72
S	3,771	9.86%	0.76		d	1,259,039	6.95%	0.53
L	3,000	7.84%	0.61		s	1,220,978	6.74%	0.51
K	2,756	7.20%	0.56		ð	1,036,575	5.72%	0.44
D	2,551	6.67%	0.52		l	987,701	5.45%	0.42
Z	1,874	4.90%	0.38		k	881,913	4.87%	0.37
P	1,864	4.87%	0.38		w	849,144	4.69%	0.36
M	1,704	4.45%	0.35		m	752,233	4.15%	0.32
B	1,050	2.74%	0.21		z	691,865	3.82%	0.29
ŋ	970	2.54%	0.20		j	529,299	2.92%	0.22
F	968	2.53%	0.20		b	529,151	2.92%	0.22
V	830	2.17%	0.17		p	516,677	2.85%	0.22
ʃ	803	2.10%	0.16		v	487,260	2.69%	0.21
G	611	1.60%	0.12		f	459,060	2.53%	0.19
W	573	1.50%	0.12		h	424,178	2.34%	0.18
dʒ	537	1.40%	0.11		ŋ	353,916	1.95%	0.15
H	379	0.99%	0.08		g	321,165	1.77%	0.14
ʧ	369	0.96%	0.07		θ	238,255	1.32%	0.10
J	353	0.92%	0.07		ʃ	190,148	1.05%	0.08
Θ	219	0.57%	0.04		dʒ	136,628	0.75%	0.06
Ð	66	0.17%	0.01		ʧ	134,397	0.74%	0.06
ʒ	57	0.15%	0.01		ʒ	9,027	0.05%	0.00
Total	38,260	100.00%			Total	18,113,860	100.00%	

differentiates two words. Minimal pairs (MP) abound in the English language due to the large number of monosyllabic words and their frequent use.

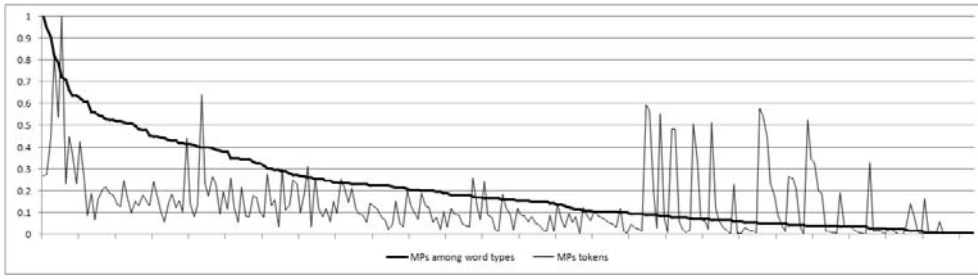
For the purposes of this study, MP analyses took into consideration primary stress but not secondary stress information. A search of the data set yielded a total of 14,418 MPs, 3,688 vowel MPs and 10,730 consonant MPs. Following Brown (1991), the frequency of occurrence of the members of each pair was added in order to compute the weight of each MP and, consequently, each contrast. The totals obtained were 21,927,775 occurrences for vowel MPs and 58,120,215 for consonant MPs. The FL of consonant MPs is superior to that of vowel MPs whether as a collection of types (1 to 0.34) or as they appear in language use (1 to 0.38).

Of the 10,000 words, 4,542 participate in at least one MP relationship. There are 15 words that form 30 or more MPs (the maximum case is 34), 300 words form 20 or more MPs, 1,062 words form 10 or more MPs, and 1,952 words form 5 or more MPs. The majority of MPs are formed by monosyllabic words (~84.1%) even though monosyllabic words account for about half (~54.3%) of the 4,542 words that participate in MPs. Out of the 2,824 monosyllabic words in the data set, ~87.3% form MPs in contrast with ~42.2%(1,731 of 4,100) of the disyllabics, ~13.7% of the trisyllabics, ~7.0% of the tetrasyllabics, ~3.8% of the pentasyllabics, and ~8.3% of the hexasyllabics (see Table 4 for reference). The FL of monosyllabic words in the formation of MPs is much higher than any other type of word.

MP analyses of the data set found 99 vowel and 254 consonant contrasts. However, not all contrasts are of equal importance for comprehension and intelligibility since some pairings are formed by segments that are highly dissimilar (i.e. /v/ and /h/). Deciding which contrasts can definitely be dismissed is not always a straightforward task since the learners' L1 has a bearing on what segments could be problematic. For example, Munro and Derwing (2006) report on the difficulty Chinese L1 learners may have with the contrast /l/ and /n/, a contrast that other English learners do not struggle with. Since we cannot anticipate all possible L1 backgrounds, this paper reports on those contrasts between segments that differ in one distinctive feature, that is, segments that are objectively similar and that are likely to be of relevance to learners regardless of their previous linguistic experience.

Before describing the results for the vowel and consonant contrasts that differ in one distinctive feature, we present in Chart 1 the FL ranking for all consonant contrasts (the smooth curve). The jagged line represents the frequency of occurrence of MPs per contrast. The y-axis represents the FL range and the x-axis represents the 254 consonant contrasts sorted so that the contrasts with more MPs precede those with less MPs (hence the gradual slope). The chart shows what we have been observing all along, namely, that there is a noticeable disparity between language as a static system where all elements have equal weight and language as a collection of utterances where elements are used with unequal frequency.

Chart 1. Consonant MPs among types and tokens



In this particular case (Chart 1), those contrasts that have the greatest number of MPs do not correspond, in general, with those MPs whose constituent words are most frequent. In other words, while some contrasts serve to distinguish a large number of words, these words hardly ever occur. Conversely, some contrasts serve to distinguish but a few words yet these words are very frequent, the high level of activity of these words in language use makes these contrasts important for language users and language learners. This observation, we feel, is one that is important to keep in mind as teachers and materials designers make decisions concerning what should receive the highest priority.

Table 7. Vowel MPs and occurrence in language use

Number of MPs					MP frequencies			
Contrast		Amount	Share	FL	Contrast		Amount	Share	FL
i	eɪ	99	15.87%	1.00	i	eɪ	965,712	24.31%	1.00
ɪ	ɛ	80	12.82%	0.81	u	oʊ	589,902	14.85%	0.61
i	ɪ	77	12.34%	0.78	i	ɪ	518,996	13.07%	0.54
ɛ	æ	71	11.38%	0.72	ɛ	æ	508,453	12.80%	0.53
ɛ	ʌ	70	11.22%	0.71	ɛ	ʌ	414,270	10.43%	0.43
ɛ	eɪ	67	10.74%	0.68	ɪ	ɛ	332,316	8.37%	0.34
u	oʊ	65	10.42%	0.66	ɔ	ʌ	248,838	6.26%	0.26
ɔ	ʌ	54	8.65%	0.55	ɛ	eɪ	190,195	4.79%	0.20
ɔ	oʊ	30	4.81%	0.30	ɔ	oʊ	172,951	4.35%	0.18
ɔ	ʊ	6	0.96%	0.06	u	ʊ	17,524	0.44%	0.02
u	ʊ	5	0.80%	0.05	ɔ	ʊ	12,978	0.33%	0.01
Total		624	100.00%		Total		3,972,135	100.00%	

As mentioned, 99 vowel contrasts were found among the 10,000 words in the data set. Of these, 11 contrasts are between vowels that share all but one distinctive feature and these are shown in Table 7. In the case of, for example, /ɛ/ and /ʌ/ the distinctive feature is [back] while in the case of /i/ and /ɪ/ the

distinctive feature is [tense]. The distinctive feature matrices employed for these analyses are based on O'Grady et al. (1993).

The *share* and *FL* values in Table 7 have been calculated in relation to only those contrasts that appear in the table rather than the total number of vowel contrasts found (the same applies to Tables 8, 9, 10). All distinctive features (except [reduced]) are represented in the data set as are all the contrasts that are distinguished by a single feature. The feature [reduced] is the exception since the segments /ʌ/ and /ə/ cannot make MPs in this transcription system. There are four contrasts distinguished by the feature [high], four more by the feature [tense], while [low], [back], and [round] distinguish one contrast each. The *number of MPs* can, then, be characterized by saying that the feature [high] has the highest FL (it accounts for ~40.0% of all MPs), followed by the feature [tense] (~28.7% of the MPs). In regards to *MP frequencies*, there is an increase in the work done by the feature [high] (~47.9%) while there is a decrease in the amount of work the feature [tense] does (~22.7%). In other words, the preferences exhibited by fluent speakers in language use highlights the necessity for adequate command of the [high] feature to distinguish segments and, consequently, words in production and processing.

Table 8. Consonant MPs and occurrence in language use

Number of MPs					MP frequencies			
Contrast	Amount	Share	FL		Contrast	Amount	Share	FL
r L	189	24.80%	1.00		t d	1,367,847	42.56%	1.00
t d	164	21.52%	0.87		r l	463,540	14.42%	0.34
p b	84	11.02%	0.44		ð d	297,359	9.25%	0.22
p F	77	10.10%	0.41		s ʃ	195,394	6.08%	0.14
k g	46	6.04%	0.24		p b	173,826	5.41%	0.13
s ʃ	45	5.91%	0.24		k g	155,282	4.83%	0.11
s z	39	5.12%	0.21		p f	153,016	4.76%	0.11
t θ	34	4.46%	0.18		v f	129,799	4.04%	0.09
s θ	24	3.15%	0.13		t θ	112,067	3.49%	0.08
v F	21	2.76%	0.11		s θ	88,772	2.76%	0.06
v b	15	1.97%	0.08		s z	38,624	1.20%	0.03
ʧ ʤ	13	1.71%	0.07		v b	33,807	1.05%	0.02
ð d	10	1.31%	0.05		ʧ ʤ	4,459	0.14%	0.00
ð z	1	0.13%	0.01		ð z	299	0.01%	0.00
Total	762	100.00%			Total	3,214,091	100.00%	

Out of the 254 consonant contrasts found in the data set, there are 14 that pair two segments sharing all but one distinctive feature. Table 8 provides FL rankings for these contrasts. As is the case with vowels, the FL values for language use *(MP frequencies)* are more steeply ranked. Note that the contrast /r/ and /ð/ is not included in Tables 8, 9, and 10 because the single feature that distinguishes this pair of segments is the major class feature [sonorant].

The data set does not contain words that contrast /θ/ and /ð/, /ʒ/ and /z/, or /ʒ/ and /ʃ/ although these are distinguished by a single feature, [voice] in the case of the first two and [strident] in the case of the last contrast. The reason is because at least one of the words that create such MPs is not used frequently enough to appear in the data set (less than 2.4 occurrences per million running words). The seven distinctive features [labial], [round], [coronal], [high], [back], [nasal], and [delayed release] cannot create contrasts on their own given the characteristics of the English consonant system.

Table 9. Consonant MPs and occurrence in language use (WI only)

Number of MPs			
Contrast	Amount	Share	FL
p B	73	21.10%	1.00
p F	65	18.79%	0.89
t D	47	13.58%	0.64
r L	42	12.14%	0.58
s ʃ	31	8.96%	0.42
k G	30	8.67%	0.41
t Θ	15	4.34%	0.21
v B	10	2.89%	0.14
v F	9	2.60%	0.12
s Θ	8	2.31%	0.11
ʧ ʤ	8	2.31%	0.11
ð D	7	2.02%	0.10
s Z	1	0.29%	0.01
Total	346	100.00%	

MP frequencies			
Contrast	Amount	Share	FL
t d	424,156	26.70%	1.00
ð d	296,959	18.69%	0.70
s ʃ	191,450	12.05%	0.45
p b	171,420	10.79%	0.40
p f	147,837	9.31%	0.35
k g	131,999	8.31%	0.31
r l	80,225	5.05%	0.19
s θ	76,976	4.85%	0.18
t θ	49,962	3.15%	0.12
v b	7,119	0.45%	0.02
v f	6,989	0.44%	0.02
ʧ ʤ	2,966	0.19%	0.01
s z	462	0.03%	0.00
Total	1,588,520	100.00%	

Of the remaining five features, [voice] distinguishes six contrasts and accounts for ~48.2% of the 762 MPs under consideration while [lateral] distinguishes one contrast and accounts for ~24.8% of the MPs, [continuant] distinguishes four contrasts and accounts for ~17.8% of the MPs, [anterior] distinguishes one contrast and accounts for ~5.9% of the MPs, and [strident] distinguishes two contrasts and accounts for ~3.3% of the MPs. In regards to *MP frequencies*, there is an increase in the work done by the feature [voice]

(~58.2%) while there is a decrease in the amount of work the feature [lateral] does (~14.4%). In other words, adequate command of the feature [voice] is more necessary than other features to distinguish segments and, consequently, words in production and processing as demonstrated by the preferences exhibited by fluent speakers in language use.

The boundaries of words, especially the initial segments, are recognized as playing a determinant role in lexical access (Bent et al., 2007; Dell and Gordon, 2003; Gow et al., 1996; Marslen-Wilson and Zwitserlood, 1989). In the next section we will provide a broad characterization of word boundaries by inspecting word-initial (WI) onsets and word-final (WF) codas. Before moving on, however, it is relevant to ask what role MPs play in relation to the challenges of word identification and comprehensibility. To this end, we have isolated those consonant MPs (and contrasts) from Table 8 that occur in WI and WF positions.

Table 9 provides a FL ranking for those contrasts that are found to have MPs in WI position. Since lexical access is known to rely on WI segments, the proposition is that these contrasts increase the chances of incorrect word identification. In other words, failure to properly articulate or process a word-leading segment may trigger the activation of a MP partner and result in a breakdown of communication.

Comparing the results shown in Tables 8 and 9, one can see that a significant amount of the MPs for the contrasts /p/-/b/, /p/-/f/, /s/-/ʃ/, and /k/-/g/ occur in WI position, ~86.9%, ~84.4%, ~68.8%, and ~65.2%, respectively. MP frequencies for these contrasts are even more striking, ~98.6%, ~96.6%, ~98.0%, and ~85.0%, respectively. It can be safely said that these contrasts exert most of their influence in WI position and are, therefore, of special relevance to word identification and those comprehensibility problems that may result if they are not properly distinguished by learners in production or perception.

Since most of the MPs are formed by monosyllabic words, it is unsurprising that most consonant contrasts are found in either WI (~45.4%) or WF position (~33.9%). Together, 79.3% of all MPs exhibit a contrast at a word boundary. Regarding MP frequencies, these MPs account for ~96.3% (~49.4% and ~46.8%, respectively) of the cumulative total (n = 3,214,091).

Table 10 isolates those consonant contrasts from Table 8 that occur in WF position. In relation to word identification, distinction of words by a single phoneme in WF position implies that two (or more) words are able candidates up to that point. That is to say, the role of that WF segment is of a last chance for correct identification when processing might have already selected a (wrong) candidate due to the higher frequency of one of the words, part of speech and other collocational information, discourse context and expectations, and so on.

Contrasts in WF position (Table 10) are dominated by /t/-/d/ and /r/-/l/ in terms of number of MPs (~60.9%) and even more so in terms of MP

frequencies (~84.6%). These two contrasts also have the greatest FL values when taking into consideration all positions (Table 8). Both these contrasts mostly occur at word boundaries (from Tables 8, 9, and 10; /t/-/d/ ~88.4% of MPs and ~99.4% of MP frequencies; /r/-/l/ ~53.4% and ~90.2%), particularly in WF position (from Tables 8 and 10; /t/-/d/ ~59.8% of MPs and ~68.4% of MP frequencies; /r/-/l/ ~31.2% and ~72.9%), implying that MPs for these contrasts that do not occur in word boundaries are highly infrequent. Since most MPs (~84.1%) are formed by monosyllabic words, this is unsurprising.

Table 10. Consonant MPs and occurrence in language use (WF only)

Number of MPs					MP frequencies			
Contrast	Amount	Share	FL		Contrast	Amount	Share	FL
t d	98	37.98%	1.00		t d	935,727	62.16%	1.00
r L	59	22.87%	0.60		r l	337,781	22.44%	0.36
s z	34	13.18%	0.35		v f	95,030	6.31%	0.10
t θ	16	6.20%	0.16		t θ	61,501	4.09%	0.07
s θ	12	4.65%	0.12		s z	35,831	2.38%	0.04
k g	11	4.26%	0.11		k g	20,594	1.37%	0.02
p F	8	3.10%	0.08		s θ	11,107	0.74%	0.01
v F	7	2.71%	0.07		p f	3,772	0.25%	0.00
s ʃ	5	1.94%	0.05		s ʃ	2,082	0.14%	0.00
p b	4	1.55%	0.04		ʧ ʤ	1,285	0.09%	0.00
ʧ ʤ	3	1.16%	0.03		p b	615	0.04%	0.00
ð d	1	0.39%	0.01		ð d	123	0.01%	0.00
Total	258	100.00%			Total	1,505,448	100.00%	

We conclude the section on MPs by observing that some researchers have gone beyond this level of explanation to suggest that contrasts where MPs are seriously imbalanced in favor of one of the members of the pair are less relevant to the computation of the FL of contrasts (and possibly to learners) than those where the frequency of occurrence of both members of MPs is balanced. Brown (1991) quotes Rischel (1962, p.18-19) as saying: "the functional load of a contrast in the text depends on the existence of minimal pairs of words that are both frequent", so that when one member is relatively infrequent, the "minimal pair can hardly be said to have any importance" (Brown, 1991, p.219). It is easy to see the logic and relevance of such an observation. However, we feel, imbalance of occurrence does not necessarily rule out the importance of a MP since, after all, the infrequent member is obscured both by its own infrequency and by the dominating frequency of a highly similar word. For these reasons, these observations have not been taken into consideration in the analyses reported here. Our study, however, included

additional analyses that explore these observations and an investigation is ongoing.

ONSETS AND CODAS

A presentation of the structure of the English syllable can benefit from first investigating onsets and codas, that is, those consonant segments that precede and follow the vowel nucleus of a syllable. In particular, it is of interest to inspect the manner and frequency with which consonants are sequenced into clusters for this is a well-known source of problems for learners in both production (Celce-Murcia et al., 1996; Gilner and Morales, 2000; Jenkins, 2000; Suenobu, 1992) and comprehension (Altenberg, 2005; Dupoux et al., 1999; Flege, 2003; McAllister et al., 1999; Tarone, 1987).

We begin with a discussion of onsets. Table 11 shows that there are 21,533 onsets in the data set (*in word types*). This number coincides with the total number of syllables in the 10,000 words since empty onsets are counted too. Taking frequency into consideration, the collection of utterances in the subcorpus contains 11,747,726 occurrences of onsets and, therefore, syllables. Note that results are always going to be influenced by the method of syllabification adopted.

Table 11. Breakdown of onsets by length

In word types					In word tokens			
Length	Amount	Share	FL		Length	Amount	Share	FL
1	12,238	56.83%	1.00		1	7,476,015	63.64%	1.00
0	6,989	32.46%	0.57		0	3,719,423	31.66%	0.50
2	2,135	9.92%	0.17		2	529,430	4.51%	0.07
3	171	0.79%	0.01		3	22,858	0.19%	0.00
Total	21,533	100.00%			Total	11,747,726	100.00%	

The onset with the highest FL in both types and tokens is a single consonant, followed by the absence of a consonant (the empty onset). Double-segment (CC) and triple-segment (CCC) clusters in onset position amount to ~10.7% of the types and, significantly, to ~4.7% of the tokens. In short, consonant clusters in onset position are relatively rare and even more so in actual speech. The implication is that, from an instructional point of view, frequent words exhibiting CC or CCC clusters in onset position may be of interest in their own right rather than as exemplars of a phonotactic characteristic that, results show, is not abundant in word formation or in language use. The most frequent of these words are: CC clusters, *from (~2,265 per million running words), three (~1,786), through (~800), still*

(~798), probably (~606), start (~454), school (~448), try (~421), etc; CCC clusters, *straight (~ 169), street (~143), structure (~83)*, etc. Patterns can clearly be seen across the onsets of these words but acquisition of the actual exemplars provides learners with the precise words that they are most likely to encounter in production and perception. Whether or not there is agreement on this point, the strongest argument is that, in any case, these are the very words that should be used to illustrate this particular characteristic of the onset of syllables precisely because of their role in language use.

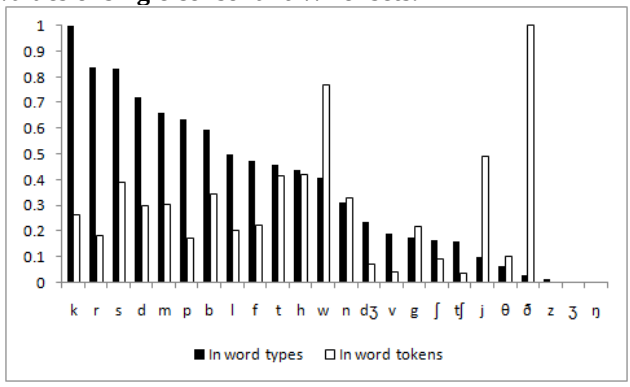
The attentive reader may have noticed that the most frequent words listed above all have the CC or CCC cluster onset in word initial (WI) position. This is unsurprising since most running words (From Table 4, ~79.4%) in the subcorpus are monosyllabic words. Inspection of the 21,533 onsets reveals that although ~46.4% of these onsets fall in WI position, they account for ~78.1% of the 11,747,726 tokens. Table 12 presents results for onsets in WI position.

Table 12. Breakdown of onsets by length (WI only)

In word types					In word tokens			
Length	Amount	Share	FL		Length	Amount	Share	FL
1	6,615	66.15%	1.00		1	6,261,142	68.24%	1.00
0	1,734	17.34%	0.26		0	2,475,842	26.99%	0.40
2	1,544	15.44%	0.23		2	422,328	4.60%	0.07
3	107	1.07%	0.02		3	15,338	0.17%	0.00
Total	10,000	100.00%			Total	9,174,650	100.00%	

The first observation is that empty onsets in types (word formation) drop by approximately half while all other kinds of onsets increase their share (Tables 11 and 12). However, the values for frequency of occurrence generally hold, with an increase for both single consonant and empty onsets. Fewer empty onsets are doing more work. In an empty onset situation, the vowel nucleus leads the word. The segment /ə/ accounts for ~25.3% of the words in the data set that start with a vowel, /ɪ/ accounts for ~21.1%, /ɛ/ accounts for 15.3%, and /æ/ accounts for ~12.8%. Together, these four segments amount to ~74.5% of all words that start with vowels. In terms of tokens, /ə/ and /ɛ/ are relatively infrequent (~6.7% and ~4.2%, respectively) while /ɪ/ and /æ/ are the most frequent of all vowels (~24.8% and 16.8%, respectively). Of interest, /aɪ/ plays a small role in word formation (~2.7%) but ranks third in terms of frequency (~14.2%).

Chart 2. FL values of single consonant WI onsets.



the consonant /k/ has the greatest FL, this value is not significantly greater than those immediately following. Token-wise, the distribution is more pronounced with the consonant /ð/ having the largest FL value.

In word formation (*in word types*), plosives account for ~39.8% of all single consonant WI onsets, while fricatives account for ~24.5%, liquids for ~14.9%, and nasals for ~10.8%. In terms of frequency of occurrence, the reorganization yields fricatives (~35.6%), plosives (~27.8%), and glides (~19.8%). Nasals maintain their share at ~10.0% while liquids drop to ~6.1%. The observations made about the general distributions of consonants (Table 6) apply to the information shown in Chart 2, that is, obstruents do more work than sonorants and voiced consonants do more work than voiceless consonants.

As mentioned, there are relatively few WI onsets made of a CC cluster and these do not occur frequently in language use. Phonotactic constraints limit which consonants can pair and which can precede and follow, that is, there are restrictions in terms of variety. In particular, analyses show that CC clusters starting with the consonants /s/, /p/, and /k/ (listed in order of frequency) amount to ~61.3% of all WI-CC onsets and account for ~54.7% of all occurrences (all segments drop their share from word formation to language use). WI-CC onsets starting with the consonants /f/, /t/, /θ/ are more active in language use than in word formation.

All CCC clusters in WI onset position necessarily start with the consonant /s/, followed by the plosives /t, /k/, /p/ and ending with a liquid or a glide. Of the five WI-CCC onsets found among the 10,000 words, the cluster /str/ amounts to ~55.1% of those found, accounting for ~69.1% of all tokens. This cluster is, therefore, used often in word formation and the words in which it is found occur frequently.

We now move on to codas. As mentioned, the 10,000 words in the data set are formed by a total of 21,533 syllables, a number that naturally coincides with the number of onsets and codas (since, again, empty onsets and codas are counted too). Similarly, the frequency of occurrence of codas is the same as

that of onsets and of syllables. Table 13 presents a breakdown of codas by length. It shows that, as with onsets, single consonant codas have the highest FL values in both types and tokens, followed by the absence of a consonant (the empty coda). Differences between onsets and codas can be observed once we start to look a bit closer.

In word formation (*in word types*), single consonant onsets account for ~56.8% of all onsets while single consonant codas account for ~66.6% of all codas. Also, empty codas (~18.3%) do a lot less work than empty onsets (~32.5%) while as CC onsets (~9.9%) do less work than CC codas (~13.21%). Quadruple (CCCC) codas exist but their presence is reduced to 5 instances. These do not occur in onset position. In terms of language use (*in word tokens*), FL values are quite similar (to those of onsets) with the exception of

Table 13. Breakdown of codas by length

In word types					In word tokens			
Length	Amount	Share	FL		Length	Amount	Share	FL
1	14,347	66.63%	1.00		1	6,939,450	59.07%	1.00
0	3,942	18.31%	0.27		0	3,574,740	30.43%	0.52
2	2,845	13.21%	0.20		2	1,130,022	9.62%	0.16
3	394	1.83%	0.03		3	103,139	0.88%	0.01
4	5	0.02%	0.00		4	375	0.00%	0.00
Total	21,533	100.00%			Total	11,747,726	100.00%	1.69

CC codas that more than double their presence. Inspection of the data reveals that this increase reflects the use of inflection suffixes.

What held true for onsets regarding their presence in WI position holds true for codas in word-final (WF) position. The numbers are, naturally, identical and ~46.4% of codas fall in WF position, accounting for ~78.1% of the 11,747,726 syllable tokens. Table 14 presents a breakdown of codas in WF position.

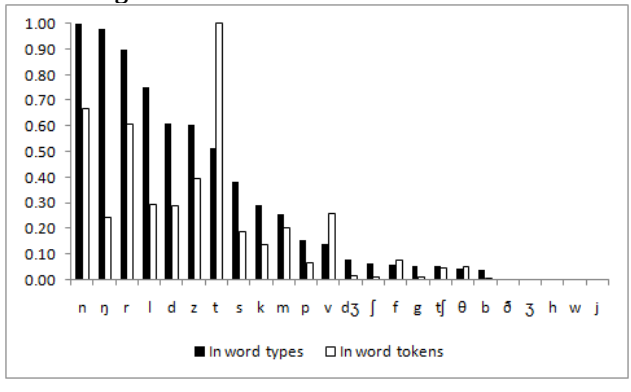
Table 14. Breakdown of codas by length (WF only)

In word types					In word tokens			
Length	Amount	Share	FL		Length	Amount	Share	FL
1	5,580	55.80%	1.00		1	4,961,903	54.08%	1.00
2	2,647	26.47%	0.47		0	3,013,293	32.84%	0.61
0	1,377	13.77%	0.25		2	1,096,209	11.95%	0.22
3	391	3.91%	0.07		3	102,870	1.12%	0.02
4	5	0.05%	0.00		4	375	0.00%	0.00
Total	10,000	100.00%			Total	9,174,650	100.00%	

As is the case with empty onsets, empty codas drop significantly when in word boundary position and, in fact, do less work than double consonant WF codas *(in word types)*. The exposed vowel nuclei that end these words and that do the most work are /i/ (~39.7% in word formation and ~36.4% in language use), /eɪ/ (~16.7% and ~14.0%), /aɪ/ (~11.2% and ~12.5%), /oʊ/ (~11.2% and ~9.3%), and /u/ (~11.1% and ~21.4%). Grouping WF empty codas, front vowels amount to ~75.6% in word formation and ~66.2% in language use while high vowels amount to ~71.5% and ~60.2%, respectively.

As shown in Table 14, single consonant WF codas have noticeably higher FL values in both word types and word tokens than any other kind of WF coda. Chart 3 shows the distribution of segments ranked according to FL in word formation so that the segment with the highest Fl value is placed left-most and the segment with the lowest FL value is placed right-most. Along the x-axis are all consonants, including those that do not occur in WF position (/h/, /w/, /j/). Note that /ð/ and /ʒ/ occur in WF position in two less frequent words each but their respective FL values are too low to be visibly appreciated in Chart 3.

Chart 3. FL values of single consonant WF codas.



From Chart 3, we can appreciate the dominant role of /t/ in language use despite ranking 7th in terms of FL in word formation. Thus, those WF-C words that end in /t/ are relatively few (n = 410) yet very frequent (1,087,942 occurrences or 12% of all running words in the subcorpus). Problems with these words and with this segment in this position will, therefore, contribute to accentedness and perhaps unintelligible speech. Again, this situation points to the advantages of selecting exemplars in instruction well (that is, of building word lists of frequent words with adequate range), so that even if a particular feature is mastered only in these exemplars, learners will be able to deal with most of the language encountered and required until mastery of the feature is generalized.

In terms of word formation, the sonorants /n/, /ŋ/, /r/, and /l/ have the highest FL values of all WF-C codas and account for ~52.1% of all words that end with a single consonant. As a class, sonorants account for ~55.8% and obstruents for ~44.3% of the 5,580 words that end with a single consonant. Regarding voicing, the large majority of WF-C codas are voiced (~77.7%). Regarding manner of articulation, nasals (~32.1%), plosives (~23.8%), and liquids (~23.7) dominate although fricatives (~18.6%) do substantial work. In terms of language use, obstruents (~55.9%) do more work than sonorants (~44.1%). Regarding voicing, WF-C voiced codas still do most of the work but the voiceless consonants increase their share. Regarding manner of articulation, plosives (~33.1%) overtake nasals (~24.4%) and fricatives (~21.6%) overtake liquids (~19.7%).

The share of work done by CC codas in word formation doubles when considering only those codas that occur in WF position (Tables 13 and 14) and, in fact, the FL value of WF-CC codas is second to that of WF-C codas. In word tokens, however, the amount of work than by WF-CC codas is only slightly greater than when considering CC codas in all positions. The implication is clear. Although most CC codas occur in WF position (2,647 out of 2,845 words) and empty codas in WF position are relatively fewer (1,377 out of 3,942 words), words ending in a vowel are substantially more frequent in use than words ending in CC codas.

Inspection of the WF-CC clusters by final segment reveals that there are only five types of WF-CC codas ending in a sonorant, accounting for ~2.0% of all WF-CC. There are 68 different WF-CC that end in an obstruent and these dominate both word formation and language use (~98.0% and ~98.4%, respectively). In word types, the WF-CC codas with the highest FL values belong to cluster types that mostly correspond to inflectional markers (/rz/ FL = 1.00, /nz/ 0.94, /ts/ 0.56, /lz/ 0.53, /ks/ 0.36, /ns/ 0.36) or that, at least some times, correspond to inflectional markers (/st/ 0.61, /nd/ 0.52, /rd/ 0.47). The exception is the cluster /nt/ which ranks third with an FL value of 0.78. In word tokens, the cluster /nd/ accounts for ~31.0% and the cluster /st/ accounts for ~9.0% of the occurrences of WF-CC words in the subcorpus. Grouping WF-CC codas by the final consonant and looking only at language use, we find that plosives account for ~71.1% and fricatives for ~26.1% of all occurrences of WF-CC words (obstruents account for ~98.0% the occurrences). Affricate, nasal, and liquid ending WF-CC are both few and infrequent.

All 59 WF-CCC cluster types include inflection markers with the exception of four cluster types (i.e., /rld/, /ksθ/, /lfθ/, and /rmθ/). Discounting these four cluster types, the plural and third person marker ends as many clusters as does the past tense marker. The 59 WF-CCC cluster types are distributed among 391 WF-CCC words and, specifically, 24 of them occur in only one word and seven occur in only two words each. The WF-CCC cluster type that does the most work in word formation is /nts/ that occurs in 84

words. When grouped by the final consonant, codas that end in a fricative account for ~77.2% of the WF-CCC words and ~61.9% of the occurrences of these words (n = 102,807) in the subcorpus. There are only five WF-CCCC words (*twelfths*, *attempts*, *lengths*, *sixths*, and *worlds*) and each corresponds to one of the five WF-CCCC cluster types (/lfθs/, /mpts/, /ŋkθs/, /ksθs/, and /rldz/, respectively).

SYLLABLES

As mentioned in the previous section, the words in the data set are composed of 21,533 syllables and these syllables occur 11,747,726 times in the subcorpus. The average length, therefore, is ~2.15 syllables per word and the average occurrence of a syllable in the subcorpus is ~545.6 times.

Naturally, not all 21,533 syllables are unique in terms of their constituent segments and the order in which these segments occur. Inspection of the data set reveals 4,600 different syllable types, that is, unique combinations of segments. Of the 4,600 syllable types, 2,650 correspond to monosyllabic words since there are exactly that many unique monosyllabic transcribed forms (2,824 monosyllabic words minus 174 duplicates from homophones). Table 15 shows a breakdown of syllable types based on their position in words. The table separates syllables based on whether their role can be confined to a specific position in the word (e.g., WF position) or they can be found in multiple positions in word formation (e.g., WI and WF position).

Table 15. Breakdown of syllable types based on their position

Syllable types					Syllable tokens			
Length	Amount	Share	FL		Length	Amount	Share	FL
Monosyllabic only	1,659	36.07%	1.00		Multiple pos (mono)	7,356,448	62.62%	1.00
Multiple pos (mono)	991	21.54%	0.60		Multiple pos (other)	2,137,536	18.20%	0.29
Final only	750	16.30%	0.45		Monosyllabic only	1,503,131	12.80%	0.20
Initial only	589	12.80%	0.36		Final only	358,741	3.05%	0.05
Multiple pos (other)	446	9.70%	0.27		Initial only	340,750	2.90%	0.05
Mid only	165	3.59%	0.10		Mid only	51,120	0.44%	0.01
Total	4,600	100.00%			Total	11,747,726	100.00%	

Several general observations can be made based on the information displayed in Table 15. First, ~68.8% of the syllable types (3,163 out of 4,600) occur always in the same position in the word. Of these, 1,659 syllable types

correspond to monosyllabic words and the remaining 1,504 to syllable types that only occur in polysyllabic words. Second, there are 991 syllable types corresponding to monosyllabic words that also play a role in word formation. Third, when considering syllable tokens, those syllable types that occur in multiple positions are far more frequent (do far more work) than those that occur in specific ones only.

The breakdown presented in Table 15 does not take into consideration stress. If primary stress is taken into consideration (effectively eliminating monosyllabic words), ~56.1% of all syllable types that can be found in WI-only position receive primary stress compared to ~5.6% of those syllable types only found in word internal positions and ~19.9% of those syllable types found in WF-only position. The remaining ~18.4%of syllable types can take several positions in the word and these syllable types amount, in terms of syllable tokens, to ~58.2% of all occurrences. Significantly, primary-stressed syllable types that occur in WF-only or internal positions are quite infrequent. Primary-stressed syllable types that occur in WI-only position amount to ~32.5% of all occurrences.

Table 16. List of syllable shapes found in the subcorpus

In word types					In word tokens			
Shape	Amount	Share	FL		Shape	Amount	Share	FL
CVC	8,060	37.43%	1.00		CVC	4,172,188	35.51%	1.00
VC	4,956	23.02%	0.61		CV	2,589,865	22.05%	0.62
CV	2,197	10.20%	0.27		VC	2,419,814	20.60%	0.58
CVCC	1,715	7.96%	0.21		V	857,400	7.30%	0.21
CCVC	1,228	5.70%	0.15		CVCC	636,256	5.42%	0.15
V	1,187	5.51%	0.15		VCC	423,509	3.61%	0.10
VCC	754	3.50%	0.09		CCVC	331,083	2.82%	0.08
CCV	529	2.46%	0.07		CCV	124,472	1.06%	0.03
CCVCC	339	1.57%	0.04		CVCCC	77,591	0.66%	0.02
CVCCC	263	1.22%	0.03		CCVCC	66,931	0.57%	0.02
CCCVC	103	0.48%	0.01		VCCC	18,624	0.16%	0.00
VCCC	91	0.42%	0.01		CCCVC	16,365	0.14%	0.00
CCVCCC	38	0.18%	0.00		CCVCCC	6,760	0.06%	0.00
CCCVCC	37	0.17%	0.00		CCCVCC	3,326	0.03%	0.00
CCCV	29	0.13%	0.00		CCCV	3,003	0.03%	0.00
CVCCCC	3	0.01%	0.00		CCVCCCC	184	0.00%	0.00
CCCVCCC	2	0.01%	0.00		CCCVCCC	164	0.00%	0.00
CCVCCCC	1	0.00%	0.00		CVCCCC	115	0.00%	0.00
VCCCC	1	0.00%	0.00		VCCCC	76	0.00%	0.00
Total	21,533	100.00%			Total	11,747,726	100.00%	

If the analysis is conducted based on secondary stress, the relative weight of WI-only and WF-only syllable types is reversed. While the amount of secondary-stressed syllable types in WF-only position is ~43.7% and that of WI-only syllable types is ~33.1%, in terms of occurrence WF-only syllable types amount to ~32.2% of the tokens while WI-only syllable types account for ~16.0% of the tokens. Again, syllable types that occur in multiple positions are the most frequent in terms of tokens (~44.3%) even though there are comparatively fewer in number (~13.3%).

The distribution of unstressed syllables is as follows: WI-only syllable types (~9.1%) and syllable tokens (~1.1%), WF-only syllable types (~36.4%) and syllable tokens (~4.2%), and word internal only syllable types (~8.9%) and syllable tokens (~0.5%). Again, syllable types that occur in multiple positions are more numerous both in terms of types (~45.6%) and, interestingly, tokens (~94.2%). It should be noted that ~70.48% of unstressed syllable types that occur in multiple position correspond to monosyllabic words.

We conclude this overview of syllables by looking at types and tokens of syllable shapes. Table 16 presents all syllable shapes found in both word formation and language use. As expected from the discussion on onsets and codas, the CVC, CV, and VC shapes are the most frequent types of syllable shapes found in the subcorpus.

Together, the CVC, CV, and VC shapes amount to ~70.7% of all syllable shapes employed in word formation and ~78.2% of all occurrences in language use. Interestingly, the VC shape does twice as much work as the CV shape in word formation but both shapes do a similar amount of work in terms of language use. Syllable shapes with CC onsets and codas follow the top three in terms of word formation but are not as frequent in language use as single vowel syllables. Shapes with CCC onsets and codas are rare in both word formation and use. This reinforces the observation made elsewhere that these types of consonant clusters may be best taught by using the few specific and actual high frequency words in which they appear.

CLOSING REMARKS

As previously stated, the purpose of this paper has been to provide an up-to-date description of spoken English relevant to teaching. The BNC spoken subcorpus has provided the language samples to do so while the analyses undertaken have made it possible to quantify segmental, sequential, and syllabic features as they occur in word formation as well as in language use. Employing the construct of functional load as a means of reference has served to highlight the relative importance of the elements within a given linguistic class.

These findings are of immediate pedagogical application in, at least, three ways. First, this description of spoken English is an alternative to intuition

worthy of consideration. Second, there are a number of cases where FL provides clear rationale for selection and sequencing of material. Third, while exemplification of any aspect of the pronunciation of English should rely on frequent words (Gilner and Morales, 2008), there are situations when the actual frequent words that exhibit a given feature are few in number. Using these words as illustrative material addresses production and perception problems even the feature itself is not learned beyond these words.

The information reported makes it possible for teachers, curriculum planners, and material designers to make informed decisions regarding what to teach and when. Moreover, researchers have now at their disposal raw data reflecting some of the phonetic characteristics of a spoken corpus of substantial size.

REFERENCES

Altenberg, E. P. (2005). The Judgment, Perception, and Production of Consonant Clusters in a Second Language. *International Review of Applied Linguistics in Language Teaching (IRAL), 43*(1), 53-80.

Anderson, S. R. (1982). The Analysis of French Schwa: Or, How to Get Something from Nothing. *Language Learning, 58*, 121–138.

Bent, T., Bradlow, A. R., & Smith, B. L. (2007). Phonemic Errors in Different Word Positions and Their Effects on Intelligiblity of Non-Native Speech. In O.S. Bohn & M. J. Munro (Eds.), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege* (pp. 331-347). Amsterdam; Philadelphia: John Benjamins Publishing.

Breitkreutz, J. A., Derwing, T. M., & Rossiter, M. J. (2002). Pronunciation Teaching Practices in Canada. *TESL Canada Journal, 19*, 51-61.

Brown, A. (1991). *Teaching English Pronunciation: A Book of Readings*. London; New York: Routledge.

Catford, J. C. (1987). Phonetics and the Teaching of Pronunciation. In J. Morley (Ed.), *Current Perspectives on Pronunciation: Practices Anchored in Theory* (pp. 83-100). Washington DC: TESOL.

Celce-Murcia, M., Brinton, D., & Goodwin, J. (1996). *Teaching Pronunciation*. Cambridge: Cambridge University Press.

Dell, G. S., & Gordon, J. K. (2003). Neighbors in the Lexicon. In N. O. Schiller & A. Meyer (Eds.), *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities* (pp. 8-37). Berlin; New York: Mouton de Gruyter.

Derwing, T. M., & Munro, M. J. (2005). Second Language Accent and Pronunciation Teaching: A Research-Based Approach. *TESOL Quarterly, 39*(3), 379-397.

Denes, P. B. (1963). On the Statistics of Spoken English. *The Journal of the Acoustical Society of America, 35*(6), 892-904.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic Vowels in Japanese: A Perceptual Illusion? *Journal of experimental psychology: human perception and performance, 25*(6), 1568-1578.

Flege, J. E. (2003). Assessing Constraints on Second-Language Segmental Production and Perception. In N. O. Schiller & A. Meyer (Eds.), *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities* (pp. 319-355). Berlin; New York: Mouton de Gruyter.

Fraser, H. (2002). *Change, Challenge, and Opportunity in Pronunciation and Oral Communication*. Paper presented at the English Australia Conference. Retrieved November 2002 from http://www-personal.une.edu.au/~hfraser/docs /HFChangeChallengeOpp.pdf.

George, H. V. (1997). *Essays in Informational English Grammar with Reference to English Language Teaching*. Victoria, AU: La Trobe University.

Gilner, L., & Morales, F. (2000). Interlanguage Development: Phonological Processes and Complexity. *Studies in International Relations, Nihon University, Department of Internationals Relations Research Institute Bulletin, 20* (3), 269-282.

Gilner, L., & Morales, F. (2008). Elicitation and Application of a Phonetic Description of the General Service List. *System, 36*(4), 517-533.

Gow Jr., D. W., Melvold, J., & Manuel, S. (1996). *How Word Onsets Drive Lexical Access and Segmentation: Evidence from Acoustics, Phonology and Processing.* Paper presented at the International Conference of Spoken Language Process (ICSLP), Philadelphia.

Hockett, C. F. (1955). *A Manual of Phonology*. Baltimore: Waverly Press.

Jenkins, J. (2000). *The Phonology of English as an International Language: New Models, New Norms, New Goals*. Oxford: OUP.

Kilgarriff, A. (1995). BNC Database and Word Frequency Lists. Available from http://www.kilgarriff.co.uk/bnc-readme.html.

King, R. D. (1967). Functional Load and Sound Change. *Language, 43*, 831-852.

Kitahara, M. (2008). Context of Oppositions for an Estimation of Phonemic Functional Load. *Journal of the Phonetic Society of Japan, 12*(1), 15-23.

Kreidler, C. W. (1997). *Describing Spoken English: An Introduction*. London; New York: Routledge.

Kreidler, C. W. (2004). *The Pronunciation of English: A Course Book in Phonology*. Oxford, UK; New York, NY, USA: B. Blackwell.

Ladefoged, P. (2001). *A Course in Phonetics*. Fort Worth: Harcourt College Publishers.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Harlow: Longman.

Macdonald, S. (2002). Pronunciation - Views and Practices of Reluctant Teachers. *Prospect: An Australian Journal of TESOL, 17*(3), 3–18.

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing Spoken Words: The Importance of Word Onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 576-585.

McAllister, R., Flege, J. E., & Piske, T. (1999). *Second Language Comprehension: A Discussion of Some Influencing Factors.* Paper presented at the Ninth annual conference on the European Second Language Association (EUROSLA 9), Lund, Sweden.

Munro, M. J., & Derwing, T. M. (2006). The Functional Load Principle in ESL Pronunciation Instruction: An Exploratory Study. *System, 34*, 520-531.

Nation, I. S. P. (2004). Study of the Most Frequent Word Families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a Second Language* (pp. 3-13). Amsterdam: John Benjamins.

O'Grady, W. D., Dobrovolsky, M., & Aronoff, M. (1993). *Contemporary Linguistics: An Introduction*. New York: St. Martin's Press.

Pulgram, E. (1970). *Syllable, Word, Nexus, Cursus*. The Hague: Mouton.

Rischel, J. (1962). On Funictional Load in Phonemics. *Statistical Methods in Linguistics, 1*, 13-23.

Setter, J., & Jenkins, J. (2005). Pronunciation. *Language Teaching, 38*(1), 1-17.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stokes, S. F., & Surendran, D. (2005). Articulatory Complexity, Ambient Frequency, and Functional Load as Predictors of Consonant Development in Children. *Journal of Speech Language and Hearing Research, 48*(3), 577-591.

Suenobu, M. (1992). An Experimental Study of Intelligibility of Japanese English. *IRAL, 30*(2), 146-156.

Surendran, D. (2003). *The Functional Load of Phonological Contrasts.* The University of Chicago, Chicago, Illinois.

Surendran, D., & Niyogi, P. (2006). Quantifying the Functional Load of Phonemic Oppositions, Distinctive Features, and Suprasegmentals. In O. Nedergaard Thomsen (Ed.), *Competing Models of Linguistic Change* (Vol. 279, pp. 43-58). Amsterdam: John Benjamins Publishing.

Tarone, E. (1987). Some Influences on the Syllable Structure of Interlanguage Phonology. In G. Ioup & S. Weinberger (Eds.), *Interlanguage Phonology: The Acquisition of a Second Language Sound System* (pp. 232-247). Cambridge, MA: Newbury House Publishers.

Yavaş, M. S. (2006). *Applied English Phonology*. Malden, MA; Oxford: Blackwell Publishing.