A Corpus-Based, Pilot Study of Lexical Stress Variation in American English

Alice Henderson

A corpus-based, pilot study of lexical stress variation in American English

Abstract

Phonological free variation describes the phenomenon of there being more than one pronunciation for a word without any change in meaning (e.g. *because, schedule, vehicle*). The term also applies to words that exhibit different stress patterns (e.g. *academic, resources, comparable*) with no change in meaning or grammatical category.

A corpus-based analysis of lexical stress variation is one way of testing the validity of surveys of speakers' pronunciation preferences for certain variants. Such surveys include Wells' surveys of British English (1999, 2008) and Shitara's survey of American English (1993). The present paper presents the results of a pilot study of American English, replicating part of Mompéan's corpus-based study of British English (2010).

In the current paper, the corpus consists of talks from the TED website (http://www.ted.com), covering the period February 2002 to June 2009. The corpus includes approximately 11.5 hours of transcribed speech (92,750 words) produced by 34 educated speakers (17 men and 17 women with an American accent of standardized variety and some traces of regional pronunciations).

To guarantee a minimum of representativeness for this pilot study, the items analyzed were found at least ten times in the corpus. The preliminary list of search terms showing lexical stress variation was based on the 261 items in the 2008 Longman Pronunciation Dictionary for which survey data was provided, lists found in a textbook on American pronunciation (Celce-Murcia et al., 1997/2007), lists in two previous studies (Shitara, 1993: Mompéan, 2010) and from anecdotal knowledge of frequent variants. Detailed results for lexical stress are provided for seven items.

The TED corpus results do not always concur with LPD data and raise interesting issues concerning the use of authentic spoken corpus data. The paper also discusses designing and carrying out corpus-based pronunciation studies.

I. Introduction

In general, phonological free variation describes the phenomenon of there being more than one pronunciation for a word without any change in meaning (e.g. *because, schedule, vehicle*). The term also applies to words that exhibit different stress patterns (e.g. *academic, resources, comparable*) with no change in meaning or grammatical category. According to Mompéan, phonological free variation may occur for a variety of reasons, which may interact: sound change, phonetic processes and cognitive or sociolinguistic/sociocultural factors, where analogy might affect lexical stress (2010). In his 2010 study, he excluded homographs and variation due to changes in grammatical category, which is entirely justifiable in a study of phonemic variation. However, in his study variation also had to be a characteristic of citation forms and therefore he excluded variation due to rhythmic, contextual influences. Applying the last criterion to a study of lexical stress variation would make it extremely difficult to find enough occurrences in naturally occurring speech but, more importantly, would exclude from analysis a potentially rich locus of variation.

One source of lexical stress variation due to rhymthic, contextual influence is that of stress shift. In Cruttenden, accentual variation confirms "the tendency in English to avoid adjacent accented syllables. It is in order to avoid the placing of primary accents on adjacent syllables that 'accent shift' occurs in phrases such as *'Chinese 'restaurant* (but *Chi'nese*) ..." (2001, 280). Rhythmic constraints can be among the most difficult for teachers to explain and for learners to acquire; it is therefore essential that they be addressed in any publication that seeks to prioritise competing pronunciations. This pedagogical reality emphasizes the need for more studies that use corpora to verify preference poll data.

This paper is a corpus-based pilot study of lexical stress variation in a corpus of modern American English. It provides a useful approach for checking the validity of surveys of pronunciation preferences referred to in the 2008 edition of the Longman Pronunciation Dictionary (LPD), and which are meant to provide "some kind of objective data regarding the relative prevalence of competing pronunciations of various words" (Wells, 2003, 215). As such this paper tries to replicate a small part of

Mompéan's 2010 broader study of free variation in British English, in relation to the LPD (2008) surveys[i].

II. Method

II.1. Data: Corpus Creation

In order to create a spoken corpus for further study, various on-line sources were explored. As variation in stress involves connected speech processes, dictionary sites and CDs were excluded because they typically provide citation-form pronunciations[ii].

A major criterion in the corpus design was that a transcription must accompany the sound file, in order to eliminate time-consuming transcribing work. However, other factors also influenced the choice of sources. For example, the Voice of America covers current affairs on its ESL/EFL site and provides transcriptions to accompany sound files which can be downloaded. A variety of American voices are used and the majority have long stretches of monologue speech. However, as these are designed for learners they tend to involve slow, careful pronunciation that cannot be seen as representative of normal, everyday connected speech. Similarly, the NOVA ScienceNow site also looked like a promising source of podcasts, as transcripts were readily available for free. Unfortunately, these tend to involve several speakers.

The final choice for this study was sound files from videos on the TED website. TED is a small nonprofit organisation in the United States "devoted to Ideas Worth Spreading". It started out in 1984 as an annual conference bringing together people from Technology, Entertainment and Design. Videos of these talks are stored on-line, along with interactive text transcriptions and subtitles in various languages provided by viewers (see Figure 1):

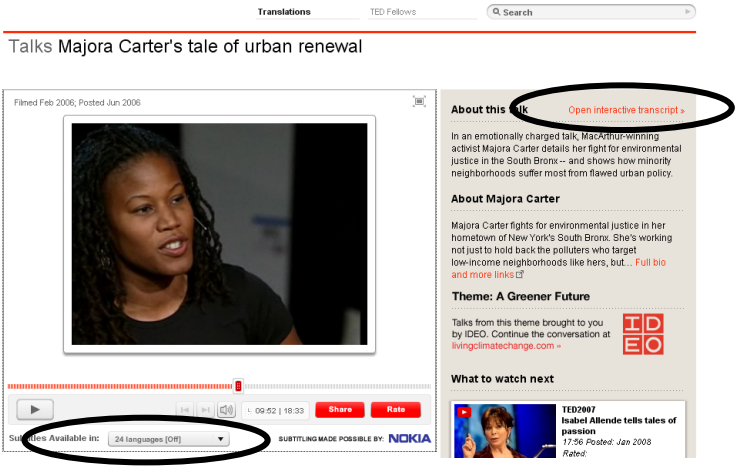

Figure 1. Screenshot of www.ted.com

The talks selected for the present study cover a variety of topics from February 2002 to June 2009. The talks are listed in Appendix A. The transcriptions range from 275 to 5150 words in length. The corpus includes approximately 11.5 hours of transcribed speech (92,750 words) produced by 34 speakers (17 men and 17 women) with an American English accent. As this study looks at variation over a range of American accents, the corpus was not limited to speakers of a Network Standard or other "standardized" form. However, given the formal, public context the selected talks are assumed to represent intelligible, educated American English, though perhaps exhibiting certain regionalisms. The speakers range in age from early 30s to early 60s and come from a host of professions. Further socio-cultural details could be found on-line, as the identity of all of the speakers is clear.

In order to extract high quality sound files from the videos, AudaCity freeware (version 1.2.6 Stable) was used. A cable simply joined the "headphone" output to the "microphone" input and "line input" was selected as the sound source in AudaCity.

Corpora come in a variety of sizes, subject to both the nature of the research question and logistical concerns (McEnery, Xiao & Tono, 2006, 72-73). McEnery and Wilson argue that "…the size of the corpus needed

to explore a research question is dependent on the frequency and distribution of the linguistic features under consideration in that corpus" (2001, 80). Larger corpora are needed for studies of lexis than for grammar, for example, because the validity of conclusions is largely dependent upon the frequency of occurrence of a word. Research which seeks to determine which pronunciation variants are most likely to occur arguably require similarly sizeable corpora, as frequency of occurrence is the determining factor in ranking variants. At first glance, the size of the TED corpus is respectable, being intermediate in size between the SEC and the WSC corpora of spoken English (Table 1):

Name of corpus	Size	Other Information
TED Corpus of American speech	92,750 words	spoken, prepared monologues
BNC (British National Corpus)	10 million words	*spoken = 10% of total 100 million words
ANC (American National Corpus)	22-100 million words	since 1990, in development
MICASE (Michigan Corpus of Academic Spoken English)	1.7 million words	university speech, through 2002
LLC (London-Lund Corpus)	250,000 words	UK, 1960-70s, monologues
SEC (Lancaster/IBM Spoken English Corpus)	53,000 words	UK, radio broadcasts, through 1987
CANCODE (Cambridge/Nottingham Corpus of Discourse in English)	5 million words	UK, interaction, through 1997
WSC (Wellington Corpus of Spoken New Zealand English)	120,000 words	NZ English, formal speech, out of a total of 1 million words, through 1998
ICE (Internat'l Corpus	potentially	spoken & written,

of English)	500,000 spoken words from each world English	out of a total of 20 x 1m words of each world English, since 1989

Table 1. Size of different spoken English orpora

However, as the results and analysis show, the small size of the corpus meant that a significant number of occurrences was not always obtained. This limited the number and the robustness of conclusions which could be drawn from the data, as is often the case in a pilot study.

II.2. Data: Search Terms

Corpus queries are often based on pre-established lists; for this study the goal is to see how these descriptive lists of pronunciation variants compare to authentic, connected speech. Such lists are often based on items found in previous research, dictionaries or textbooks. Using five such sources, a preliminary list of almost 400 potential search terms was compiled:

- 52 items from Mompéan (2010),
- 261 items in the 2008 LPD for which survey data was provided and where variable stress would be expected,
- 37 items from Shitara's 1993 opinion poll of American word stress variation,
- numerous items listed in Celce-Murcia et alia's textbook on teaching pronunciation (1997/2007),
- 9 items from anecdotal/personal knowledge of frequent variants, eg. *development, academic*.

None of these sources could be used as the sole search list, because a preliminary analysis of the TED corpus did not reveal enough occurrences. In line with Mompéan (2010), items were only included in the study if they occurred ten or more times in the corpus, giving a final list of eight items:

- 2 items from Mompéan (2010): *complex (12 occurrences), economic (20)*
- 3 items from the LPD (2008): *Chinese* (61), *individual* (17)/*individuals* (12), *Japanese* (11).
- 1 items from Shitara (1993): *create* (44)
- 1 item from Celce-Murcia (1997/2007): *necessarily* (12)
- 1 item from anecdotal/personal knowledge: *research* (12)

Items were included even if fewer than ten speakers produced them. This is a major drawback of data that is not produced in a controlled, laboratory setting; it is not always possible to collect enough occurrences of lexical items, nor is it always feasible to control for intra-speaker variation by getting enough occurrences from different speakers. Table 2 contains the final list of words studied:

Chinese	create	individual	research
complex	economic	Japanese	necessarily

Table 2. Lexical items studied in the TED 2002-2009 corpus

II.3. Speakers

Thirty-four speakers were chosen: 17 females, 17 males. Their accents were classified as American, based on features such as the presence or absence of rhoticity and typical segmental inventories described for General American English; six native speakers of English were also asked to confirm whether or not speakers were native or non-native speakers of American English, regardless of regional accent. American English is defined as in the LPD as the accent spoken by most Americans "…..who do not have a noticeable eastern or southern accent" (LPD, 2008, xx). One Canadian speaker, Steven Pinker, was excluded because of his nationality and his accent, which is a mixture of Canadian and GAE features.

II.4. Procedure

The interactive transcription of each talk was copied into an Excel file which included: a) the speaker's name and background; b) the URL where the audio file is available; c) the title of the speaker's talk; d) the length in minutes/seconds of the talk; e) the number of words of the talk; and f) the dates the talk was "performed" and posted. Each sound file was downloaded and then carefully listened to in order to correct mistakes in the transcriptions.

Analysis involved four steps: locating the target words in the written corpus using the freeware concordancer ANTConc (Anthony, 2007); listening to the relevant sound file on-line; noting each occurrence along with the speaker's name; determining which variant was produced. The variant was initially identified by the author. When a firm identification was not possible, items were inspected spectrographically using PRAAT, a freeware speech analysis tool developed by Boersma and Weenink (2008). An attempt to use external raters failed, due to faulty editing of sound files and other design issues. Future research will correct this error.

III. Results & Analysis

For some of the items, several forms (eg. plural, past tense) were found in the corpus; the term "word family" in Table 3 reflects this reality:

Word Family	n° of Occurrences	n° of Speakers
CHINESE	61	2
COMPLEX	10	6
CREATE create (44) creates (6) created (20) recreate (3)	74	25
ECONOMIC	19	7

INDIVIDUAL individual (16) individuals (12)	26	12
JAPANESE	11	4
NECESSARILY	12	9
RESEARCH research (12) researcher (2) researchers (5)	19	12
Total	**232**	**x**

Table 3. Number of occurrences in the word list generated by AntConc

Different words in each word family were included in the analysis, even though only four of them provided ten or more occurrences for ten or more speakers: CREATE, ECONOMIC, INDIVIDUAL and RESEARCH. The results for CHINESE were not analysed, as they were skewed by there being only two speakers.

The analysis of the occurrences shows that the TED speakers do not always concur with the LPD's listed pronunciations for General American English (GAE), shown in Table 4:

Headword (n° Tokens / n° Speakers)	LPD Dictionary Pronunciations	
	RP	*GAE*
CHINESE (61/2)	ˌtʃaɪˈniːz stress shift possible	ˌtʃaɪˈniːs less commonly; stress shift possible
COMPLEX (10/6) *{noun (1)}* *{adj. (9)}*	ˈkɒmpleks √ ˈ**kɒmpleks** kəmˈpleks ˈkɒmpleks	ˈkɑːmpleks √ ˌkɑːmˈpleks kəmˈpleks ˈkɑːmpleks stress shiftpossible (only indicated for GAE ?)
CREATE (74/25) create(s) (51)	kriˈeɪt kriːˈeɪt ˈkriːeɪt	
created (20)	ˈeɪtɪd ˈeɪtəd	ˈeɪt̬əd
recreate (3)	ˌriːkriˈeɪt	
ECONOMIC (19/7)	√ ˌiːkəˈnɒmɪk stress shift possible (only indicated for RP?) ˌekə -	√ ˌiːkəˈnɑːmɪk ˌekə - - ˈnɑːmɪk
INDIVIDUAL (26/12) individual (13) *{noun 6, adj. 7}* individuals (13)	√ ˌɪndɪˈvɪdʒu ˌəl stress shift possible - ˈɪndə - -ˈvɪdju ˌz	
JAPANESE (11/4)	ˌdʒæpəˈniːz, stress shift possible	
NECESSARILY (12/9)	√ ˌnesəˈserəli ˌnesɪˈserəli ˈnesəsr ˌəli - sɪs - - ˌli	
RESEARCH (19/12) research (12) *{noun11, adj.1}* researcher(s) (7)	√ riˈsɜːtʃ rəˈsɜːtʃ √ riˈsɜːtʃə rəˈsɜːtʃə	√ riˈsɜːˈtʃ ˈriːsɜːˈtʃ √ riˈsɜːˈtʃˈr ˈriːsɜːˈtʃˈr

Table 4. LPD Dictionary pronunciations of the items for RP (Received Pronunciation) and General American English (GAE).

In Table 4, items in the middle have the same pronunciation in both RP and GAE. LPD conventions apply:

- italic /*ə*/ = sound sometimes optionally omitted
- raised /ʳr/ = sound sometimes optionally inserted
- /‿/ = possible compression of adjacent syllables
- /ɾ̰/ = alveolar tap, usually voiced, like in AmEng *city*

The √ symbol indicates the LPD recommended main pronunciation.

The results for each word family are analysed in more detail in the following sections. Despite this paper's focus on stress variation, one phonetic process – compression – is mentioned, as it affects the number of syllables and often lexical stress.

III. 1. Complex

In the adjectival form of the bisyllabic word *complex*, variation is commonly expected but the LPD (2008) preferences for American English showed a distinct preference (73%) for stress on the second syllable. Interestingly, this pattern was only found twice in the nine adjectival occurrences, and from two different speakers: *incredibly complex* and *no matter how complex they are*. Table 5 shows the other seven occurrences from four speakers which are stressed on the first syllable, the opposite of the variant proposed by the LPD:

Speaker	Search item in context
Tulley	*actually 'complex things made by other*
Tarter	*find more 'complex 'signals*
	to find faint, 'complex 'signals that our
Roach	*in the 'complex 'sensory-motor action*
Boston	*many other 'complex 'human motions*
	conform to the 'complex 'topological shape
	to deal with this 'complex to'pology, various

Table 5. Occurrences of *complex* in context, per speaker

All except the last example appear to be cases of stress shift.

III. 2. Create

For the item *CREATE*, no occurrences of stress shift were found, even though the LPD lists this as possible. According to the LPD, 87% of respondents preferred to stress the second syllable; this was the case in most of the 74 occurrences over 25 speakers. Compression seems to be occurring in a few cases, so *create* sounds like *crate*. This may or may not be due to regional variation. Unfortunately, at this stage it is impossible to say precisely how many occurrences concur with the LPD, because the external raters showed far too much variation in their judgments. Further studies will examine this in detail, and external raters will be given better designed stimuli and instructions.

III. 3. Economic

The LPD gives the main pronunciation with stress on <no> and 11 of the 19 occurrences follow this pattern. Stress shift is not mentioned as a possibility in GAE and yet six cases were found in one speaker (Pine), as shown in Table 6:

Search item in context
the predominant 'economic 'offering
this progression of 'economic 'value
a new level of 'economic 'value
becoming the predominant 'economic 'offering
are the 'economic 'offerings you are providing
think about the 'economic 'value they have

Table 6. Stress-shifted occurrences of *economic* in context

As no other speakers produced the collocations *economic +value* or *economic +offering*, it is impossible to know whether or not collocational

factors influence the likelihood of stress shift, but future research could look into this.

Two other cases of stress shift occurred in two other speakers: *caring about 'economic 'factors* and *is an 'economic 'tipping point*. Finally, Table 7 shows four examples which did not exhibit stress shift:

Speaker	Search item in context
Carter	*for environmental and eco'nomic 'justice*
Alcorn	*Most of the eco'nomic 'models are built about 'social-eco'nomic 'movements*
Rosenda le	*kinds of eco'nomic 'forces*

Table 7. Occurrences of *economic* in context, per speaker

Whether or not these cases of shift represent speaker-specific idiosyncracies or regional variations, they are unpredictable cases; the speakers could have shifted the stress because a word with primary stress follows.

III. 4. Individual

Stress shift is indicated by the LPD as possible for the word *individual*, but none of the 26 occurrences in the present study display this. Compression, however, was found (regardless of grammatical category) in 6 of the 12 speakers.

III. 5. Japanese

For the item *Japanese* the LPD lists one pronunciation and stress shift as possible. In the 11 occurrences in the TED corpus, several examples of stress shift (regardless of grammatical category) were found over three speakers (Table 8):

Speaker	Search item in context

Wallace	*direction that ‘Japanese ‘toilet technology*
Baraniuk	*languages like ‘Chinese, ‘Japanese and Thai*
Lee	*‘Japanese ‘Chinese food*
	all the ‘Japanese ‘bakers who introduced
	Chinese food and ‘Japanese ‘foods,
	the ‘Japanese ‘immigrants came
	something that is ‘Japanese to being
	locked up all the ‘Japanese during World War
	invented by the ’Japanese, popularized
	sort of like a ‘Japanese ‘guy coming

Table 8. Stress-shifted occurrences of *Japanese* in context, per speaker

Wallace’s shift to word-initial stress is predictable, but not all of the other examples can be explained by stress clash avoidance: for example, Baraniuk’s *languages like ‘Chinese, ‘Japanese and Thai* but also Lee’s *something that is ‘Japanese to being something that is ‘Chinese* and *invented by the ’Japanese, popularized by the ‘Chinese*. The latter two can be attributed to contrastive stress, as the extended context shows. However, the speaker could just as easily have maintained initial stress and expressed contrast.

In one case stress shift actually resulted in stress clash: *It was a Japa’nese scientist who first undertook …* It is difficult to ascribe this to the discursive context. The stress pattern is used in a context where contrast is not being signalled, as the preceding text is about the vegetation where Bonobos frequently live:

> *The wild Bonobo lives in central Africa, in the jungle encircled by the Congo River. Canopied trees as tall as 40 meters, 130 feet, grow densely in the area. It was a Japanese scientist who first undertook serious field studies of the Bonobo, almost three decades ago.*

However, a low speech rate may explain this shift. This is scripted monologue which accompanies a video clip from a documentary film that the TED speaker showed, so the film-speaker was probably not at a loss and searching for words.

III. 6. Necessarily

Speech rate may also explain the compression found in the occurrences of the word *necessarily*. The LPD only provides preference data for British English, finding that 68% prefer primary stress on the third syllable *nece'ssarily* and 32% prefer initial stress. This is close to the 25% (3/12) of occurrences with word-initial stress in the TED corpus of American English (Table 9):

Speaker	Search item in context
Lee	*who ate rice would' necessarily bring down*
	couldn't 'necessarily be fraud,
Abrams	*you wouldn't 'necessarily think of when*

Table 9. Word-initial stressed occurrences of *necessarily* in context, per speaker

Analysis with PRAAT showed that compression may have occurred in two of the very fast speakers, Wallace and Powers. However, external raters' judgments for these two were extremely varied. This not only reinforces the case for including speech rates in corpus-based studies but also confirms the well-known difficulty some individuals have in perceiving syllables and/or stress.

III. 7. Research

The final item, *RESEARCH*, seems to reflect national and socio-economic influences. According to the LPD:

> the ˈsɜːtʃ//ˈsɜːrtʃ form appears still to predominate in universities, although ˈriːsɜːtʃ//sɜːrtʃ has increasingly displaced it in general usage both in Britain and in America. Some speakers may distinguish between the verb .'. and the noun '. . . (2008, 683).

The LPD preference poll of British English found 80% in favour of word-final stress in the word *research*, the figure rising to 95% among university teachers. Conversely, for American English the LPD found a preference for word-initial stress (78%). Table 10 shows the 4 of the 19 occurrences from the TED corpus that do not have word-initial stress, including two occurrences of *researchers*:

Speaker	Search item in context
Benyus	*mainly about re'search in biomimicry.*
Tarter	*generously supported this re'search.*
Wallace	*from some re'searchers at Stanford*
	that these re'searchers did MRI brain

Table 10. Word-final stressed occurrences of *research* in context, per speaker

None of these can be attributed to clash avoidance. Thus, the results in the TED corpus (15/19 or 79%) confirm the LPD results for American English.

IV. Conclusion

Phonological free variation, or variation in the pronunciation of a word without any change in meaning, also applies to words that exhibit different stress patterns with no change in meaning or grammatical category. Such variation may occur for several reasons, of which phonetic processes, sound change and cognitive or sociolinguistic/sociocultural factors. Which of these variants to prioritise is a recurring problem in modern dictionaries and the use of pronunciation preference survey polls can be one solution. However, another solution could be corpus-based studies, as they can provide greater quantities of more objective data, with a corresponding increase in the validity of those predictions and perhaps a reduced "logistical cost". Consequently, this paper has provided some initial results from a corpus-based pilot study of spoken American English, partly a replication of an earlier study of British English by

Mompeán (2010). Authentic connected speech from the TED corpus was used to study lexical stress variation, including that due to the rhythmic phenomenon known as stress shift. Mompéan's study of free phonological variation was much more extensive (2010) but excluded stress shift; given the current study's focus on lexical stress variation the influence of stress shift was actively sought it out. This proved productive, as it allowed speech rate and larger discursive context to be proposed as factors influencing stress variation.

In general, evidence from the TED corpus confirms some but not all of the preferences in the LPD pronunciation polls. In direct contradiction to the LPD, seven of the 9 occurrences of *complex* are stressed on the first syllable; all seven also seem to be cases of stress shift. The items *create* and *individual* showed compression but no stress shift.

Analysis of the results for two other items, *economic* and *necessarily*, raised the possibility that two other factors might play a role in stress shift. Eight occurrences of *economic* showed shift despite it not being mentioned in the LPD and six of the occurrences from one speaker hint at the possible influence of collocational knowledge. This speaker used the collocations '*economic* '*value* and '*economic* '*offering*; it is not impossible that collocational factors affect the likelihood of stress shift, and future research could look into this. Secondly, the LPD does not provide any data for American preferences for *necessarily*. Notwithstanding, 25% of the examples in the TED corpus use word-initial stress, which is not very different from the LPD's finding of 32% for British English. Speech rate might help to explain the compression found in several occurrences of *necessarily*.

For the item *research*, the results from the TED corpus (79%) confirm the LPD results (78%) for American English. However, it must be emphasized that given the small size of the corpus and the under-representation of several items, none of the statistics can be used to generalize about lexical stress in American English. Moreover, the LPD data reflect the preferences of people from various social backgrounds, which is also true for the TED corpus. TED speakers tend to be well-spoken, articulate individuals, with varying levels of academic qualification. This may skew the results for *research*, where the LPD poll shows different preferences for university teachers and other respondents.

Perhaps the most interesting item is *Japanese*, as 3 of the 11 occurrences cannot be explained by stress clash avoidance and one shift creates a clash. The first two occurrences reveal the speaker's desire to contrast two nationalities but the third might result from a low speech rate. As compression and other connected speech processes are more likely to occur when the number of unstressed syllables and the overall speech rate are increased, perhaps stress shift becomes less predictable when speech rates are lower. An objective measure of speech rate may need to be included in corpus-based studies of spoken language such as this.

Despite all the insights provided and the data obtained, it should be borne in mind that the present study is only a pilot study and, as such, has severe limitations. It was impossible to control the number of speakers and occurrences, so certain items are under-represented, which means that no claims can be made about the relative prevalence of free variants. Age differences were not explored but a large, diachronic corpus could potentially address this. Likewise, regional differences were glossed over, assuming that one General American English exists with shared recognizable tendencies. Finally, verification by external raters, which can be useful, was not possible and would have to be integrated into further work on these issues.

Directions for future research include addressing all those issues in further corpus-based studies or controlled production tasks. Nonetheless, this pilot study confirms that such corpora can be usefully designed to verify survey data. English is a living language and preferences are bound to evolve. This is a major argument in favour of using a large corpus (which can be easily updated to track diachronic change) in addition to survey data, in order to decide which pronunciation(s) to prioritise not only in dictionaries but also in language teaching.

The use of such polls in pronunciation dictionaries gives teachers and learners access to explanations about current usage. They can then organise that knowledge into rules which help them to predict lexical stress patterns. For example, if General American English (GAE) speakers tend to final-stress bisyllabic French loan words, then *'garage* is probably a British pronunciation and *ga'rage* is probably GAE. This ability to predict gives learners more autonomy, which is the goal of much teaching: independent application of appropriate knoweldge in new contexts.

Similarly, easy access to digital resources in today's world means that it is no longer adequate to take at face value statements such as "Speakers

of Canadian English tend to stress the second syllable in words x, y and z." Teachers have the ability to collect data for themselves and/or access data via on-line journal subscriptions, forums, etc. This enables them to analyse first-hand what is happening, for example, in American English today.

Bibliography

Anthony, L. (2007). ANTConc (Version 3.2.1) [Computer Program]. Retrieved June 2007, from http://www.antlab.sci.waseda.ac.jp/software.html.

AudaCity. (2009). (Version 1.2) [Software]. Available from http://www.audacity.sourceforge.net

Boersma, P. & Weenink, D. (2008). Praat: doing phonetics by computer (Version 4) [Computer program]. Retrieved January 20, 2008, from http://www.praat.org/.

Celce-Murcia, M, Brinton, D.M. & Goodwin, J.M. (1997/2007). *Teaching pronunciation: A reference for teachers of English to speakers of other languages.* Cambridge: Cambridge University Press.

Cruttenden, A. (2001). *Gimson's pronunciation of English* (6th edn). London: Arnold.

Free Dictionary. [Website]. http://www.thefreedictionary.com/squishy Accessed June 2009.

McEnery, A. & Wilson, A. (2001). *Corpus linguistics* (1st edn 1996). Edinburgh: Edinburgh University Press.

McEnery, A., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* London: Routledge.

Merriam-Webster Dictionary. [Website].. http://www.merriam-webster.com/help/audiofaq.htm. Accessed January 2009-February 2010.

Mompéan, J.A. (2010). A corpus-based study of phonological free variation in English, in Henderson, A.J. *English Pronunciation: Issues & Practices, Conference proceedings of EPIP I, June 3-5, Université de Savoie, Chambéry, France.* Chambéry, France: Presses de l'Université de Savoie. (forthcoming).

NOVA ScienceNow. [Website]. http://www.pbs.org/wgbh/nova/sciencenow/ Accessed March 2009.

Shitara, Y. (1993), "A survey of American pronunciation preferences". *Speech Hearing and Language 7,* 201-232. Available at www.phon.ucl.ac.uk/home/wells/shitara.pdf

TED. [Website]. http://www.ted.com. Accessed January 2009-February 2010.

Temperley, D. (2009). Distributional Stress Regularity: A Corpus Study. *Journal of Psycholinguistic Research, 38,* 75-92.

Trudgill, P. & Hannah, J. (2008). *International English: A Guide to the varieties of Standard English*. London: Hodder Education.

Voice of America. [Website]. http://www.manythings.org/voa/rss/ Accessed June 2009.

Wells, J.C. (2008). *Longman Pronunciation Dictionary*. 3rd edition, Harlow, England: Pearson-Longman.

Wells, J.C. (1999). "British English pronunciation preferences: a changing

scene". *Journal of the International Phonetic Association 29 (1)*, 33-50.

Appendix A

TED corpus: List of speakers and URL

Speaker's Name	URL : http://www.ted.com/talks
Pete Alcorn	pete_alcorn_s_vision_of_a_better_world.html
Benjamin Wallace	benjamin_wallace_on_the_price_of_happiness.html
Ray Anderson	ray_anderson_on_the_business_logic_of_sustainability.html
JJ Abrams	j_j_abrams_mystery_box.html
Richard Baraniuk	richard_baraniuk_on_open_source_learning.html
Dan Barber	dan_barber_s_surprising_foie_gras_parable.html
Michelle Obama	michelle_obama.html
Elizabeth Gilbert	elizabeth_gilbert_on_genius.html
Dave Eggers	dave_eggers_makes_his_ted_prize_wish_once_upon_a_school.html
George Smoot	george_smoot_on_the_design_of_the_universe.html
Noah Feldman	noah_feldman_says_politics_and_religion_are_technologies.html
Janine Benyus	janine_benyus_shares_nature_s_designs.html
Majora Carter	majora_carter_s_tale_of_urban_renewal.html
Stewart Brand	/stewart_brand_on_squatter_cities.html
Robert Neuwirth	robert_neuwirth_on_our_shadow_cities.html
Mae Jemison	mae_jemison_on_teaching_arts_and_sciences_together.html

Gever Tulley	http://www.ted.com/speakers/gever_tulley.html
Rob Forbes	rob_forbes_on_ways_of_seeing.html
Joseph Pine	joseph_pine_on_what_consumers_want.html
Mike Rowe	mike_rowe_celebrates_dirty_jobs.html
Deborah Scranton	deborah_scranton_on_her_war_tapes.html
Jenny 8. Lee	jennifer_8_lee_looks_for_general_tso.html
Nancy Etkoff	nancy_etcoff_on_happiness_and_why_we_want_it.html
Jill Bolte Taylor	jill_bolte_taylor_s_powerful_stroke_of_insight.html
Philip Rosedale	the_inspiration_of_second_life.html
John Markoff	john_markoff_on_newspapers.html
Penelope Boston	penelope_boston.html
Catherine Mohr	catherine_mohr_surgery_s_past_present_and_robotic_future.html
Sylvia Earle	sylvia_earle_s_ted_prize_wish_to_protect_our_oceans.html
Samantha Power	samantha_power_on_a_complicated_hero.html
Alisa Miller	alisa_miller_shares_the_news_about_the_news.html
Jill Tarter	jill_tarter_s_call_to_join_the_seti_search.html
Susan Savage-Rumbaugh	susan_savage_rumbaugh_on_apes_that_write.html
Mary Roach	mary_roach_10_things_you_didn_t_know_about_orgasm.html

Endnotes

[i] It is unclear which data for American English are taken from Vaux's 2002 polling figures and which data stem from discussions with Dauer or the work of Shitara, both of which Wells used in preparing the 1999 edition (LPD, 2008, x-xi).

[ii] The Free Dictionary seems to use one man's live, human-being (not Text-To-Speech) voice to pronounce individual words in American English. Clicking on the flags gives voices which are definitely TTS, but clicking on the megaphone symbol next to the word usually gives the voice of one man; there is enough intonational variation to believe that this is not TTS. The Merriam-Webster on-line dictionary also has audio pronunciation and they are recorded by real human beings (e-mail confirmation July 27, 2009). The CD which accompanies the 2008 edition of the LPD has both RP and GAE pronunciations of headwords recorded by real-human beings but, like several other on-line dictionaries, it does not provide models of all items under each headword.