Making Historical Dictionaries by Computer

Júlia PAJZS, Budapest, Hungary

Abstract
The paper examines the difficulties encountered when compiling a historical dictionary from scratch. The well-known unabridged dictionaries were mainly made by traditional methods, but today it seems obvious that computer corpora should be used for new projects, as well as for updating existing dictionaries. Through the example of the ongoing project of the "Historical Dictionary of Hungarian"[1] I would like to draw attention to the specialities of historical dictionaries and the limitations of using computerised corpora for compiling them.

1 Historical dictionaries

Although the greatest historical dictionaries were created during the late 19th and early 20th centuries, an increasing number of representatives of lesser known languages still feel the need to record the changes of their vocabulary in unabridged, "OED-like" dictionaries. Several nations were not in a position to promote their own languages, rather they were supposed to forget it and assimilate. While people increasingly tend to use English as a common language, they are not willing to forget their mother tongue, they seem to become more aware of the fact that 'small languages' can only keep their identity if they are accurately described in dictionaries and grammar books.

When reading the success story of the corpus-based dictionaries like COBUILD, CIDE and other thoroughly revised ones [LDOCE3 1995], enthusiastic cultivators of language might have the misapprehension that all that is required is to feed a large historical corpus into a computer, press a button or click a mouse, and out comes the ready-made historical dictionary. Naturally enough this idea is not so much cherished by lexicographers, rather by the representatives of publishing houses and other institutions which are likely to finance large, monolingual dictionaries.

What makes the main difference between up-to-date corpus-based dictionaries and traditional unabridged ones? A historical dictionary supposedly contains more or less every word which was ever written in the period covered by it, and the senses follow each other in order of their occurrence. Each sense and subsense is illustrated by several examples, which are again in chronological order. The earliest and latest attested occurrences of a word in a given sense is a major point, which is thoroughly up-dated in the revised versions of these dictionaries. As [Landau 1989: 71] writes on the great model: "The OED not only provides a historical record of the development of meaning of each word, with illustrative quotations and definitions for each sense. It also shows the changes in spelling, the different forms each word assumed during its history. It gives by far the most complete and authoritative etymologies that existed up until that time, a body of information that is still unchallenged as a whole. The divisions of sense are the most detailed and exacting, the definitions the most precise and clearly substantiated, of any English dictionary." In addition to this "a large dictionary is first class reading. Murray's would be as good a companion on a desert island as a man could hope for, as apart from the history

of the words, the quotations are endlessly entertaining in themselves. It is like having all the birthday books and literary calendars ever written rolled into one" quoted by [Considine 1998: 580].

Most of these dictionaries were prepared by using traditional slips as sources. The project for the "Trésor de la langue française" was one of the first to be based mainly upon a computerised corpus, combined with a traditional citation collection. Although the use of slips seems to be hopelessly outdated, they are more appropriate for this kind of dictionary from several points of view. The readers who collected them were intelligent persons who picked up only those quotations which attested a new, interesting, unusual meaning of a word. In a computer corpus, on the other hand, several words have thousands of occurrences, some of which can be really new or interesting, but it is hardly possible to realise them in a huge concordance.

2 Corpus use and entry compilation

2.1 Corpus collection for historical purposes

The best way of collecting a corpus is still debated. There are many reasons to use what are called "opportunistic corpora" in which everything is collected that is available in electronic format. You can also try to prepare a more balanced and representative corpus by throwing away some parts of the available texts and adding new ones [Sinclair 1991] and creating what is a "monitor corpus". Representativity in itself is sometimes questioned [Biber/Conrad/Reppen 1998]. It is certainly a lot quicker, easier and cheaper to maintain an opportunistic or a monitor corpus and, if large enough, it might even be adequate for a dictionary of the present day. However, while preparing a corpus for a historical dictionary one must be meticulous. The selection of the corpus is easiest when the vocabulary to be covered is closed, as for example in the case of the Dictionary of Old English. The closer you are to the living language, the more difficult it is to choose the texts to be recorded. If you decide to make a corpus of small excerpts for the sake of representativity, as we have done for the "Dictionary of Hungarian", you must be aware that recording itself will be rather complicated, slow, and the result will be still far from being perfect. One always has the feeling that so many other texts should have been included, and it is really difficult to decide when to finish the collection (if ever). For long term projects it can also be a problem if one continues to maintain the corpus during compilation of the dictionary: the last volume will contain quotations from earlier or later sources than the first one.

To find the correct compromise between a thoroughly selected representative corpus and an opportunistic corpus is not easy either. One can choose only among the texts which are already available in electronic format and decide that the dictionary will only cover the vocabulary of these sources. However, electronic texts usually do not contain any reference to the page number of the printed version for obvious reasons. Historical dictionaries have considered the exact bibliographic reference of each quotation inevitable so far. This means that in order to be able to use the available electronic texts, they have to be supplied with the page numbering of a specified printed version. Although it is much less work than keyboarding, one might argue that the traditional notion of philological thoroughness should be reconsidered. The main point in giving the page number was to make it possible for the users of the dictionary to find the

actual text in a book. Once the source corpus itself is constantly available through the internet, browsing of the larger context of the quotation is more feasible on line. If we insist on including the page numbering in the corpus we loose the possibility of simply updating the corpus with newly available texts.

Whilst deciding the way of collecting the corpus and its planned size (if there is a final planned size) you must also take into account the problems caused by too rare and too frequent words. After lemmatisation of the 17 million running-word Hungarian corpus we found approximately 180,000 possible headword lemmas. More than half of them only occurred once, while about 10,000 occurred hundreds of times, and these covered about 70% of the whole corpus. Only about 55,000 (less than one third) seemed to be attested by a "comfortable" amount of examples, neither too few nor too much. The large amount of occurrences can only raise problems if the compilers insist on the traditional method of reading every single quotation to make sure they did not leave out a new or interesting sense of the word. From the above numbers it can also be concluded that historical corpora should be a lot larger to contain enough entries. On the other hand, the larger we make the corpus the problem of handling the "too frequent" words becomes more and more serious.

The traditional historical dictionary very accurately contains the first and latest occurrences of each sense of the words. To be able to order the concordance for the date of writing this must be recorded in the corpus in a retrievable format. In the case of a corpus containing several different texts this again necessitates a meticulous philological work. If the date is recorded properly, the first and latest occurrences of a character string can be searched relatively easily, but it is not so simple to match them to actual senses of words. Even if the first and latest quotations can be matched to each sense by fastidious lexicographic work, one must be aware that these were only the first and latest examples in the corpus, but not in the whole language or not even in the period which was aimed to be covered.

2.2 Analysis and retrieval of the corpus

In order to be able to search words, not just character strings, it is necessary to apply some kind of analyser or tagger tool before retrieval. Although most of these tools claim that they are language independent, it only means that as soon as the morphology of the language is described in the format required by the tool, it is able to analyse or tag your language. The main difference between tagging and analysis is that taggers usually only supply the running words with part of speech codes and some inflectional information, and the analysers actually segment the word into stem and suffix(es). So while a tagger can identify that '*says*' is a verb in present tense, third person singular, the analyser can segment it and identify '*say*' as the verbal root and '*s*' as the suffix. While English morphology is relatively simple, some languages, among them Hungarian, have a highly complex morphology. That was the reason to develop the Humor morphological analyser shortly after the beginning of the dictionary project [Prószéky 1996]. Since it is also used as a spell checker, it is continuously revised. It is able to recognise and analyse quite complex words, even when the stem of the word changes. It can be efficiently used mainly for current texts, but it could correctly identify a large part of the texts written in the 19^{th} century as well. The same tool was used/tested for some other languages, but for the real working version an exact morphological database is necessary, which contains the stems

and possible suffixes of the language, supplied with an accurate morphological code. As most of the analyser tools, this one was also developed for current texts. In the case of historical corpora one must be able to recognise earlier words, archaic forms as well. For this, special morphological databases should be created, which simply cannot be merged with the databases of the current texts. So, for example, according to current Hungarian orthography when the word *asszony* 'woman' is followed by the instrumental suffix *–nyal*, one of the *y*-s should be omitted, and written: *asszonnyal* . The current spell checker and analyser should not allow it to be written as *asszonynyal*, although it was often spelled this way in earlier texts. There are also several old suffixes which do not exist any more, or not in the same format, and which are either not recognised by the analyser or misinterpreted. A modular analyser tool, which recognises the correct words written during many centuries can hardly be developed within the framework of a dictionary project. Rather it should be made in separate projects, where historical linguists, computational linguists and lexicographers can co-operate efficiently.

Given an accurate morphological database of the language, one can also choose some other methods for retrieval of the possible headword lemmas. Among others, the Intex© [Silberztein 1999] software can index the running texts according to lemmas. For this, a database has to be created which contains all of the possible inflected forms of the language (a DELAF dictionary). This seems to be a feasible approach for many languages with a simple morphology. It was successfully applied to several languages: Bulgarian, French, Italian, Serbian for example. We are planning to test it for Hungarian as well. In our case we will have to face special difficulties, because of the huge number of possible inflected forms. For testing the method itself, we will first try to use it on the more frequently occurring forms of the most frequent words. The main advantage of using this tool is that the analysis and retrieval can be managed in one step. During the very quick index process the program creates a full word list and afterwards one can look at every occurrence of each word or word combination in various sizes of context. Regular expressions can also be used for retrieval so linguistically relevant information is made available in this way.

Since many inflected words are ambiguous, some taggers are also supplied with a disambiguator tool. The most efficient ones usually work with different kinds of statistical methods, for example the HMM which was developed and used in the Multext and Multext-East Copernicus projects, or the [Brill 1994, 1995] tagger which was originally tested on English but is more or less successfully used for other languages as well. There are some attempts to use linguistically more intelligent solutions by the help of local context grammars. Among others the Intex© software has a module in which it is relatively easy to write simple local context rules and test their effect on the corpus right away. Local context rules were also tested on the Hungarian corpus [Pais /Pajzs 1998], by using regular expressions written in Perl. Some statistical approaches were also tested [Meggyesi 1999], [Oravecz 1998]. For the time being statistical methods seem to be more accurate, but if there are good syntactic and semantic analysers for a language, one can expect much better results by using them.

The available corpus retrieval software is usually language independent. We started to use the Open Text© SGML retrieval software several years ago, when it was a pioneer tool. Since that time several more linguistically oriented programs were developed, and some of them are available from universities or research centres by a simple agreement, if they are used for research purposes. (e.g.: the DBT concordance program made in the University of Pisa, the Corpus Word

Bench program from the University of Stuttgart). Since they are available free for researchers, they are often not very easy to handle for new users, they might not be documented and supported well enough. Therefore it is not too simple to test several of them for one's own language, in order to be able to decide which one is the very best for your purpose, especially at the start of a brand new dictionary project, when you are not quite sure yet what you will need from the corpus. To make the proper choice even more difficult, the hardware and software environment must be changed every 3-4 years usually, which does not necessarily mean that the old and very much liked tools will still work on them. For a long term project it is usually more advisable to try to purchase the very best software and hardware environment at the beginning of the project and try to stick to it as long as possible. It is also worthwhile to choose a well-known software enterprise to support the project, rather than trying to make everything with a seemingly inexpensive in-house staff.

Most of the retrieval tools are only able to search the words, but hardly any of them can help you to distinguish the different senses of those words. If you already have an on-line electronic dictionary or even better, a real lexical database, the differentiation of the senses can be greatly helped by semiautomatic methods. Some interesting suggestions in this field were already made by [Clear 1994], [Atkins 1996], [Ooi 1998]. In her paper, Atkins envisaged a "dictionary of the future" where you could easily search for the semantic features of the words. (For example, verbs which express movements, or even slow or quick movements or movements made by typical actors etc.) For this, a lexical database must be created which contains information on the semantic and grammatical properties of the words not so much in human readable but rather in "computer digestible" format. She suggested to use Fillmore's frame semantics for this purpose, but this of course is only one of the possible methods for this task. Ooi describes the Datr lexical knowledge representation language as an alternative solution to record semantic databases, and he also shows some specimen lexical entries based on corpora. As soon as you have a semantically coded lexical database for at least the core vocabulary of a language, it is much more easy to improve methods for finding either typical quotations for already known senses or to guess the appearance of a new meaning. Statistical observation can also help to realise new meanings, again [Ooi 1998:144] mentions the Z-score method to measure collocational strength. Clear's idea on distinguishing senses of quotations was also based upon the frequency of the collocates of the words.

2.3 Compilation of the dictionary entries

The compilers of the "Trésor de la langue française" not only used the computerised corpus but they had access to several millions of traditional dictionary slips as well. The lexicographers were also supplied with the full bibliography of the entry and received a photocopy of the same word in other dictionaries. This made it possible to integrate all former knowledge into the Trésor. Even with this method I can hardly imagine how they could cope with the entries with thousands of examples but they must have managed somehow as the dictionary was completed and published.

When the collected corpus is believed to be sufficiently large and representative of the targeted vocabulary, the actual dictionary writing can be started. No matter how large the corpus is, you will very soon realise that it is never really large enough for covering everything you originally

intended. In that phase you might either decide to compromise with what you actually have or to enlarge the corpus infinitely.

To check the coverage of the vocabulary of the corpus, one can make a list of the words occurring in the corpus either by the retrieval tool, or by a purpose built tool. For the Hungarian corpus we have prepared the headword lemma list by the combination of several programs. After analysing the text with the Humor program, we reproduced the possible entries. We have also added the date of first and latest occurrences of the words. This list contained more than 180,000 elements, but after its hand validation some entries were erased, which were either keyboarding errors or misinterpretations made by the analyser. Now we are able to compare this list with the headword entries of other dictionaries, which are already available in electronic format, and we can see more clearly what is missing from our corpus and how to enlarge it further. With the aid of this list it is easier to decide which entries should be included in the dictionary. The corpus based list is now being merged with the headword lemmas of the traditionally collected dictionary slips and other monolingual dictionaries. In the headword list of the letter 'A' the number of headwords has doubled after this operation. (Which means that there are roughly twice as many possible headwords in the old archive, than in the corpus.) On the other hand, the corpus contains more than twice as many headwords as the current monolingual dictionary of Hungarian (180.000 vs. 72.000).

The frequency list of the possible headwords along with the date of their first and latest occurrences is to be published in electronic format. An additional advantage of this format is that not only the fields mentioned above can be retrieved but the endings of the words as well. This is especially useful for finding the last part of compounds and derivational suffixes. During the correction of the list we have also received valuable information on the typical errors made by the analyser which will help us to maintain the morphological database used by this software.

For compiling the dictionary articles a detailed style manual must be prepared. It is advisable to make several types of draft entries before preparing the final manual, in order to see what is desirable and feasible. Today it is also a necessity that traditional lexicographers and computational experts should work in close co-operation. For the computerised format of the entries, it is now natural to use SGML/XML markup. Using TEI guidelines for customising your own DTD is a great help. My own experience agrees with [Veronis/Tutin 1998]: the TEI guidelines can be best used as ideas for the possible tags. It is much more convenient to use the tag names suggested by it so that your database conforms to other electronic dictionaries. Recently, more and more SGML tools are equipped with a TEI DTD, so one can save plenty of work in designing it from scratch. After making the style manual along with the suitable DTD, one must choose an SGML editing tool. This choice is becoming increasingly difficult, because there are already several of them on the market. Similarly to the retrieval software, you might choose something cheap or even free (like emacs under linux) but it will probably not be very user friendly and it might make the lexicographers' task more difficult than essential. For years we have been looking for something affordable and convenient to use, but we have not managed to find the ideal solution so far. If a publishing house has plenty of money, the best solution is to purchase a complex integrated SGML toolset which can handle the corpus, the dictionary entries under compilation and the maintenance and retrieval of the existing entries in a professional way.

When the hardware and software environment is settled, the lexicographers are burdened by the task of actually writing the entries. Day by day they have to cope with words with either hardly

any occurrence at all, or with several hundreds and thousands. To illustrate this phase I examined the English word *coach,* which is an international loan word coming from the Hungarian *kocsi*. The word still exists in both languages, but the main meanings have diverged. In the OED2 these were the main meanings of the noun, (the dates of the earliest and latest quotations are in parenthesis):

1.a A large kind of carriage: in the 16^{th} and 17^{th} centuries usually a state carriage of royalty or people of quality (still occasionally used, as e.g. the Lord Mayor's coach) now, usually, a large close carriage with four wheels, with seats inside, and several outside, used for public conveyance of passengers. (1556-1841)
1.b ...a supplementary or extra coach, beside the usual service (1732-1802)
1.c Sometimes used for the passengers by a coach (1840)
1.d A railway carriage (1832-1948)
1.e A single-decker bus (1923-1955)
1.f Economy or tourist class, on a passenger aircraft (1949-1985)
2 *Naut* An apartment near the stern of a man of war, usually occupied by the captain. (1660-1850)
3.a *University colloq.* A private tutor who prepares a candidate for an examination (1848-1878)
3.b One who trains others for an athletic contest, esp. a boat-race. (1885-1888)
3.c A tame bullock or horse used as a decoy in catching wild cattle or horses *Australian* (1873)

After consulting the OED2 I searched the word in the Cobuild corpus. I was surprised to see that most of the occurrences belonged to sense *3.b* of the noun or the corresponding verbal sense. Out of 120 concordance lines only 27 belonged to some other sense, usually to sense *1.d* or *1.e* (bus or railway carriage).

Seeing this, I became curious of how the new corpus based dictionaries could cope with this fact.

In COBUILD 1987 the entry was already reorganised:

1.1 A large motor vehicle which carries passengers on long journeys by road, used in British English.
1.2 A vehicle carrying passengers that is part of a train, used in British English.
1.3 An enclosed vehicle on four wheels pulled by horses in which passengers used to travel. Coaches are still used for ceremonial events.
2 If you coach someone,
2.1 you train them in a particular sport;
2.2 you give them special teaching especially in order to prepare them for an examination.
3. A coach is also
3.1 someone who trains a person or a team in a particular sport;
3.2 someone who gives people special teaching, especially in order to prepare them for examinations.

In CIDE there are two entries:

coach VEHICLE a long road vehicle on which people travel
A coach is also an old fashioned carriage pulled by horses, now used mainly in official and royal ceremonies.

coach TEACH to give special classes on sports or a school subject esp. privately, to one person or a small group.

In COBUILD 1999 the entry is thoroughly reorganised according to the order of frequency of the senses:

1. A coach is someone who trains a person or team of people in a particular sport.
2. When a trainer coaches a person or a team, he or she helps them to become better at a particular sport.
3. A coach is someone who gives people special teaching in a particular subject, especially in order to prepare them for an examination.
4. If you coach someone, you give them special teaching in a particular subject, especially in order to prepare them for examination.
5. A coach is a large comfortable bus that carries passengers on long journeys, used mainly in British English
6. A coach is one of the separate sections of a train that carries passengers; used mainly in British English.
7. A coach is an enclosed four-wheeled vehicle pulled by horses, in which people used to travel. Coaches are still used for ceremonial events.

The original meaning of the word became the very last sense, for obvious reasons. I agree with the editors, it helps the users of the dictionary greatly, if the most frequent senses are at the beginning of the entry. We can also realise that the definitions of COBUILD 1999 have became even more readable and well arranged than before. In the *1.1* sense of COBUILD 1987 it took me some time to realise that a coach is simply a kind of bus in British English, in the later version we can see this immediately from the definition no**.** *5*.

I suspected that the CDAE 1999 should place this sense even further in the entry, because it is based upon a corpus of American English. My hopes were well grounded, here are the definitions from CDAE:

coach TEACHER (esp. in sports) a person who is responsible for managing and training a person or a team.
A coach is also an expert who trains someone learning or improving a skill, esp. one related to performing.

coach PART OF VEHICLE the less expensive sections of an aircraft that most people sit in.
A coach is also one of the separable parts of a train.
A coach is also a kind of old-fashioned vehicle pulled by one or more horses.
(Br) A coach is a BUS.

So the British sense has been shifted to the very end of the entry. The original ‘old-fashioned vehicle’ meaning thus became the last but one. We can also see that the ‘tourist class of the aircraft is more often used in American English.

What has happened to the original Hungarian word *kocsi* simultaneously? According to the *Magyar értelmező kéziszótár* Concise Dictionary of Hungarian [Juhász et al. 1972] the first meaning is the original one:

1. Négy keréken járó lófogatú személyszállító jármű.
   ‘A four-wheeled vehicle pulled by horses carrying passengers’
2. [Kisebb, könnyű] szekér.
   ‘small and light wagon’
3. Kézi v. gépi erővel mozgatott kisebb szállítóeszköz, kézikocsi, gyermekkocsi stb.
   ‘A small vehicle of transport moved by hand or machine’.
   babakocsi ‘babycar’
4. (*Vasúti*) ∼ : (v.) teher v. személykocsi
   ‘railway carriage’
5. *biz* Gépkocsi, autó
   informal ‘car’
6. (jelzőként) amennyi egy kocsira ráfér
   ‘(as adjective) the quantity which can be carried by one vehicle’
7. *Műsz* Gépnek, szerkezetnek kerekeken, görgőkön, ide-oda mozgó alkatrésze.
   ‘A part of a machine which makes a shuttle-movement’

Nowadays the most frequently used meaning is number *5.*, which was labelled as informal in 1972. Nobody would label it in this way anymore, this is one of the most common ways of talking about a car (the most frequent alternative is *autó*, and we rarely use *gépkocsi* in normal circumstances). In the Hungarian corpus *kocsi* occurred 3054 times. The earliest quotation is from 1805[2], the last is from 1992. According to the data of the corpus, the car was first called *autó* (from 1908 to 1992; number of occurrences: 940), and *gépkocsi* (from 1909 to 1992; number of occurrences: 178). The first usage of *kocsi* in the sense *car* was found in two different texts from 1932. In order to find this first occurrence I did not read the 3054 quotations, rather tried to narrow my guess, so it is possible that there were some earlier examples for this meaning. The found example was: *A főügyész úgy érezte, hogy tartozik állásának azzal, hogy az orvos miatt autóba üljön. Csak amikor már bent ült s a kocsi elindult, akkor jutott eszébe, hogy semmi pénz sincs nála, most mi az ördögöt fog csinálni.* The attorney general felt that he should take a car for the sake of the doctor. Only when he was already in the car and it started had he realised that he did not have any money, what the hell he should do about it?’. The reason why I noticed that this occurrence must have meant the car was that its synonym *autó* appeared in the preceding sentence. So instead of trying to read thousands of examples I could have searched for *kocsi* near *autó* or *kocsi* near *gépkocsi* and would have found the very same quotation. Likewise to select quotations for sense number *1.* one can search the occurrences of *kocsi* near *ló* ’horse’, for sense number *4.* one can look for *kocsi* near *vonat/vasút* ’railway’. In neither case can one make sure to find the very first and latest occurrences of the given sense, but it is possible to select enough quotations for each or most senses relatively quickly and efficiently. When a dictionary

project arrives at the phase of actual entry compiling based on the given corpus, it is vital to think over the original concept. This is perhaps the last moment to decide if the requirements of the traditional historical dictionary can be met at all by using the available corpus. For the sake of producing the dictionary in a reasonable time it might be inevitable to find a compromise between the ideal and the realistic versions.

3 Conclusion

Historical dictionaries compiled recently have no alternative but to use computer corpora, similarly to other up-to-date dictionaries. At the same time, however, the requirements set by the traditional historical dictionaries should be thoroughly reconsidered, especially in the case of projects starting from scratch today. Instead of trying to copy the great ancestors, today's lexicographers should make a better use of the possibility of modular designing: computers enable them to make the compilation in several steps. One can start by collecting a corpus, then making a word list out of it, linking the word list to an existing dictionary (if there is any in electronic format), then revising the existing dictionary based on the corpus data in several phases. It can be feasible to revise first the words which are currently being used, then prepare the definition of archaic words. When any well defined part is completed (say, for example, an up-to-date, one volume dictionary) it should be published in printed form as well, while the computerised version can be continuously developed further, and made more and more similar to the traditional historical dictionary, if required.

Notes

[1]The Project for the Historical Dictionary of Hungarian is supported by the Hungarian National Science Foundation Number: T30297/1999-2002.

[2]The historical corpus only contains texts from 1800to 1992 at present.

References

Atkins, B.T.S. (1996). Bilingual Dictionaries: Past, Present and Future. *EURALEX'96 proceedings* University of Göteborg, Göteborg, pp.515-546.

Biber, D. – Conrad, S.- Reppen, R.: (1998). *Corpus Linguistics.* Cambridge University Press, Cambridge.

Brill, E. (1994). Some Advances in Rule-Based Part-of-Speech Tagging. In: *Proceedings of the 12th AAAI '94.* Seattle Wa.

Brill, E. (1995). Unsupervised Learning of Disambiguation Rules for Part-of-Speech Tagging. In: *Proceedings of the 3rd Very Large Corpora Workshop.*

Clear, J. (1994). I Cant See the Sense in a Large Corpus *COMPLEX' 94 Proceedings.* Research Institute for Linguistics, Budapest, pp. 33-22.

Considine, J. (1998). Why do large historical dictionaries give so much pleasure to their owners and users? *EURALEX '98 Proceedings.* University of Liège, Liège, pp. 579-587.

Juhász, J. et al. (1972). *Magyar értelmező kéziszótár* Concise Dictionary of Hungarian. Akadémiai Kiadó, Budapest

Landau, S. (1989). *Dictionaries,* Cambridge University Press, Cambridge.

Meggyesi, B. (1999). Improving Brills pos tagger for an agglutinative language. *ACL '99 Proceedings*

Ooi, V.B.Y. (1998). *Computer Corpus Lexicography.* Edinburgh University Press, Edinburgh.

Oravecz, Cs. (1998). Disambiguation of suffixal structure of Hungarian words using information about part of speech and suffixal structure of words in the context. GRAMLEX report.

Pais, J. – Pajzs, J.(1998) Using local rules for disambiguation of homographs in Hungarian corpora. *Proceedings of the EURALEX '98 Conference.* University of Liège, Liège, 1998. pp. 239-248.

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

Pajzs, J. (1997) Synthesis of results about analysis of corpora in Hungarian. *LinguisticæInvestigationes* XXI-2 John Benjamins, Amsterdam . pp 349-365

Prószéky, G.-Tihanyi, L. (1992). A Fast Morphological Analyser for Lemmatizing Corpora of Agglutinative Languages. In: *Proceedings of COMPLEX '92.* Research Institute for Linguistics, Budapest, pp. 275-278.

Prószéky, G. (1996). HUMOR - A Morphological System for Corpus Analysis. In: *Proceedings of the first TELRI Seminar in Tihany.* Budapest, pp. 149-158.

Silberztein, M. (1999). INTEX Tutorial Notes *COMPLEX '99 Proceedings* Research Institute for Linguistics, Budapest, pp. 121-151.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford University Press, Oxford.

Tutin A. and Veronis J. (1998). Electronic Dictionary Encoding: Costumizing the TEI Guidelines *EURALEX '98 Proceedings.* University of Liège, Liège, pp. 363-374.

agraph*Dictionaries cited

CDAE (1999) *Cambridge Dictionary of American English* (Landau, S.) Cambridge University Press, Cambridge

CIDE (1995) *Cambridge International Dictionary of English* (Procter, P.) Cambridge University Press, Cambridge

COBUILD (1987) *Collins COBUILD English Dictionary* (Sinclar, J., Hanks, P. et al.) Harper Collins Publishers, London

COBUILD (1999) *Collins COBUILD English Dictionary* (Sinclar, J., Hanks, P. et al.) Harper Collins Publishers, London

LDOCE3 (1995) *Longman Dictionary of Contemporary English,* (Summers, D.) Longman, London.

OED2 (1992) *Oxford English Dictionary on CD-ROM* version 1.01. Oxford University Press, Oxford, AND software B.V. Rotterdam