COLING 82, J. Horecký (ed.)
North-Holland Publishing Company
© Academia, 1982

EMPIRICAL DATA AND AUTOMATIC ANALYSIS

Ferenc Papp

Slavic Department
University L. Kossuth
Debrecen
Hungary

The purpose of the present paper is to show the usefulness of (1) the computer processing of the manifold data of lexicographic works; and (2) the normal and reverse alphabetized concordances compiled on the basis of different texts.

1. More than ten years have passed since the Reverse-alphabetized Dictionary of the Hungarian Language (see Papp 1969) was published as the first major document of the computer processing of A Magyar Nyelv Értelmező Szótára (the Explanatory Dictionary of the Hungarian Language). In the course of this work quite a lot of grammatically relevant information rendered by the cca. 60 thousand entry words of the source dictionary was registered, e.g. how many stems the entry word consists of; whether it has an affix and where it is; what style it represents, what etymological explanation it is given, and mainly: what important morphological and (with verbs) syntactic characteristics it has, which part of speech it belongs to, etc. On the basis of the coded morphological features the automatic synthesis of the forms of these 60 thousand words becomes directly possible and the automatic analysis of Hungarian texts can be carried out indirectly. (NB. it is always the synthesis that the processing of dictionaries makes directly possible, see, e.g. Zaliznjak's outstanding work [1977] which affords an excellent opportunity for the morphological synthesis of the cca. 100 thousand Russian entry words listed in it.) Below I am going to show a direct and concrete potential for application (1.1) and I also wish to demonstrate what general results can be gained from the possibility of the flexible and many-sided computer processing of a great number of words, indicating in what way the results gained can influence the whole strategy of our automatic analysis (1.2).

1.1. Since 1981 we have been participating in the realization of the Hungarian part of the MULTILEX programme elaborated by the Vsesojuznyj Centr Perevoda (Moscow). In our case the immediate purpose of this programme is to make it possible for an expert without any knowledge of Hungarian to look up certain technical terms; a further step requires a rudimentary morphological analysis of the Hungarian text. The user underlines the word which may occur in any of its grammatical forms in the text, the programme carries out the analysis (removes the endings and finds the verbal prefixes - see below, 2.3, etc.). Taking this future analysis into account now we register certain pieces of morphological information for each entry word of the Hungarian-Russian computer dictionary we are currently preparing. This is not a trivial task if one considers that in the highly inflective, agglutinative Hungarian language an inflectable word has not only one or two (cf. English: *table* – *tables*; – *read* – *reading* – *reads*, etc.) or even about ten (e.g. Latin, Russian) paradigmatic forms, but hundreds of them. In theory it has long been known and now it is also shown by the reverse alphabetized concordances that will be spoken of later that this unbelievably large number of forms contrast with prepositions in our neighbouring Indo-European languages. Thus, whereas in English, Russian, German, etc. certain grammatical elements are "comfortably" separated by spaces, in the agglutinative Hungarian language they became, so to say, merged with the word stem. It is not the lack of a space to denote them which causes the principle difficulty. Being within a single word form they can undergo different changes themselves and can also cause different changes in the word stem itself depending on the grammatical nature of the lexeme and on vowel harmony, etc. To solve this purely practical task a mass of empirical data is needed. It is precisely the reverse alphabetized concordances that furnish them, but the data gained from dictionaries are also indispensable.

1.2. It was at a conference in Prague some years ago that I gave an account of the general results gained from the study of the material rendered by the Hungarian dictionary mentioned above. The matter in question is that a large group of Hungarian nouns derives one of their important forms, namely the third person possessive form, in a way that seems to be rhapsodic, sometimes inserting a "j" element and sometimes not: *ablaka* 'his/her window' – *barackja* 'his/her apricot'; *türelme* 'his/her patience' – *filmje* 'his/her film', etc. This possessive form was registered with each entry word, so we had the opportunity to group the tens of thousands of nouns of our dictionary from the greatly varying points of view, keeping in mind the main question of "with "j" – without "j"". Thus, the two examples quoted above were used to illustrate the point that the appearance of the "j" element could not depend directly on the final consonant: in the first pair of examples both stems ended in "k" and in the second in "m"; it could not even depend on the final consonant group, because in the second pair of examples ""j" – not "j"" was preceded by "lm" in both cases; but, of course, we nevertheless had such a sorting made. The investigation of the full noun stock of the dictionary in different combinations led to the following conclusion: the seemingly rhapsodic appearance or omission of "j" can easily be explained if we suppose that the Hungarian language as a natural code is structured in such a way that automatic analysis can be carried out using a minimal vocabulary. Concretely: if the non-possessive stem is typical and "well-formed" from the viewpoint of the whole vocabulary, i.e. if it has a frequently occurring end, the "j" element does not appear after it in the possessive form; and the other way round: where the stem would not be recognized automatically because it has a rare ending and therefore the bare "a/e" ending would be linked automatically with the stem, "j" emerges, so to say, in order to stop this and to emphasize the end of the bare stem. Thus in the first pair of examples quoted above "j" did not appear after the bare final "k" because this is frequent for the end of stems in Hungarian; whereas "j" appeared after the "bad" "ck" group, which is not a typical stem end in Hungarian. Our second pair of examples is also subjected to the same rule, although in a slightly different way: the word *türelme* did not require "j" because a

typical change of the stem of a productive Hungarian suffix is hidden in it (the bare nominative is türelem, here the stem was automatically emphasized by the türelem/türelm- opposition), whereas filmje required a special denotation of the stem by "j", there being no ^{x}filem/film opposition. It must be added that the behaviour of suffixes was diagnostic from our point of view. If the suffix is productive, there is no "j" after it; if it is not productive or it is not a nominal suffix but e.g. an adjectival one and is used with a noun only occasionally, the appearance of "j" is more or less necessary. It is "more or less" so because it represents the linguistic manifestation of a regularity that is practically a statistical one; what the several examples of instability and parallel forms are explained by is that the linguistic instinct is not a computer, it is not always possible for a whole community to decide unequivocally whether a stem end is frequent or rare, whether a suffix is productive or not. From this point of view the behaviour of the different historico-etymological strata is characterisitic. The nearer we move towards the younger loanwords the more frequently "j" appears: a loanword often has a "wrong", "atypical" end. But, of course, we can only say "more frequently", "in many cases", etc.: e.g. after a great number of words ending in the easily perceptible -um, -(t)-or had been borrowed from Latin, they gradually became "good", i.e. recognizable and so they did not necessarily have to take the special sign "j". A similar phenomenon can also be observed with compound words. A root word having a "wrong" end requires "j" - but if it is often used as the second part of a compound, it becomes something like a suffix, "we have got used to it" at the ends of compounds and this is why the "j" will sooner disappear from there. This is easily noticeable even on the basis of a simple reverse alphabetized list, as the root word often having a morphological code with "j" stands in the first place there and it is immediately followed by the compounds in which this root word is the second part, and in many cases its code has already no "j". (NB. when the coding was going on this regularity was not even guessed at, so this theoretical consideration could not have influenced the coders, in the case of root words they were compelled to take over the corresponding code of the source dictionary. Of course, no one could see these compounds in one group before the publication of the reverse alphabetized dictionary!)
All this, however, was nothing more than a plausible hypothesis supported by evidence from the dictionary. Its real confirmation could be achieved by the study of texts. Reverse alphabetized concordances based on Hungarian linguistic material afford a very good opportunity to do this.

2. For the last couple of years a number of normal and reverse alphabetized concordances have been made at the L. Kossuth University on the basis of English, French, Swahili and mainly Hungarian and Russian texts. Relying upon the material rendered by the last two languages, we are going to show what kind of empirical data can be provided for the analysis. Properly made concordances from texts in different natural languages have features of their own. Thus, it is clear that a normal Swahili concordance works the other way round in the sense that the material is divided into different groups according to the grammatical indices; that in a French concordance it is not expedient to print running words consisting of three or fewer letters. These technical details will not be discussed here, we only note that in Hungarian concordances the article "a" which makes up a comparatively high percent of running words in the

different stylistic strata has been left out of consideration.

2.1. Having made a reverse alphabetized concordance from Hungarian newspaper texts consisting of approximately 26 thousand running words we can arrive at the following conclusions. Of the 64 phonemes in contemporary Hungarian only 49 actually occur at the ends of words, half of all the word ends being occupied by the first five ot these (the percentage number of these phonemes and their occurrence at the ends of words in our material: /t/ 13, /k/ 12, /n/ 9, /s/ 9, /a/ 7). As it will be proved in the next section - a comparison of this with Russian data - this division shows a situation very similar to that in Russian. One should say the agglutinative character of Hungarian becomes clear within this from the quantitative point of view: within the different final phonemes large blocks with the same long agglutinative ending group can be seen. Thus, 18 % of all words ending in /t/ are made up by those ending in /et/ (non-possessive acc. sing.; verbs 3rd pers. sing.), 10 % by those ending in /át/ (possessive acc. sing., non-possessive acc. sing., verb. verbal prefix), etc.; the end /k/ has final groups, sometimes containing as many as three phonemes: /nak/ and /nek/, each taking up 11 % (dat./gen, verb 3rd pers. plural); the same can be observed with the end /n/: (ban) - (ben) (32 % altogether - inessiv) and so on. All this suggests that morphological analysis in Hungarian should be started at the ends of words: much useful grammatical information is concentrated there. These final clusters of two, three or four phonemes, of course, are not completely homogeneous, but the number of words to be analysed in another way is insignificant. Thus, e.g. the acc. sing. forms ending in /ot/ make up 3 % of those ending in /t/ and there are only two running words among them in which this is in the form of nom. sing. (the two occurrences of the lexeme *állapot*); or to mention another example: among the dozens of occurrences of the final quadruplet /ének/ (3rd pers. poss. dat/gen) to be analysed on the basis of the same principle only two running words can be found: *békének* and *versikének*, used as the non-possessive dat/gen of the lexemes *béke* 'peace' and *versike* 'little verse'

We have already mentioned above that the final empirical evidence for our hypothesis on the possessive /j/ was provided by these reverse alphabetized concordances. It was found that of all the words ending in /a/ 11 % ended in /ja/ mainly owing to this possessive form, whereas the words ending in /je/ did not make up 3 % of those ending in /e/: both the contemporary Hungarian vocabulary and the contemporary texts with their frequent /a/ end and less frequent /e/ end will sooner require the special denotation of the end of the stem with /j/ immediately before the /a/. Other Hungarian texts presented a very similar picture of the division of final phonemes, especially concerning /a/-/e/ at the ends of words. Thus, e.g. there was not a single noun with the ending /je/ in this possessive form among the thousands of nouns of the approximately 20 000 running words of "Toldi" (an epic poem written by János Arany in the middle of the last century); it goes without saying that at the same time a number of them took the endings /a/, /ja/ and /e/ in this grammatical form: the various possessive forms in their sum total proved to be even more productive than the plural ones. (By the way, all this testifies that Hungarian texts can be considered to have been "contemporary" from this point of view since at least the middle of the last century.)

Concerning analysis let us make one more essential remark in connection with the /ja/ ending. Forms like barackja 'his/her apricot' can already be well differentiated in the nominal declension, but the same /ja/ ending has created a new homonymy at the ends of words: the 3rd person singular forms of velar verbs take the same ending in their objective conjugation: e.g. this form of the verb vág 'to cut' is vágja 'he/she cuts (it)'. The prominence of the /ja/ ending can be explained by this fact as well, which at the same time makes our evidence weaker: in the case of palatal harmony there is another ending (cf. néz 'to look at': nézi 'he/she looks at (it)' and not the expectable ^{x}nézje or something like this). The whole morphology of Hungarian, however, is dominated by a particular feature: namely that no difference is made between the parts of speech: the /m/ at the ends of words refers to the first person of verbs, nouns, pronouns, etc.; the /k/ refers to some kind of plural. This, of course, makes morphological analysis based on the word end more difficult: how practical it is to know that in Russian endings containing the element /y/ (ye, yx, ymi, etc.) belong to an adjective; that the overwhelming majority of verbal word ends (eš', et, em, ete, etc.) is charateristic only of verbs, etc. (It is interesting to note that English, a language with an extremely poor system of endings and hardly comparable with Hungarian from this point of view, shows a similar indifference towards parts of speech and even grammatical meaning: it is only the simple /s/ that forms the plural of nouns, the 3rd person singular of verbs and even the genitive of nouns; such a polysemy, of course, could hardly be imagined in Hungarian.)

2.2. Here are the five most frequent final phonemes of "Onegin" containing about 22 000 running words, the percentage number is indicated in brackets: /j/ (10), /i/ (10), /a/ (9), /o/ (8), /e/ (8); 45 % of the running words end in one of these phonemes. Within the most frequent word ends, however, one can find fewer final pairs (not to speak of triplets or quadruplets), and what is important is that if an ending can still be brought into prominence, it can bear many various and incoherent functions. Thus, e.g. in this material the word end /ej/ makes up more than one fifth of all the 480 running words ending in /j/. Within the limits of this material the following proportions have been stated (100 = 480): 1. pronouns like ej, sej, vsej (48 %), 2. gostej-type genitive plural (14 %), 3. poslednej -type adjectival forms (10 %), 4. nočej-type genitive plural (10 %), 5. lenivej-type comparative forms (8 %), and so on. The remaining 10 % are spread over a dozen functions (parts of speech, grammatical cases, moods of verbs, etc.). It should be noted that none of the most frequent five types enumerated here is homogeneous from the point of view of grammatical analysis, cf. especially the 1. and the 3. with their grammatical polysemy. Some of the more "fortunate" endings as "ij", may have only half as many functions, but even in this case the mass of empirical data yielded by a reverse alphabetized concordance may be indispensable to make the analysing algorythm as exact and elegant as necessary.

2.3. It was quite clear, even in the early stages of mechanical translation, that the separable Hungarian verbal prefixes would present a special problem for the analysis. (Thus, the very first step of the very first Hungarian-Russian MT-algorythm in word-finding was the search for verbal prefixes that might have been separated, cf. Mel'čuk 1958, 231.) According to the testimony of the concordances, this problem is rare, though it does exist structurally and cannot be neglected. In the newspaper concordance mentioned above, not more than 362 "separated" verbal prefixes have been found, hardly less than 1,5 % of all the running words. Further, there were only five (meg, el, ki, fel, be) of about fifty possible verbal prefixes that occured separately in three quarters of all these cases.

Normal concordances also provide rich material for problems of word order of separable verbal prefixes, negation etc., i.e. the topicalization of Hungarian sentences. (Topicalization is of the utmost importance from the viewpoint of the analysis of Hungarian sentences, cf. É.Kiss 1981; at the same time this question, if treated en masse, i.e. by using a computer, can hardly be approached in a way other than the one mentioned.)

REFERENCES:

[1] É.Kiss, K., Structural Relations in Hungarian, a "Free" Word Order Language, Linguistic Inquiry 12.2 (1981) 185-213.

[2] Mel'čuk, I.A., O mašinnom perevode s vengerskogo jazyka na russkij, Problemy kibernetiki 1 (1958) 222-264.

[3] Papp, F. (ed.), Reverse-alphabetized Dictionary of the Hungarian Language (Akadémiai Kiadó, Budapest, 1969).

[4] Zaliznjak, A.A., Grammatičeskij slovar' russkogo jazyka. Slovoizmenenie (Izdatel'stvo "Russkij jazyk", Moskva, 1977).