Studia Anglica Posnaniensia XXVIII, 1994

PRACTICAL AUTOMATIC PHONEMIC TRANSCRIPTION SYSTEMS*

MICHAŁ JANKOWSKI

Adam Mickiewicz University, Poznań

In this paper I discuss the following issues connected with managing pronunciation in dictionaries and other reference publications:

– automatic transcription from orthography,
– converting transcription from one transcription convention to another, and
– automatic transcription of multi-word items.

1. The background

The programs I designed to perform these tasks form a package that I like to call "Pronunciation editor's toolbox". I have used this software with relatively high degree of success for a few years, adding new features as I go along. The initial incentive to design a general, multi-purpose and language-independent transcription editor's toolbox came in the form of a commision from a local publisher to develop a series of phrasebooks for Polish tourists. The series had been planned to eventually cover the whole of Europe (over twenty languages from English, German, French, Italian, through Dutch and Portugese, to Danish and Finnish to name just a few) and the USA. So far ten phrasebooks have come out and several are in various stages of preparation. Other projects that have benefited from the toolbox to a certain degree include the publication of two larger and several smaller bilingual dictionaries for the Polish market.

---

* This paper is based on a presentation at the ALLC-ACH '92 joint conference of the Association for Literary and Linguistic Computing and the Association for Computing in the Humanities held at Christ Church, Oxford in April 1992.

2. The contents of the pronunciation editor's toolbox

The pronunciation editor's toolbox includes the following programs:

- A language-independent index generator, which produces lists of word tokens for a given text, with optional frequencies and line references.
- A cluster generator, which generates a list of clusters along with frequency data for a given list of words.
- A rule-based transliterator, which:
  - generates transcriptions for a given list of words on the basis of a set of rules, and
  - converts transcriptions from one convention to another on the basis of a set of rules.
- Phrase transcriber, which transcribes multi-word strings using a transcription look-up list.
- Format converter, which converts the final document to the format specified by the publisher.
- Other software:
  - ChiWriter – flexible, graphics-mode word processor
  - Superkey – keyboard reassignment TSR tool
  - grep (part of Borland's C++ package), uniq (my own imitation of UNIX's uniq), sort (my own language-independent imitation of UNIX's sort)

3. The stages of transcription editing

3.1. The word index

As soon as the foreign side of a phrase book is keyboarded, an index of the traget language word tokens is obtained with the index generator program. Language independence is achieved by defining the alphabet separately for each language. Foreign characters are represented in my system by ASCII codes from 146 upward. The 18 Polish characters occupy the codes from 128 to 145. The language which had the most non-standard characters on top of the Polish 18 was Czech with its 30 non-standard characters. The index may include single word tokens for further processing, and/or frequency information and line numbers for future reference during the proofreading and editing stages. The average total number of word tokens in the text of the phrases is approx. two thousand.

3.2. The spelling-to-sound rules

The list of unique word tokens is then fed to the cluster generator program and an index of letters and letter clusters is generated giving individual letter frequencies and letter cluster frequencies, including word initial and word final positions. This information together with any available reference material describing spelling-to-sound rules for a particular language is submitted to the author, whose job at this stage is to prepare a subset of spelling-to-sound rules for the phrasebook. The format of the rules was especially designed to match the format of spelling-to-sound rules given in general reference books such as dictionaries and grammars, and to be easy to understand and work with for the authors of the phrasebooks, not all of whom have been linguists. Additional, more detailed information might come from research on phoneme-grapheme alignment such as that reported by Véronis (1988) and other research quoted therein, although in the case of the phrasebooks there has been no need to go beyond the generally available reference material.

3.3. From spelling to sound

A list of transcriptions of the word list is then produced automatically on the basis of the rules, possibly with a trial-and-error testing stage with the use of the transliterator program. The idea of the transliterator originally came from a character transliterator available under UNIX as the tr command. My transliterator works with multiple-character production rules much like the Markov algorithm (cf. Tremblay and Sorensen 1984) and is context sensitive.

The production rules consist of a grapheme sequence, with an optional context, a corresponding phoneme sequence, and a pointer advance index. The program uses a deterministic one-pass algorithm which tries to match the longest substring of the input string with the longest rules first, shortening the substring after each unsuccessful scan of the rules. When a match is found, the corresponding phoneme sequence is copied to the output and the input pointer as advanced by the value of the index. The matching cycle is then repeated. To facilitate the handling of prefixes, suffixes, endings, etc., word initial and word final contexts are also supported.

The list of transcriptions of the target language word tokens, which is never 100 % error free, is then proofread and possibly manually supplied with word stress marks, which are not handled by the transliterator. The correctness ratio of it has never been specially calculated. Therefore, I can only say that its success rate has varied from very high in the case of Hungarian, Czech and Bulgarian (almost 100 %), through satisfactory in the case of Italian, to disappointingly low in the case of Swedish and Danish. In addition to the general irregularity of the spelling of some of these languages, the reasons for the low rate may also have been connected with my complete ignorance of a particular language, possibly coupled with a poor understanding of the system on the part of the author of the respective phrasebook. In any case, as the objective of the project has not been to achieve a linguistically complete description but a successful commercial product, the only consideration is that poor rules make the job longer and more painstaking.

Here are some examples of the handling of the letter <c> in different contexts in the Italian-to-IPA rules:

```
cia,tʃ,1
cie,tʃ,1
cio,tʃ,1
ciu,tʃ,1
ce,tʃ,1
ci,tʃ,1
ch,k,2
cq,kk,2
c,k,1
```

The following example are the rules involved in the processing of the German word *schließfächer* ([ʃliːsfeçə]):

```
sch,ʃ,3
l,l,1
ie,iː,2
ß,s,1
f,f,1
äch,eç,3
er$,ə,3
```

where the "$" character marks the word final position.

English, by the way, has never been tried on this system, although theoretically it would have worked to a certain degree. The list of English transcriptions I used for the English phrasebook and other English language publications originally cames from the disk version of the dictionary of computing terms I co-authored. Naturally, I have kept the list and I have since regularly edited and enlarged it. The original English transcription list was keyboarded manually with some use of a macro facility long before the idea of the transliterator was conceived.

The automatization of the grapheme-to-phoneme transition in Polish, on the other hand, is possible and was described in a comprehensive study by Steffen-Batogowa (1975).

3.4. The choice of transcription system

The author of a phrasebook was free to choose any transcription system for the preliminary list – usually a system used in a dictionary that he or she likes – whatever is considered easier to proofread. The final system in the phrasebook, however, is simplified to suit the Polish speaking, low-language-awareness user and incorporates Polish spelling of foreign sounds whenever possible. Occasionally, a special symbol is used for a particularly distant sound, based on a Polish spelling representing the closest sound. The transliterator program is used extensively at this stage to convert the pronunciation list from the preliminary system to the final simplified system. As the final transcription has to be tested, the conversion may have to be performed several times.

Similarly, in the case of English-Polish dictionaries and glossaries, publishers request a particular system to be used and considering the number of considerably different systems available, this may create a problem for the pronunciation editor. I have recently used my English transcription bank to supply transcriptions to a number of smaller dictionaries and glossaries published in Poland. In each case I used the transliterator to make adjustments to the transcriptions stored in the transcription bank so that they would conform to the system specified by the publishers. I also used the transliterator to adjust the transcription in the USA version of the Polish-English phrasebook (Jankowski, Nadstoga and Sawala 1991) and the Polish semi-bilingual version of the Chambers Concise Usage Dictionary (Schwarz and Seaton 1985, Schwarz, Seaton and Fisiak 1990) published in 1990.

Here are some examples of the IPA-to-simplified-Polish rules used in the production of the Polish-English phrasebook (Jankowski, Nadstoga and Sawala 1990):

```
ʌ,a,1
ɑː,aː,2
aɪ,aj,2
aʊ,ał,2
ɪ,y,1
iː,i,2
ɔɪ,oj,2
ɒ,o,1
.
.
.
v,w,1
ʃ,sz,2
tʃ,cz,2
ʒ,ż,2
dʒ,dż,2
etc.
```

The following are some examples of the adjustments in the Chambers Concise before publication in Poland:

```
a,æ,1
aː,aː,2
ai,aɪ,2
au,aʊ,2
ei,eɪ,2
i,ɪ,1
iə,ɪə,2
oi,oɪ,2
```

u,ʊ,1

3.5. Automatic transcription of phrases

As soon as the final list of the target language word-transcription correspondences is completed, all the phrases in the phrasebook can be automatically transcribed word by word. As the procedure does not handle phrase stress and/or word boundary assimilations, the transcription may have to be further edited manually by the author of the phrasebook (as it was in the case of French). In the case of the dictionary of computer terminology mentioned above, I used an earlier version of the phrase transcriber program, which also handled abbreviations.

4. The future

The obvious direction in which the rule-based transliterator could develop is for it to

- accept more condensed, general rules,
- handle word stress, syllable boundary, and related issues

The first problem is easy to solve by providing UNIX-like conventions such as this rule abbreviation

[aeiou]b

to mean ab, eb, ib, ob, and ub or even macro definitions such as

#define VOWELS aeiou

to be used in general rules, as in

VOWELSb

to mean ab, eb, ib, ob, and ub.

However, considering the mechanical, brute-force nature of the algorithm itself, the second problem does not seem easy to solve at all. The program has served its purpose well, but from the linguistic point of view it seems to be in a dead end.

On the other hand, the idea of devising a universal transcription system for a master transcription bank from which various transcriptions could be generated on demand seems more attractive and realistic. A system like that might for example be prepared to generate both British and American English in a variety of transcription systems when supplied with the appropriate set of rules.

One simple and obvious example of a transcription in such a universal system might be

cycle ['saikL]

where "L" represents the so-called 'syllabic l', a phenomenon that has almost as many different representations as there are different systems. From there it would be easy to convert to any of the following transcription systems used in dictionaries:

['saɪkl] (Hornby's OALD, Hornby 1989)
['saɪkᵊl] (Longman DOCE, Proctor 1991)
['saɪkᵊl] (Well's Longman Pronunciation Dictionary, Wells 1990)
['saɪkə^{0}l] (Collins COBUILD, Sinclair 1987)
['saɪkl̩] (English-Polish Dictionary of Computer Science, Marciniak and Jankowski 1991)
etc.

On the other hand, this would not be possible if the source transcription was stored in the bank as

['saikl]

The development of a system like that for my English transcription bank thus seems to be a natural next step.

5. Conclusion

The pronunciation editor's toolbox does not solve all of the problems that are encountered at the various stages of development of reference publications which provide pronunciation information. However, the software makes the job much easier, less time consuming and less frustrating, as it takes away much of the mechanical, "boring" aspect. Thanks to the software, transcriptions that are created and verified can be reused and continuity can be established from publication to publication with minimum error and maximum consistency. The toolbox proves especially useful in low budget projects such as the phrasebook series, where apart from the author of a pharasebook and myself, only two other people are involved prior to final typesetting and printing: a general series editor and a keyboarder/technical editor.

REFERENCES

Hornby, A.S. 1989. *Oxford advanced learner's dictionary of current English*. Oxford: Oxford University Press

Jankowski, M., Nadstoga, Z. and Sawala, K. 1990. *Wielka Brytania. Informacja Turystyczna. Rozmówki Polsko-angielskie. Mini-słownik* [Great Britain. Polish-English Phrasebook]. Poznań: Atena.

Jankowski, M., Nadstoga, Z. and Sawala, K. 1991. *USA. Informacja Turystyczna. Rozmówki polsko-angielskie. Mini-słownik* [USA. Polish-English Phrasebook]. Poznań: Atena.

Marciniak, A. and Jankowski, M. 1991. *Słownik informatyczny polsko-angielski* [English-Polish dictionary of computer science]. Warszawa-Poznań: PWN.

Proctor, P., ed. 1987. *Longman dictionary of contemporary English*. Burnt Mill: Longman Group Ltd.

Schwarz, C.M., Seaton, M.A. and Fisiak, J. 1990. *English dictionary for spȩakers of Polish*. Toronto-Poznań: Kernerman Publishing Inc. - SAWW.

Schwarz, C.M. and Seaton, M.A. 1985. *Chambers concise usage dictionary*. Cambridge-Edinburgh: Chambers.

Sinclair, J. ed., 1987. *Collins COBUILD English language dictionary*. London: HarperCollins Publishers.

Steffen-Batogowa, M. 1975. *Automatyzacja transkrypcji fonematycznej tekstów polskich* [Automatization of phonemic transcription of Polish texts]. Warszawa: PWN.

Tremblay, J-P. and Sorensen, P.G. 1984[2]. *An introduction to data structures with applications*. New York: McGraw-Hill Book Company.

Véronis, J. 1988. "Phoneme-to-grapheme correspondences". *Computers in literary and lingustic research – Literary and linguistic computing 1988. Proceedings of the fifteenth international conference*. Paris-Geneve: Champion – Slatkine. 373-83.

Wells, J.C. 1990. *Longman pronunciation dictionary*. Burnt Mill: Longman Group Ltd.