EFL Wordstation

Włodzimierz Sobkowiak

School of English, Adam Mickiewicz University, al.Niepodłeglosci 4, 61-874 Poznan, Poland

KEYWORDS: machine-readable dictionaries, EFL

E-MAIL: sobkow@hum.amu.edu.pl
FAX NUMBER: (48-61) 523-103
PHONE NUMBER: (48-61) 528-820

0. Introduction

There is an amazing variety of machine-readable English dictionaries (MRDs) now available in terms of required hardware, software platform, design, size and purpose. Yet, there are relatively few MRDs which were custom-made for learners of English as a foreign language. The Collins Cobuild and Longman Interactive readily come to mind, of course, and it is obvious that from the point of view of EFL they are much more useful than, say, the second edition of the OED on CD-ROM. Even these excellent resources, however, share the two possibly most damaging weaknesses of all EFL-oriented MRDs: **L1-insensitivity** and **access-inflexibility**.

In this paper I will describe an English-Polish MRD now in preparation in the School of English, Adam Mickiewicz University, Poznan, Poland. This MRD avoids the flaws mentioned above, and thanks to its radically innovative design deserves the proud name of **Wordstation**.

1. What is available?

Monolingual MRDs, just like other, technologically more traditional EFL resources are made for the "generic" learner, i.e. the learner of an indefinite L1 background. By this token, they cannot take into account those factors whose exclusion is perceived by most EFL teachers and methodologists (including the undersigned) as inadmissible, for example the notorious L1-to-L2 transfer on all levels of language structure. Even the existing bilingual English-Polish MRDs, of which there are a few, are only nominally L-1 sensitive. Most of them are hastily prepared word lists, with practically no phraseology, grammatical or usage advice and thesaurus functions, not to mention such niceties as phonetic transcription[1] or sense subcategorization. Usually no morphological normalization is built in, so that the EFL user must first reduce *children* to *child* him/herself, only to find the following entry (in one English-Polish MRD): *child node, child process, child task*, duly translated into Polish computerese.

Another deficiency in English-Polish MRDs is their inflexibility. While some monolingual English MRDs allow wildcard and definition (or reverse) searches, this is a feature conspicuously absent from Polish EFL MRDs. More specifically, but very importantly in this context, among all the off-the-shelf English MRDs known to me only the OED2 provides for (poorly implemented) phonetic access, whereby users can search for words through their pronunciation, an indispensable feature if learners are to be convinced that the foreign language which they are acquiring is mostly spoken. Thus, the only type of dictionary access offered in most existing MRDs remains the thoroughly traditional spelling lookup, which is a rather obvious waste of the available technology.

2. Multi-access in the EFL workstation

Following the recommendations of the more far-sighted computer lexicographers (e.g. Calzolari 1989 or Knowles 1990), we have decided to offer the user of the Wordstation the maximum number of access paths to our 60-thousand-wordform English-Polish MRD. Thus, the user will not only be able to look up a word as s/he would do in a hard-copy dictionary (and see it on screen in comparable graphical layout), but also formulate arbitrarily elaborate queries yielding carefully filtered word-lists, which can then be used for practice, drilling, material-writing, testing, or simply serendipitous browsing.

Starting with the lowest level of language structure: **phonetic access** (see Sobkowiak 1994 a,b,c). All pronunciation-related information built into the dictionary will be readily available: (a) the phonetic transcription of the British and American pronunciation of each headword and wordform, (b) segmental length in letters and phonemes, (c) accentual pattern (primary and secondary stress), (d) syllabic length, boundaries and C/V structure, (e) some EFL-wise important distinctive features like voicing in obstruents or vowel tenseness.

All this phonetic information will be made available to the user from a simple and easily customizable menu so that queries are possible of the type: "What English words have two syllables with a syllabic nasal at the end?", or "Give me all trisyllables with primary stress on the penult", or "What words retain unstressed vowels or diphthongs in unstressed syllables?". On the other hand, users will be able to reduce the phonetic display at will should they not currently be interested in the pronunciation of words looked up. This reduction can be quite radical, to the point of avoiding any phonetic information whatsoever. This principle applies equally to other types of lexical information in this MRD.

Second: **frequency access**. Word frequency has only recently been recognized as an important lexical datum worthy of inclusion in a learner's

dictionary (see the recent line of corpus-based learner's dictionaries and resources published by Collins and Longman (The former based on the COBUILD corpus, the latter on the Bank of English corpus). In our Wordstation both written (printed) and spoken word frequency will be provided, so that the learner will be able to (using simple syntax and menus) formulate queries like the following: "Which common English words (say, among the first thousand in rank) are relatively [2] more frequent in writing than in speech?". Exact frequency data will be available if required (unlike in the Collins and Longman dictionaries), so that material-developers and EFL researchers may use these on top of such global frequency codes as rare.

Third, as much of the **morphosyntactic information** will be accessible for active searching as is practically implementable. Derivatives and inflectional forms will of course be listed, but properly linked to their headwords, and the user can decide how s/he wants to view them. Other morpho-syntactic information encoded and potentially exploited in lexical searches will be (unsurprisingly): (a) part-of-speech tagging with subcategorization where appropriate and useful (verb transitivity, tense-form and 3rd ps sg, noun plurality, adjective grading, zero-derivation, etc.), (b) compound flag to allow easy retrieval of compounded forms, which must be differentiated from hyphenated strings, fore-stressed polysyllabic nouns and multi-word sequences with separating spaces, (c) important idiosyncrasies in the morphological and syntactic behaviour of words, for example *pluralia tanta* or the fact that some transitive verbs (e.g. *lack, survive, thank, deserve, undergo*, etc., cf Kjellmer 1992:341) are practically never passivized in natural English, both areas extremely problematic for EFL learners.

Thus, in addition to phonetic searches or in combination with them the user will be able to formulate queries like: "Which verbs are not regularly inflected for past tense from their base form?" (most of the so-called 'irregular verbs'), or "List all nominal binomial compounds with a deverbal first term" (e.g. *looker-on*). The query language and user interface will allow this level of sophistication without the requirement of the elaborate metalanguage or grammatical jargon, as in the examples above.

Fourth: **phraseology, idiom, style**. Currently no comprehensive contextual exemplification is envisaged as part of the Wordstation. In a bilingual dictionary this function can to some extent be fulfilled by translation. Yet, typical English collocations will normally be listed for two categories of words: (a) those in some respect irregular, idiosyncratic or collocationally restricted (e.g. the noun *abandon*, practically untranslatable into Polish when outside the phrase *with (reckless) abandon*), and (b) most of the so-called 'function' words, especially articles, prepositions and some pronouns. Additionally, acronyms and abbreviations will be expanded, again with idiosyncrasies marked (e.g. *AAA* is *Amateur Athletic Association* in GB, but *Automobile Association of America* in the US).

This access mode will give the EFL user some idea of the collocability of the most troublesome words, another area which is commonly regarded as one of the most difficult in EFL instruction. The realization that a given word or class of words do not exhibit collocational idiosyncrasy is an important EFL datum in its own right.

Fifth: **meanings**. Three avenues of semantic access will be implemented: (1) through Polish translation, which may be thought of as a direct analogue of definition search mentioned above, (2) through about 60 semantic field labels, like *tools, army, animals, space, bedroom, colours*, added to nouns with high concreteness ratings, and (3) through a thesaurus facility containing common synonyms and antonyms of the word, where relevant. Using all this information the user will be able to formulate queries like: "Which synonyms of *aberration* have a medical meaning (semantic field *medicine*)?". Answer: *abnormality, delusion, dementia, derangement, hallucination, insanity, mania, psychosis.*

3. L1-sensitivity in the EFL Wordstation

Let us start the discussion of EFL MRD L1-sensitivity with errors. It is very useful, but by no means sufficient, in an EFL MRD to append — as the Longman Interactive does — a note of 'common error' committed by 'generic' learners of English in connection with a given lexical entry. There may well be a common denominator for EFL errors made by learners of varying language backgrounds, but this is a relatively restricted area that needs to be complemented by a careful treatment of L1-specific errors.

In our Wordstation Polish interference in the Polish-English interlanguage (or Polglish, for short) is fully taken into account. First, on the level of pronunciation, we have designed a **simplified Polglish phonetic transcription** (see Sobkowiak, in press) whereby beginning-level users can enter English words as they hear them (phonetic access) and the way they would spell them in Polish, rather than in the somewhat cryptic IPA transcription or its derivate. For example, entering *szoł* (with the slashed Polish l) would correctly retrieve *show* [ʃɜu], with all its associated lexical information. By using Polglish transcription in this manner we are making the dictionary user-friendly and L1-sensitive on two counts: (a) to Polish spelling, and (b) to Polish pronunciation errors.

This transcription is bound to be inaccurate phonetically and massively ambiguous, of course. These deficiences are mitigated by the fact that it will only be used heuristically, i.e. to help active searches,and not representationally: proper IPA representation will appear on screen (if desired by the user). Fuzzy search programming (as in spell-checkers) will take care of excessive or zero hits.

Another way in which L1-sensitivity is manifest in the phonetic stratum of the dictionary is the **phonetic difficulty index**. The index contains two fields: the first is a numerical tag carried by each wordform which indicates its approximate difficulty for the Polish learner; the latter is a code of the actual phonetic difficulty/ies present in the wordform. The index has been derived through a combination of algorithmic and manual tagging and is a result of years of EFL teachers' experience. The difficulty index can be used in a number of ways in the actual word searches and queries. First, it will caution the user as to the high phonetic difficulty of the word currently displayed. Second, the index can be used in direct queries of the type: "Which words of this or that semantic/morphological category are particularly difficult phonetically?", or "Give me the phonetically difficult words of the first 1000 (spoken) frequency rank". Third, because the index contains information about the exact nature of the difficulty involved, it will allow the user to investigate it directly through listing words with this same difficulty present, for example: "If *radio*, which I am now having on screen is pronunciation-wise difficult for Poles because they tend to reduce the second-syllable vowel to a glide /j/ and the whole word to a bisyllable, give me more words with this phonetic problem in them". An exemplary answer to such a query (listed in the order of frequency): *ratio, enthusiasm, appreciate, studio, embryo, associate, abbreviate, negotiate, cardiac, video, dissociate, kiosk, mediate, deviate, humiliate, pistachio, stereo.*

To account for L1-motivated errors originating on higher levels of linguistic structure explicit notes will be taken in the body of the given lexical record of common morphosyntactic and semantic difficulties associated with it, such as (a) translational pseudoequivalents ('false friends'; E(*actual*) P(*aktualny*) 'topical'), (b) homonymy and homophony, (c) homography, (d) all other common Polish errors and problems. The treatment of the last category will be informed — apart from the anecdotal, experiential and intuitive data — by the analysis of the corpus of English essays written by Polish students and collected in our School as part of the International Corpus of Learner English (ICLE) Project coordinated from Louvain, Belgium (cf. Granger 1993). The Polish component of the corpus, which is currently being collected, edited, tagged and parsed, will contain about 200,000 words of argumentative and expository writing at an advanced academic level of English proficiency.

Finally, all the available lexical information contained in the Wordstation will be dynamically used as a database for the CALL-like **vocabulary exercise component** which we plan to build into the package. The scope and detail of data as presented above will make it possible to design a variety of exercise types, including multiple choice ("Which of these objects is not a musical instrument: *banjo, bass, baton, cello*?"), matching ("Connect these Polish words with their American English equivalents: ..."), classification ("What category do these objects belong to: *leek, carrot, potato, peas*?"), identification ("What is the English word for *mały domek mysliwski*?"), elicitation ("Write five words beginning with the prefix *con-*"), ordering ("Order the following words from the most to the least common: ..."), etc.

We believe that with this CALL facility our MRD will transgress the limits of the multi-access L1-sensitive dictionary, and will become a fully fledged EFL Wordstation.

Notes

1 The CD-ROM technology is quite recent in Poland so there are thus far no dedicated multimedia speaking dictionaries of English. Some such facilities appear as parts of certain advanced CALL packages.

2 Relatively is an important word here, of course, as in absolute terms the predominance of speech over writing is such that (in general English) no written word is ever more frequent than its spoken equivalent.

Bibliography

Calzolari, N. 1989. "Computer-aided lexicography: dictionaries and word data bases". In I.S. Batori, W. Lenders and W. Putschke (eds). 1989.*Computational linguistics.* Berlin: Walter de Gruyter. 510-519.

Granger, S.1993."The international corpus of learner English". In J. Aarts, P.de Haan & N. Oostdijk (eds).1993. *English language corpora: design, analysis and exploitation.* [ICAME 13 proceedings].Amsterdam: Rodopi. 57-71.

Kjellmer, G. 1992. "Grammatical or nativelike?". In G. Leitner (ed.). 1992. *New directions in English language corpora.* Berlin: Mouton de Gruyter. 329-344.

Knowles, F.E.1990."The computer in lexicography". In Hausmann et al. (eds). 1990. *Wörterbücher. Dictionaries. Dictionnaires.* Berlin: Walter de Gruyter. 1645-1672.

Sobkowiak, W. 1994a."Phonetic-access dictionaries in TEFL: from vision to project". *Nordlyd 21: 33-41.*

Sobkowiak, W. 1994b. “Phonetic-access dictionaries with L1-based simplified transcription”. Poster presented at the 6th EURALEX Congress, Amsterdam 30.8.-3.9. 1994.
Sobkowiak, W. 1994c. “Beyond the year 2000: phonetic access dictionaries (with frequency information) in EFL”. System 22.4: 509-23.
Sobkowiak, W. (in press). “Radically simplified phonetic transcription for Polglish speakers|”. In S.Puppel & R.Hickey(eds). Festschrift for Professor Jacek Fisiak on His 60th birthday. Berlin: Mouton.