Phonetically Controlled Definitions?

Włodzimierz Sobkowiak
School of English, Adam Mickiewicz University, Poznań, Poland
al.Niepodleglosci 4, 61-874 Poznan, Poland
sobkow@amu.edu.pl

Abstract

A phonostatistic analysis of pronunciation difficulty in the 88495 definitions of the British edition of *Macmillan English Dictionary for Advanced Learners* (MEDAL) is performed. The L1-sensitive (Polglish) Phonetic Difficulty Index (PDI) of each word in every definition is automatically calculated, summed and averaged, as is the definition word-length (mean=11.91). Selected definitions are compared across three dictionaries: MEDAL, CALD and LDOCE4. Some defining vocabulary (DV) phonostatistics are also provided for the three dictionaries (PDI frequency distributions and means). Appeal is made for more control over definition phonetics, which is claimed to be important, especially for beginning and intermediate learners, who tend to subvocalize in reading.

1. Introduction

In my Euralex 2002 contribution (Sobkowiak & Kuczyński, 2002) I investigated the phonetic difficulty of defining vocabularies (DVs) in two EFL dictionaries: LDOCE and CIDE, concluding that they "are, after all, significantly phonetically easier than the frequency-matched portions of the reference lexicon, here treated as chance level", even if "This does not necessarily show that editors of the two learners' dictionaries of English have exercised effective control over this important aspect of the defining vocabulary's structure". The results of that study contradicted the working hypothesis which was that DVs "are on the whole not phonetically easier than the 'ordinary' lexicon of English".

Some of the underlying assumptions of my research were that:

- EFL learners "in their majority [...] want to find out about the meaning of an unknown word", which in monolingual EFL dictionaries would require them to read the definition(s) and/or example(s),
- as DVs are carefully controlled by dictionary makers (or so they claim), "the choice of the deployed vocabulary is quite crucial for the learner's vocabulary acquisition as a whole",
- "if lexicographers are aiming at global 'user-friendliness' of their defining vocabulary, they should certainly also make it phonetically friendly",
- "phonetically difficult words in dictionary definitions will tend to impede the reading and understanding process, particularly in those learners who continue to vocalise or articulate subvocally in silent reading. And, according to Gibson and Levin (1980:342): 'it is perfectly certain that the inner hearing or pronunciation, or both, of what is read, is a constituent part of reading by far the most of people, as they ordinarily and actually read'".

Of course, it is one thing to measure the average phonetic difficulty of the dictionary's DV, but another to ascertain the phonetic difficulty of the actual definitions put together from the given DV repertoire. It is, after all, the definitions which are read by the learners, not the word-lists. And, reversing the perspective: lexicographers' command over the phonetics of their DV lists (even if it could be demonstrated) does not automatically imply that they yield comparable control over the end-product, the definitions. Thanks to the courtesy of Bloomsbury Publishing Plc (I am especially grateful to Michael Rundell, who talked to Bloomsbury on my behalf and negotiated the conditions), which supplied me with the full computer-readable list of definitions used in their MEDAL (Rundell, 2002), I could extend my Euralex 2002 project to cover the actual definitions (of one of the leading EFL monolingual dictionaries on the market), rather than only DVs.

In what follows a brief look is first cast at how MEDAL DV's phonetic difficulty compares to the two dictionaries investigated before (LDOCE and CIDE); whereupon a phonostatistical analysis of MEDAL definitions is attempted. Some global measures are presented as well as a very limited comparison of some definitions across the three dictionaries. Both the size and scope of this contribution, as well as the unavailability of full reference data on other EFL dictionaries explains the severe limitations of this analysis.

2. Phonetic Difficulty of DVs Revisited

The measure of phonetic difficulty used to evaluate DVs in my 2002 Euralex contribution is described in detail in Sobkowiak (1999). Briefly, the idea of the index is that it is a global numerical measure of the phonetic difficulty of the given English lexical item for Polish learners. The algorithm was applied to the OALDCE word-list (see Mitton, 1986 and 1992), to generate the phonetic difficulty index (PDI) with a range between 0 and 10, mean 2.24, and standard deviation 1.5. The PDI values were then copied for DV words in the respective lists. The empirical validity of the partly intuitively assigned PDI scores was tested and confirmed in my unpublished paper (Sobkowiak, unpublished).

Starting with MEDAL definitions I decided to first revisit the DV phonetic difficulty issue by running the respective 2002 tests on MEDAL DV and comparing the results with those for the other two dictionaries. Notwithstanding some methodological problems, such as partial incompatibility of the three DVs (mostly due to different size and lemmatization schemes), frequency normalization against the BNC-derived list of commonest English lemmas (see Kilgarriff, 1997 and http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html; last accessed 26th February 2004) and others, the overall results are rather clear, presented in Table 1 below (the few top PDI scores were dropped).

The nonparametric (highly non-normal distributions) Mann-Whitney rank test (see Butler, 1985, chapter 8.2) shows that while the DVs of LDOCE and CIDE are indeed significantly phonetically easier than chance (see Sobkowiak & Kuczyński, 2002 for the operational definition of chance in this context), MEDAL's DV is not, even if mean PDI is almost identical across the three dictionaries. As this contribution mostly concerns definitions rather than simply defining vocabularies, I will not proceed any further in the discussion of the latter here.

PDI	LDOCE3	CIDE	MEDAL	OALDCE
0	545	494	507	460
1	550	547	553	546
2	388	400	412	441
3	293	328	293	319
4	152	168	165	167
5	48	51	54	52
6	31	24	27	27
N	2015	2015	2175	2015
mean PDI	1.64	1.66	1.68	1.73
rank test	-2.76	-3.60*	-1.32	

* Calculated over the 1715-item simplex CIDE DV subset

Table 1. Frequency distribution of PDI in LDOCE3.

3. Phonetically Controlled Definitions?

As mentioned above, phonetic control over DV does not necessarily imply one over definitions themselves. It is legitimate to ask how such definitions fare with respect to pronunciation. There are potentially many different ways to approach such a question, of course. One could, for example, ask if there is any correlation between the defining vocabulary's deployment frequency in definitions and its phonetic difficulty index (there is none in MEDAL). Or phonetic sandhi (inter-word) phenomena could be studied in definitions as largely independent from DV choice, but potentially important, especially for the more proficient learners, who do not read text word-by-word any more. The overall phonostatistical pronunciation profile of dictionary definitions compared with ordinary text (of different genres and levels of difficulty) could be another promising line of attack. In what follows I can only hope to achieve a much more modest aim, i.e. to retrieve some general descriptive statistics of phonetic difficulty of one dictionary's definitions, hopefully representative of a larger sample of EFL monolingual learners' dictionaries. As I am not aware of any research so far published in this area, I tend to think of this contribution as tentatively opening it for metalexicographic inspection. In this context the preliminary nature of this research can perhaps be excused.

3.1. Preparation of data

The definition files supplied by Bloomsbury counted 93042 records altogether, 88495 of which used in the British edition of MEDAL. Only the latter were used for all calculations. The text of all definitions was first tokenized. As expected, some problems arose at that stage, mostly concerning non-alphabetic symbols, capitalization and typographic errors. These were resolved by hand, yielding 1,053,629 words altogether. Each word was then looked up in the phonetically transcribed and PDI-tagged lexical database derived from OALDCE, as mentioned above. In its current form it counts 85430 wordforms. Of all the one million tokens 2892 types were not found in my lexical database. These included: 1004 numbers (dates, cardinals, ordinals, Roman), 370 proper names, 122 acronyms, 640 hyphenated compounds, 756 others (rare words, recent neologisms, taboo/slang, Americanisms, French, Latin, etc.). I arbitrarily assigned PDI=3 to them, with the exception of those hyphenated compounds, which were found after decomposition into simplexes and scored their respective PDIs, such as *grey-brown* (PDI=0+1). These compounds were counted as two words; those which could not be resolved – as one. The number of words, the global PDI (sum of all word-PDIs) and the mean PDI (global PDI divided by the number of words) were recorded for each definition.

3.2. Analysis

The global PDI mean for the 88495 British MEDAL definitions is 1.55, with standard deviation 0.45. There are 337 definitions with PDI=0, most of them (212) being single words, and only seven counting more than 3 words, e.g. *felt in an extreme way* (*exquisite*), *very simple in design* (*primitive*) or *an exact point in time* (*instant*). The mean number of words in a definition is 11.91, s.d. = 6.2. Mean definition PDI appears to be unrelated to its length measured in words.

These numbers have little sense, of course, without any reference data. On analogy with DV phonetics, one could try a comparison with some normalized 'ordinary' (properly operationalized) English text and/or with other dictionaries. Either of these is impractical, however: the former because it is not clear which text genre would be stylistically compatible with dictionary definitions; the latter because access to complete definition files of standard EFL dictionaries is severely restricted. In this situation I decided to try a stopgap measure: comparing MEDAL definition phonostatistics with (a) a random text lifted from the internet and (b) selected definitions of Longman and Cambridge monolingual EFL dictionaries.

For the former I took Diane Nicholls' short text on "What is learner English", which opened the first issue of *MED Magazine*, the monthly webzine of the Macmillan English Dictionary (http://www.macmillandictionary.com/MED-Magazine/Sample-Issue/01-language-interference-learner-english.htm; last accessed 26th February 2004). The text counts 1698 words in 57 sentences, which were treated like MEDAL definitions for the purposes of this calculation. The average record word-length is of course much higher than in MEDAL, almost 30 words per sentence. The mean sentence PDI is 1.92. Considering that Nicholls' essay is on a rather technical topic and written for teachers of EFL rather than learners, the small (?) mean PDI difference of 0.37 is certainly intriguing, although obviously no statistical significance at all can be claimed for these figures.

Comparison with other dictionaries would prima facie seem more reasonable. I decided to try the following test: a few phonetically difficult MEDAL definitions are compared against Longman and Cambridge definitions of the same keywords to phonostatistically case-study the choice of words in all three. "Phonetically difficult MEDAL definitions" was tentatively operationalized as "at or beyond 2.58 standard deviations from the PDI mean (p<.01)", i.e. at or beyond PDI=1.55+2.58*0.45=2.71. There were 1131 definitions meeting this criterion. Fourteen of these had word-length 12, i.e. almost exactly mean value for MEDAL. All fourteen were compared with their CALD (Cambridge Advanced Learner's Dictionary; Gillard, 2003) and LDOCE4 (Summers, 2003) sense-adjusted equivalents. Three examples with highest MEDAL PDI scores appear in Table 2.

word	MEDAL	CALD (=CIDE2)	LDOCE4
breathe in	to take other substances into your lungs through your nose or mouth (3.1)	to move air into and out of the lungs (breathe; 2.0)	to take air into your lungs (2.3)
surrogate	someone or something that replaces another person or thing as their representative (3.1)	replacing someone else or used instead of something else (1.4)	a person or thing that takes the place of someone or something else (1.8)
underwear	clothing that you wear next to your skin under your other clothes (3.1)	clothes worn next to the skin under other clothes (2.9)	clothes that you wear next to your body under your other clothes (3.2)

Table 2. Selected definitions in the three dictionaries, with their PDI scores

This comparison is not meant, of course, to prove that MEDAL's definitions are on average phonetically harder than in the other two dictionaries. This was not the point of the exercise. But it certainly is interesting that cross-dictionary differences of up to 1.7 PDI can arise for some definitions (*surrogate*). Even definitions of very similar wording, like those of *underwear*, can vary in terms of phonetic difficulty. With the mean word-length of the fourteen definitions standing at 11.83 (CALD) and 10.92 (LDOCE4), the average definition length of the three samples is comparable, while the mean PDI is 1.9 and 2.2, respectively. This shows that, at least to a certain extent, definition PDI is an independently manipulable variable, which – all other things being equal – could come under active lexicographic control.

4. Conclusions

This study has emphatically no pretence to a comprehensive and conclusive phonostatistical treatment of dictionary definitions at large, for reasons which were briefly alluded to above. Its findings are at best provisional and suggestive. There is no previous research to refer to, and the availability of data is far from satisfactory. Yet, the obtained results seem to promise a new and fruitful metalexicographic perspective on dictionary writing and use, and definitions are in the centre of either. Phonolexicographic interest so far has been restricted exclusively to the phonetic transcription field in the entry's microstructure. In a series of papers and in my 1999 book I have demonstrated that this little field is more complex than normally believed. It is now time to widen the perspective.

For example, it is not unthinkable to imagine EFL electronic dictionaries dynamically adjusting their definitions to the learner's needs and requirements, not only in terms of their lexical scope (DV) and syntax, but also in terms of pronunciation (see deSchryver, 2003 for this and other lexicographer's dreams). If *thorough* is among the phonetically hardest lexical items for Polish EFL learners (PDI=7), why not use a substitute in definitions (*complete* has

PDI=1) adjusted for pre-intermediate learners? Or at least why not reduce the definition incidence of *thorough*, which now stands at 22 in MEDAL?

As mentioned above, the somewhat mechanical extrapolation of lexical phonetic difficulty onto that of whole sentences and text, as practised in this contribution, is but a stopgap measure, which should quickly give way to more subtle instruments of gauging and scoring PDI, ones taking into account not only inter-word phonetics, but also subphonemic pronunciation problems, such as aspiration, lateral velarisation, vowel length/timbre variation or palato-alveolar articulations, as well as fast-speech phenomena, such as vocalic reductions and losses, cluster simplifications, stress switches, consonantal assimilations and coalescence. Similarly, the phonetic difficulty index itself in its current shape is only an ad-interim tool which, to be fully reliable, would have to be derived from careful inspection of errors and perceived problems of learners with different mother tongues and at different levels of proficiency. A project like this, while certainly feasible, has not been attempted yet.

The field is now open for study. Considering, on the one hand, that vocabulary continues to be in the focus of EFL teaching and learning, and on the other, that new learners' dictionaries are now published almost monthly, perhaps it might not be too optimistic to expect that phonolexicographic research will now significantly grow in volume?

References

Butler, C. S. 1985. *Statistics in Linguistics*. Oxford: Blackwell.

Gibson, E. J. and Levin, H. 1980. *The Psychology of Reading*. Cambridge, Mass.: The MIT Press.

Gillard, P. (ed.) 2003. *Cambridge Advanced Learner's Dictionary* (CALD). Cambridge: Cambridge University Press.

Kilgarriff, A. 1997. 'Putting Frequencies in the Dictionary'. *International Journal of Lexicography 10.2.*135-55.

Mitton, R. 1986. 'A Partial Dictionary of English in Computer Usable Form'. *Literary and Linguistic Computing 1*. 214-15.

Mitton, R. 1992. 'A Description of a Computer-Usable Dictionary File Based on the Oxford Advanced Learner's Dictionary of Current English'. [bundled with the software]

Rundell, M. (ed.) 2002. *Macmillan English Dictionary for Advanced Learners* (MEDAL). Oxford: Macmillan Education.

de Schryver, G.-M. 2003. 'Lexicographers' Dreams in the Electronic Dictionary Age'. *International Journal of Lexicography 16.2.* 143-99.

Sobkowiak, W. 1999. *Pronunciation in EFL Machine-Readable Dictionaries*. Poznan: Motivex.

Sobkowiak, W. (unpublished). 'Subjective Phonetic Difficulty of English Words to Polish Learners: Does Frequency Matter?'. [http://elex.amu.edu.pl/~sobkow/diffind2.doc]

Sobkowiak, W. and Kuczyński, M. 2002. 'Phonetics and Ideology of Defining Vocabularies' in A. Braasch and C. Povlsen (eds.), *Euralex'2002 Proceedings*. Copenhagen: Center for Sprogteknologi. Vol 2. 495-502.

Summers, D. (ed.) 2003. *Longman Dictionary of Contemporary English* (4th ed., LDOCE4). Harlow: Longman.