Włodzimierz Sobkowiak, School of English,
Adam Mickiewicz University

Phonetic Transcription in Machine-Readable Dictionaries

Abstract

In MRDs the representative/informative function of phonetic representation is complemented by its indexical function. MRDs can be searched phonetically using any available information on the pronunciation of words: phonemic structure, stress pattern, syllable boundaries, segmental length, etc. The design of MRD phonetic transcription should be sensitive to the requirements imposed by this unorthodox function. The user trying to locate a word by its (transcribed) sound or generating a list of words meeting phonetic criteria will use transcription actively: clarity and ease of use ('friendliness'), consistency and grapho-phonemic bi-uniqueness are essential in this connection.

Three widely used systems of MRD phonetic transcription are discussed from the point of view of their indexical search function. Their relative merits are compared and deficiencies pointed out.

1. Phonetic dictionary access

The main function of phonetic transcription in traditional dictionaries is representational/informative. Together with spelling, etymology, syntactic categorization and meaning(s), it **represents** yet another facet of the word's linguistic identity. It **informs** and **advises** about how the word is/should be pronounced. This function invariably assumes that the reader has already located the word using the orthographic code, which serves as practically the only search key in dictionary look-up. This remains true even in pronunciation dictionaries: one needs to know the spelling of the word to look up its pronunciation.

Exceptions to this principle are found in thesauruses and rhyming dictionaries. In the former (some aspect of) meaning can be used to locate a word, and the latter (e.g. Walker 1890 or Fergusson 1985) organize the vocabulary list according to some kind of reverse phonetic order, so that word pronunciation can be used as a retrieval key to some extent. This last proviso has to do with the fact that rhyming dictionaries (a) are by their nature restricted to rhyming words, (b) are a-tergo (reverse) lists, and (c) are heavily dependent on spelling (Walker included many eye-rhymes, for example).

The only dictionary of English written before the advent of computers which was organized truly phonetically ("arranged alphabetically by pronunciation") and allowed the user to locate words by their pronunciation directly was that by Michaelis & Jones (1913). Among the many uses of phonetic dictionaries envisaged by Michaelis and Jones was that "A person hearing a word for the first time, and being in doubt as to how it should be spelt, may ascertain the current spelling by looking it out in a phonetic dictionary" (Michaelis & Jones 1913:vii). Thus, for the first time in English lexicography phonetic transcription was deployed to actively function as a retrieval key: using the IPA phonetic representation of a word the user could enter the dictionary and find its orthographic rendering (or renderings in case of homophones). For a number of reasons, however, the revolutionary idea of the **phonetic-access dictionary** failed as a lexicographic and commercial venture. One reason must have doubtless been the relative inflexibility of the Michaelis & Jones resource: full-fledged IPA transcription with its exotic-looking symbols, no reverse and morphological indexing, etc. Other reasons may have been the reluctance of foreign learners to "learn much by simply reading down the lists of words beginning with, say, **j-** or **rai-**" (Michaelis & Jones 1913:vii) or the simple overestimation by the authors of "the rapidly growing interest now being taken in English pronunciation" (Michaelis & Jones 1913:i).

Computers changed the scene completely. In a machine-readable dictionary (MRD) any information at all can function indexically. In the second edition of the OED on CD-ROM, for example, searches can be conducted over (a) head entries, (b) etymology, (c) quotations, (d) pronunciation, and (e) the entire text of the dictionary. In such **multi-access** (or, as the fashion has it, **hyper-reference**) dictionaries, like in Michaelis & Jones, phonetic transcription serves both representative and **indexical** purposes. Because MRDs are so much more flexible than paper-based traditional dictionaries, the latter function of transcription gains in importance. Words can now be searched by their left-to-right and right-to-left phonetic representation, wildcards can be used to skip irrelevant or uncertain fragments, lists of words can be generated meeting complex phonetic criteria. EFL applications of such procedures are perhaps particularly salient (cf. Sobkowiak 1994a and b).

While the usefulness of phonetically accessible lexicons, dictionaries and word lists is generally acknowledged, and they are more and more widely used in theoretical and applied linguistics and language teaching, the implications of the changing function of MRD phonetic representation for the design of MRD phonetic transcription seem to pass

unnoticed among phoneticians and lexicographers alike. In what follows I will briefly examine a few issues relevant in this context.

2. MRD phonetic transcription

There are many considerations which inform the design of a phonetic transcription system having to do mainly with the choice of: (a) the dialect and phonostylistic level to be represented, (b) an underlying phonemic theory, (c) the 'width' of transcription, (d) contextual and suprasegmental effects (e.g. 'linking-r', secondary stress, syllable boundary) to be accounted for, (e) typographical concessions made for the benefit of the user, and others. I will largely ignore the first four issues here (some of them are nicely summarized in Wells 1985) and concentrate on the fifth. In particular, I will be interested in how some MRD phonetic transcription systems relate to the indexical function of phonetic representation in MRDs.

For the purposes of this paper I will look at three such systems:

- that used in the Oxford Advanced Learner's Dictionary of Current English (OALDCE) by A.S.Hornby, computer-readable version prepared by Roger Mitton (see Mitton 1986 and paper accompanying the resource) and available from the Oxford Text Archive and other sources,
- that used in the second edition of OED on CD-ROM, and
- SAMPA, the recently arising standard elaborated and fostered by the Speech Assessment Methods consortium (see Wells 1987 and Wells et al. 1992; compare also http://www.phon.ucl.ac.uk/home/sampa/home.htm).

In Table 1 phonetic representation of the more transcriptionally controversial RP English phonemes is shown. For comparison, the most widely used respelling symbols (Webster) and IPA symbols are also listed.

IP	as in:	OALDCE	OED	SAMPA	WEBSTER
θ	*thin*	T	T	T	th
ʃ	*she*	S	S	S	sh
ð	*then*	D	D	D	dh
ʒ	*vision*	Z	Z	Z	zh
ʧ	*chin*	tS	tS	tS	ch
ʤ	*June*	dZ	dZ	dZ	j
i	*see*	i	i:	i	ee, ē
ɪ	*sit*	I	I	I	i
e	*ten*	e	E	e	e
æ	*hat*	&	X	{	a
ʌ	*cup*	V	V	V	u
ɑː	*arm*	A	A:	A	ä
ɔ	*got*	0	Q	O	ô
o	*saw*	O	O:	o	
u	*too*	u	u:	u	
ʊ	*put*	U	U	U	oo
ə	*ago*	@	@	@	ə
əː	*fur*	3	3:	3	ur
aɪ	*five*	aI	aI	aI	ī
aʊ	*now*	aU	aU	aU	ou
ɔɪ	*join*	oI	OI	oI	oi
eɪ	*page*	eI	eI	eI	ā
əʊ	*home*	@U	@U	@U	ō
ɪə	*near*	I@	I@	I@	
eə	*hair*	e@	E@	e@	
ʊə	*pure*	U@	U@	U@	

Table 1. Some MRD phonetic transcription systems

The most immediately noticeable discrepancy between IPA and the three MRD systems is that the latter are restricted by the current computer and telecommunications technology to the so-called (lower) ASCII characters, i.e. those available on a standard computer keyboard (without overstriking or super- and sub-scripting, i.e. without any diacritics). While there are in principle no limits to what characters computer monitors can **show**, the inflexible keyboard remains the main tool for data entry, as far as phonetic transcription is concerned. This means, for

example, that key presses may be echoed on screen as proper IPA symbols, as happens in the OED.

The hundred available ASCII characters can be deployed in many different ways, of course, but there are three principles which are universally adhered to, at least in theory: (a) the system must be IPA-fashion grapho-phonemically bi-unique (one symbol – one phoneme), (b) lower-case alphabetic symbols should have the same values as in IPA, and (c) non-IPA symbols should be mnemonic as far as possible. The first principle serves the representational correctness function while the latter two are meant to make the systems 'user-friendly'. This friendliness or ease of use helps the user not only to interpret phonetic representations of words located through spelling (or other codes) but also to search and find words via their (typed-in) phonetic transcription in what I called above phonetic access. Mnemonic simplification, then, serves the indexical function of MRD transcription.

Notice, incidentally, that the simplification is staunchly transcription-based, rather than spelling-based, as is usual in respelling systems, of which Webster is an example. That is, the mnemonic value of a symbol depends on its relation to the standard (IPA) transcription of the sound (/S/ for /ʃ/, for example), rather than to its standard orthographic representation (/sh/ for /ʃ/). I am not aware of any research which would show that the former approach is representationally or indexically superior to the latter. On the other hand, there are situations where the opposite appears to be the case, e.g. in some popular lexicographic and didactic applications (see Sobkowiak in press).

Where the three MRD transcription systems differ most is – unsurprisingly – the representation of vocalic length and quality. The use of the length mark continues to be a hotly debated issue regardless of MRDs (compare Wells 1985 with Mepham 1978, for example) and the lexicographic practice varies widely. The editors of the OED on CD-ROM decided to follow Gimson's (1980) lead in allowing considerable redundancy in the representation of what is sometimes called the tense-lax vowel contrast. The five colon-marked vowel symbols would be unique without the length mark in this system. Because of this, the introduction of this mark may be indexically inconsequential in searches for words containing the five vowels. A command of the sort "give me words with /A/" (like *arm*) will work fine. However, in locating individual words rather than in list generation ("give me the word pronounced /Am/") forgetting the colon will lead to failure. If vowel length is an underlying phonemic feature of English (Wells 1985:50), native speakers will tend to forget about it while entering symbols in MRD phonetic access. There is dispute over whether foreign learners of

English are better off with the 'chroneme' symbol or without it. In this situation, the OED's choice appears to be (at least indexically) counterproductive.

While the usage of capital letters for lax vowels and lower-case letters for tense vowels appears to be somewhat of a standard, a different treatment can be observed in Table 1. SAMPA is the most consistent here, with both OALDCE and OED opting for a non-standard representation of the two /o/ vowels. The use of /0/ (zero) and /Q/ is rather counterintuitive here (if not entirely countermnemonic, due to letter-shape similarities) and about the only rationale behind it which I can think of is the desire to keep the representation of the two simple /o/ vowels different from that of the vowel functioning as part of the IPA/oi/ diphthong, which is indeed indexically desirable. The price to pay in inconsistency and counterintuitiveness, however, is high.

The discussion of the MRD phonetic representation of vowel length and diphthongs leads to the issue of digraphs, i.e. those cases where two adjacent letters stand for one phoneme.[1] Following the IPA tradition digraphs are used in the three systems of Table 1 to transcribe diphthongs and affricates. This is representationally legitimate, if not quite uncontroversial: such transcription seems to accept the view that diphthongs and affricates are regarded as phonetically and/or phonemically bisegmental. It is also unproblematic indexically in single-word look-up. However, it is rather seriously annoying in list generation where phonetic access proceeds in terms of structural conditions rather than fully specified word representation. Because the symbols of digraphs are singly non-unique (represent more than one sound), using them in such searches leads to confusion. It is rather obvious that an unsuspecting user issuing a command like "list all these words which end in /I/" will not be prepared for a load of word-final /oi/'s, /ai/'s and /ei/'s. In this situation fronting diphthongs would have to be explicitly excluded (as would /iə/ if the search was for /I/ anywhere in the word). Thus, in order to preserve complete bi-uniqueness both affricates and diphthongs would have to receive consistently monographemic representations. While /C/ and /J/ are obvious candidates for the former, it is more difficult to find mnemonic monographs for the eight RP diphthongs. In my own MRD work I have been using the digits from 1 to 8.

Monographemic representation of diphthongs carries another advantage: it avoids the indexical indeterminacy in triphthong symbolization. While few users of MRDs would have doubts about where the syllable boundary is in *fire, power, payer, mower* or *employer*, a search for words with high-centring diphthongs (/iə/ and /uə/) will yield these words and dozens of other triphthong entries on top of the desired lot.

This is clearly the deficiency of most transcription systems allowing digraphs, including the orthodox IPA. Yet, in the strictly representational approach to transcription this 'triphthong-segmentation' indeterminacy is repaired by general phonotactic rules and principles (hiatus avoidance) as well as by the trivial fact of left-to-right scanning, whereby the closing diphthongs are segmented first in triphthongs. It is the flexibility of MRD phonetic access that reveals the problem fully and identifies it as an obstacle to univocal indexicality.

3. Conclusion

The potential indexical use of MRD phonetic transcription imposes special requirements on the design of the system, from the underlying phonemic decisions to the shape of the symbols. With the spread of popular applications in computer lexicography, such as the OED on CD-ROM or the many diskette-based mono- and bilingual dictionaries of English now in existence, the importance of the active use of phonetic code for lexical access will grow. So will the relevance of issues briefly discussed in this paper.

Note

1 In standard non-MRD transcriptions the meaning of 'digraph' is somewhat different; see, for example, Roach 1992:31.

References

Fergusson, R. 1985. *The Penguin rhyming dictionary.* Harmondsworth, Penguin.

Gimson, A.C. 1980[1962]. *An introduction to the pronunciation of English.* London, Edward Arnold.

Mepham, R. 1978. "Why not simplify our phonetic transcription?", in: *Zielsprache – Englisch,* Volume 4, pp. 8–11.

Michaelis, H. & Jones, D.1913. *A phonetic dictionary of the English language.* Hannover, Carl Meyer.

Mitton, R. 1986. "A partial dictionary of English in computer usable form", in: *Literary and Linguistic Computing*, Volume 1, pp. 214–15.

Roach, P. 1992. *Introducing phonetics.* Harmondsworth, Penguin.

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

Sobkowiak, W. 1994b. "Beyond the year 2000: phonetic access dictionaries (with frequency information) in EFL", in: *System*, Volume 22, Number 4, pp. 509–23.

Sobkowiak, W. (in press). "Radically simplified phonetic transcription for Polish speakers", in: S. Puppel & R. Hickey (eds) *Festschrift for Professor Jacek Fisiak on His 60th birthday,* Berlin, Mouton de Gruyter.

Walker, J. 1890 [1775]. *The rhyming dictionary of the English language.* London, Routledge & Sons, Ltd.

Wells, J.C. 1985. "English pronunciation and its dictionary representation". In R. Ilson (ed.) *Dictionaries, lexicography and language learning*, Oxford, Pergamon Press, pp. 45–51.

Wells, J.C. 1987. "Computer-coded phonetic transcription", in: *Journal of the IPA*, Volume 17, Number 2, pp. 94–114.

Wells, J.C. *et al.* 1992. "Standardized Computer-Compatible Transcription", Esprit Project 2589 (SAM), Doc. no. SAM-UCL-037. London, Department of Phonetics & Linguistics, UCL.