Computational work on the Student's Illustrated Dictionary of Hungarian and the Computational study of its vocabulary

GÁBOR G. KISS

The Publishing House and Printing Press of the Hungarian Academy of Sciences is going to publish the Students' Illustrated Dictionary of Hungarian (henceforth: SIDict) before Christmas 1992. This dictionary for pupils (age 10-16) is being written in the Research Institute for Linguistics by a team of 7 scholars. According to the plans, the dictionary is going to contain 14.000 vocabulary entries as well as 80 pages of colored illustrations (BÍRÓ, 1991).

When comparing it with dictionaries of similar character for other languages, we can say that as regards its vocabulary and complexity SIDict has an intermediate position between the *OXFORD Basic English Dictionary* (10.000 entries) and the *LAROUSSE débutants* (20.000 entries) (OXFORD, 1981; LAROUSSE, 1986).

Parallel with the publication of this volume by Christmas 1992, a floppy-disc version for IBM PC/AT will made available for pupils, lexicographers as well as experts interested in further developments. The size of the dictionary will be 2.8 Mbyte.

The Publishing House and Printing Press of the Hungarian Academy of Sciences has arranged, in order to speed up the work, that the writing and the preparation of lexical entries for printing should go simultaneously, with a minimal delay. This was achieved by using computers. This way the Publishing House has achieved that not much after the preparation of the last dictionary entry the camera-ready set material of the SIDict will available.

When preparing the SIDict, IBM PC/ATs and XTs working under DOS help the writers of the dictionary in the following ways:

1. Compiling the dictionary entry list:

We used computers when compiling the dictionary entry list. We typed an amplified dictionary entry list into the computer. The amplified entry list based on a preliminary selection contains 16.000 lexical entries. * denotes compound lexical entries, e.g. *nyak*leves, idő*álló*. The dictionary entry list was created by WordPerfect. Using the SORT command, we ordered the dictionary entry list in an alphabetical order according to the last element of compounds. Thus we got lists like

```
1) esküdt*szék,     7)      tan*szék,      1)  film*szerű,
2)  forgó*szék,     8)  törvény*szék,      2) közép*szerű,
3)  hinta*szék,     9)      úri*szék,      3) sport*szerű,
4) iskola*szék,    10) villamos*szék,      4)  szak*szerű,
5)  pohár*szék,    11)  zongora*szék,
6)    szó*szék,
```

This list was used when preparing the last version of the dictionary entry list, i.e. when reducing the preliminary list containing 16.000 entries to the final version of 14.000 entries.

2. Typographical preparation of SIDict by the help of computers:

The Publishing House and Printing Press of the Hungarian Academy of Sciences works on FERRANTI printing equipment and composing system. In the previous years the possibility of a novel input to these typographical equipment was designed and elaborated. Through this input coded texts containing ASCII characters can be entered in the composing system. The codes refer on the one hand to the letter types, and on the other to the special characters used in the dictionary. Codes are given, as it can be seen from a sample below, in brackets < >:

```
normal letter type  = <1> [ xxxxxxxx ]
italics letter type = <2> [ xxxxxxxx ]
bold letter type    = <3> [ xxxxxxxx ]

tilde clinging to a suffix [ ~xxx ] = <50>
tilde with a comma         [ ‘xxx ] = <51>
tilde standing alone        [ _~_ ] = <52>

... = \q\q\q
* = <66>
| = <6>
```

The dictionary entries of SIDict were stored in the computer using these codes. An example for the dictionary stored for the printing press:

```
<5><41>híre-hamva<40> szragos fn (csak esz 3. személyben,
ált. alanyesetben)
<4><2>A tegnap leesett hónak ma már <52> sincs, <1>nyoma
sincs, eltűnt.

<5><41>híres<40>
<4><3>I. <1>mn <2><50>ek, <50>t <1>v. <2><50>et, <50>en
<3>1. <2><52> ember, író, politikus, színész:
<1>kiemelkedő tulajdonságairól, teljesítményéről jól
ismert (= hírneves). <2><52> regény, dal, város:
<1>közismert, nevezetes. (biz) <2>Nem valami <52> ez a
dolgozat, <1>gyenge minőségű, középszerű. <3>2.
<1>(pejor) <2>Megint itt vannak a <50>, <1>rossz hírű (=
hírhedt) <2>barátaid.
<3>II. <1>fn <2><50>ek, <50>et, <50>e <1>(pejor)
```

```
<1>Kétes hírű személy. <2>Hol az a <50>?
<1>(Megszólításban:) <2>No, te <50>!

<5><41>híresség<40> fn <2><50>ek, <50>et, <50>e
<4><3>1. <2>Apjának <50>e, <1>híres volta <2>nemcsak
segítette, akadályozta is pályáján. <3>2. <2>Az előadásra
meghívták az összes <50>et, <1>híres embert.
```

The SIDict was stored on the computer in WordPerfect files. In this work we utilized the vast macro-programming possibilities of the WordPerfect program to a great extent (KISS, 1991). We assigned the most common codes consisting of several characters to keyboard macros. This had a twofold advantage: on the one hand, typing and storing was quicker and, on the other, the character sequences of the codes appeared without any misspellings in the texts. E.g.

```
Alt key + o letter = (pej) [=(ironic)]
Alt key + b letter = <1>(=
Alt key + g letter = <1>v. <2> [<1> or <2>]
Alt key + a letter = <40> ige <2> [=\<40> verb <2>]
```

The writers of the dictionary asked us to enable them to check the texts, which had been stored in the computer's memory, in a non-coded form. The authors wanted to see and correct the text in the same form as it will be printed out in its final version. We were not able to have all parts of the dictionary printed out several times for proof-reading (this was due, among other factors, to constraints of time and budget). Thus before having the text printed for the authors, we used a series of WordPerfect macros entitled PREPA.WPM to convert the text full of codes to the common typographical form. In the following we exemplify this by showing the program listing of the macro that converts entries starting with <2> to italics.

```
Macro:       Action
File:        C:\WP51\MACROS\PREPA2.WPM
Description: A <2> kód utáni szöveg italizálása
```

```
{DISPLAY OFF}
{Home}{Home}{Up}
{LABEL}kezdet~
{Search}<2>{Search}
{Block}
{Search}<{Search}{left}
{Font}22
{ON NOT FOUND}{GO}vege~~
{GO}kezdet~
{LABEL}vege~
{Home}{Home}{Up}
```

By using the macro-program series we got the following printed version for the coded text seen above:

híre—hamva szragos fn (csak esz 3. személyben, ált. alanyesetben)
A tegnap leesett hónak ma már ~ sincs, nyoma sincs, eltűnt.

híres
I. mn *~ek, ~t* v. *~et, ~en*
1. *~ ember, író, politikus, színész:* kiemelkedő tulajdonságairól, teljesítményéről jól ismert (= hírneves). *~ regény, dal, város:* közismert, nevezetes. (biz) *Nem valami ~ ez a dolgozat*, gyenge minőségű, középszerű. **2.** (pejor) *Megint itt vannak a ~*, rossz hírű (= hírhedt) *barátaid.*
II. fn *~ek, ~et, ~e* (pejor)
Kétes hírű személy. *Hol az a ~?* (Megszólításban:) *No, te ~!*

híresség fn *~ek, ~et, ~e*
1. *Apjának ~e*, híres volta *nemcsak segítette, akadályozta is pályáján.* **2.** *Az előadásra meghívták az összes ~et*, híres embert.

These is the final form of SIDict from the Publishing House and Printing Press of the Hungarian Academy of Sciences before last correction:

láb fn *~ak, ~at, ~a*
1. Ember, szárazföldi állat járásra való végtagja, ill. ennek alsó része; lábfej. *Töri a cipő a ~át. ~a kel:* eltűnik, nyoma vész; *nagy ~on él:* pazarló, költekező életmódot folytat. **2.** Tárgynak az(ok) az oszlopszerű része(i), amely(ek)en áll. *Az asztal, a zongora ~a.* **3.** Hegy(ség), halom stb. alsó része. *A hegy ~ánál áll a ház.* **4.** Régi, ill. külföldi hosszmérték: kb. 30 cm. *Hat ~ magas.*
lábad ige *~ni* (vál)
Könnybe ~ a szeme, könnyes lesz.
lábadoz|ik ige *~ni*
Súlyos betegségből gyógyulóban van.
lábal ige *~ni* **lábol**
Sekély vízben, sárban, hóban – lábát nehezen emelgetve – jár, megy. *A pocsolyában ~. A pocsolyát ~ja.* (= gázol)
labanc fn *~ok, ~ot, ~a* (tört)
1. A kurucok ellen harcoló császári zsoldos. **2.** (pejor) Habsburg-párti magyar.
lábas[1] mn *~ak, ~at, ~a*
Lábon, talapzaton álló. *~ óra. ~ ház:* árkádos.
lábas[2] fn *~ok, ~t, ~a* **lábos**
Alacsony, kétfülű főzőedény.
lábatlankod|ik ige *~ni* (biz)
Másnak útjában van, ténfergésével zavarja, láb alatt van.
lábazat fn *~ok, ~ot, ~a*
Bútornak, építménynek, falnak, szobornak stb. a legalsó, sajátosan kiképzett része.
lábbeli fn *~k, ~t, ~je*
Az öltözéknek a lábon viselt tartozéka: cipő, csizma, szandál, papucs stb.
ra. **2.** Fölfelé nyíló fedelű, alacsony bútor ruhanemű, élelmiszer stb. tárolására. *Tulipán(t)os ~:* régi parasztházak festéssel (ritkábban faragással) díszített jellegzetes bútordarabja. **3.** Nagyobb doboz.
ladik fn *~ok, ~ot, ~ja*
Lapos fenekű csónak. *„Általmennék én a Tiszán ladikon, Ladikon, de ladikon"* (népdal).
lagúna fn *~k, ~t, ~ja*
Homok- v. korallszigetekkel (részben) elzárt sekély tengerrész. *A ~k városa:* Velence.
lagzi fn *~k, ~t, ~ja* (nép, biz)
Lakodalom.
lágy
I. mn *~ak, ~at, ~an*
1. Könnyen formálható, laza szerkezetű, sokszor nedvességet tartalmazó (= puha, ≠ kemény). *~ tojás:* héjában hígra főzött. *Az ólom ~ fém. A ~ szárú növénynek nincs fás része.* **2.** *~ víz:* kevés ásványi sót tartalmazó. **3.** Nem határozott, nem éles, nem erős, nem éles körvonalú. *~ vonások, ~ hang, ~ fém, ~ hullámok, ~ szellő.* **4.** Szelíd, gyöngéd, engedékeny. *~ szíve van, ~an cirógat.*
II. fn *~ak, ~at, ~a*
Vminek a ~a: a lágyabb része. *A feje ~a:* a koponyacsontok találkozásának kisgyermekkorban porcos része. 66 *Benőtt a feje ~a:* megkomolyodott.
lágyék fn *~ok, ~ot, ~a*
A hasfalnak alulról a csípőkig terjedő, háromszög alakú része.

3. Editing SIDict by using DATABASE I:

The finished dictionary entries of SIDict were placed continuously into a textual data base, which will be referred to DATABASE I. The GREP program was used to look up any character sequences in their contexts which the editors and/or authors of SIDict requested. This way this we gave an effective help both to the editor in chief as well as to the authors to have dictionary entries in the SIDict that are uniform in their form as well as content. In the following we shall illustrate this by two concordances made by the program GREP. In case of the first, we show a section of a dictionary entry classified as "private" [=(biz)], while in the second we demonstrate "idioms" quoted in the dictionary entries. The code for idioms is <66> appearing as * on the printer.

EXAMP.001

nyakleves fn (biz)
Kapott egy ~t, a nyakára v. tarkójára mért ütést.
nyakra—főre hsz
(biz) ~ *hívogat:* újra meg újra.
nyargal ige *~ni*
2. (biz) *Hova ~sz* (= rohansz, szaladsz) *sebesen?*
nyavalyog ige *~ni* (biz)
1. *Tavaly sokat ~tam,* betegeskedtem.
nyel ige *~ni*
4. (biz) *A kocsi ~i a kilométereket,* gyorsan halad.

EXAMP.002

hó\1 fn *havak, havat, hava*
* *Fehér, mint a ~:* tiszta fehér.
holló fn *~k, ~t, ~ja*
* *Ritka, mint a fehér ~:* nagyon ritka.
holt II. fn *~ak, ~at, ~ja*
* *Nem volt se ~, se eleven:* nagyon megijedt.
homlok fn *~ok, ~ot, ~a*
* *Nincs a ~ára írva:* nem látszik rajta.
homok fn -, *~ot, ~ja*
* *~ba dugja a fejét:* nem hajlandó tudomást venni a veszélyről.
* *~ra épít(i terveit, elméletét):* bizonytalan alapra.

4. SIDict as a dictionary data base: DATABASE II

Through the computerized typographical preparation the text body of SIDict was placed into a computer data base called DATABASE I. However, we were compelled to carry out two significant modifications on the data base created during the typographical preparation of SIDict in order to get a computerized version of the dictionary that suits also

the purposes of teaching and linguistic research. Thus by modifying DATABASE I we developed DATABASE II.

Before placing the text of SIDict written in DATABASE I into DATABASE II, we had to carry out the following two modifications:

1) Modification (changing the tilde)

Within DATABASE II it is no longer practical to have the entry heading substituted by the tilde. It is one of the characteristics of the Hungarian language that the last vowel of some stem words are changed when a suffix is added to the stem. In such cases we use tilde with a comma for denoting the stem variant of the suffixed dictionary entry e.g.

```
gólya    (+k,+t,+a) = gólyák, gólyát, gályája       = ~k, ~t, ~ja
gondola  (+k,+t,+a) = gondolák, gondolát, gondolája = ~k, ~t, ~ja
görbe    (+k,+t,+a) = görbék, görbéket, görbéje     = ~k, ~t, ~je
ige      (+k,+t,+a) = igék, igéket, igéje           = ~k, ~t, ~je
```

The "tilde" and the "tilde with a comma" was changed into the lexical entry i.e. its appropriate form variant after the elaboration of the algorhytm by a program written in C language.

2) Modification (morphological analysis)

It seems to be practical to place the dictionary corpus of SIDict DATABASE II that has been morphologically analyzed and lemmatized. This is due, on the one hand, to the fact that the Hungarian language has an agglutinative structure and, on the other to the stem variation mentioned above. The morphological analysis and lemmatization was carried out by the program HUMOR developed by László Tihanyi and Gábor Proszéky.

In the forthcoming we demonstrate this by the showing the previous detail analyzed by the program HUMOR:

```
<5><41>híre-hamva [FN]<40> %szragos %fn (csak [HA] esz
[FN] 3. személy [FN] ·ben [INE], %ált. alanyeset [FN]·
ben [INE]) <4><2>a [DET] tegnap [FN] & le [IK]· esik
[IGE]≈es· ett [Me3]& hó [FN]· nak [DAT] ma [FN]& már [HA]
híre-hanva sincs [IGE], <1>nyom [FN]· a [PSe3] sincs
[IGE], eltűnt [MN]&.

<5><41>híres [MN]<40>
<4><3>I. [SZN]&. <1>%mn <2>híres%ek, híres%t <1>%v.
<2>híres%et, híres%en
<3>1. <2><52> ember [FN], író [FN], politikus [FN]&,
színész [FN]: <1>ki [NM]· emelkedő [FN]
tulajdonság [FN]·ai [PSe3i]·ról [DEL], teljesítmény
[FN]·é [PSe3]·ről [DEL]& jól [HA] ismert [MN]& (=
hírneves [MN]). <2><52> regény [FN], dal [FN], város
[FN]: <1>közismert [MN], nevezetes [MN]. (biz [HA])
<2>Nem [HA]& valami [NM]& <52> ez [NM] a [DET] dolgozat
[FN], <1>gyenge [MN] minőség [FN]·ű [UKEP], középszerű
```

```
[MN]. <3>2. <1>(%pejor) <2>Megint [HA] itt [HA] van
[IGE]·nak [t3] a [DET] híres, <1>rossz [MN] hír [FN]·ű
[UKEP] (= hírhedt [MN]) <2>barát [FN]·aid [PSe2i].
<3>%II. <1>%fn <2>híres%ek, híres%et, híres e [NM]&
<1>(%pejor) <1>Kétes [MN] hír [FN]·ű [UKEP] személy [FN].
<2>Hol [HA] az [DET]& a [DET] híres? <1>(Megszólítás
[FN]·ban [INE]:) <2>No [ISZ], te [NM] híres!

<5><41>híresség [FN]<40> %fn <2>híresség%ek, híresség%et,
híresség%e
<4><3>1. <2> apa [FN]≈Ap·já [PSe3]·nak [DAT] híressége
[NM]&, <1>híres [MN] volta [FN]& <2> nemcsak [KOT] segít
[IGE]·ette [TMe3]&, akadályoz [IGE]·ta [TMe3]& is [KOT]
pálya [FN]≈pályá·já [PSe3]·n [SUP]. <3>2. <2>Az [DET]&
előadás [FN]·ra [SUB] meg [IK]·hív [IGE]·ták [TMt3]& az
[DET]& összes [MN] híresség%et, <1>híres [MN] ember
[FN]·t [ACC].
```

Presently we are currently evaluating retrieval programs that could be used the to query dictionary DATABASE II. The simplest solution would be to use GREP, however we are still making some experiments with programs like KAYE and WordCruncher which serve more purpose and thus can be used in a more comfortable way.

As the dictionary data base has already been analysed morphologically and is full of codes, therefore it is necessary the filter the output resulting from the individual searches through a special program filtering out the codes and transforming the text into a form similar to that of a "normal dictionary".

5. Analysing the vocabulary of SIDict

We want to carry out the analysis of the SIDict vocabulary, on the basis of DATABASE II, in two directions:

5.1. The relationship between lexical entries and words occuring in the dictionary (lexemes):

We want to explore how the 14.000 envisaged lexical entries relate to the words occuring in the the text body od SIDict. Points of view for the research:

a) Are there any words in the text body that are not lexical entries?

b) Are there any lexical entries that occur only in their own lexical record?

Of course we analyse the vocabulary of definitions and sentence examples within the text body separately.

5.2. Classifying the lexical entries and their meaning nuances:

SIDict is made for pupils and, according to the plans, contains words of the basic stoek of vocabulary. This is the reason why we think it worth examining what is the proportion of so-called "classified" lexical entries in comparison to the 14.000 lexical entries and what proportion of the meaning nuances of the lexical entries was classified. Lexical entries classified in the dictionary are to be found on the margin of the base vocabulary. (The classifications were the following: (nép) [= dialectal], (biz) [= family usage, colloquial],

(pej) [= pejorative], (rég) [= archaic], (tréf) [= jocular], (vál) [= refined]). The results of these examinations will yield objective data as regards the size of a planned basic vocabulary treasury. It seems to be most likely that the size estimated by Júlia Pajzs (about 10.000 to 12.000 lexical entries) will be justified (PAJZS 1991).

With the computer-stored data base of SIDict, lexicographers, linguists and those who are interested in such topics will have for the first time a computational explaining dictionary of Hungarian. Due to the students's dictionary character of SIDict we think it can be utilized excellently for educational purposes. The experiences gained while preparing it will be extremely useful in the future when working on more voluminous computational dictionaries e. g. *A magyar irodalmi és köznyelv nagyszótára (1533—1990)* [= the Great Literary Dictionary of Hungarian] (KISS — PAJZS, PAJZS 1988, PAPP — HEXENDORF).

Bibliography:

BÍRÓ ÁGNES: *A Képes Diákszótár módszertani kérdései* [= Methodological questions in connection with the Students' Picture Dictionary]. Az I. magyar alkalmazott nyelvészeti konferencia, Nyíregyháza 1991. május 3—4. Előadások: 56—61. oldal.

KISS GÁBOR: *A Word Perfect szövegszerkesztő programozási lehetőségeinek felhasználása szövegek szótárszerű feldolgozásának előkészítésében — Bemutatva a Vizsolyi Biblia négy Evangéliumán* [= Utilizing the programming possibilities of Word Perfect when preparing dictionary-like elaboration of texts - illustrated on the Four Gospels of the Bible of Vizsoly]. Az I. magyar alkalmazott nyelvészeti konferencia, Nyíregyháza, 1991. május 3—4. Előadások: 392—404.

KISS LAJOS — PAJZS JÚLIA: *A magyar irodalmi és köznyelv nagyszótára (1533—1990)* [= The Great Dictionary of Hungarian Literary and Everyday Language (1533—1990)]. Magyar Nyelv 1989. LXXXV. évf. 2. szám. 129—136. oldal.

LAROUSSE DÉBUTANTS, 20.000 Mots, Direction de René LAGANE, Libaire Larousse, 1986.

OXFORD BASIC ENGLISH DICTIONARY. Edited by Shirley Burridge, Oxford University Press 1981.

PAJZS JÚLIA: *Creating a Historical Dictionary of Hungarian with the Aid of Computer.* BudaLEX '88. Proceedings, Papers from the EURALEX 3rd International Congress, Budapest, 4—9 September 1988. T. Magay and J. Zigány (eds.). p. 559—563.

PAJZS JÚLIA: *A Debreceni Tezaurusz egyik felhasználási lehetőségéről: a magyar nyelv számítógépes alapszókincstára* [One of the possibilities to utilize the Debrecen Thesaurus: the basic vocabulary collection of Hungarian on computers]. Könyv Papp Ferencnek, Tanulmánygyűjtemény Papp Ferenc 60. születésnapjára. Szerkesztette: Hunyadi László, Klaudy Kinga, Lengyel Zsolt, Székely Gábor. Kossuth Lajos Tudományegyetem Debrecen, 1991. 343—348.

PAPP FERENC — HEXENDORF EDIT: *Magyar szókincs a könyvnyomtatástól napjainkig — számítógépre tervezve* [= Hungarian vocabulary since the invention of book-printing till the present - designed for computers]. Magyar Tudomány 1985. XXX. évf. 1. szám 36—40. oldal.