M. Füredi
(Institute for Transport Sciences, Budapest, Hungary)

«*LANGUE*» AND «*PAROLE*» IN FREQUENCY DICTIONARIES

The author characterizes some earlier lexicographic research done in Hungary involving considerable philological efforts and sorting machines (later computers), and shows some interesting results about unexpected closeness in some characteristics of lexemes (as a dictionary) and their usage in real texts. As the late professor L.R. Zinder used to tell his students: «We hear what we are used to hear». It is important to study the language in its process, in its functioning, as the statistical properties are relevant not only in the characteristics of text, no matter how interpreted, but also on all our psycholinguistic behaviour, hearing and composition capabilities, etc.

In Hungary one line of the efforts in quantitative linguistics was to make «machine-aided» or «computerised» variant of existing dictionaries, to reflect some statistical properties of the Hungarian «*langue*».

One of the first research of this type was F. Papp's *a tergo* dictionary[1], based on the 58 323 lexemes of the Hungarian Explanatory Dictionary (HED), compiled in the fifties and published between 1959–1962. He and his colleagues inserted some new, linguistically relevant codes (so called «Debrecen usage») besides those used in the HED, and sorted the dictionary by 40 different orders, using punch-cards and IBM-type electro-mechanical sorting machines.

Another big effort of Hungarian linguists was the preparation of a Hungarian Frequency Dictionary[2] (HFD), initiated originally by

[1] *Papp F.* A magyar nyelv szóvégmutató szótára [A Reverse-Alphabetized Dictionary of the Hungarian Language]. Budapest, 1969.

[2] *Füredi M., Kelemen J.* A mai magyar nyelv szépprózai gyakorisági szótára (1965 – 1977) [A Hungarian Frequency Dictionary of Modern Fiction

J. Kelemen in the sixties, the work on which began only in 1975, having been revised many times. The undertaking originally aimed to reflect the Hungarian linguistic usage, the Hungarian «*parole*», by preparing a five-genre (later only four-genre) frequency dictionary, each genre consisting of 500 000 running words following the design of frequency dictionaries by A. Juilland and A. Zampolli. Detailed manual coding was done by hundreds of university students, later controlled by their teachers all over Hungary's universities and colleges, everywhere at the chair of Hungarian language. The final, so called super-revision was done by linguistic experts at the Research Institute for Linguistics (RIL) of the Hungarian Academy of Sciences (HAS). This work – due to financial problems at the HAS – was completed for only one genre, the fiction. Manual coding (including homonymy, part of speech, tense, mood, singular or plural, possessive suffix characteristics, collocations etc.) and its systematic controls were finished as early as 1976, headed by M. Füredi, then computer input and sorting followed with higher level logical controls. All work for the fiction was finalised in the middle of 1983, including a rather complicated photo type-setting program done by Gy. Visontay and M. Füredi in SIMULA'67 programming language at the computer centre of the HAS. This was the first Hungarian dictionary prepared by real computers (IBM mainframes).

Reports and partial results were given at conferences in Szombathely (Hungary), Eisenstadt, Hamburg, Prague, Syktyvkar, Vienna.[1] Among other interesting facts **it was shown that the**

(1965 – 1977)]. Budapest, 1989. On both dictionaries see: *Papp F., Füredi M.* Применение ЭВМ в изучении лексики венгерского языка // Nyelvtudományi Közlemények. Vol. 86. 1984. № 2. P. 409–413.

[1] *Füredi M.* A short account on the Frequency Dictionary of Modern Hungarian Fiction // COLING'82. Prague, 5–10 July, 1982. P. XXX; *Füredi M.* A mai magyar szépprózа statisztikai vizsgálata [Statistical investigation of modern Hungarian fiction] // IV International Congress of Hungarian linguists: «A magyar nyelv rétegződése». Szombathely, Hungary, 23–26

distribution of subclasses of Hungarian vowels (velar – palatal, labial – illabial, open – closed) show a rather close correlation (in some cases even statistically significant one) in a dictionary of lexemes and in a dictionary of word-forms weighted by frequency.

The published version of the HFD contains only the most frequent **3410 lemmas** with all their word-forms occurred in texts, and a separate list of those **4898 word-forms**, where the modified frequency according to Juilland's formula (in our case, as a novelty, used within one only genre: the texts were mechanically divided for 5 subsets of approximately 100 000 running words each), reached 10,00 or more. In **a linguistic corpus of 508 008 Hungarian running words**, after the exclusion of proper nouns, **the number of different word-forms found was 91 471 representing 33 169 lemmas**. This seems to be unexpectedly low, considering an agglutinative language with a rather rich suffixation capabilities.

The mentioned two types of dictionary, each with its own coding systems gave a good insight into the linguistic structure of Hungarian words, and – at the same time – into the very usage of word-forms of an agglutinative language. The post-life of both dictionaries was (and still is) promising, due to successful coincidence of several rather

---

August, 1983. P. XXX; *Füredi M.* Frequencies of Hungarian affricates // Nyelvtudományi Közlemények. Vol. 86. 1984. № 2. P. 337–340; *Füredi M.* Összesített adatok a szépprózai gyakorisági szótárról (Igék és igeszármazékok) [Summarised data about the Hungarian frequency dictionary (Verbs and verbal derivatives)] // Magyar Nyelv. Vol. LXXXII. 1986. № 2. P. 190–198; *Füredi M.* A linguistic data-base for the Hungarian lexicography // Congress of Hungarology. Vienna, September 1986. P. XXX; *Füredi M.* Vowel frequency in Hungarian // Studia Uralica. Band 4. Studien zur Phonologie und Morphonologie der uralischen Sprachen. Akten der dritten Tagung für Uralische Phonologie. Eisenstadt, 28 Juni – 1 Juli 1984. Wien, 1987. P. 97–113; *Füredi M.* Multi-level analysis of Hungarian literary texts // IV International Symposium «Uralische Phonologie». Hamburg, 4–8 September 1989. P. XXX.

strange and unexpected anecdotical events. The lemmas of both dictionaries were united into one huge data-base, and, parallelly, a special linguistics-oriented data-base handling system was elaborated by L. Éltető[1] on the (then new) IBM 3031 computer of the HAS, as early as in 1984–1985. A new field of CV-structure was also introduced into the lexeme data-base by A. Kornai[2] in 1985. The resulting SZOTA1R data-base consists now of about 80 000 Hungarian lemmas, and since that time it has been used at the RIL and by many linguists worldwide. This united Hungarian database can be accessed for research purposes, free of charge. As far as both dictionaries on their own had been declared *public domain* early enough, they could have also been incorporated into the Hungarian spelling checker programs, like that one sold by the Hungarian firm *Morphologic* for such program packages as different versions of *InDesign*, *PageMaker*, *QuarkXpress* and, naturally, the widespread versions of *MS Office 95-2002*.

There are some other dictionaries worth mentioning, like some Hungarian poets' dictionaries (S. Petőfi, Gy. Juhász), a newspaper and magazine frequency dictionary based on texts oriented for children and young people, lemmatised «on the fly» by only one person (Szeged, 1986), and last, but not least the Hungarian Academic Dictionary (HAD) under preparation at the RIL, originally planned to use 10 million running words from the beginning of Hungarian book printing until modern times, and now having a corpus of about 150 million running words. The HAD was approved by the HAS as early as on February 28, 1984, and the undertaking was headed at that time

[1] *Éltető L.* Új adatbáziskezelő rendszer VM/CMS alatt [A new data-base handling system under VM/CMS] // Információ – Elektronika, 1985. P. XXX.

[2] *Kornai A.* Szótári adatbázis az akadémiai nagyszámítógépen [A dictionary data-base on the mainframe computer of the Academy] // Műhelymunkák a nyelvészet és társtudományai köréből. II. – Budapest: MTA Nyelvtudományi Intézete, 1986. június.

by academician F. Papp. Results having been published about HFD gave good estimates for the number of different word-forms in larger texts, not known earlier.

Naturally, during the past 20 years new opportunities emerged to use special purpose computer programs for morphological, syntactic and even semantic analysis. The linguistic work nowadays is much more attractive than in our «stone» age. It was instructive for me to reread my own article[1], which mentioned – besides our «normal» *text-based* approach – also the *individual free association* method of building such kind of dictionaries. The first part of this article would have to be basically rewritten nowadays, needing to reflect new achievements in linguistics, mainly in automated morphological analysis, developed database handling and processing of textual data, much improved hardware and software capabilities of personal computers.

* * *

A frequency dictionary of all word-forms, found and extracted from a corpus is an intermediate step between the *"langue"* and *"parole"* representation of the given corpus in statistical terms. A *"langue"* representation reflects – in our opinion – only the lemma-frequencies (the frequencies at the level of the dictionary entries), while the *"parole"* representation is considered to be the corpus itself, the text under consideration. The dictionary of different word-forms is to be found somewhere in-between.

For the analytic type English language a FD can be (and so was many times) built on the basis of only word-forms because of the lack of too many inflectional forms, as far as no complicated cases and word-forms exist in nouns and verbs. Let us consider the Czech (7

[1] *Füredi M.* Hogyan készíthetnénk ma gyakorisági szótárt? [How could we compile a frequency dictionary today?] // Studia Russica XII. Budapest, 1988. P. 442–448.

cases for nouns), Russian (6 cases for nouns) or German (4 cases for nouns) languages. Different verbal forms and other parts of speech can be simply enumerated, too. In Russian, there are no too many different word-forms either for nouns to be considered (even if we count also the homonymic cases of second Accusative and second Genitive), which could cause problems to our computers even for larger text corpora, considering the recent development in storage and computing capacities, or in general, hardware computing power and 4th generation software possibilities.

Now let us turn our attention to the agglutinative type Hungarian language.

The famous Hungarian descriptivist L. Antal[1] states 17 different cases for Hungarian nouns. As an agglutinative language, the possessive forms are also part of the NP structure within the same «word». If we count also singular and plural forms, that is alltogether all non-recursive Hungarian noun-forms (by the way, some recursivity exists not only in Hungarian, but, although rarely, even in Russian, see: *прапра.. прабабушка*), than we get a huge number of different potential word-forms for one and the same Hungarian lexical unit, approximating the number 5 000.

The Hungarian version of HunSPELL program accounts for 4 936 different noun-forms, 4 041 adjectival forms and 59 verbal forms.

Much earlier the German linguist W. Veenker[2] counted around 3 020 different endings and ending-combinations.

How this richness of different word-forms works in the real texts, and how is it reflected in the HFD?

---

[1] *Antal L.* The Hungarian case system // Nyelvtudományi Értekezések № 29. Budapest, 1961. P. XXX.

[2] *Veenker W.* Verzeichnis der Ungarische Suffixe und Suffixkombinationen // Mitteilungen der Societas Uralo-Altaica. № 3. Hamburg. YEAR

The texts tagged for the HFD (compiled from 258 text samples of contemporary fiction from different authors, 2 000 running words each) contained 508 008 running words, and the 91 471 different word-forms could be reduced to 33 169 lemmmas.

For the most frequent common noun in the HFD we found three magnitudes less different word-forms in a half-million running words real texts than they exist potentially in the statistical population. For less frequent nouns this difference seems to be even more astonishing. This fact for such a big (at that time) Hungarian corpus was not known before 1983. This finding should be somewhat similar for all agglutinative languages as Finnish, Estonian, Mordvin, Turkish or even Japanese.

Taking a closer look, we find that in the HFD the most frequent common noun «*ember*» meaning 'man' had only 51 different word-forms in the tagged text samples of half a million running words, and one of the most frequent verb «*néz*» meaning «look» had 46 different word-forms. As we go down on the frequency list of lexemes, the number of different word-forms sharply decrease even for the highest frequency layers.

If we would like to represent all word-forms in a statistically relevant way for only the most frequent Hungarian common noun «*ember*» we would need a corpus of several magnitudes higher.[1]

Some conclusions

(1) **For agglutinative languages the sampling method remain relevant**, but we shall draw our attention more closely to the stems and develop better morphological analysis, disambiguation and lemmatisation process. It was not a chance, why many researchers had chosen text samples of 2 000 or 3 000 running words with an overall

[1] *Kornai A.* How many words are there? // Glottometrics. 2002. № 4. P. 61–86. Without quoting his argumentation, A. Kornai states that vocabulary size V(N) tends to infinity as $N \to \infty$.

covering of 500 000 running words (A. Juilland, A. Zampolli). For most quantitative linguistic purposes the exactness offered by this size seems to be enough, and, at the same time, the individual texts and/or authors do not bias the final results. Naturally, for deeper lexicographic description much wider samples are needed for every language, and we have to study and analyse more attentively the *hapax legomena*. That is why, essentially for agglutinative languages, sampling methods shall be used together with more intricate techniques, as e.g. detailed and good morphological analysis and tagging, being the basis of higher level (syntactical, semantic) analyses. Works are begun in this direction not only by Hungarian linguists and by the appropriate scientific institutions, but also by enthusiastic mathematicians and engineers at large in all over Hungary (e.g. see the Hungarian WordSword project at **www.szoszablya.hu**).

(2) At the same time **we shall consider the widely known fact, how deeply genre characteristics influence the frequency results at all linguistic levels**. In this sense the distilled Hungarian Web Corpus (WordSword, see **www.szoszablya.hu**) is not uniform, because inside this huge web corpus we can find scores of genres: a lot of descriptions (e.g. web sites of scientific institutes or universities), discourses (discussion fora), scientific articles (popularization of science) and other, more web-specific genres, to be defined later more precisely.

E.g., intuitively, a discussion forum uses much wider set of linguistic tools, and also much more different word-forms, than a conference presentation or a web-published scientific article. The now lacking genre characteristics tagging and equal sample sizes of different genres would help us to compare the results in a statistically strict and relevant way, although the first task is not so easy to automate, than quickly sweeping web pages.

In the Hungarian FD of fiction
First *X* lemmas covering *Y%* of texts

X	*Y%*
1	10.93
5	16.62
10	20.37
20	25.36
30	28.52
40	30.83
50	32.65
60	34.15
80	36.78
100	38.86
110	39.75
115	40.16

Dr. Füredi Mihály
KTI, Institute for Transport Sciences
Budapest, Hungary
Librarian, head of the Documentation and Information Centre of KTI
University doctorate in General Linguistics, Eötvös Loránd University, Budapest, Hungary
E-mail: **furedi@kti.hu**