Judit PAIS and Júlia PAJZS, HAS Institute for Linguistics

Using local rules for disambiguation of homographs in Hungarian corpora[1]

Abstract

The historical corpus of Hungarian contains about 20 million running words at the moment. To be able to retrieve the occurrences of the lexemes, a morphological analyser programme was developed which is able to segment the running words and identifies the lexeme and the suffixes. Over 30% of the running words can have more then one correct analysis. Therefore we are aiming to develop methods for automatic disambiguation of the analysed text. This paper desrcibes an attempt for disambiguation by using local rules.

Keywords: corpus analysis, disambiguation, lexeme retrieval

1. Introduction

The historical corpus of Hungarian was collected to serve as the resource of the Historical Dictionary of the Hungarian Language which is being compiled in the Department of Lexicography and Lexicology of the Research Institue for Linguistics. The dictionary should cover the vocabulary of the last two centuries, including the last decades of the eighteenth century as well, similarly to the *Trésor de la langue française*. The sample texts of the corpus were carefully selected by literary historians so that they can represent the vocabulary of these centuries. The size of the already computerized corpus is about 20 million running words at the moment, and it is still being enlarged and maintained.

As soon as we started to computerize the corpus we decided to develop an automatic lemmatizer for our running texts. Nowadays most of the corpus based projects include some kind of analyser or tagger tool, but at the mid-eighties we were one of the first to design and apply a morphological tool. The reason for this is that Hungarian morphology is very complex compared to that of the indo-european languages. Our language can be mainly characterized as agglutinative. Most of the words can be followed by a complex set of suffixes, and some of the stems change before certain suffixes. The suffixes can also change form when followed by some other suffixes. Since our aim was to be able to retrieve the lexemes from the corpus we had to find a way to identify them. For example the word *alszik* 'sleep' can have several inflected forms, like *aludtam, alvó, alszol*; the correct segmentations of these are the following: *alud+tam, alv+ó, alsz+ol*. The actual form of the root is always different; therefore in an unanalysed corpus we had to search for all occurrences of the words starting with *alud, alv, alsz*. In some cases when we search for all the words starting with a specified string we get several results which do not belong to the same dictionary entry. For example, when we search all the words starting with *áll* 'stand [V] or chin [N]', we get all the occurrences of words like *állandó* 'permanent', *állam* 'state', *állapot* 'condition', *állít* 'declare'. In the lemmatized corpus, however, we can search all occurrences of each inflected form of the verb *alszik*, or all inflected or not inflected occurrences of the verb *áll* without getting surplus data.

The HUMOR analyser program - which was developed by MorphoLogic Ltd first for Hungarian and later for other languages - is able to segment the running words into lexeme (root) and suffixes (Pajzs 1991, Prószéky - Tihanyi 1992, Prószéky 1996). The result of the analysis is very similar to the result of taggers, which are nowadays widely used in corpus based projects. The main difference is that the analyser not only identifies the root and supplies it with the part of speech code and suffix code, but it also specifies the boundary of the morphemes. (E.g. while the word *kitüntetett* is tagged by the ISSCO tagger as *kitüntet\ [IK][IGE][Me3]* IK=Verbal Prefix, IGE=Verb, Me3=Past Tense 3rd ps sing. the analyser segments it into the following morphemes: *ki[IK]+tüntet[IGE]+ett[Me3].*) The advantage of the more detailed analysis is that the result cannot only be used for lexeme oriented retrieval of the text but for different kinds of linguistic and morphosyntactic researches as well.

2.1. Disambiguation of the analysed text

A large component of the running words can have more than one analysis. The number of alternative solutions depends on many factors: the number of tags in the tagset, the level of analysis, and the number of homonymous and homographic entries in the vocabulary of the given language. In the Hungarian corpus after the application of the HUMOR analyser 30% of the running words have at least two analyses. The process of choosing the correct solution from the possibilities is the disambiguation of the text. For Hungarian three different methods for disambiguation were tested.

- Statistical part of speech tagger based on the Hidden Markov Model (HMM).
- Disambiguation by syntactic analysis of the sentence.
- The use of local rules in the context for disambiguation.

Each method was tested on samples taken from the corpus. The HMM tagger for Hungarian was first implemented and tested in the framework of the MULTEXT-EAST[2] project. This project made available the ISSCO tagger for all the languages involved in the MULTEXT and the MULTEXT-EAST projects. The major advantage of this process is that the tagger can be trained on untagged data. After testing the program on the common corpus of the MULTEXT-EAST (Orwell: 1984) it was also tested on the historical corpus of Hungarian in the framework of the GRAMLEX project by Csaba Oravecz (Oravecz 1998). The major difference in using the tagger for the Orwell corpus and for the historical corpus is that a wordform lexicon was prepared for the Orwell corpus conforming to the MULTEXT lexical specification. This procedure cannot be followed in the case of an unconstrained corpus, since Hungarian is a highly inflective language. Therefore a different practice had to be followed, namely to provide on-line morphological analysis during the tagging process. For testing the HMM tagger on the historical corpus, a set of tools was developed to convert the corpus into the format required by the tagger, to integrate the morphological analyser into the toolset, and to create all the necessary sub results (file of ambiguity classes, probability matrices etc.). In order to use this toolset on the whole corpus it would be necessary to hand-validate a relatively large and representative sub corpus (at least 200,000 running words). For the texts from the 19th and early 20th century the analyser has to be further developed before its result can be disambiguated by the HMM tagger. The current rate of unrecognized words in these texts is much too high (10%), and we have to reduce it radically before using this statistical method for disambiguation.

The theoretically most promising method developed for disambiguation is a syntactic parser for Hungarian. A program was prepared (Novák 1997) and tested on some samples of the corpus. The aim of parsing is to disambiguate multiple analyses in the sentences as a result of syntactic analysis. In the GRAMLEX framework a first version of the program has been implemented and a simple grammar has been written for it. The input of the parser is a text analysed by HUMOR. The program uses chart parsing techniques for the analysis and it produces the successful parses as output. It works with a grammar that is based on hand-written rules and all successful analyses are output. The current version of the program is able to analyse some not highly complex sentences and to eliminate the impossible morphological segmentations as a result of parsing. However, in order to be able to apply it on the historical corpus further research must be carried out.

2.2. Using local rules for disambiguation

The idea of disambiguation by local rules was already suggested in earlier researches (Hearst 1991, Laporte 1994, Mohri 1994). Based on these proposals an examination was carried out in the framework of the GRAMLEX project for three languages: French, Italian and Hungarian. At the first stage the local rules were expressed by the help of regular expressions for each of the languages. Later on a new module was integrated into the INTEX text retrieval software which is able to handle local grammars. The grammars can be described by regular expressions and by graphs as well (Laporte, Monceaux 1997). The main advantage of this system is that the result of disambiguation is very easy to check, so one can correct the rules as soon as they are applied on the corpus. It is also modular, so any number of rules can be added and/or corrected without changing the effect of the earlier local rules. So far it has been applied to French and Italian.
Some typical examples of local rules created for Italian:

— *The determiner/pronoun alternatives are determiners when they are directly followed by adjectives or nouns.*
— *The determiner/pronoun alternatives are pronouns when they are followed by prepositions, verbs or they are at the end of the sentence.*

The local rules which were first formulated for Hungarian were very similar to the Italian. For example:

— *If the alternative is (az[NM] 'pronoun' ¦ az[DET] 'defininite article') and the next word is a noun starting with a vowel, choose the article.*
— *If a noun is followed by an adverb/postposition alternative, choose the postposition.*
— *If there is another verb in the clause omit the verb from the verb/noun, verb/adjective alternatives.*
— *If there is no other verb in the clause, choose the verb from the verb/noun verb/adjective alternatives.*

These and some additional rules were applied on the whole text of *Orwell: 1984*, and the result was hand checked afterwards. On the whole this attempt proved to be rather promising: more than half of the ambiguities were filtered out by the programme correctly 94% of the time. (The summary table of this test is presented in the appendix.) The most problematic rules were the ones choosing or omitting the verb among the alternatives. The theoretical background of this error has many causes. In Hungarian it is possible to create sentences without a verbal predicate, therefore the underlying presupposition in these rules (leave one verb in each clause) is already not quite correct. We presumed however, that statistically this solution would be appropriate in most of the cases. Since sentences not containing a verbal predicate are all in the present tense third person indicative mood, it was feasible to suppose that most of our texts which are samples from literature would rather contain sentences in past tense. On the whole the test did not contradict this supposition: the rules choosing the verb among the alternatives were correct in over 90% of the decisions, while the verb omitting rules were incorrect in 42%.

In the next stage of the research we were trying to improve the rules treating the verb/noun alternatives. In order to get better results we have started to gather data for a Hungarian verb valency database. The first list contained the valencies of those frequently occurring verbs which are homographic either in their base or inflected forms. (For example the word *él* is a verb 'live' and a noun 'edge' and its suffixed form *élet* can be the accusative form of the noun or another noun 'life'.) Several regular expressions were written including these verbs with all of their required case suffixes. The current version of the program chooses the verb among the alternatives whenever it finds the required case ending either directly before the verb or after it in the same clause.

Different kinds of rules were also added to the program, some of them using morphosyntactic properties of the context (e.g.: If the word *szemét* accusative of 'eye' or 'rubbish' is followed or preceded by a verb in a definite conjugation choose the accusative of 'eye'), others trying to use semantic features (e.g.: If the word *nyír* 'cut' or 'birch-tree' is followed or preceded by another noun meaning tree, choose 'birch-tree'). These should be considered as attempts toward a multi-level analysis of the texts. The result of these is also quite promising, in most of the cases the appropriate solutions were chosen by the programme.

After thorough testing, the program was run on the whole analysed corpus. Since at this stage our aim was to gain an analysed corpus with only one segmentation for each running word, after the disambiguation a post-processor programme was used which simply chose the first solution in those cases, where the disambiguator could not match any of the regular expressions.

Example

```
kilenc[SZN]=Kilenc                              kilenc[SZN]=Kilenc
perc[FN]+cel[INS]                               perc[FN]+cel[INS]
nyolc[SZN]                                      nyolc[SZN]
{ óra[FN] ¦ ó[MN]+ra[SUB] }                     { óra[FN] ¦ ó[MN]+ra[SUB] }
{ múlt[MN] ¦ múlt[FN] ¦ múl[IGE]+t[Me3] }, /    { múlt[MN] ¦ múlt[FN] ¦ múl[IGE]+t[Me3] },/
kigyúl[IGE]+t[Me3] #3.01                        { kigyúl[IGE]+t[Me3] ¦ kigyúlt[MN]}
a[DET]                                          a[DET]
víz[FN]                                         víz[FN]
alatt[NU] #7.2                                  { alatt[NU] ¦ alatt[HA] }
a[DET]                                          a[DET]
tűz[FN] #2.1                                    { tűz[FN] ¦ tűz[IGE] } /
és[KOT]                                         és[KOT]
sűrű[MN]+bb[FOK]                                sűrű[MN]+bb[FOK]
lesz[IGE]=le+tt[Me3] #5.2                       { lett[FN] ¦ lesz[IGE]=le+tt[Me3] ¦ lett[MN] }
a[DET]                                          a[DET]
{ parti[MN] ¦ parti[FN] }                       { parti[MN] ¦ parti[FN] }
fűz[FN] #4.4, /                                 { fűz[FN] ¦ fűz[IGE] }, /
hogy[KOT]                                       hogy[KOT]
az[DET] #1.1                                    { az[NM] ¦ az[DET] }
árnyék[FN]                                      árnyék[FN]
%közészorúlt.                                   %közészorúlt.
kilenc[SZN]=Kilenc perc[FN]+cel[INS] nyolc[SZN] óra[FN] #100 múlt[MN] #100,
```

/ kigyúl[IGE]+t[Me3] #3.01 a[DET] víz[FN] alatt[NU] #7.2 a[DET] tũz[FN] #2.1
/ és[KOT] sũrũ[MN]+bb[FOK] lesz[IGE]=le+tt[Me3] #5.2 a[DET] parti[MN] #100 fũz[FN] #4.4, / hogy[KOT] az[DET] #1.1 árnyék[FN] %közészorúlt.

In the example above you can see the output of the analysis (left column of the upper part), the result of disambiguation (right column of the upper part) and the output of the post-processor (lower part). The numbers preceded by '#' are the identification numbers of the rules, the work of the post-processor is identified by #100. The analyser puts an '%' before the unrecognized words. The number of decisions made by the post-processor greatly depends on the actual text. In the texts taken from the 19th century the number of unrecognized running words is already relatively large (10%), therefore the regular expressions which can be employed on the well-analysed texts cannot be used there in many cases. In the more modern texts, where the percentage of unrecognized words decreases (3%), the regular expressions can be used quite frequently, and usually the decision made with their help is correct. The example above is a stanza of a poem from 1934 (author: Miklós Radnóti). This shows the ratio of decisions made by the disambiguator and the post-processor in modern texts. Of the 21 running words of the sentence, 9 had more than one possible analysis, 6 were disambiguated by the regular expressions correctly, and 3 were left for the post-processor. The overall result of this process is a usable lexeme oriented concordance, which serves the needs of the lexicographers in a much more efficient way than the "raw", unanalysed concordances.

The total number of ambiguous word forms in the whole (17 million running words) corpus was 5,716,333. The local rules were applied in 2,226,767 cases and the first analysis was chosen in 3,489,333 cases. These results show that we have to further develop the local rules. We are also aiming to work out a method for combining local rules with statistical data: instead of choosing the first possible solution in the post-processing phase we would use statistical algorithms for those cases where the local rules are not applicable.

2.3. The retrieval of the analysed and disambiguated text

Let's compare the result of the search of the above mentioned *áll* in the unanalysed and the analysed corpus

Unanalysed corpus:
áll 59 157 occurrences
1790-1828 FAZEKAS MIHÁLY: AZ ÉRZÉKENYSÉGEK ÉNEKBEN
Irigylem kis virág *állapotodat*, / Boldogul letört ág, szép halálodat. **'condition'**
1790-1828 FAZEKAS MIHÁLY: AZ ÖRÖM TÜNDÉRSÉGE
Látván, hogy e mind õ érte van, s érettem, / Érzém, mely nagy lelkem *állattársim* felett, / Kikbe ily bölcset a fõ bölcs nem lehellett. **'animal mate'**
1790-1828 FAZEKAS MIHÁLY: NYÁRI ESTI DAL
/ Addig a menny földre hinti / Balzsamának harmatit, / S új erõvel áldva inti / Új örömre *állatit*. A mocsáros nép kuruttyol, / Prüccsög a sok kis bogár, / **'animal'**
1790-1828 FAZEKAS MIHÁLY: ESMERKEDÉS A CSILLAGOS
n egymáshoz, egynehány pedig helyét változtatja, azt csak a figyelmezõ veheti észre. Az *álló* csillagok saját fényekkel ragyognak, melynél fogva úgy **'standing'**

Analysed corpus:
áll[IGE] 25 052 occurrences
1790-1828 FAZEKAS MIHÁLY: ESMERKEDÉS A CSILLAGOS
és ez a Fiastyúkkal és a Kos nagyobbik csillagával egyenlõ háromszegben *áll*. **'stand'**

1790-1828 FAZEKAS MIHÁLY: ALEDDIN NAPKELETI POGÁNY

akarnám építeni, az igazságtalan keresmény nem lehet állandó, mert ki *állhatja* ki a megkárosodott szegény átkát, én legalább magamra nem vonom. **'endure'**

1790-1828 FAZEKAS MIHÁLY: ALEDDIN NAPKELETI POGÁNY

mellé teszi, és ismét jegyzõkõnyveit forgatja. Aleddin némán csak ott *állott*, míg az öreg mintegy történetbõl felvetette a szemét, kérdvén, hogy mit **'stand'**

1790-1828 FAZEKAS MIHÁLY: ALEDDIN NAPKELETI POGÁNY

- Isten hozott - felel Aleddin -, az én ajtóm nyitva *áll* minden idegen elött, mennyivel inkabb pedig olyanok elõtt, **'stand'**

The number of occurrences of the searched string has significantly decreased. As the example sentences show, in the first search only one example was actually relevant among the four, while in the analysed version only one example belonged to a quite different word *kiáll* 'endure'. This could only be mixed among the occurrences of *áll* because the *ki* verbal prefix can be separated from the verbal stem. If the lexicographer had to select the quotations for the entry *áll* by using the unanalysed corpus (s)he would have to go through more than 30,000 quite irrelevant examples, which only happen to begin with the same character string. It is also possible to search for the noun *áll* 'chin' in the analysed corpus. Although the decision of the disambiguator is not always correct, the lexicographer gets much less unnecessary or irrelevant data than before when (s)he could only search in the unanalysed corpus.

The user interface of the text retrieval tool hides the analysed form of the text from the user, because it would be rather difficult to read. However, when one asks for the larger context of the searched word, the analysed form of the text is also shown.

Example:

```
<eg><cit> Jer néz [!] a' Balatont, mikor a' nap' reggeli lángja / Tükrözetén reszket' s mikor a' hold fénnye alatt ég. / Nézd a' kék hegyeket mint állnak sorba körülte / Mellyeken a' Nectár tsorog és az öröm dala harsog.</cit>
<bibl>
<wdate> 1799-1802 </wdate>
<pubDate>1979</pubDate>
<author>BERZSENYI DÁNIEL</author>
<pubTitle> BERZSENYI DÁNIEL ÖSSZES MŰVEI I. KÖLTŐI MŰVEI</pubTitle>
<p>28</p>
</bibl>
<id>1900054019</id></eg>
elemezve: 'analysed version'
jer[ISZ]=Jer néz[IGE] [!] a[DET] ' Balaton[FN]+t[ACC] , mikor[HA] #100 a[DET] ' nap[FN] ' reggeli[MN] #100 láng[FN]+ja[PS] / %Tükrözetén reszket[IGE] ' s[KOT] mikor[HA] #100 a[DET] ' hold[FN] %fénnye alatt[NU] #100 ég[FN] #4.4 . / néz[IGE]=Néz+d[TPe2] a[DET] ' kék[MN] hegy[FN]+ek[PL]+et[ACC] mint[KOT] áll[IGE]+nak[t3] #3.1.1.1.4 sor[FN]+ba[ILL] kör[FN]+ülte[FN] #100 / %Mellyeken a[DET] ' %Nectár %tsorog és[KOT] az[DET] #1.1 öröm[FN] dal[FN]+a[PSe3] harsog[IGE].
```

The upper part contains the quotations supplied with all the necessary bibliographic data and the SGML tags which are used in the dictionary entry. The lexicographer can simply copy this to the entry under compilation and (s)he only has to erase the surplus parts of the quotation.

2.4. Sense distinction by using local rules

Similarly to the disambiguation process, in some cases it is possible to try to discriminate different senses of a word by using the morphological features of the context of the word.

This method was tested on some homographic entries which also has several senses. We will illustrate this through a not highly complex example:

fűz ige [V] 'stitch, sew' **1.** vki vmit vmire/vmibe 'sg to/on sg':
FŰZ [FN],[FN+ACC],[FN+SUB/FN+ILL]: *cérnára fűzi a gyöngyöt,* 'she stiches the pearl on a thread' *koszorúba fűzi a fokhagymát*, 'she sews the garlic into a wreath'
2. 'tie, attach' vmi vkit vkihez 'sg to sb':
FŰZ [FN],[FN+ACC],[FN+ALL]: *barátság fűzi őket egymáshoz* 'they are attached to each other'
fűz fn 'willow'

The different senses of this verb require different case endings, so it is relatively easy to differentiate the senses from the context. The regular expressions for this choice:

```
$word =~s/\{ (fu3z\[FN\]\S* )\| (fu3z\[IGE\]\S* )\} (.*fa\[FN\])/$1 #3.20.01 $3/;
$word =~s/(fa\[FN\].*) \{ (fu3z\[FN\]\S* )\| (fu3z\[IGE\]\S* \})/$1$2 #3.20.02/;
$word =~s/(\[ALL\]\S* \}? *~ )(\{ fu3z\[FN\]\S* \| )(fu3z\[IGE\]\S*) \}/$1$3\<snu\>2\.\<\/snu\> #3.20.1.1/;
$word =~s/(\{ fu3z\[FN\]\S* \| )(fu3z\[IGE\]\S* )\} (.*\[ALL\])/$2\<snu\>2\.\<\/snu\> #3.20.1.2 $3/;
$word =~s/(\[ILL\]\S* \}? *~ )(\{ fu3z\[FN\]\S* \| )(fu3z\[IGE\]\S*) \}/$1$3\<snu\>1\.\<\/snu\> #3.20.1.5/;
$word =~s/(\{ fu3z\[FN\]\S* \| )(fu3z\[IGE\]\S* )\} (.*\[ILL\])/$2\<snu\>1\.\<\/snu\> #3.20.1.6 $3/;
$word =~s/(\[SUB\]\S* \}? *~ )(\{ fu3z\[FN\]\S* \| )(fu3z\[IGE\]\S*) \}/$1$3\<snu\>1\.\<\/snu\> #3.20.1.7/;
$word =~s/(\{ fu3z\[FN\]\S* \| )(fu3z\[IGE\]\S* )\} (.*\[SUB\])/$2\<snu\>1\.\<\/snu\> #3.20.1.8 $3/;
$word =~s/(\[ACC\]\S* \}? *~ )(\{ fu3z\[FN\]\S* \| )(fu3z\[IGE\]\S*) \}/$1$3 #3.20.1.3/;
$word =~s/(\{ fu3z\[FN\]\S* \| )(fu3z\[IGE\]\S* )\} (.*\[ACC\])/$2 #3.20.1.4 $3/;
```

An attempt to use semantic criteria for differentiation was also added in this case: if this word is preceded or followed by the noun *fa* 'tree', we choose the meaning 'willow' among the alternatives. Of course these kinds of rules could only work really well if we had a full semantic database in the background with all the words meaning a kind of tree supplied with the same semantic feature. The other expressions choose the verb from the alternatives when it is directly preceded or followed by the required case ending in the same clause. If the ending is [ILL] or [SUB] it puts the <snu>1.</snu> tag after the word, when the ending is [ALL] sense number 2 is chosen.

Example sentences:
e1s[KOT]=E1s ce1rna[FN]=ce1rna1+ra[SUB] fu3z[IGE]<snu>1.</snu> #3.20.1.7, mint[KOT] a[DET] csereboga1r[FN]=cserebogar+at[ACC], e1s[KOT] zu2mmo2g[IGE]+o2k[e1] %ko2rbe-ko2rbe-ko2rbe.
És cérnára fűz, mint a cserebogarat, és zümmögök körbe-körbe-körbe.
'And he stiches me to a thread like a may-bug, and I am buzzing around'.

szi1v[FN]=Szi1v+em[PSe1]+ben[INE] szent[MN] #100 e1rzelem[FN] e1l[IGE] #5.3, mely[NM] maga[NM]=Maga1+hoz[ALL] fu3z[IGE]<snu>2.</snu> #3.20.1.1 e1s[KOT] hatalmas[MN] ero3[FN]+t[ACC] ad[IGE] a[DET] sors[FN] minden[NM] #100 csapa1s[FN]+ai[PSe3i]+nak[DAT] #100 elvisele1s[FN]+e1[PSe3]+re[SUB].
Szívemben szent érzelem él, mely magához fűz, és hatalmas erőt ad a sors minden csapásainak elviselésére.
'My heart is filled with a saintly emotion which ties me to you, and gives an enormous power to endure all of the strokes of fate.'

As these examples show, in some cases it is possible to identify the different senses of the dictionary entries by examination of the context of the word, the morphological categories occurring in the context and/or semantic features of the word. This process can be done during pre-processing of the corpus, before indexing and retrieving. However, a part of the sense discrimination can be more efficiently done during retrieval of the quotations: from the analysed text the lexicographer can retrieve the co-occurrence of the examined word with other words and/or morphologic classes.

3. Conclusion

Different methods for disambiguation of the analysed corpus of Hungarian are being tested. Each method has several advantages, but they still need to be further improved. Writing local rules for disambiguation is relatively simple and straightforward, easy to test and to apply. They can be used not only for homograph separation but in some cases for sense distinction as well. The effectiveness of the process needs thorough checking. Until it can be done on a large scale, the users of the analysed and disambiguated corpus have to be aware of the possible errors made by the procedure. Syntactic analysis would be the most elegant and adequate method from the theoretical point of view. However, it requires a research work exceeding the possibilities offered by the undergoing projects. Statistical taggers are widely used on English, French and other languages, and as our experiment shows, it is not impossible to use them on such highly inflected languages as Hungarian. The toolset for this method has been prepared, and as soon as a large enough sample corpus can be hand validated it can be used as a training corpus for the statistical tagger. Here it has to be mentioned that another statistical tagger is being tested on parts of our corpus, namely the Brill tagger.

The ultimate solution would be a combination of the above described methods. In the case of the corpus, the realistic procedure could be to combine methods 1 and 3, to apply local rules where possible, and for the rest to use the statistical tagger. For this we should rewrite our local rules and keep only those which are close to 100% correctness. The result of this should be in a format usable by the statistical tagger. The final decision among the remaining possibilities would be made by the statistical method.

4. Notes

[1] The research was carried out in the framework of GRAMLEX Copernicus project. No.: 621 1995-1998). The project for the Historical Dictionary of Hungarian is supported by the Hungarian National Science Foundation No: 014798 1995-1998.

[2] MULTEXT and MULTEXT-EAST are Copernicus projects (No: 0106, 1994–1997)

5. References

Ahlswede, T. E. (1993) Sense Disambiguation Strategies for Humans and Machines. In: *Making Sense of Words Corpora Proceedings of the Ninth Annual Conference of the UW Centre for the New OED and Text Research.* Waterloo, 75-88.

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.*

Hearst, Marti A. (1991) Toward Noun Homonym Disambiguation Using Local Context in Large Text Corpora. In: *Using Corpora Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research.* Waterloo, 1-22.

Justeson John S.-Katz S. M. (1993) Principled Disambiguation: Discriminating Adjective Senses with Modified Nouns. In: *Making Sense of Words Corpora Proceedings of the*

Ninth Annual Conference of the UW Centre for the New OED and Text Research. Waterloo, 57-74.

Laporte, E. (1994) Experiences in Lexical Disambiguation Using Local Grammars. In: *Proceedings of COMPLEX '94.* Budapest, 163-172.

Laporte, E., Monceaux, A. (1997) Grammatical disambiguation of French words. GRAMLEX report.

Mohri, M. (1994) Syntactic Analysis by Local Grammars Automata: an Efficient Algorithm. In: *Proceedings of COMPLEX '94.* Budapest, 173-178.

Novák A. (1997) Disambiguation of suffixal structure of Hungarian words using information about part of speech and suffixal structure of words in the context. GRAMLEX report.

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.

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.* Waterloo, 129-136.

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

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

Silberztein, M.: (1997) The lexical analysis of natural languages. In: *Finite-State Language Processing.* Cambridge Mass./London, 175-203.

Appendix: Rules tested on the Orwell corpus

Rules using the tag [DET]

1.1, #1.2 If the alternative is *{az[DET] ¦ az[NM]}* 'definite article' 'pronoun' followed by a noun or adjective beginning with a vowel choose *az[DET]* 'definite article'.

#1.3, #1.4 If the *{az[DET] ¦ az[NM]}* is followed by a *[DET]* 'art' choose the *[NM]* 'pronoun'.

#2.0 If the *[DET]* is followed by a not suffixed *{FN ¦ MN}* 'noun, adj' alternative and it is followed by another *FN* 'noun' choose the *MN* 'adj' from the alternative.

#2.1, #2.2 If the *[DET]* 'art' is followed by a *{FN ¦ IGE}* 'noun verb' or a *{FN ¦ HA}* 'noun, verb', or a *{FN ¦ ISZ} 'noun, interjection'* alternative choose the *FN* 'noun'.

#2.3, 2.4 If the *{nem[FN] ¦ nem [HA]}* 'noun, adverb' is preceded by a *[DET] 'art'* choose the *FN* 'noun', otherwise the *HA* 'adverb'.

Rules for choosing between verbs and other possible alternatives

#10.1 If the alternative is *{van[IGE]=vagy ¦ vagy[KOT]}* 'verb, conjunction', and it is the first word of a sentence or a clause, choose *vagy[KOT]* 'conjunction'.

#10.2 If the alternative is *{mert[KOT] ¦ mer[IGE]+t[Me3] ¦ mert[MN]}*, 'conjunction, verb, adj.' and it is the first word of a sentence or a clause, choose *mert[KOT]* 'conjunction'.

#3.01, #3.02, #3.03, #3.04 If the verbal part of the *{IGE ¦ FN}* or *{IGE ¦ MN}* alternative is in past tense third person singular, and the alternative is followed by the auxiliaries *volna* or *lesz* or a participle, choose the verb.

#4.1, #4.2, #4.3, #4.4 If there is another verb anywhere else in the sentence, omit the verb from the *{FN ¦IGE} {MN¦ IGE}* alternatives

#5.1, #5.2, #5.3 If there is no other verb in the clause, choose the verb from the alternative.

Various

#6.1, #6.5, #6.2, #6.3, #6.7, #6.4 If there is a verb ending with *-ik*, and the same verb without *-ik* and a noun in accusative follows or precedes it, choose the verb without *-ik*, othewise the one with *-ik*.

#7.1, #7.2 If *FN* noun is followed by a *{NU ¦ HA ¦IK ¦ FN}* 'postposition, adverb, verbal prefix, noun' alternative choose the *NU* 'postposition'.

#8 If the *{jobb[FN] ¦ jol[MN]=jo+bb[KFOK]}* alternative 'right, better' is followed by the nouns *kéz* or *láb* or *boka* 'hand, leg, ankle' choose the jobb[FN] 'right noun'.

#9.1, #9.2 If the *{szemelt[FN] ¦ szem[FN]+el[PSe3]+t[ACC]} 'rubbish' 'eye in accusative'* is followed or preceded by a verb in definite conjugation choose *szem[FN]+el[PSe3]+t[ACC]*.

Summary of the application of the rules

	Correct	Wrong	Total
#1.1	1.765	1	1.766
#1.3	105	2	107
#2.0	90	8	98
#2.1	56	1	57
#2.4	1.212	0	1.212
#10	232	0	232
#3.0	159	19	178
#4.	431	313	744
#5	5.364	283	5.647
#6	158	4	162
#7	367	1	368
#8	1	0	1
#9	7	0	7
Total:	**9.956**	**632**	**10.588**

The sequence of the rules is the same as the sequence of the actual application of the rules within the program.