PDI revisited: lexical coocurrence of phonetic difficulty codes[1]

Włodzimierz Sobkowiak
IFA UAM, IN PWSZ w Koninie

0. Abstract

A lexical database of English originating from the machine-readable version of OALDCE has been annotated with two types of Polglish pronouncing difficulty tags: (a) aggregate numerical measures of grapho-phono-lexical difficulty, the Phonetic Difficulty Index (PDI), with a range of 0-10, and (b) qualitative PDI codes of specific grapho-phonetic difficulty, selected from a list of 57 such items, derived empirically from teaching experience. Some PDI codes (i.e. some grapho-phonetic difficulties) tend to cooccur in English words with higher than chance frequency. Words identified in this manner as 'knots' of the difficulty network can then be used for EFL teaching, dictionary making, natural language processing, linguistic experimentation and theory building, etc. Further applications of PDI-code cooccurrence are foreseen, such as building a lattice of qualitative phono-lapsological lexical equivalence within the English lexicon.

1. Introduction

In my Mikorzyn'04 workshop (Sobkowiak 2004), I presented the genesis, rationale, the main functionalities and the then current implementation of the Phonetic Difficulty Index (PDI). I briefly listed the PDI bibliography and speculated on the potential didactic applications of PDI in the EFL context. In discussing the latter, I paid particular attention to the recently developed qualitative PDI tagging, originally promised in passing in my 1999 book: "Third, because the index in its full version will contain information about the exact nature of the phonetic difficulty involved, it will allow the user to investigate it directly through listing words with this same difficulty" (Sobkowiak 1999: 215). I argued that such qualitative tagging, alone or in combination with the original quantitative tagging of aggregate numerical difficulty measure, carries a great potential for (electronic) EFL/FLT lexicography in particular, as well as language teaching/learning and applied linguistics in general. Some of this potential has been demonstrated in my contribution to the Workshop in Assessing the Potential of Corpora, at the 35th Poznań Linguistic Meeting, in May 2004. I argued there that: "on top of automatic phonetic transcription of raw text, which is now conceptually and technologically rather trivial, a sophisticated L1-sensitive automatic phonetic annotation is feasible, with a variety of EFL-related functions, in particular text/sentence selection" for psycholinguistic and pedagogical experiments (Sobkowiak, in press).

---

[1] I am grateful to Dr. Przemek Kaszubski and Dr. Robert Lew for discussing Mutual Information with me. Errors are my own.

Because I briefly summarized all of these developments in my Mikorzyn paper, and because space alloted to this contribution is limited, I will not dwell upon the structural, functional and applicational details of PDI in its shape as of a year ago. What I propose to do in this paper is to develop the "recent research" section of the Mikorzyn handout, the material which did not make it into the published text of the proceedings. To fully understand the following discussion the reader will need access to the complete list of PDI difficulty codes, which also appeared in the Mikorzyn handout, but was not printed in the conference volume. This list is reproduced in Appendix 1.

2. Coocurrence of phonetic difficulty codes

2.1. PDI in brief

The algorithm assigning PDI numerical difficulty tags and qualitative difficulty codes (see Sobkowiak 2004 for an excerpt of the algorithm listing) was run over the machine-readable *Oxford Advanced Learner's Dictionary of Current English* (OALDCE) word-list (see Mitton 1986 and 1992). It generated the PDI range between 0 and 10, with a mean of 2.45, and standard deviation 1.5. The list currently counts 85431 (unlemmatised) records and 25264 lemmas. On top of the global numerical rating of phonetic difficulty, the PDI algorithm now assigns 57 qualitative difficulty codes taken from the list reproduced in Appendix 1. The following, for example, are all four lemmas with PDI=9 in the database, with codes on the right:

Table 1. Examples of words with PDI=9, with their difficulty codes

word	PDI codes
entourage	bgsGNQT13
misbehaviour	bgACORV13
undervaluation	EJQSTX123
undistinguishable	vEHJQTX23

By looking up the phonetic difficulty list in Appendix 1 the reader will be able to ascertain exactly which potential Polglish pronunciation problems have been recognized for each of the four words. The list has no pretense to being definitive, of course; it is simply an interim, doubtless personally biased, codification of general professional wisdom and experience among EFL teachers in Poland (see Sobkowiak 2004 for further disclaimers and suggestions for improvement).

2.2. The internal structure of the phonetic difficulty list

As can be seen in Appendix 1, the list of difficulties, while claiming no strict phonetic organization, is loosely divided into three sections: (a) mostly spelling and morphology, (b) mostly pronunciation, and the convenient (c) others. The PDI algorithm is now being thoroughly revised, including the difficulty list, but the shape of it reproduced here is exactly the same as that presented a year ago in Mikorzyn, including some controversial PDI decisions and assignments. It might, for example, be justifiably argued that codes such as <m> and <o> could/should be discarded, as they only account for potential pronunciation problems of very low

incidence. Many such observations could be made, and the PDI list can doubtless be improved in many respects.

What is of more consequence for this contribution is the observable correlation between some difficulties. Some orthographic and phonemic sequences are necessarily associated by force of English grapho-phonemic equivalences. Similarly with links across the spelling and 'other' sections of the list. Take code <s>, for example: if word-final spelling <age_> occurs in the lemma, which is not phonetically realized as /eɪdʒ_/, this is bound to correlate highly with code <U>, tagging post-alveolar affricates. The two difficulties are of course different in terms of their linguistic source, lapsological characteristics and frequency in the English lexicon, but they are of necessity highly correlated, nevertheless. Notice parenthetically that the association is not balanced: code <s> implies code <U>, but not necessarily vice versa. On the basis of these observations one could predict that there will be a sizable set of English words where both difficulties occur together. As a matter of fact, there are exactly 332 such records in my lexical database, which accounts for 87.1% of all <s>-code words and 4.3% of all <U>-code words, including such frequent items as: *advantage, encourage, image, language, manage, village*. As a matter of interest: the 13% words spelled <-age>, but pronounced without an affricate, are all French /-ɑʒ_/ borrowings, including *garage* as the most textually frequent item.

Upon careful analysis of the difficulty list, many such phonetically trivial inter-difficulty links could be discovered. What I propose to do in the rest of this paper, however, is to see if one could use some statistical tools to discover coocurrence patterns going beyond the phonetically trivial, which would at the same time exhibit some statistical significance. The general question to pose will be: what are the most frequent coocurrence patterns of phonetic difficulties within English words? In order to answer this question, a statistical measure of coocurrence is needed which could be derived automatically from the PDI-tagged lexical database, as described above. The following section will briefly introduce this statistic.

2.3. Mutual Information

Mutual Information (MI) "compares the probability of observing *x* and *y together* (the joint probability) with the probabilities of observing *x* and *y independently* (chance). If there is a genuine association between *x* and *y*, then the joint probability *P(x,y)* will be much larger than chance *P(x) P(y)*, and consequently *I(x,y)*»0. If there is no interesting relationship between *x* and *y*, then *P(x,y)* ≈ *P(x) P(y)*, and thus *I(x,y)* ≈ 0. If *x* and *y* are in complementary distribution, then *P(x,y)* will be much less than *P(x) P(y)*, forcing *I(x,y)*«0" (Church & Hanks 1990:23). Computationally, MI is usually expressed as: $MI(x,y)=\log_2\{P(x,y)/P(x)*P(y)\}$, i.e. on a logarithmic scale. While there are many complications and unresolved controversies concerning the arithmetic, statistical and linguistic issues in the use of mutual information scores, I will steer away from them because: (a) they have no place in this contribution, (b) nothing in the results below crucially depends on them, and (c) I have no pretense to creatively developing information theory. The interested reader will find a lucid introduction in Church & Hanks 1990, Church et al. 1991, Stubbs 1995, and sources listed therein.

From the informal Church & Hanks exposition above it will be readily understood why MI is a convenient tool (with all the mentioned provisos) for discovering significant patterns of PDI code coocurrence within words. The *x* and *y* events are of course occurrences of specific PDI codes: these can be counted *independently*, i.e. with reference to the whole 'population' of such codes in

my lexical database (see Appendix 1: summed incidence of all counted difficulties = 208953), or interdependently, i.e. in terms of their lexical *coocurrence*. The two counts can then be compared yielding the MI score, calculated as above. If the codes cooccur in words significantly often, compared to chance, there is reason to suppose that there is some 'magnetic' force of phonetic (?) attraction drawing them together.

Notice, additionally, that: (a) the MI algorithm is computationally rather simple, and can be easily run over my lexical database containing the PDI codes, (b) there are statistical methods to filter out spurious results due to varied code incidence in the database (see Stubbs 1995 for the *T-value*), and (c) MI can accommodate not only pairs of variables (bigrams), but also n-grams with an arbitrary value of n. With reference to (c), notice, however, that both computationally and conceptually, matters get rather complicated (see Banerjee & Pedersen 2003:377); hence I will not venture beyond trigrams in this paper.

2.4. Coocurrence of PDI code bigrams in the lexical database

The full list of all PDI n-grams was derived from my lexical database in January 2004 by Mr. Grzegorz Krynicki, to whom I hereby express my gratitude. Generally speaking, there are 208953 single PDI codes (monograms) in the database. The number of wordforms with PDI=0, i.e. without any difficulty codes attached to them, is 7265 (8.5% of all words). The incidence of particular codes taken singly (monograms) is provided in Appendix 1. The basic statistics of the n-grams for n>1 are presented in Table 2.

Table 2. Some basic statistics of PDI code n-grams

n	tokens of attested n-grams	top n-gram	top n-gram frequency
2	254638	J1	16409
3	202205	JN1	6466
4	113873	aJN1	996
5	47270	JNTY3	260
6	14598	JNTY13	117
7	3348	JNTY123	17
8	561	bdJKNQU1	6
9	64	aAJKQT123	2
10	4	abAJKQT123	2

Notice that without using any further statistical machinery, some interesting phono-lapsological observations about the English lexicon can be made on the basis of raw frequency data. First, short schwa (code <J>) is expectedly the most frequent difficulty, not only singly, but also in coocurrence with a variety of other pronouncing problems. Second, the most common PDI n-grams with a given n tend to 'inherit' some difficulty coocurrences from n-grams with n-1. This is very obvious for the 2-7 series, whereas the most common 8-gram introduces some new elements: codes <K> and <Q> (long schwa and vowel nasalization). Third, some coocurrences are quite pervasive, for example <J1>, which is the most common bigram, and occurs in all top-frequency n-grams in Table 2, with the exception of the 5-gram <JNTY3>.

This last observation also demonstrates rather persuasively that there are limits to insights obtained from raw frequency data alone. Certainly, considering that both <J> (short schwa) and <1> (British pronunciation different from American) are independently very frequent in the database (see Appendix 1), their frequent coocurrence in English words need not be statistically (and hence phonetically) interesting. It is at this point that the subtler statistical tools, such as MI can be used to some advantage. MI compares the events' independent and combined frequencies; the higher the latter with respect to the former, the higher MI. Thus, we can expect that the raw-data salience of independently frequent codes, such as <J>, <N> or <1> and their combinations will be somewhat attenuated once MI is calculated. This is exactly what we observe in Table 3, which contains nineteen top-MI bigrams (out of 959 different bigrams) in my database[2]. Actual phonetic difficulty labels are copied from Appendix 1 for ease of reference.

Table 3. Bigrams with highest MI

	bigram code	**PDI difficulty**	**frequency**	**MI**
1.	OR	<pre-voiced /dɪs/ or /mɪs/> + <voiced obstruent + /s/ or /s/ + voiced obstruent>	133	5.60
2.	VX	<<able_> in head and not /eɪbl_/> + <word-final syllabic sonorants>	485	5.39
3.	X4	<word-final syllabic sonorants> + <ical_> in trisyllabic-plus adjectives – stress>	133	5.38
4.	cr	<<ei> in word> + <<gh_> or <ght_> in head>	81	5.38
5.	bD	<<ur> in word> + </ʊə/>	484	4.95
6.	FU	</tʃt_/ or /dʒd_/> + < post-alveolar affricates>	518	4.48
7.	sU	<<age_> in head and not /eɪdʒ_/> + < post-alveolar affricates>	332	4.29
8.	w9	<<ey_> in head and not /eɪ_/> + <proper noun>	76	4.18
9.	HI	<velar nasal> + </ŋ/+V with no /g/>	141	4.09
10.	bK	<<ur> in word> + < long schwa>	1135	4.08
11.	dC	<<eo> in word> + </ɪə/>	140	4.07
12.	23	<more than 5 syllables> + <secondary stress>	727	4.01
13.	aW	<compound> + <stop geminates>	123	3.92
14.	ST	</ueɪ/ or /ieɪ/> + <post-alveolar fricatives>	212	3.36
15.	TX	<post-alveolar fricatives> + <word-final syllabic sonorants>	1650	3.36
16.	bg	<<ur> in word> + <<ou> in word>	692	3.24
17.	30	<secondary stress> + <abbreviation, incl. acronym>	198	3.18
18.	O3	<pre-voiced /dɪs/ or /mɪs/> + <secondary stress>	420	3.14
19.	AC	<linking /r/> + </ɪə/>	748	3.01

Because MI is a measure which does not "allow for significance to be assigned to their value" (Banerjee & Pedersen 2003:376), a cut-off point of MI=3 has customarily been accepted on the strength of empirical evidence (see Church & Hanks 1990, Clear 1993 and Stubbs 1995). There are 53 bigrams above this threshold in my data, 43 of which show T-value (Stubbs 1995) higher than 2, which ensures that extremely low-frequency bigrams are not considered, even if their MI is relatively high. Nineteen of these appear above, arranged by decreasing MI.

[2] The list has been adjusted to filter out extremely low-frequency bigrams by two methods, (a) by multiplying MI by log frequency (Kilgarriff & Tugwell 2001:189) and (b) by counting T-value with a cut-off point at 2 (Stubbs 1995).

Many of these coocurrences are admittedly rather phonetically trivial. The top three MI scores, for example are clearly cases of PDI criteria overlap, and as such could be predicted from the PDI difficulty list itself, after some scrutiny. In such cases MI can be used as an empirical test of PDI criteria validity: should <V> not correlate with <X>, there would be something wrong with the PDI list. For phono-lapsological and didactic applications, however, it is the other high-MI cases that are much more interesting. Notice, in particular, <cr>, <w9>, <aW>, <ST>, <TX>, <30>, <O3> and <AC>. These correlations are arguably non-trivial in that they are not an automatic consequence of the structure of the PDI phonetic difficulty list.

Thus, there is no a priori reason why <ei> and <-gh(t)_> should cooccur with higher than chance frequency in English spelling, even if it is not particularly hard to think of a few words which do exhibit both strings: *eight, freight, height, sleigh(t), weigh(t)*. Some others, however, may not come to mind so easily: *aweigh, inveigh, neigh*. Or take <O3>: /mɪs-/ and /dɪs-/ in pre-voice position cooccur with secondary stress; this may 'feel' post-hoc obvious in view of the English phonotactic fact that the two prefixes often occur in longish words. But notice that: (a) it is easy to justify the high-MI cooccurence of some PDI codes post-hoc, but it may be much harder to predict them, (b) there are many words shorter than, say, five syllables, which exhibit the <O3> bigram: *dismount, misdate, misdeal, misdeed, misgive, misguide, misjudge, mislay, mislead, misname, misread, misrule, misuse* are all bisyllables with a secondarily stressed prefix. The reader is invited to try to make up lists of English words for the other phonetically non-trivial bigrams: <w9>, <aW>, <ST>, <TX>, <30> and <AC>. Some English lemmas illustrating these are listed in Appendix 2.

2.5. Coocurrence of PDI code trigrams in the lexical database

Now consider briefly a few top-MI trigrams appearing in my lexical database. Data presentation below will be analogous to that in section 2.4. Of all 5109 trigram types, 1590 are those where MI>3 and T-value>2, i.e. showing both significant coocurrence (by MI), and satisfactorily high frequency (by T)[3]. Table 4 contains some of the phonetically most salient trigrams. Only such trigrams are listed which do not directly 'inherit' their high coocurrence from that of their three bigrams contained therein and listed in Table 3 (compare <STX>, MI=4,90, with <ST> and <TX>) .

Table 4. Trigrams with highest MI

	bigram code	PDI difficulty	frequency	MI	some example words
1.	BC2	</eə/> + </ɪə/> + <more than 5 syllables>	43	9.43	*authoritarian, disciplinarian, egalitarian, equalitarian, humanitarian, latitudinarian, nonagenarian, octogenarian, parliamentarian, septuagenarian, sexagenarian, totalitarian, utilitarian, valetudinarian, veterinarian*
2.	UX4	<post-alveolar affricates> +	43	8.87	*archaeological, biological, geographical, geological,*

[3] This condition excludes, for example, <HJ1>, with MI=1.41, the trigram mentioned in my Mikorzyn handout, i.e. before I applied the significance statistics to PDI n-gram calculations.

		<word-final syllabic sonorants> + <<ical_> in trisyllabic-plus adjectives – stress>			*ideological, psychological, sociological, technological*
3.	vEX	<<able_> in head and not /eɪbl_/> + </ʌ/> + <word-final syllabic sonorants>	111	8.24	*adjustable, clubbable, come-at-able, comfortable, culpable, cultivable, insufferable, justifiable, lovable, punishable, recoverable, sufferable, touchable, vulnerable* (<un>-prefixed words were filtered out)
4.	vQX	<<able_> in head and not /eɪbl_/> + <vowel nasalization> + <word-final syllabic sonorants>	97	8.19	*considerable, constable, uncomfortable, understandable, unreasonable*
5.	vX3	<<able_> in head and not /eɪbl_/> + <word-final syllabic sonorants> + <secondary stress>	95	7.72	*biodegradable, certifiable, disagreeable, identifiable, justifiable, realizable, reconcilable, specifiable, utilizable* (<un/in/ir>-prefixed words were filtered out)
6.	TY3	<post-alveolar fricatives> + <non-word-final syllabic sonorants> + <secondary stress>	583	6.94	*evolutionary, insufficient, organizational* (<-tions> words were filtered out)
7.	bDJ	<<ur> in word> + </ʊə/> + <short schwa>[4]	201	6.72	*assure, bureaucrat, insurance, mature, obscure, rural, tourism*

As, by now, the rationale, the calculation details, the functionalities and the provisos of PDI n-grams should be well understood, I will abstain from further analysis of trigrams in Table 4 at this point.

2.6. Potential applications

Some of the potential uses of PDI coocurrence n-grams, as presented in this paper, derive from those of simple PDI tags (monograms), as foreseen in my 1999 book and the ensuing series of papers summarized in my Mikorzyn contribution. These include primarily (phono)lexicographic and pedagogical. On top of this, "outside of the narrowly defined EFL arena, the potential of the PDI appears to be the greatest in (a) phonetics and phonology, (b) (meta)lexicography and lexicology, (c) lexical psycholinguistics, (d) contrastive and corpus linguistics, (e) natural language programming, especially automatic speech recognition (ASR), speech synthesis (TTS), and their applications in machine translation (MT), robotics, data mining, abstracting, and the like" (Sobkowiak 2004).

In this last context, for example, consider the following (pedagogically rather trivial) remark of some leading experts in the field: "A cross-reference mapping of linguistic features for each language is therefore desirable in order to make predictions about what kinds of difficulties a student is likely to have. One solution to this problem is to train specific models to detect mispronounced phonemes based on the phonetic properties of both the mother tongue and the

[4] Notice that the diphthong is coded monosegmentally in my database, so that <short schwa> is not part of it.

target language" (Deroo et al., 2000). A quantitative L1-sensitive PDI measure, i.e. the aggregate numerical index of difficulty would seem to be extremely useful as an expert knowledge database to train automatic speech recognizers and tutors. A qualitative monogrammatic PDI tagging would obviously be even more beneficial, because the machine would now know what kinds of pronunciation errors to expect. But a suitably coded lexical database with PDI n-gram tagging, like I have presented here, would obviously be best of all. The recognizer/tutor could now concentrate on those lexical items where pronunciation problems characteristically cooccur, making them ideal test-beds for tweaking and perfecting the software. All this is true, ceteris paribus, of human teachers, tutors, testers, materials-designers, as well as learners themselves.

As far as 'pure' phonetics and phonology are concerned, I would like at this point to support a methodological point forcefully presented by Stubbs in his 1995 paper on collocations and semantic profiles. Says Stubbs:

> Chomsky's (1957, 1965) rejection of induction, by machines or humans, is still widely assumed to be valid. In his attack on American structuralism, he rejects the concept of "discovery procedures". But he provides no real arguments against such methods, merely stating that linguistic theory is not "a manual of procedures", and asserting that there are simply no practical and mechanical ways of extracting a grammar from a corpus of utterances [...] I have emphasized throughout that no procedures can ever be entirely automatic. We always start with intuitions about what is interesting to study, and intuition re-enters, in designing procedures and in interpreting findings. But, given such caveats (which apply to any study of anything), quantitative procedures can identify lexical sets largely on the basis of the frequency and distribution of lexical items in a corpus, leaving the human analyst to discard a few irrelevant collocates which the procedure throws up (due to the idiosyncratic content of corpora), and to interpret the resulting lexical sets.

By analogy, I am ready to claim after Stubbs (but not ready to argue here for lack of space) that quantitative procedures, like the PDI n-gram cooccurence calculations presented above, can automatically identify (candidates for) phonetically, phonotactically, phonologically and lapsologically interesting (phono-)lexical sets, facts, rules, regularities, generalizations, tendencies and observations, i.e. the very groundwork of models and theories in the area of pronunciation teaching. While "the human analyst [would indeed have] to discard a few irrelevant collocates [i.e. n-gram cooccurences – WS] which the procedure throws up", like I did collecting data for Tables 3 and 4, the potential of algorithms such as PDI n-gram counting as a theory construction tool is in my opinion beyond reasonable doubt.

3. In place of conclusions

To follow the Mikorzyn tradition, I will close with a glimpse at my PDI-related work in progress. Consider the following perspective: lexical items in my database differ in terms of their PDI in two respects: quantitative and qualitative. First, some words are PDI-harder from others; second, most words, possibly with the same numerical difficulty index, display difficulty code subsets different from all other words. It would be interesting to see how words are located on the PDI-code similarity/equivalence scale. In particular: is it possible to have pairs of phonemically different words which would be identical in terms of their PDI codes? Such words would not only exhibit the same phonetic difficulty level, but they would also be qualitatively phono-lapsologically identical. Would such pairs be interesting/useful in any phonetic sense? Some preliminary answers to such questions appear to be affirmative. Consider, for example, the following two words: *lightning-conductors* and *scandalmongers*. Despite their obvious phonemic non-equivalence, they are identical in terms of PDI, both quantitatively (PDI=7) and

qualitatively: <aEHJN13>. One can thus expect identical pronouncing problems to arise for the learner in processing these words (both for human-producer and machine-recognizer, by the way). This kind of strong qualitative phono-lapsological equivalence holding for word-pairs and subsets in the English lexicon may also have other, equally interesting phonetic ramifications. These might be the theme of my contribution to the sixth phonetics teaching conference in the year 2006.

Bibliograhy

Baker,M., G.Francis & E.Tognini-Bonelli (eds). 1993. *Text and technology. In honour of John Sinclair*. Philadelphia: Benjamins.

Banerjee,S. & T.Pedersen. 2003. "The design, implementation, and use of the ngram statistics package". In A.Gelbukh (ed.). 2003. 370-81.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, Mass: MIT Press.

Church,K., W.Gale, P.Hanks & D.Hindle. 1991. "Using statistics in lexical analysis". In U.Zernik (ed.). 1991. 115-64.

Church,K.W. & P.Hanks. 1990. "Word association norms, mutual information, and lexicography". *Computational Linguistics* 16.1. 22-9. (http://acl.ldc.upenn.edu//J/J90/J90-1003.pdf)

Clear,J. 1993. "From Firth principles. Computational tools for the study of collocation". In M.Baker et al. (eds). 1993. 271-292.

Deroo,O., C.Ris, S.Gielen & J.Vanparys. 2000. "Automatic detection of mispronounced phonemes for language learning tools". Proceedings of the 6th International Conference on Spoken Language Processing, Beijing, China, October 16-20, 2000. (http://tcts.fpms.ac.be/~ris/postscript/icslp00_1.ps.gz)

Gelbukh,A. (ed.). 2003. *Computational Linguistics and Intelligent Text Processing*. 4th Int Conf, CICLing 2003, Mexico City, February 16-22, 2003. Heidelberg: Springer-Verlag.

Kilgarriff,A. & D.Tugwell. 2001. "WASP-Bench: an MT lexicographers' workstation supporting state-of-the-art lexical disambiguation". Proceedings of MT Summit VII, Santiago de Compostela. 187-190. (http://www.itri.bton.ac.uk/~David.Tugwell/mt.rtf)

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.

Sobkowiak,W. 1999. *Pronunciation in EFL machine-readable dictionaries*. Poznań: Motivex.

Sobkowiak,W. 2004. "Phonetic Difficulty Index". In W.Sobkowiak & E.Waniek-Klimczak (eds). 2004. 102-107.

Sobkowiak,W. (in press). "Automatic phonetic annotation of corpora for EFL purposes". Paper presented at the Workshop in Assessing the Potential of Corpora, 35th Poznań Linguistic Meeting, May 20, 2004.

Sobkowiak,W. & E.Waniek-Klimczak (eds). 2004. *Dydaktyka fonetyki języka obcego.* Zeszyt Naukowy Instytutu Neofilologii Państwowej Wyższej Szkoły Zawodowej w Koninie nr 3. Konin: Wydawnictwo PWSZ w Koninie.

Stubbs,M. 1995. "Collocations and semantic profiles: on the cause of the trouble with quantitative studies". *Functions of Language* 2.1. 23-56. (http://www.uni-trier.de/uni/fb2/anglistik/Projekte/stubbs/ cause.htm)
Zernik,U. (ed.). 1991. *Lexical acquisition: exploiting on-line resources to build a lexicon*. Hillsdale,N.J.: Lawrence Erlbaum Associates.

Appendix 1: PDI codes used in the lexical database with their incidence (in brackets)

(a) mostly spelling and morphology

a -- compound	(11148)
b -- <ur> in word	(3145)
c -- <ei> in word	(623)
d -- <eo> in word	(427)
e -- <ow> in word	(1609)
f -- <au> in word	(991)
g -- <ou> in word	(3992)
h -- <aw> in word	(582)
i -- <lk_> in head	(132)
j -- <mb_> in head	(117)
k -- <mn_> in head	(36)
l -- <alm_> in head	(32)
m -- <gm_> in head	(9)
n -- <stle_> in head	(83)
o -- word=<mn>	(2)
p -- word=<ps>	(65)
q -- word=<al>C; C<>l	(129)
r -- <gh_> or <ght_> in head	(534)
s -- <age_> in head and not /eɪdʒ_/	(381)
t -- <ate_> in head and not /eɪt_/	(238)
u -- <ative_> in head and not /eɪtɪv_/	(163)
v -- <able_> in head and not /eɪbl_/	(510)
w -- <ey_> in head and not /eɪ_/	(278)

(b) mostly pronunciation

A -- linking /r/	(4787)
B -- /eə/	(1129)
C -- /ɪə/	(3337)
D -- /ʊə/	(851)
E -- /ʌ/	(8394)
F -- /tʃt_/ or /dʒd_/	(518)
G -- interconsonantal /ʊ/, but not <oo>	(1419)
H -- velar nasal	(10044)
I -- /ŋ/+V with no /g/	(141)
J -- short schwa	(32192)
K -- long schwa	(3639)
L -- voiced apico-dental	(724)
M -- voiceless apico-dental	(1803)
N -- final voiced obstruent	(31427)
O -- pre-voiced /dɪs/ or /mɪs/	(790)
P -- /əʊ/CCV	(784)
Q -- vowel nasalization	(7612)
R -- voiced obstruent + /s/ or /s/ + voiced obstruent	(594)
S -- /ueɪ/ or /ieɪ/	(496)
T -- post-alveolar fricatives	(7132)
U -- post-alveolar affricates	(7631)
V -- glottal fricative /h/	(4267)
W -- stop geminates	(125)
X – word-final syllabic sonorants	(3862)
Y – non-word-final syllabic sonorants	(2893)

(c) others

1 -- british<>american	(31710)
2 -- more than 5 syllables	(750)
3 -- secondary stress	(10351)
4 -- <ical_> in trisyllabic-plus adjectives -- stress	(141)
5 -- <ic_> in bisyllabic-plus adjectives -- stress	(477)
7 -- <ary_>/<ory_>/<ery_> in bisyllabic-plus heads	(717)
8 -- contraction of pronoun with verb, e.g. <you've>	(38)
9 -- proper noun	(2589)
0 -- abbreviation, incl. acronym	(363)

Summed incidence of all counted difficulties	**208953**
Sum of all wordforms in the database	**85431**
Average PDI per wordform	**2.45**

Appendix 2: Some phonetically interesting PDI bigrams

bigram code	PDI difficulty	example words
w9	<<ey_> in head and not /eɪ_/> + <proper noun>	*Audley, Audrey, Barnsley, Batley, Beverley, Bewdley, Bexley, Bingley, Birtley, Bletchley, Bromley, Buckley, Burley, Burnley, Canvey, Canvey Island, Chorley, ...*
aW	<compound> + <stop geminates>	*about-turn, aide-de-camp, backcloth, blackcurrant, blood-donor, boat-train, book-keeping, bookcase, bookclub, brickkiln, cart-track, cock-crow, dirt-track, flattop, freight-train, frock-coat, ...*
ST	</ueɪ/ or /ieɪ/> + <post-alveolar fricatives>	*appreciate, associate, aviation, continuation, deviation, differentiate, evaluation, fluctuation, initiate, negotiate, radiation, recreation, situation, valuation, variation*
TX	<post-alveolar fricatives> + <word-final syllabic sonorants>	*action, application, association, attention, condition, decision, education, election, especial, financial, information, international, operation, population, position, production, relation, section, situation, social, special*
30	<secondary stress> + <abbreviation, incl. acronym>	*am, asap, cc, cf, cif, cm, cp, eg, et al, et seq, eta, etd, ie, lbw, loc cit, mod cons, mpg, mph, op cit, pm, qv, rpm, sae, wpb, wpm*
AC	<linking /r/> + </ɪə/>	*appear, atmosphere, barrier, beer, behaviour, career, carrier, clear, dear, disappear, ear, engineer, familiar, fear, gear, hear, here, interfere, ...*