Wading through Letter A:The Current State of the Historical Dictionary of Hungarian

Júlia Pajzs
Research Institute for Linguistics
Hungarian Academy of Sciences
H-1068 Budapest Benczúr u. 33, Hungary
pajzs@nytud.hu

Abstract
The Historical Dictionary of Hungarian project has been underway for some years. The first stage consisted of the collection and analysis of a 25 million word Historical Corpus of Hungarian. Work has now begun on the dictionary itself. It will contain approximately 100,000 entries in eight volumes. Each sense of each headword will be illustrated with actual corpus examples, including bibliographic references to the sources. The period to be covered is 1772-2000. The present paper summarizes lessons learnt while working on the letter *A*.

Introduction

The Historical Dictionary of Hungarian project has been underway for some years. We collected and analysed the Historical Corpus of Hungarian (Pajzs 1991, 1997, 2000, 2002) over a number of years, while at the same time experimenting with ways of compiling draft entries. After extensive debate about what the dictionary should contain, the number of projected volumes, as well as the historical span to be covered, we compromised on an eight-volume dictionary, with some 100,000 entries, covering the period between 1772 and 2000. The entries are based not only on the electronic corpus, but also include old-fashioned, hand-written dictionary slips from earlier attempts. Each sense of each headword will be illustrated by corpus examples, including bibliographic reference to the sources. The design restricts the maximum number of examples from the sources: only the earliest quotation for each sense of the headword is to be included. The decision restricts the number of examples per sense to one, which is normally the earliest quotation found. Since this strict rule has proved too restrictive, I suggested including some "hidden" quotations, to be available only in the electronic version of the dictionary.

The manual for the preparation of the dictionary was published last year (Ittzés 2002), and in the same year the actual writing of the entries finally got underway. By mid-2003 we reached letter *B*, which was a historical moment in the life of this extremely long-term project. In the present paper I would like summarize our experience in producing the entries for the letter *A*. This does not, sadly, mean that every single entry in *A* has been finalized: from the 4800 headword lemmas selected for letter *A* over 97% were finalized by spring 2004. The analysis is based on these entries.

1. Selection of the Headword Lemmas

The selection is based on several (re)sources. First, the major Hungarian monolingual dictionaries are consulted. Most of the 75,000 headword lemmas of the one volume monolingual dictionary of Hungarian (Pusztai 2003) will be present in our dictionary as well. With the inclusion of a selection of words from the nineteenth century historical dictionary of Hungarian (Czuczor–Fogarasi 1862) some archaic or infrequent words can be saved for posterity. Certain words chosen from the dictionary of foreign words (Bakos 2002), a distinct Hungarian genre, were quite frequently used in the nineteenth century, others have recently revived due the influence of Europeanization and globalisation

Naturally, we are adding many hitherto unregistered words, in so far as the restricted scope of the work makes this possible. A number of interesting ancient and/or rare words have been found in the corpus (eg. *ágyszék* `'sofa'`, *akaszték* `'loop (on clothing)'`, *alamuszta* `'laisy, clumsy'`). Neologisms, too, are included.

2. The Structure of the Entries

The structure of the entries is compiled in an XML DTD, based on the TEI recommendations. TEI tags are used whenever a corresponding item was found in the recommendations, while for items not present in the standard a short English-based tag set was prepared. The DTD was based on our style sheet, prepared in cooperation with the editors. In cases where the lemma is classified as more than one part of speech a block called 'sensegroup' <*sengr*>. Instead of presenting the whole DTD, here I show only the skeleton structure:

entry (head, (sengr*|sense*))
sengr (sense+)
sense (def, cit, (subsen*))

The element referred to as <subsen> is a subsense, or a shade of meaning. In the printed version these will be marked by letters of the alphabet (a, b, c, etc.).

A sample entry is presented in Figure 1.

árnyalat fn

1. 'vmely szín tőle enyhén eltérő változata': ||**1a.** 'vmely színnek sötétebb, ill. világosabb fokozata': A hegy alján végig zúdul a [...] mindig ezer zöld *árnyalatban* játszó kristály tiszta Traun (1877 Wohl Janka, Wohl Stefánia C4560, 74) | Az öltönyökben "ő" a kék színt szereti, annak minden *árnyalatát* (1880 Jókai Mór C2301, 146). || **1b.** 'vmely színnek egy másikkal kissé keveredő változata': {Színeik [ti. a halaké] minden *árnyalatot* felöleltek} (1947 Szabó Lőrinc 2000629171, 57) | [Werner] haja hirtelenszőke {és olyan a bajusza is,} némi rőtes *árnyalattal* (1957 Rónay György C3638, 5).
2. *(átv is)* 'kis / csekély / jelentéktelen különbséget / eltérést mutató változat, fokozat': egy nő' tetszelgésének ezer [...] *árnyalatai* vannak (1847 Kemény Zsigmond C2599, 49) | a buzgóság közt is van ezer *árnyolat* (1852 Kemény Zsigmond C2593, 205) | [Clermontné] egy *árnyalattal* tartózkodóbban fogadta, mint máskor (1895 Abonyi Árpád C0473, 84) | senki sem érti és érzi meg a költői nyelv minden *árnyalatának* varázsát az első hallásra (1918 Babits Mihály C0696, 222) | a vad világ többé nem oly sivár: Egy *árnyalattal* tűrhetőbb (1928 Tóth Árpád C4230, 146) | Soltész Anni anyós-figurája [...] gondoskodik a komikum erősebb *árnyalatairól* (1964 Magyar Nemzet márc. 8. C4801, 10) || **2a.** *(kissé rég)* 'politikai nézet, irányvonal, áramlat': Voltak *árnyalatok* a nyilvánitott nézetekben, de azon időben még csak *árnyalatok*, ellentétek nem (1860-1862 Kossuth Lajos ?, ?) | {Öten vitatkoznak s} mindegyiknek más a politikai *árnyalata* (1889 Justh Zsigmond C2463, 14) | a függetlenségi párt két *árnyalatra* szakadt (1900 Budapesti Napló szept. C0057, 6).
3. *(ritk)* 'árnyék': (1843 Athenaeum C0027, 360) | az élet oly sok *árnyalattal* bir, s csak egyetlen fénye van, és ez a szeretet (1872 Bajza L, 11).
Vö. CzF.; ÉrtSz.; TESz. *árnyék*; ÉKsz.
Szócikkíró: *Kéthely Anna*; első változat kelte: 2003. 07
Szerkesztő: *Ittzés Nóra*; szerkesztés kelte:
Cédulák száma: . Korpuszbeli adatok száma: 591.
A szócikk állapota: 1.

Figure 1: The entry **árnyalat** 'nuance'

The quotations in braces {} are the so called "hidden examples", only to be included in the electronic version of the dictionary. Sometimes only parts of the quotations are hidden.

The result of the statistical analysis of the completed entries is given in Table 2.

	A Subsense	B No subsenses	A over B	Total	
1 sense	997	914	1,09	1911	52%
2 senses	627	350	1,79	977	27%
3 senses	323	109	2,96	432	12%
4 senses	163	24	6,79	187	5%
>4 senses	164	15	10,93	179	5%
				3686	100 %

Table 2: Statistical analysis of the completed entries

A comparison with the structure of largest monolingual dictionary of Hungarian (Bárczi et al 1959-1962) shows that the major features are quite similar: over half of the words have only one meaning, either with or without a submeaning, a quarter of the words have two meanings, while the number of words having four or more meanings is around 10 per cent.

3. The Illustrative Quotations

The citations are chosen from a wide variety of (re)sources: the 25-million-word Historical Corpus of Hungarian, some 6 million traditional dictionary slips, and a collection of prose fiction published on CD-Rom,

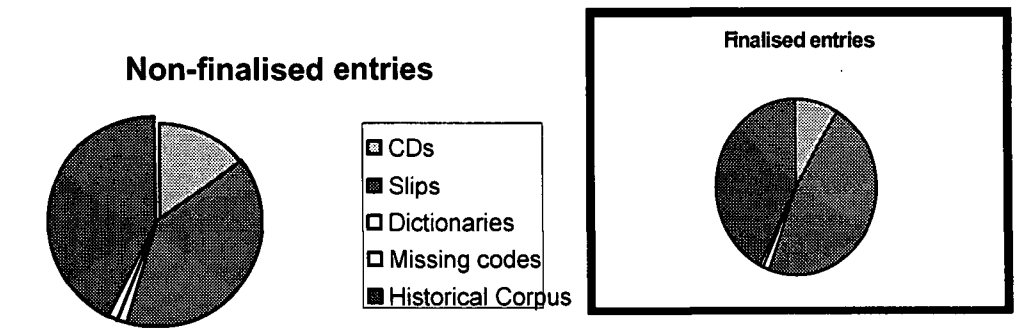

Figure 2: The use of different resources

newspaper materials and the major monolingual dictionaries. The actual use of these resources is illustrated in Figure 2. We may note that the use of the CD resources is more frequent in the non-finalised (i.e. the prepared, but as yet unedited and uncorrected) entries than in the final versions. This may be due partly to the fact that the unedited entries contain

a far higher number of less carefully chosen quotations than the final versions, because the majority of the lexicographers in our team are new or fairly new to the field and thus not yet fully confident in their choice of the most suitable illustrative quotation. Another reason is that our CD resources are continuously expanding and thus the number of quotations from these is growing as we proceed with the work.
Although the number of illustrative quotations is fairly restricted, by using the "hidden" examples it is possible to display the uses of the given meaning during the whole period covered by the dictionary. Figure 3 shows the distribution of the quotations by date.

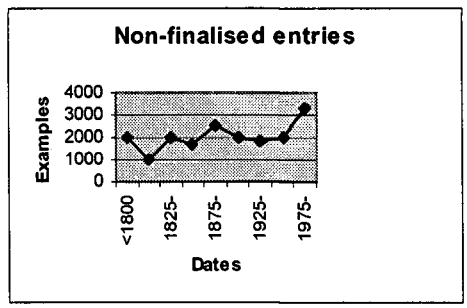

Figure 3: The distribution of the quotations by date

As we are concentrating on the earliest quotations for each sense, the graph has a local maximum at the starting point of the time interval, and the absolute minimum at the next quarter-century, where the number of hidden examples is also limited. A local maximum can be seen in the last quarter of the 19th century due to two reasons. On the one hand, the lexicographers try to illustrate the senses as used in the late 19th century, and on the other hand, the most frequently used CD resources (a standard multi-volume 19th encyclopaedia, the *Pallas*, and the complete works of two well-known Hungarian authors, Jókai and Mikszáth) were also produced in that period. The surprisingly large maximum in the final quarter of the whole period is the result of similar factors: the most frequently used resource is a CD containing a daily newspaper from the period 1994-1998 (*Magyar Hírlap*). A comparison of the two graphs reveals that the graph of the final version entries has a slightly narrower range than that of the non-finalised entries.

Figure 4. presents the graph of the earliest quotations, showing a similarity to the former graphs up to the end of the 19th century. For later periods there is a nice and smooth convergence.

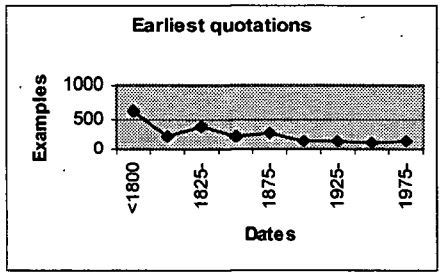

Figure 4: The graph of the earliest quotations

In spite of inequalities in the representation of certain authors and sources, on the whole the use of the various sources is fairly balanced: the examined 25,000 sentences were taken from over 4,000 different extracts from over 2,000 different authors (the 25-million-word corpus is derived from about 2,500 authors). Continuous monitoring is helping us identify errors to be avoided in the future.

Naturally, in the dictionary database any word in the illustrative sentences can be searched for. Contrary to the tradition of most English monolingual dictionaries, our lexicographers do not especially try to avoid sentences containing well-known proper names. It is both exciting and instructive to search for some of Hungary's best-known figures, as this forces us to consider our chosen example from a different angle. For example, we had to confront a rather ambiguous image in the case of one of Hungary's greatest poets Endre Ady (1877-1919), who was notorious for his provocative behaviour. It may well be advisable not to over-represent these events of his life in the largest monolingual dictionary of Hungarian, which will provide a very important perspective on Hungarian culture as a whole. It is also revealing to search for the names and events of the times before the fall of communism in Hungary in 1989. For example, the name of Hungary's ruler, János Kádár, who was in power for more than 30 years (1956-1988), occurs in the illustrative sentences more than 20 times, mainly in unfavourable contexts, and it often collocates with the words *apa, atya* "father", illustrating that he was considered a father figure, rather than a "big brother". One of my favourite quotations used as a partially hidden illustrative example in the entry *apa* is the following:

> *Amikor kissrác voltam, azt mondták körülöttem. Horthy apánk. Az idősebbek még úgy mondták, Ferenc Jóska apánk. Aztán jött Rákosi apánk. Ment is. Aztán jött Kádár, és amikor már falun azt hallottam, hogy Kádár apánk, akkor én mentem (Konrád György: Agenda. Kerti mulatság, 1989.)*
>
> 'When I was a lad, I heard about our father Horthy [governor of Hungary between the two World Wars]. The elderly still said: our father Franz Joseph [Habsburg emperor and king of Hungary before WWI]. Then came our father Rákosi [communist party leader in the 1950s]. Then he went. Then came Kádár, and as soon as I heard in the countryside: our father Kádár, it was me who left'.

The author of this quotation is the dissident writer György Konrád, silenced for more than fifteen years because of his political views.

Future Plans

The first volume of the dictionary, containing letter *A-B*, is due to appear in 2005 or 2006. The electronic version of the dictionary will be published alongside the printed version, containing the additional hidden examples and the database of the bibliography of the (re)sources. We are also considering the inclusion of some additional information in the database version. One idea is to add some specification of the domain of the sense, so that one can search for all the words which refer to colours, or names of flowers, or vehicles, etc. The hierarchic specification still has to be worked out, and the feasibility of this proposal must be tested on a sample of entries. Our international advisor Anna Braasch suggested using the SIMPLE ontology for this task. The application of the framenet approach is also being considered. Another very interesting idea was proposed by our other international advisor Peter Sherwood, to include the English equivalent of the meanings in the database version. This would certainly enlarge the number of the possible users of the future dictionary, but it would necessitate a corresponding bilingual dictionary project. Since our major problem is the lack of adequate financial and human resources, these ideas might remain in the realm of dreams.

Conclusion

In spite of being a computational lexicographer, during the long years of corpus collection, correction and analysis, I often felt that it would have been more sensible to compile the dictionary in the more traditional, tried and tested, way, namely by using only the old archive dictionary slips and the good old paper-pencil-eraser method (especially after some major virus infections and other dramatic computer disasters). However, as we survey the almost- finalised database for the letter *A*, it is tremendous to be able to check and search the already prepared entries. It is definitely worth the effort, after all.

Acknowledgements

The project is financed by the following grants: OTKA T042705, NKFP 5/142, OKTK 2004., TKI 2003-2006, and mainly by the Hungarian Academy of Sciences.

References

Bakos F (2002) *Idegen szavak és kifejezések szótára* 'Dictionary of Foreign Words' Akadémiai Kiadó, Budapest.

Bárczi G, Országh L (chief eds.) (1959-1962) *A magyar nyelv értelmező szótára I.-VII.* 'The explanatory dictionary of Hungarian' Budapest, Akadémiai Kiadó

Czuczor G, Fogarasi J (1862) *A magyar nyelv szótára I.-VI.* 'The dictionary of Hungarian' Pest, Emich Gusztáv Magyar Akadémiai Nyomdász

Ittzés N (2002) Az Akadémiai nagyszótár szerkesztési szabályzata 'Style sheet for the Academic Dictionary of Hungarian' *Mutatványok az Akadémiai Nagyszótárból.* MTA Nyelvtudományi Intézet, Budapest, 12-98.

Pajzs J (1991) The Use of a Lemmatized Corpus for Compiling the Dictionary of Hungarian *Using Corpora Proceedings of the 7th Annual Conference of the OUP & Centre for the New OED and Text Research.* Waterloo, University of Waterloo, p. 129-136.

Pajzs J (1997) Synthesis of results about analysis of corpora in Hungarian. *Linguisticæ Investigationes* XXI/2: 349-365

Pajzs J (2000) Making Historical Dictionaries by Computer. *Proceedings of EURALEX 2000.* Ulrich Heid ed. University of Stuttgart, Stuttgart, 2000. p. 249-259.
Pajzs J (2002) A Corpus Based Investigation of Collocations in Hungarian *Proceedings of EURALEX 2002* Center for Sprogteknologi, Copenhagen, 2002. p. 831-840.
Pusztai, F. & al. (2003). *Magyar értelmező kéziszótár* 'Concise Dictionary of Hungarian.' Akadémiai Kiadó, Budapest