FERENC PAPP

AUTOMATIC ANALYSIS OF HUNGARIAN TEXTS AND LINGUISTIC DATA

1. First of all I would like to give an account of the practical experience gained in the course of *processing* the about 60,000 or so entries of a Hungarian unilingual (explanatory) dictionary *(A magyar nyelv értelmezö szótára,* vv. I-VII, 1959-1962). In this case by " text " we mean this non-natural corpus, that is the sum total of the entries of the dictionary; and by linguistic data the information given below.

1.1. Naturally the head-word itself is a *linguistic datum* as well, as long as it is processed according to its component letters, the phonemes represented by them, and its sounds. Furthermore we coded also the following data in explicit form beside each head-word:

(1) complexity (how many roots does the lexeme consist of),
(2) homonymy,
(3) part of speech (in case of conversion each possible form class denoted),
(4) the number of meanings,
(5) style,
(6) morphological data (for words capable of inflexion),
(7) strongly required government (only for verbs),
(8) etymology,
(9) suffix,
(10) length (this latter datum, of course, was coded not by us; it was decided and punched on the card in the proper place by the processing machine).

1.2. It is a natural concomitant of the automatic processing that we can get back any combination of these linguistic data so as to draw *conclusions* from it; and that in order to answer a particular question we can ask for special outputs. In the following account I am going

to set forth such questions of theoretical or practical character and the answers given to them.

1.2.1. Some of the conclusions issued simply from the different *mechanical arrangements* of the material. So we got data as regards the length of the lexemes included in the examined dictionary, (though these data must be viewed critically: in a larger dictionary there would be presumably a greater number of longer words for any language (F. PAPP, 1970), as regards their distribution according to the number of their meanings (presumably this index too depends, however, on the size of the dictionary and in a very large dictionary it would come close to 1 meaning per lexeme (F. PAPP, 1970)). As a result of simple, mechanical arrangements, it was interesting to see the entire material of the dictionary grouped according to complexity – derivation, all the homonyms collected, the elements belonging to the same parts of speech (in the case of conversion all the possible combinations of parts of speech) put together, etc.

1.2.2. On the basis of these simple arrangements a project has been started in the course of which the 60.000 entries and the morphological information concerning them have been put into the memory of the machine and have been used as the basis on which natural Hungarian texts can be analysed. (The director of the project is Gy. Hell, Budapest Technical University). We are still at the beginning of this work, but there is reason to suppose that a mechanical dictionary having at its disposal the appropriate grammatical information renders an automatic analysis of optimum efficiency possible.

1.2.3. The information given by the dictionary led to some further synchronic and diachronic conclusions. Let me quote two examples:

1) Certain forms of the noun (a possessive form and others formally similar to it) have long been a problem in Hungarian. In these forms the suffixes occur sometimes with a *j* element, sometimes without it: *lába* " his/her foot " (from *láb* " foot ") without a *j* element and *combja* " his/her thigh " (from *comb* " thigh ") with a *j* element. The appearance of the *j* seems extremely arbitrary: as our pair of examples indicates, it sometimes does and sometimes does not occur even after the same consonant (in this case after *b*); in spite of the fact that the Hungarian system does not favour consonant clusters, it

rather appears as a third or fourth consonant after a final consonant group fairly inconvenient in itself (like in this case the *-mb*: *comb*), than after a single consonant at the end of a lexeme, etc. After the analysis of several special arrangements of the material it turned out that the *j* serves actually to emphasize the unusual, rare, young, etc. stem-endings; that is why it appears after for example unusual consonant combinations. This treatment of the stem-ending (the special emphasis given to it) throws light on the one hand on the behaviour of the stem in an agglutinating system; on the other hand on the attitude of the speaker and the listener, who – it seems in this case – seek to do the automatic analysis with the help of a minimal dictionary. Without the *j* element the listener would be turning to his inner dictionary all the time while doing the analysis: this way, however, he either meets a typical, frequent, etc. stem-ending, which therefore he recognizes easily; or a special *j* element helps him to divide the word (by showing where the comparatively independent suffix-cluster of the agglutinating system begins).

2) The analysis of the 6000 unquestionable *root-words* included in the dictionary (and of the 4000 or so elements that can be regarded in a wider sense as root-words too) led to several results. So for instance it came to light that vowel-harmony which is of central importance to the system of the Hungarian language, has undergone an interesting change during the one thousand years of the history of the Hungarian language we can directly trace back. While in the Finno-Ugrian and ancient Turkish etymological strata the velar and palatal stems are represented half-and-half; the chronologically successive etymological strata (such as Slavic, German, etc.) contain more and more velar elements; as regards the Latin elements, which belong to the youngest ones, the great majority of them is velar. The data indicate as well that vowel harmony holds its own in the system of present-day Hungarian as well and solves easily, almost without any uncertainty the vocal adaptation of the loanwords which have been streaming continuosly into the language for thousand years.

We also had made a detailed investigation of phoneme-statistics about the different etymological strata of the root-words (with the ODRA 1204 machine of the Kossuth Lajos University, programmed by P. Jékel), and then also placed the different etymological layers in a 66 dimension space depicting the phonemes. The multi-dimensional space proved to be very manageable. Besides revealing trivial truths such as " it is the layer of the inner, Hungarian developments that is

nearest to the Finno-Ugrian layer", or "the Latin and Neo-Latin layers are the farthest ones from the Finno-Ugrian layer", it gave about these distances certain valuable quantitative indices as well on the basis of which the model could be applied to other fields where indeed it has already brought really new results (cf. 2.1.2).

2. We have prepared *natural Hungarian texts* as well for automatic processing. At the time of writing, a text of half a million phonemes (the complete poetic works of E. Ady), has already undergone the description set out below; further texts of similar size from the complete works of 20th century and earlier Hungarian poets will be put into the machine in the near future. The *general aim* of this work is naturally to have a thoroughly controlled, uniformly punched Hungarian text-store consisting of millions or tens of millions of phonemes. A computational text-store like that can serve as the basis of automatic processings of different levels and of concrete purposes. In the following some of these will be described; especially the one that has already been most fully realized.

2.1. These texts have been processed first *on the level of phonemes*. For this purpose a program has been made that

a) changes texts written with present-day Hungarian orthography from letters into phonemes;

b) identifies and counts the phonemes, produces tabular and other indices from them as outputs;

c) places the different texts in a multi-dimensional space according to the results measured alongside the phonemes, and measures their distances observed in this space.

In the following we illustrate by only a few examples what we expect from a processing like this and what we have already got from it.

2.1.1. A thorough phoneme-statistical work done on a much larger corpus that the previous ones might disprove certain prejudices. So it has been a fairly widespread belief that a large proportion of vowels in a text is beautiful. (In connection with this e. g. the Italian language is often praised for its high ratio of vowels). In respect of at least the examined Hungarian texts, however, it has turned out that the proportion of vowels depends on the average length of the running words: the longer words are more vocalic than the shorter ones;

therefore the poetic texts containing naturally more short words are less vocalic than the different technical texts.

2.1.2. We have been measuring the distances of the different texts in the multi-dimensional space. Up to the present our most valuable result is that the first juvenile volume of the above-mentioned E. Ady, which Hungarian *literati* and aestheticians consider not to be characteristic of the later poet, is really so far from the other volumes, from his *oeuvre* – specifically as regards its tone, that it has to be stated from a pure mathematical aspect also that it takes place in quite a different part of space non-characteristic of the author. The results we are expecting just now concern the relation of a poet's own poems and his translations; what are the acoustic distances among his own volumes and translated volumes of a poet who translated a lot and wrote a great number of poems of his own as well; what is the situation if he translated a rather bulky volume twice in his life. (That is what happened to the 20th century L. Szabó who translated Shakespeare's *Sonnets* both at the beginning and towards the end of his poetic life: we are going to measure the distance of these two complete volumes from each-other, from the original Szabó-volumes of the same periods, and from his other volumes of translations as well as from other contemporary Hungarian poets).

2.2. The punched material may offer us even farther-reaching possibilities for the near and the distant future.

2.2.1. The easiest and most natural step forward will be to compile also sound-statistics after the phoneme-statistics, especially as there is an excellent theoretical preliminary study to it (J. LOTZ, 1972), which has already been applied to the computer (T. ARKWRIGHT, A. KEREK, 1972).

2.2.2. Naturally we are going to advance towards the compilation of poetic concordance dictionaries as well as towards the morphological and syntactical processing of the given material.

*

REFERENCES

T. ARKWRIGHT, A. KEREK, *The Mechanical Conversion of Hungarian Script to Phonetic Notation*, 1972 (manuscript).

J. LOTZ, *Script, Grammar and the Hungarian Writing System*, Budapest, 1972.

F. PAPP, *Sur quelques lois statistiques du vocabulaire*, in *Actes du X^e Congrès International des linguistes*, Bucarest, 1970, pp. 471-475.