An attempt to develop a lemmatiser for the Historical Corpus of Hungarian

Gabriella Kiss and Júlia Pajzs
Department of Lexicography and Lexicology
Research Institute for Linguistics, Budapest
{gkiss, pajzs}@nytud.hu

For the project of the Historical Dictionary of Hungarian a carefully selected representative corpus was collected (24.5 million running words). The texts were chosen from three centuries. A morphological analyser programme was successfully run on the modern texts, but the analysis of the earlier texts was problematic. In our paper we will describe a method for the conversion and analysis of archaic texts, without losing the original word forms. The Historical Dictionary project itself, and the analyser developed for the modern text, will also be reported briefly.

1. The Historical Dictionary project and its corpus

The project first started in the late 19th century with the collection of old fashioned dictionary slips. The idea was to compile an OED-like dictionary which covers the vocabulary of the Hungarian language from 1772 up to the time of collection. (There exists an historical dictionary for the former period : Czuczor G & Fogarasi J 1862). The collection of the slips continued until about 1960. From time to time there was an attempt to compile some draft entries based on this collection, but somehow or other these experiments always happened to fail. Possibly these repeated failures were partly due to the lack of an adequate personality in charge of a project at this scale (Hutás, 1974). Without a dedicated and convinced chief editor not even a one volume dictionary can ever be completed, not to mention an OED-like several volume reliable historical dictionary.

During the years 1950-1960 a fairly valuable seven-volume monolingual dictionary was prepared using the slips collected for the historical dictionary (Bárczi & Országh 1959-1962). It contains some illustrative quotations, but not necessarily to each meaning. Bibliographic information is not supplied with the quotations (only the author's name in abbreviated form is given in some cases). It was not published as a historical dictionary, but as an explanatory one, and it still continues to be the largest existing Hungarian explanatory dictionary (Országh 1960). In the late 60's a one-volume abridged version was produced on its basis. It is still in print, and is now being revised to be published in 2002. This revised version is also based on newly but traditionally collected data, not on a corpus.

In 1985 the Hungarian Academy of Sciences decided to start the historical dictionary project all over again, based on a computerised corpus to be compiled first. The collection method seems rather naive these days: it was a combination of old fashioned collection and corpus building. At that time nobody in Hungary had any experience about efficient corpus collection methods, or about realistic expectations from a corpus of a given size or type. At the outset we planned to collect a corpus of 10 million running words. To establish the source material for a sound historical dictionary, several small excerpts were carefully chosen by the literary historians of each period. Since the short sample texts usually contained only some book pages we found that optical character recognition was not really efficient for this task. Therefore the selected text parts were keyboarded manually, which, of course, took a very long time (keyboarding was finished at the very end of 2000), and, despite repeated controlling, many keyboarding and other kinds of errors still exist.

Currently the corpus consists of 24.5 million running words, with a majority from the 20th century (16 million words), 6.8 million words from the 19th century and 1.7 million from the late 18th century. The size of the samples is varied: since every poem and each part of a book is considered a separate sample, the number of different samples is quite large: over 21,000. The average sample length is around 1200, but there are also extremely short samples (e.g. a very short poem of two words), and there are surprisingly large ones: a text of 34,812 words is the current maximum. Although the literary experts tried to be as objective as possible and the selection was reviewed by different experts, the extremely overrepresented works obviously happened to be their favourite ones, and some very well known authors and/or works were simply left out for some reason. Most of the texts are different kinds of prosaic texts (prosaic fiction: 31%, other kinds of prose: 51%, poetry: 8,5%, drama: 5,7%). In order to supply exact philological data for the planned historical dictionary, bibliographic data and the estimated (or sometimes exact) year of production were recorded for each sample text. This information along with the data of keyboarding, updating etc. are stored in the header of each sample file in SGML format. When we started keyboarding SGML did not exist, so for a long time special codes were used to mark paragraph ends, stanzas, quotations, etc. As soon as we learnt about SGML (in 1987) we started to convert the earlier material into this format, and after a while keyboarding contitued in it, there are, however, still too many formal errors in the texts keyboarded previously. (This is a very serious point to consider for projects starting these days: one should try to rely on XML and TEI or other standard recommendations as strictly as possible, and should not think that it is always possible to correct the errors easily by a good conversion programme.)

The first draft dictionary entries based on the corpus were prepared seven or eight years ago, when the corpus was only about half as large as it is today. When we first realised that we can not make reliable historical dictionary entries based solely on the corpus, we hoped to be able to fill in the obvious gaps either by the traditional slips or by the enlargement of the corpus. Recently, we had to cope with the fact that, although it is theoretically possible to compile a more or less historical-like dictionary from the available sources, this would require a despairingly long time and/or several times larger staff and facilities than we can hope for. (Just to give you a hint of our possibilities: right now nine full time compilers are working on the project, but only three of them have some experience in writing a dictionary, and none of them have a formal education in lexicography. And this is about the largest team we can expect for the future, as well.) Faced with this situation, we are about to redesign the whole project. At this turning point there are three alternative plans: one is to further improve the historical corpus, and forget about dictionary writing, at least for the near future. The other is a less maximalistic historical dictionary than the one planned originally, but it is sill a very traditional historical approach, which wishes to build a dictionary that includes only the chronologically first quotation to each known sense of the words (either from the slips or the corpus), chiefly due to space considerations (this plan imagines an eight-volume dictionary). Our personal view, which is shared by most in our team, is to produce a good, up-to-date one volume dictionary, at least as a first step, with an electronic version including many additional facilities. This, in many respects, would rather resemble the modern corpus-based monolingual English dictionaries (COBUILD, CIDE, LDOCE, or even the new OALD) more than the OED. We would like to make clear-cut entries with easily understandable definitions, synonyms, antonyms, hyponyms, etc. The main specificity of this dictionary compared with the above mentioned modern English ones would be the illustrative quotations not only from the current part of the corpus, but from the earlier periods, as well. In the electronic version the exact philological reference of the citations would be supplied, as it is done in traditional historical dictionaries. The electronic version could also contain many more examples from the corpus, plus several additional properties — like the generation of the inflected forms of the words (which is a much more difficult task and therefore even more necessary for Hungarian than for English). We believe that with a brand new and up-to-date concept a really good modern dictionary could be compiled, combining the advantages of modern corpus based dictionaries and traditional historical ones. We are convinced that it could be produced in a realistic period of time even with a relatively small team of compilers, and this product would find its market.

2. The morphological analysis of the corpus

2.1. The analyser programme

The HUMOR analyser programme was developed by the MorphoLogic Ltd in co-operation with our department in the late 80's (Pajzs 1991, Prószéky 1996, Prószéky & Kis 1999). The first version was based on the entries of the above mentioned seven-volume dictionary, which were supplied with morphological codes. Encoding was further improved during the development of the programme. Now it contains a complex classification including information about the types of suffixes that can follow the given entry and the matching suffix variant. Hungarian morphology is extremely complex: it is mainly agglutinative, several suffixes may follow each other, at least in theory. Based on the large analysed corpus we have evidence that the highly complex combinations practically never occur (Pajzs & Papp 1998). The combination of two suffixes is quite frequent, but due to vowel harmony, if the first suffix is a back variant, the next suffix must also be back, therefore the actual number of real suffix combinations is not that large. The information on the possible forms of the suffixes was stored in the databases used by the analyser. The databases of the entries and the suffixes are disjunct, and the programme only checks whether the elements actually found in the text can be matched according to the information given in the databases. It uses the unification method for choosing the correct solution(s). When the analyser finds a possible match, it can also identify the root of the word, even when the actual form of the root is different from the entry (e.g. the original root *ló* ’horse’ becomes *lov* in front of the suffixes, and its analysed version contains both the original lemma in front of the „=” sign and the actual root after it: *ló[FN]=lov+aink[PSt1i]+nak[DAT]* ‘for our horses’).

This analyser programme was applied on the corpus in several steps. First it was only tested on the contemporary part of the corpus, later on the 19th century part, but this test already raised numerous problems. Hungarian orthography was standardised only in the late 1930’s, therefore the earlier texts contain many alternative orthographic possibilities. Since the analyser is also used as the engine of a Hungarian spell-checker, the old (currently unacceptable) alternatives could not be included in its databases. Vowels in some words which now should be spelt with long accents used to be spelt either with short or with long accents (even within the same text). Several compounds which are now written in one word used to be written in separate words or with a hyphen. So when we applied the analyser on the texts from the 19th century, only 90 per cent of the words were recognised by the programme, while in the 20th century part 95 per cent were recognised. Naturally, the recognition of the analyser is not always correct, sometimes the words are analysed as non existing compounds or as strange derivates. When the programme finds several possible analyses it outputs each variant. For the disambiguation of these alternatives a local rule-based programme was tested and run on the whole 19th-20th century corpus (Pajzs 1997, Pajzs 1998). Although the result of this attempt was far from the expected correctness rate, because the texts were very varied, its outcome was a usable lemma-oriented corpus, where one could directly search the entries without enumerating every possible form of the word or having too much surplus data. (So for example if you try to search the word *ad* ’give’ in the non-analysed corpus, you will get every word which happens to start with the same character string, while if you search it in the analysed version, you can explicitly define that you only wish to search the verb *ad*.)

When using the analysed corpus (of 17 million running words at that time) we realised some of its drawbacks. In the meantime the keyboarding of the late 18th century texts also started which meant a new range of problems.

The orthography of that time hardly resembles that of modern texts. Some typical problems:

- In the earlier prints there were several characters which are not used any more. For example, instead of the short and long vowels *ö,õ* or *ü,û,* standard nowadays, there used to be several different accents in between, which represent either short or long vowels. As the historical linguists considered these specialities important, we were bound to keep this information somehow when keyboarding the texts. As these characters were used sometimes instead of the short, sometimes instead of the long variant, they could not be converted directly to their current form.
- There were several suffixes, old root forms and words that are not used anymore.
- Some of the still existing words used to be spelt in completely different ways: words which are now written in two words sometimes (not always!) used to be written in one word, others the other way round.
- Since there was no standardised orthography at all, the very same consonant phonemes were sometimes spelt differently even in various occurrences of the same word within the same text, as e.g. *lly* or *lyly* or *jj* or *lj* for [j]. The *s* letter was often represented by something like a or , and the letter *z* also had an archaic version. To make things even more complicated, the phoneme now spelt as *zs* was spelt as a single *s* or its old forms or with the combination of old *s* with old *z* or with any other combinations of these.

All this information was kept during keyboarding. The special characters were represented by a combination of letters and digits. This way we could keep all the required information in an easily convertible format (so the representation of the characters did not depend on any operation system or the facilities of any given word processor). Not only the archaic characters, but the current accented characters are also represented in this way, chiefly for the sake of portability. (The letters of the English alphabet and the digits are always the same in every code table.) When, however, we intend to retrieve the old words together with the new corpus, we either leave the problem of searching of the different possible forms for the lexicographers who should be able to retrieve the concordances as fast as possible, or we must find a way to standardise the old words, while keeping the original old form. For this aim we designed a special format:

Keyboarded form

Honnat-is nem kis bos43zs43zusa1g-te1tellel bu20no20s43u20lnek azok, kik tellyes43se1ggel

Current ASCII form

Honnan is nem kis bosszúságtétellel bûnösülnek azok, kik teljességgel
‘where from those persons will be punished by a not small annoyance, who are completely...’

Lemmatized form

```
<w><o>Honnat</o>                <t>honnan</t>             <a>honnan[HA]</a></w>
-
<w>                             <t>is</t>                <a>is[KOT]</a> </w>
<w>                             <t>nem</t>               <a>nem[MOD]</a> </w>
<w>                             <t>kis</t>               <a>kis[MN]</a></w>
<w><o>bos43zs43zusa1g</o>       <t>bosszu1sa1g</t>        <a>bosszu1sa1g[FN]</a></w>
-
<w>                             <t>te1tellel</t>         <a>te1tel[FN][INS]</a></w>
<w><o>bu20no20s43u20lnek<o>              <t>bu3no2su2lnek</t>
<a>bu3no2su2l[IGE][t3]</a></w>
<w>                             <t>azok</t>              <a>az[NM][PL]</a></w>
,
<w>                             <t>kik</t>               <a>ki[NM][PL]</a></w>
<w><o>tellyes43se1ggel</o>      <t>teljesse1ggel</t>     <a>teljesse1g[FN][INS]</a></w>
```

In the field tagged by <o> the original, old version is kept if it is different form the current spelling. In the field tagged by <t> either the converted version of the original old form, or the token as it was found in the text if it does not differ from the current norms, is kept. If there are more than one possible and analysable conversions each one is kept separated by a ’|’. If the analyser is able to recognise the token, the analysed version is kept in the field marked by <a>. If it could not find the correct analysis a special tag NE ’not analysed’ is given. If there are more than one possible analyses each one is given. The analysed (or rather tagged) version consists of the recognised lemma, the part of speech code and the suffix codes. The superfield word is marked by the <w> tag.

The main advantage of this format is that we can keep both the original archaic form as it occurs in the text, the normalised tokens and the lemmatised version. For the time being we do not intend to disambiguate the analysed version. The retrieval interface takes care of finding the lemma in the analysed field first. If the searched word is not present in this field, then the token field or even the original field can be searched. After the search the result is displayed either from the original field (when there is one) or from the token field if there was no original version. Of course this storing format is very redundant, it requires plenty of disk space, but nowadays hard disks are becoming less and less expensive. We have been using the Open Text SGML text retrieval software, but we are considering to switch to a more modern and efficient software tool. From this tagged version we can easily convert the text to any other format preferred by different tools.

2.2. The conversion of the archaic forms

The regularly occurring orthographic and grammatical alterations were aimed to be converted at this phase.

2.2.1. Old characters and variant orthography

The variant spellings of the letter *s* , were keyboarded as *s41* and *s43*, respectively, and then those can be converted to *s*.

There are several strange accented forms for the letter *u* and *o*, which were keyboarded as *u20, o20, u23, o23, u24, o24* etc. and can be converted to *ü ö,* û and *õ*, respectively, according to current orthography. If there are many of these old characters in one word, each of the possible combinations must be generated. The HUMOR programme tries to analyse each version and outputs every seemingly correct analysis to the analysed field, and from the converted tokens only those having a corresponding analysis will be kept.

Other archaic character combinations:

The phoneme currently represented by the letters *cs* used to be written by *ts*
The phoneme currently represented by the letter *c* used to be written by *tz* o*r cz*

The long versions of the digraphs are now spelt by repeating only the first character of the digraph, earlier it was variable. (E.g. a suffixed form of *asszony* ’woman’ is spelt *asszonnyal* ’with a woman’, while in the archaic texts it could either be *aszszonynyal*, *asszonynyal* or *asszonnyal*, and, of course, any of the *s*-es or *z*-s could have been old ones.)

2.2.2. Phonological rules regularly appearing in the texts

Consonants often lengthened in intervocalic position. Those which are represented by digraphs had to be handled separately.

old: *segittõ* (keyboarded as: *s43egitto24*), new: *segítõ*
old: *gyilkossa* (keyboarded as: *gyilkos34s43a*), new: *gyilkosa*
old: *tallyigába* new: *talyigába*
old: *fénnyiben*, new: *fényében*

Spelling according to pronunciation

old: *akarattyán*, new: *akaratján*
old: *tilcsa* or *tiltsa*, new: *tiltja*
old: *tanúji*, new: *tanúi*
old: *eladgyák*, new: *eladják*

Vocals lengthened before *l, n*

old: *mozdúlásra* new: *mozdulásra*
old: *múnkái*, new: *munkái*
old: *óldva*, new: *oldva*

The phoneme now spelt *zs* sometimes used to be spelt *s*

old: *strásának*, new: *strázsának*
old: *désa*, new: *dézsa*

2.2.3. Morphological variants

The third person plural possessive suffix was often spelt -*jok,* which is now -*juk*.

old: *búzájok* new: *búzájuk*
old: *hazájok* new: *hazájuk*

The third person singular possessive suffix also had a variant form -*ok,* which is now -*uk*.

The verbal causative derivational suffix -*ít* sometimes used to be written as -*it*.

The use of long and short variants of the same vowels was much less regular than it is today.

Some words had variant root forms, which are not used anymore.

2.3. The process of corpus analysis

The programme which made the above described conversions and tried to analyse the converted tokens was only one module of the process.

2.3.1. The first module is a PERL programme which picked the running words from the corpus. Its output was ordered by the 'sort' and 'uniq' unix commands. The result is a file containing 1,467,230 different tokens.

2.3.2. The HUMOR analyser was run on this list. (The number of analysed words after this phase was 896,153). After the analysis the unanalysed words were separated from the output. The number of unanalysed tokens was roughly half a million at this phase.

2.3.3. On the unanalysed list a converter programme was run, which contains a series of PERL regular expressions based on the above scetched grammatical rules. It converts the possible variant forms of the non- recognised words, and then tries to analyse them with the help of the HUMOR programme. If it finds at least one analysable version, the corresponding token and the analysed version is outputted in the format described in 2.1. If there is no analysis even after the conversion, one possible converted form is still kept in the token field, and in the analysed field the code „NE" marks the missing analysis.

2.3.4. The output of 2.3.2 is merged with the output of 2.3.3. The result is put into an ACCESS database, which contains the original running word as it was found in the corpus in the first field and the analysed version in the format described in 2.1 in the second field. The database is indexed for the first field.

2.3.5. A programme runs on the whole corpus that reads and copies the header of the text files to the analysed text files, then reads the running words one by one, searches them in the first field of the database created in 2.3.4, and outputs the result from its second field. The rest of the information found in the text (SGML tags, punctuation etc.) is copied to the analysed version without any change. In some sample texts there are notes from the original texts. They are kept separately at the end of each sample, and are copied into the end of the analysed file.

2.4. Evaluation of the lemmatisation method

With the software package described in 2.3 the whole corpus has been analysed. Here are some examples for the problems of the conversion and identification from 18th-century samples.

2.4.1. Some examples of successful analysis after the conversion

```
<w><o>kapkodgyon</o><t>kapkodjon</t><a>kapkod[IGE][Pe3]</a></w>
```

The current form is *kapkodjon* ‘imperative of ‘snatch’’. The written from, based on the pronounced assimilated word, contained the letters *dgy* instead of the current *dj*. Since it is a regular deviation, it was correctly converted and analysed.

```
<w><o>eggy</o><t>egy</t><a>egy[DET]|egy[SZN][NOM]</a></w>
```

The current form is *egy* ’one’. Again it is a regular deviation, so it was handled correctly.

```
<w><o>melly</o><t>mely</t><a>mely[NM][NOM]</a></w>
```

The current form is *mely* ’which’. Regular, correct.

```
<w><o>alats43ony</o><t>alacsony</t><a>alacsony[MN][NOM]</a></w>
<w><o>s43zaba1su1</o><t>szaba1su1</t><a>szaba1su1[MN][NOM]</a></w>
<w><o>ts43ats43ogni</o><t>csacsogni</t><a>csacsog[IGE][INF]</a></w>
<w><o>kits43al</o><t>kicsal</t><a>kicsal[IGE][e3]</a><t>kicsal</t><a>kicsal[IGE][e3]</a></w>
<w><o>kapts43olatok</o><t>kapcsolatok</t><a>kapcsolat[FN][PL]</a><t>kapcsolatok</t><a>kapcsolat[FN][PL]</
a></w>
<w><o>vis43zontag</o><t>viszontag</t><a>viszontag[HA]</a></w>
```

In these words only the characters had to be converted (*s43* to *s*, *ts* to *cs*). After the conversion the analysis was correct.

```
<w><o>tis43zta1talansa1gokat</o><t>tiszta1talansa1gokat</t>
                                        <a>tiszta1talansa1g[FN][PL][ACC]|tisztatalansa1g[FN][PL][ACC]</a></
                          w>
```

After the character conversion the analysis is only partially correct: the lemma *tisztátalanság* ’impurity’ is correctly identified, but the suffix *-ok* used here is an old form of the possessive suffix *-uk*, and not the current plural suffix *-ok*. The second analysis is a misinterpretation: the supposed lemma is *tisztatalanság*, which is a supposed derivation of the word *tiszta+talan+ság* ’clean’+privative+adjective-to-noun nominal suffix, but an actually non-existing word. The accusative is correctly identified at the end of the word. Although this analysis is not quite correct, the main point is to give the good lemma, and it is also given there.

Usually the conversion was the most successful when only one or two old characters had to be converted and the root or the suffixes were same as the current form.

2.4.2. Some examples for unsuccessful analysis after the conversion .

```
<w><o>o2s43zvekapts43olva</o><t>o2szvekaptsolva</t><a>NE</a></w>
```

The current form of this word is: *összekapcsolva* ’connected’. In order to be able to recognise this, the variant lemma *összve* ’together’ must be added to the database of the entries. Then the conversion of *s43* to *s*, and that of *ts* to *cs* would be sufficient for the recognition of this word.

```
<w><o>eggyu2gyu2</o><t>egyju3gyu3</t><a>NE</a></w>
```

The current from of this word is *együgyû* ‘foolish’. The attempt to convert the *ggy* to *gyj* was not successful, and also only one of the short *ü*-s should have been replaced by the long *û*.

```
<w><o>s43zo2me1rmetes</o><t>szo2me1rmetes</t><a>NE</a></w>
```

This is an archaic version of an obsolete word *szemérmetes* ’coy, prudent’, nowadays only used ironically as a common saying originating from a well known mid-19th century ironic epic poem.

```
<w><o>gyu24lekezo24tt</o><t>gyu2lekezo3t</t><a>gyu2lekezo3[FN][ACC]|gyu2lekezo3[MN][ACC]</a></w>
```

The current word would either be *gyülekezet* ’assembly (noun)’, or *gyülekezett* ’past tense of ‘gather’ (verb)’ or *gyülekezõt* ’present participle of ‘gather’ (verb) with an accusative suffix’. In the current example the second case would have been the correct choice, but it was not among the converted forms, because the formerly frequently occurring *-ött* variant form of the *-ett* current suffix is not included in the suffix database. In the ancient texts many *ö*-s occurred where today *e*-s are written, because this used the be a frequent pronunciation and orthographic variant. Nowadays it is rather just a regional alternative, and appears in written form only occasionally.

```
<w><o>o24rizko2dgy</o><t>o3rizko2dj</t><a>NE</a></w>
```

The current form is õ*rizkedj* ’the imperative of ‘refrain from’’. The conversion was partially correct (*dgy* to *dj*), again the alternative -*öd* form of the -*ed* suffix should have been included in the suffix database.

<w><t>mennyen</t><a>menny[FN][SUP]</a></w>

The current form is *menjen* ’the imperative of ‘go’’. The given analysis is mistaken for the superessive of the noun *menny* ’heaven’, and seems to be analysed in the first analyser phase (2.3.2), so it does not appear among the words to be converted.

<w><o>Tragyo24dia1nak</o><t>tragyo3dia1nak</t><a>NE</a></w>

The current form is *tragédiának* ’dative of ‘tragedy’’. The letter *g* is not often replaced by *gy* in the old text, and *é* is also rarely replaced by *ö/õ*. So there was no conversion rule for this word.

<w><o>Tana1ts43be1li</o><t>tana1csbe1li</t><a>tana1csbe1l[FN][IKEP][NOM]</a></w>

The current form is *tanácsbeli* ‘belonging to the council’. Although the conversion was partially correct (*s43* to *s, ts* to *cs*), the suffix -*béli* was not converted to -*beli*, and the converted token was mistakenly recognised by the analyser as *tanács+bél+i* ’council’+’intestine’+noun-to-adjective derivative suffix.

2.4.3. An analysed example sentence

Old keyboarded form:

Hallom s43ivednek foha1s43zkoda1s43ait, e1rtem azon bu1tsu1t, melly ne1ked a Kira1lyne1to3l adatott, de erro2l ma1skor les43z s43zo1llanunk, s43okkal nyomos43s43abb gondok foglalnak el lelku2nket.

Current ortographic ASCII form:

Hallom szívednek fohászkodásait, értem azon búcsút, mely néked a Királynétól adatott, de errõl máskor lesz szólanunk, sokkal nyomósabb gondok foglalnak el lelkünket.

A sentence with a similar meaning in current Hungarian:

Értem, mennyire fáj a szíved amiatt, hogy búcsut kell venned a királynétól, de errõl majd máskor beszéljünk, most sokkal fontosabb gondjaink vannak.

‘ I sympatise with your feelings on saying farewell to the queen, but we have to talk about it later, because we have much more urgent problems to solve at the moment.’

```
+ <w>                          <t>hallom</t>
   <a>hallik[IGE][Te1]|hall[IGE][Te1]|hall[FN][PSe1]|hallom[FN]</a></w>
-  <w><o>s43zivednek</o>       <t>szivednek</t>       <a>NE</a></w>
+  <w><o>foha1s43zkoda1s43ait</o><t>foha1szkoda1sait</t><a>foha1szkoda1s[FN][PSe3i][ACC]</a></w>
,
+ <w>                          <t>e1rtem</t>
   <a>e1n[NM][CAU][]|e1rt[IGE][Te1]|e1r[IGE][Me1]|e1r[IGE][TMe1]
                                                      |e1rt[MN][PSe1]|e1rik[IGE][Me1]|e1rik[IGE][TMe1]</a></w>
+  <w>                         <t>azon</t>            <a>azon[NM]|az[NM][SUP]</a></w>
+  <w><o>bu1tsu1t</o>          <t>bu1csu1t</t>        <a>bu1csu1[FN][ACC]</a></w>
,
+ <w><o>melly</o>              <t>mely</t>            <a>mely[NM]</a></w>
+  <w>                         <t>ne1ked</t>          <a>te[NM][DAT]</a></w>
+ <w>                          <t>a</t>               <a>a[DET]</a></w>
'
- <w><o>Kira1lyne1to2l</o>     <t>kira1lne1to3l</t>  <a>NE</a></w>
+  <w>
   <t>adatott</t><a>adatott[MN]|adat[IGE][Me3]|adatik[IGE][Me3]|ad[IGE][MUV][Me3]</a></w>
,
+ <w>                          <t>de</t>              <a>de[HA]|de[KOT]</a></w>
+ <w><o>erro2l</o>             <t>erro3l</t>          <a>ez[NM][DEL]</a></w>
+ <w>                          <t>ma1skor</t>         <a>ma1skor[HA]|ma1skor[FN]|ma1s[FN][TEM]|
                                                      ma1s[NM][TEM]|ma1s[MN][TEM]</a></w>
+ <w><o>les43z</o>             <t>lesz</t>            <a>lesz[IGE][e3]</a></w>
- <w><o>s43zo1llanunk</o>      <t>szollanunk</t>      <a>NE</a></w>
,
+  <w><o>s43okkal</o>          <t>sokkal</t>          <a>sokk[FN][INS]|sok[SZN][INS]</a></w>
+ - <w><o>nyomos43s43abb</o> <t>nyomosabb</t>       <a>nyomos[MN][FOK]</a></w>
+ <w>                          <t>gondok</t>          <a>gond[FN][PL]</a></w>
+ <w>                          <t>foglalnak</t>       <a>foglal[IGE][t3]</a></w>
-
+ <w>                          <t>el</t>              <a>el[IK]</a></w>
+ <w>                          <t>lelku2nket</t>      <a>le1lek[FN][PSt1][ACC]</a></w>
```

The sentence contained 22 running words. Those of them which are marked by a '+' sign at the beginning of the column are analysed correctly, or rather, at least one of the given analyses is correct. Those which are marked with '-' are not analysed correctly. There is one with a '+-' sign, the analysis is nearly correct, but not completely (the current lemma would be *nyomós* strong, powerful' not *nyomos* 'trail, trace+noun-to-adjective derivative suffix', which is just a supposed but non existing form).

Those words which were recognised as archaic versions had a first field marked with the <o> </o> tags. There were 10 words like this in the above sentence. Out of these 6 were correctly analysed after the conversion. Altogether 18 words out of the 22 were analysed correctly in this example.

2.5. Statistical evaluation of the results

Cent	Running words A	Analysed words after 2.3.2 B	Analysed words after 2.3.5 C	Archaic forms D	Analysed archaic forms E	Not analysed words F	Percentage of not analysed words
18th	1,689,735	1,433,100	1,727,697	379,898	294,597	231,319	13.6%
19th	6,839,688	6,666,168	6,873,935	339,326	207,767	684,328	10%
20th	1,615,5007	16,117,793	16,159,742	169,404	41,949	798,562	4.9%

Note that the number of analysed words is sometimes larger than that of the running words, because in many cases the words have more than one supposedly correct analysis. The ratio of unanalysable words in the texts from the 18th century was 22.48 per cent (1-[(A-D)/A]) before using the conversion programme. After the use of the above described conversion rules most of the formerly unrecognised words yielded an analysis (E/D=0.775). Although there still remained a much larger number of unanalysed words than either in the 19th or 20th century texts, it is clear, that this algorithm had the most successful effect on the targeted part of the texts: while the ratio of recognised words was raised by 22.49 per cent (E/[A-D]) in the 18th century part of the corpus, in the 19th century part the improvement was 3.1 per cent, while in the 20th century part 0.2 per cent. It might be surprising that the above described rules could have been used in the 20th century texts at all, but remember that Hungarian orthography became standardised only in the 1930s. Naturally, in many cases the attempted conversions are misleading, so the resulting analysis sometimes has nothing to do with the correct recognition of the lemma.

We have also examined the number of rules employed during the conversion, we have given an identification number to each group of rules and this number was put into the resulting database to an additional field. Using this field we could investigate the effectiveness of the rules. In most archaic words only one conversion rule was employed (42.57 per cent), in 12 per cent two rules were employed, three rules were used only in 2.39 per cent. The largest number of different rules employed for the same word was 7, but it was done for 6 different running words only (0.001 per cent). We are planning to study this database further.

3. Conclusion and further research

The attempt to be able to search lemmata in historical texts has proved to be promising. Our purpose at this stage was to handle the difficulties raised by a diachronic corpus in a relatively straightforward way. Instead of preparing a completely different analyser programme for the archaic texts we have been trying to find regular alternations and convert the archaic words to forms as similar to the recent ones as possible. Although plenty of problematic cases remain unsolved, we still believe that we are on the right track. From the preliminary results described above it is possible to draw new conclusions: whenever we look at the current output, we can find some new regularities which can be added to the programme, thus improving the ratio of correct analysis. The often occurring irregularities can also be added to one of the databases used by HUMOR, called the inclusion database. This is a very simple database where you can add a running word in the first field and put its correct analysis in the next column (so for example the frequent archaic word forms like *mennyen*, *vala* can be added here). HUMOR also uses a similar list for excluding some incorrect analyses, so whenever we find an error among the analyses given by the programme, we can just eliminate them by adding them to the exclusion database. The further development of the conversion rules combined with the careful use of HUMOR's inclusion/exclusion databases can make the lemmatisation process of the archaic texts much more accurate.

References

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

Benkõ L (ed) 1991–1992 *A magyar nyelv történeti nyelvtana I–II.*, 'The historical grammar of Hungarian' Budapest, Akadémiai Kiadó.

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

Elekfi L 1994 *Magyar ragozási szótár* – 'Dictionary of Hungarian Inflections', Budapest, Research Institute for Linguistics

Kiefer F (ed) 2000 *Strukturális magyar nyelvtan 3. Morfológia*, 'A Structural Grammar of Hungarian 3. Morpology'. Akadémiai Kiadó, Budapest

Olsson, Magnus 1992 *Hungarian Phonology and Morphology*, Lund, Lund University Press,

Országh L 1960 Problems and Principles of the New Dictionary of the Hungarian Language. *Acta Linguistica* X/3-4. Budapest, Research Institute for Linguistics of the Hungarian Academy of Sciences, pp 211-273.

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, pp 129-136.

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

Pajzs J, Papp F 1998 Statistical Examination of the Hungarian Noun Paradigm *Proceedings of ALLC/ACH* Debrecen, Lajos Kossuth University pp 89-93.

Papp I *Leíró magyar hangtan*, 'Hungarian descriptive phonology' Budapest, Tankönyvkiadó, 1966.

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

Prószéky G (1996). HUMOR – A Morphological System for Corpus Analysis. *Proceedings of the first TELRI Seminar in Tihany*. Budapest, Research Institute for Linguistics pp 149–158.

Prószéky G, Kis B 1999 Agglutinative and Other (Highly) Inflectional Languages. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* College Park, Maryland, USA pp 261–268

R. Hutás M 1974 Az Akadémiai Nagyszótár történetének vázlata (1898-1952) 'The brief history of the Unabridged Dictionary of Hungarian' *Nyelvtudományi Közlemények*. LXXV. Budapest, Akadémiai Kiadó, pp 447-465.