PHONTRNS: AN ALGORITHM FOR THE TRANSCRIPTION OF FRENCH TEXT INTO PHONETIC SYMBOLS BY A COMPUTER

GEORGETTE SILVA — BRUCE PRATT*

The purpose of this paper is to report a computer method for transcribing orthographic French into phonetic symbols. The procedure was developed at Monash University, Australia. Although similar schemes have been reported for English[1] and Swedish,[2] and work is in progress at Bonn on a system for transcription of German,[3] as far as we are aware none has yet been published for French.

The system outlined in this paper is designed to accept text punched on paper tape in "running text" form. Where the computer's internal code does not include letters with orthographic signs such as accents, substitute symbols mut be introduced at the typing stage. Other than this no precoding is required. The contents of the paper tapes are transferred to magnetic tape, proofread and corrected. The text is then ready for transcription.

Automatic transcription of modern French involves the resolution of a two-fold problem: on the one hand, certain sounds can be represented by a number of spellings; on the other, the same letter or group of letters is sounded in different ways in different words: in each specific case the sound must be defined in context in such a way that it is accurately identified.

This problem is greatly simplified if the text is first divided into syllables. This is done by computer, syllables being separated by blank spaces and words by asterisks, which provide convenient reference points during the process of transcription.

Since the computer cannot recognize phonetic symbols each sound is represented by a two digit number code: further, certain pairs of consonants have been allotted separate numbers to facilitate reference to them during processing.

Some combinations of letters regularly generate complex sound patterns and many difficulties can be eliminated by coding these before transcribing the remainder of the text. These groups have been arranged in lists according to the number of letters they contain, each item being accompanied by its pronunciation expressed in the number code refered to a moment ago.

* Monash University Clayton, Victoria, Australia.

[1] Reported by Francis F. Lee of M.I.T.

[2] Reported by Siv Engstrøm.

[3] At the Institut für Phonetik und Kommunikationsforschung, Bonn University.

The syllabized text is scanned on successive cycles in groups of four, three, two and one characters. Each time a group is formed it is compared with the list of like length. If a match is found, the numeric code is substituted for the group of letters found in the input text. On completion of each cycle an increasing proportion of the text has been encoded. This system provides for those groups of letters which never vary in pronunciation, and also permits some discrimination between different pronunciations of a given letter combination through use of precoding on early cycles. However this is not enough in itself. To ensure adequate transcription of those letters and groups of letters which vary in pronunciation from word to word, the basic procedure has been overlaid with an elaborate system of checks of adjacent letters, which in effect define the pronunciation of these groups in terms of adjacent sounds. Alternative codings are provided in a set of lists parallel to the main lists.

Our aim has been to develop a simple and flexible system for machine transcription, our concern being with the observance rather than the formulation of phonetic rules. On each successive cycle the problems facing the programmer can be stated afresh, as a certain number of possibilities have been eliminated: the only criterion used in deciding whether or not to accept a procedure has been the degree of accuracy it allows.

The real test of the system is therefore its accuracy. We are happy to report fewer than 50 errors on a text of more than 14,000 sounds, that is, an error rate of 0.4 %, and we offer suggestions for rapidly locating and correcting these.

Output can be provided in any form desired, and some examples can be found in our recently published book.[4] Some institutions may be hampered by the absence of print chains carrying phonetic symbols, but this would not greatly restrict the system's usefulness, as all significant operations are executed using the number code. It would seem that all that part of the phonetician's work which involves the statistical analysis of the language, and which up to the present time has had to be carried out manually by teams of research assistants on relatively small samples of the language, can now be performed on vast quantities of data. Studies that come to mind are the relative frequency of segmental phonemes, word length as measured by phonemes and syllables, canonical forms of consonant-vowel sequences arranged by frequency, consonant cluster distribution, transitional probabilities of phoneme sequences, entropy and redundancy values and so on.

In conclusion, we wish to emphasize that our aim has been to place in the hands of phoneticians a tool which will greatly increase the possible scope of their investigations. It is a highly flexible tool, open to considerable refinement and to modification according to particular modes of defining phonemes. We hope you will find it useful.

[4] Bruce Pratt and Georgette Silva. *Phontrns: a Procedure which uses a computer for transcribing French text into phonetic symbols.* Monash University. 1967.

DISCUSSION

Valdman:

The program described here is based on the fallacy that French spelling is a sort of phonological transcription. In fact it is a morphonemic notation which provides information about form and substance. Therefore if we wish to devise a transcribing device whose output includes spontaneous speech, we shall need to provide it with higher level information i.e. information about syntactic structure and morphological form.

Gsell:

J'ajouterai peu de mots à ce que vient de dire Mr. Valdman, puisque le système graphique français n'est ni phonologique, ni phonétique, ni même morphologique. Je demanderai seulement à Madame Silva:

1. quel est le pourcentage d'efficacité de son système
2. s'il a été testé égalemet sur la transcription de la prose.

Ayant fait travailler des chercheurs à Grenoble sur la transcription automatique, j'avais estimé que la structure du français écrit était tellement différente de celle du français oral qu'il fallait traiter les deux formes de l'expression comme 2 langues différentes et réaliser un programme de type traduction automatique avec morphologie du français oral et programme d'analyse syntaxique.

Silva:

ad Gsell: The system was tested on ‚Le Jeune Parque' by Paul Valéry, a poem of some 500 lines. It is with some confidence that we report an accuracy of over 99%.

The version of the program suitable for the transcription of prose is not yet completed.