THE MECHANICAL CONVERSION OF HUNGARIAN SCRIPT TO PHONETIC NOTATION

THOMAS ARKWRIGHT AND ANDREW KEREK

1. INTRODUCTION

This paper characterizes the structure of a procedure that converts (transduces) Hungarian orthographic texts into phonetic (broad phonetic) notation, using IPA symbols. The end goal of the project is a longer procedure that uses input from an optical text scanner to produce tape recordings of synthesized speech as output. The procedure characterized here has been implemented on an IBM 360/75 Snobol4 interpreter, and similarly on a CDC 6400; it consists of an ordered series of context-free and context-sensitive rewrite rules operating on a string of characters (the text). A few truly idiosynchratic cases aside (*lesz* → [lɛss]), the program that implements the procedure contains no ad hoc rules. It is free from manual processing. Thus, it is a perfectly explicit algorithm that follows the general phonological processes expressed by the order-dependent rewrite rules. The input is punched manually onto cards, and three Hungarian diacritical marks not found on the 29 keypunch are replaced by three arbitrary vowel-following symbols (—¬, <, >). Since processing proceeds card by card, there is no limit to the length of the text string. The string on each card is parsed into substrings (words), and these are then processed sequentially. Word processing may be grossly described as follows: each graphemic symbol is changed to its corresponding IPA symbol (through an alphanumeric code), and the resulting new string is then filtered through a series of phonological rules. The output of the text can be formatted in IPA symbols (using an IPA typeball on a 2741 terminal), or in Hungarian characters that correspond one-to-one to the appropriate IPA representation; this is illustrated by the processing of the word *kezdték* (a form of the verb for 'begin'):

THE SEQUENTIAL PROCESSING OF 'KEZDTE—¬K' FROM INPUT, (I), TO OUTPUT, (XII):

(I) KEZDTE—¬K → (ENTER THE SET OF GRAPHEME CONVERSION RULES) (II) KEZDT#07K → (III) KE#38DT#07K (IV) KE#38#18T#07K → (V) KE#38#18#35#07K → (VI) K#02#38#18#35#07K → (VII) #26#02#38#18#35#07K → (VIII) #26#02#38#18#35#07#26 → (ENTER THE

SET OF PHONOLOGICAL RULES) (IX) #26#02#38#35#35#07#26 → (X) #26#02#34#35#35#07#26 → (XI) #26#02#34#35#07#26 →...ONE FUNCTION CHANGES THE ALPHANUMERIC STRING TO A PHONETIC STRING... (XII) /KESTE—¬K/.

2. BACKGROUND

Some reports of schemes for spelling-to-sound conversion have been of the armchair variety (Venezky 1966, Lee 1967); other reports describe highly accurate algorithms that have been tested on computers. However, to our knowledge, all such programs depend upon a certain amount of manual processing of the data (Kučera 1963, 1964), or on special circumstances such as their application to verse (Silva 1968). Kučera's work is particularly admirable because it shows that by virtue of the preprocessing, highly accurate statistical studies of 'phoneme' distribution can be executed. It is clear, however, that previous studies represent a purely functional approach (e.g., to achieve consistent transcriptions of copious amounts of text for statistical studies of phone distributions and frequencies; typological and stylistic studies; talking computers; reading machines for the blind; etc.). We are proposing that another dimension can be imposed upon such algorithms, namely, to give ordered expression to the general phonological constraints that are implicit in the transcription process, once the graphemic characters have been assigned to their machine-coded counterparts. The dependence of our algorithm upon order can be seen in the example above: if degemination (XI) applies before voicing assimilation (IX, X), the output ([kɛstte:k]) is wrong. This is a general statement about the phonological component, true of countless other words (*mondta*, *küldte*, etc.). We are not aware of any previous attempt to express general phonological constraints in such transcription procedures. Silva frankly uses lookup tables, and Kučera's (1963) program is clearly not a sequentially conceived algorithm (cf. his discussion of Cz. *pěna*, *zpěv*, pp. 43 ff). Of course both programs use order heuristically, as any program must, and these comments suggest a recommended framework for future research, rather than a criticism of these pioneering studies. None of the advantages of the functional approach are lost when grapheme-to-phoneme algorithms are conceived as perfectly explicit grammars which simultaneously express phonological constraints in the language and assign a string of classificatory values to a transduced string of graphemes.

3. CORPUS

Using references (such as Lotz 1969, Papp 1966, Varga 1968), dictionaries, and native informants, we compiled what we believed was a sufficiently representative list of words, in the sense that a procedure which could convert this entire list to IPA could (in principle) properly convert all other items in a Hungarian dictionary to

IPA. The initial corpus consisted of words whose orthography was not in one-to-one correspondence with IPA ('complex' words), such as *méhben* → [me:bεn], as well as corresponding orthographically 'simple' words, such as *méhek* → [me:hεk]. This corpus included many words which clearly represented the same process (*kezdték* → [kεste:k] and *kezdte* → [kεstε]); wherever possible, all but one of these repetitious forms were eliminated. In the case of a word with more than one acceptable pronunciation (*egyszer* → [εt⁀st⁀sεr] or [εcsεr]), one alternative was chosen. The final corpus had 125 words. It should be noted that some words, such as *kétszer* → [ke:t⁀st⁀sεr] and *szebbtől* → [sεptœ:l], show multiple deviations of the output from the orthography (as opposed to *méhben* above, which shows only one). Further complexity is illustrated by words showing multiple deviations (such as *sokkban* → [ʃɔgbɒn]) whose IPA notations are identical to that of words with different graphemic spellings (*sokban* → [ʃɔgbɒn]). A similar case is *fáradtság*, *fáradság* → [fa:rɒt⁀ʃt⁀ʃa:g]. Notice that these and other such cases *(ronts/roncs)* show that spelling is not always recoverable from sound.

4. MARGINAL PHENOMENA

The program described above does not correctly account for three classes of problem words. Two of these classes, foreign words (*technika*, *ortodox*, etc.) and some proper nouns (*Pálffy*, *Kossuth*, *Svájc*, etc.), being of no immediate linguistic interest, were excluded from our corpus. The third class of problem words involved morpheme boundaries in prefixed forms *(meg + gyón)* and compounds *(ház + sor, vad + zerge)*; they failed when the rule ordering caused the symbols at the boundary ('+') to be incorrectly interpreted as one multigraph[1] (e.g., *g+gy* as /ɟɟ/; *z+s* as [ʒ]; *d+z* as [dz]), instead of correctly interpreting the symbol immediately preceding the boundary in each case as a hengraph. These failures represent, and are notable exceptions to, a small class of problem words that are successfully handled in the majority of the cases (*mázsa*, *egészség*, *díszszemle*, *hattyú*, *meggyből*, etc.) by general rules with an optimal order of application.

Importantly, some 95% of a challenging test corpus is successfully handled without any appeal to morphological, syntactic, and lexical information. It is interesting that the correct transduction of the remainder of the corpus *(ház+sor; köz+ség, lúd+zsír, szét+szór, vad+zerge, meg+gyón, millió)* requires that such information be available. We are now developing algorithms for the morphological and syllabic segmentation of the word, and for accentuation, so that this conversion procedure can be applied beyond the bounds of the word, and so that additional phonetic detail can be supplied. Hopefully, these studies will enable us to generate relatively detailed underlying phonological forms and phonetic forms, with accuracy comparable to those of the project just described. This future work will profit from statistical studies of phone sequences based on our recent transduction of words in some randomly selected texts. The accuracy of this transduction was far in excess of 99% (one error per 5020 phones).

[1] The Hungarian alphabet consists of (a) a set of multigraphs (*cs*: [tʃ], *dz*: [dz], *dzs*: [dʒ], *gy*: [ɟ], *ly*: [j], *ny*: [ɲ], *sz*: [s], *ty*: /c/, *zs*: [ʒ]), and (b) a set of hengraphs (*a*: [ɑ], *á*: [a:], *b*: [b], *c* [ts], … *z*: [z]).

Arkwright: Department of Linguistics
McGill University
Kerek: Miami University

BIBLIOGRAPHY

(no author)
1959 *A Magyar Helyesírás Szabályai*, 10th edition (Budapest, Akadémiai Kiadó).
(no author)
1959 *A Magyar Nyelv Értelmező Szótára* (Budapest, Akadémiai Kiadó).
Bánhidi, Z., Z. Jókay and D. Szabó
1965 *Learn Hungarian*, 2nd edition (Budapest, Tankönyvkiadó).
Cress, P., P. Dirksen and J.W. Graham
1970 *Fortran IV with Watfor and Watfiv* (Englewood Cliffs, Prentice Hall).
Cooper, F.S.
1966 "Toward a High Performance Reading Machine for the Blind", in *Human Factors in Technology*, ed., E.M. Bennett *et al.* (New York).
Engström, S.
1969 "*Överföring från ortografisk till fonematisk svenska med hjälp av datamaskin*", *FOA Reports* 3, 4:1-17.
Griswold, R.E., J.F. Poage and I.P. Polonsky
1968 *The Snobol4 Programming Language* (Englewood Cliffs, Prentice Hall).
Kučera, H.
1963 "Mechanical Phonemic Transcription and Phoneme Frequency Count in Czech", *International Journal of Slavic Linguistics and Poetics* VI:36-50.
1964 "Statistical Determination of Isotropy", *Proceedings of the Ninth International Congress of Linguists* (The Hague, Mouton: 713-721).
Lee, F.F.
1967 "Automatic Grapheme-to-Phoneme Translation of English", *Journal of the Acoustic Society of America* 41:594.
Lotz, J.
1969 "The Conversion of Script to Speech as Exemplified by Hungarian", *The Linguistic Reporter, Supplement* 23 (October):17-30.
(no author)
1970 *McGill University Computing Centre Operating System Users' Guide.*
Papp, I.
1966 *Leíró Magyar Hangtan* (Budapest, Tankönyvkiadó).
Silva, G.
1968 "Phontrns: an Automatic Orthographic-to-Phonetic Conversion System for French", *Computers in the Humanities* 2:257-265.
Ungeheuer, G. and Kästner
1966 "Untersuchung zur Transformation Deutscher Schrifttexte in entsprechende Phonemtexte mit Hilfe elektronischer Rechenmaschienen", Forschungsbericht (Institut für Phonetik und Kommunikationsforschung).

Vanderslice, R.
1968 "Synthetic Elocution, Considerations in Automatic Orthographic-to-Phonetic Conversion of English with Special Reference to Prosodic Features", *UCLA Working Papers in Phonetics* 8.

Varga, G.G.
1968 *Alakváltozatok a Budapesti Köznyelvben* (Budapest, Akadémiai Kiadó).

Venezky, R.L.
1966 "Automatic Spelling-to-Sound Conversion", in *Computation in Linguistics*, eds. P.L. Garvin and B. Spolsky (Bloomington, Indiana University Press).