Attila Novák, Viktor Nagy, Csaba Oravecz (2003): Corpus assisted development of a Hungarian morphological analyser and guesser. In: Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.). Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16. UCREL, Lancaster University, 583-590

Corpus assisted development of a Hungarian morphological analyser and guesser

Attila Novák*°, Viktor Nagy°, Csaba Oravecz°

{novak,nagyv,oravecz}@nytud.hu

***MorphoLogic Ltd., Budapest**

°Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest

1 Introduction

Computational processing of highly inflectional languages – that typically feature a huge number of possible word forms – relies upon an efficient morphological analysis. For this, a comprehensive morphological analyser is needed, which cannot be replaced by a simple lexicon lookup since such a lexicon should contain all word forms for the language and would be computationally intractable. Instead, an analyser should work in tandem with a base form lexicon. The analyser should have the capability of analysing all inflectional, productive derivational and compounding phenomena and it should also be capable of doing base form reduction.

Morphological processing of huge corpora inevitably faces the problem of a large number of word forms whose base form is not listed in the analyser's lexicon so they cannot be analysed. In order to cope with the problem of unknown words in the corpus, a combined method can be applied featuring symbolic constraints and statistical information. The paper will describe and empirically investigate how this method can be put into practice and utilised to improve on the output of the morphological analysis. In section 2 we will give a brief description of the analyser tool as it was originally developed. Section 3 will discuss the symbolic guesser module while section 4 will describe the data used in the experiments. In section 5 we will present the experiments carried out under different settings and show how overgeneration of the guesser module can be tamed by using information from word form and suffix statistics gathered from a huge corpus. Conclusions and suggestions for further work will end the paper in section 6.

2 The morphological analyser

Although morphological analysis is the basis for many NLP applications, especially for highly inflective languages, morphological analysers as separate NLP tools have received limited attention in the literature, most of them being commercial products[1]. The morphological analyser, called HuMOR ('High speed Unification MORphology'), which is used for tagging Hungarian corpora was also developed by a Hungarian language technology company, MorphoLogic (Prószéky and Kis, 1999). It performs a classical 'item-and-arrangement' (IA) style analysis (Hockett, 1954). The input word is analysed as a sequence of morphs, each having (i) *a surface form* (that appears as part of the input string), (ii) *a lexical form* (i.e. the 'quotation form' of the morpheme) and (iii) *a category label* (which may contain some structured information or simply be an unstructured label). The lexical form and the category label together more or less well identify the morpheme of which the surface form is an allomorph. The analyser produces flat morph lists as possible analyses, i.e. it does not assign any internal constituent structure to the words it analyses. The reason for this is that it contains a regular word grammar, which is represented as a finite-state automaton. This is clearly much more efficient than having a context-free parser and it also avoids most of the irrelevant ambiguities a CF parser would produce. Although it has a finite-state word grammar component, HuMOR does not belong to the family of two level finite-state tools which have become the standard for morphological analysis.

Facing an input word form, the program performs a search for possible analyses. It looks up morphs in the lexicon the surface form of which matches the beginning of the input word (and later the beginning of the yet unanalysed part of it). The lexicon may contain not only single morphs but also morph sequences. These are ready-made analyses for irregular forms of stems or suffix sequences, which can thus be identified by the analyser in a single step, which makes its operation more efficient. In addition to assuring that the requirement that the surface form of the next morpheme must match the beginning of the yet unanalysed part of the word

[1] It is, however, worth noting that with the ongoing development of annotated resources for languages with complex morphology that present a problem for simple lookup methods, more and more research has focused on this domain (cf. eg. the Xerox morphological tools (Karttunen, 2000) and for an independent implementation see eg. Alegria et al. (2001)).

(uppercase-lowercase conversions may be possible) is met, two kinds of checks are performed by the analyser at every step, which make an early pruning of the search space possible.

On the one hand, it is checked whether the morph being considered as the next one is locally compatible with the previous one. On the other hand, it is examined whether the candidate morph is of a category which, together with the already analysed part, is the beginning of a possible word construction in the given language (e.g. suffixes may not appear as the first morph of a word etc.). The global word structure check is performed on candidate morphs for which the local compatibility check has succeeded. Possible word structures are described by an extended finite-state automaton (EFSA) in the Humor analyser. A sample output of the analyser is illustrated in Figure 1.

```
analyser>lehetőségekben
lehetőség[S_FN]+ek[I_PL]+ben[I_INE]
lehető[S_MN]+ség[D=FN_PROP]+ek[I_PL]+ben[I_INE]
lehet[S_IGE]+ő[D=MN_MIF]+ség[D=FN_PROP]+ek[I_PL]+ben[I_INE]
lesz[S_IGE]=le+hető[D=MN_HATO]+ség[D=FN_PROP]+ek[I_PL]+ben[I_INE]
```

FN = N	MIF = Present participle	S_ = stem
MN = Adj	PROP = Adj to N deriv. suffix	I_ = inflectional suffix
IGE = Verb	INE= inessive case	D_ = derivational suffix
HATO = Modal passive suffix	PL = plural	

Figure 1: Output of the morphological analyser

Morphs are separated by + signs from each other. The representation of each morph begins with its lexical form which is followed by the category label in brackets `[]`. If the surface form differs from the lexical form, it is printed following an equal sign `=`, otherwise it is omitted. The category labels given in the example are preceded by a prefix (separated by an underscore `_`) which identifies the morphological category of the morpheme (`S`: stem, `D`: derivational suffix, `I`: inflectional suffix). In the case of derivational affixes, the syntactic category of the derived word is also given.

The very detailed morphological analyses produced by the analyser are not needed for such tasks as corpus tagging or parsing. For these tasks a lemmatiser is used, which identifies the lexical form and category of the stem of words and the category of all inflectional suffixes (i.e. it identifies the lemma and the morphosyntactic features of word forms). Compound members and derivational suffixes are normally considered part of the lemma and they do not appear as independent items in the output.

All the analyses listed above are conflated into a single analysis by the lemmatiser, as shown in the example below.

```
lemmatiser>lehetőségekben
lehetőség[FN][PL][INE]
```

Figure 2: Output of the lemmatiser

Various versions of this analyser have been in use for over a decade now. Although the analyser itself proved to be an efficient tool, the format of the morphological database which it uses turned out to be problematic. The most important problem with the analyser was that there were no special tools for creating and maintaining the redundant data structures which make up the morphological database. The descriptions were hardly readable for humans, and practically unmodifiable in any consistent way. This situation was remedied by creating an environment which facilitates the creation of the database. In the new environment, the linguist has to create a high level human readable description which contains no redundant information and the system transforms it into the redundant representations which the analyser uses.

3 The symbolic guesser module

No robust and wide coverage corpus annotation tool set can exist without some means of handling linguistic items uncovered by the initial knowledge of the system. Typical stochastic word level annotation tools, statistical POS taggers for example, normally use some built-in suffix guessing algorithm. However, as Oravecz and Dienes (2002) point out, these methods do not behave well on a language with such a varied morphology as Hungarian. For this reason, the guesser module in the present scenario is built around a symbolic partial word form analyser which can generate more intelligent hypotheses on possible lemma-plus-suffix sequences than the generic suffix guessing algorithms built into statistical POS taggers.

In contrast to some other approaches, the data used by the symbolic guesser is not produced by an automatic generalisation process over a huge set of analysed word forms (cf. eg. Daciuk, 2000). Due to the agglutinative nature of Hungarian, open class words have so many possible suffixed forms that it is not feasible to generate the thousands of forms for thousands of stems just to generalise over the results afterwards. Instead, the database was created by applying the grammatical description normally used for the creation of the regular morphological analyser to a set of possible stem endings representing all open class stem types.

The guesser can identify all inflectional suffix sequences of open categories (nouns, adjectives and verbs). Some very productive derivational suffixes have been included as well. Many of the word forms the program is likely to encounter are of foreign origin with an irregular orthography, which poses a special problem in Hungarian where suffixation is primarily determined by the phonological shape of the stem which is not reflected by the orthographic form of these words in any consistent way. For this reason, a number of constraints, observed when creating the database of the regular morphological analyser (e.g. vowel harmony), had to be relaxed or discarded. Other phonological and orthographic constraints on suffixation which are not violated even by stems of irregular orthography are directly encoded in the data and are checked by the guesser.

Since unknown words in general tend to belong to productive inflectional and derivational paradigms the hypothesis space produced by the guesser can effectively be reduced in the first place by considering only these paradigms in the partial analysis. However, a problem may arise with compounds where the reason for the lack of successful analysis is that the first compound member is unknown to the analyser while the second is a member of a closed stem class (e.g. exhibiting a non-productive stem alternation), or the compound is an instance of a non-productive compound construction not listed in the lexicon of the analyser. Since nominal compounding is very productive in Hungarian, the guesser inevitably encounters such words. A possible solution is to include in the partial analyser all closed class stems which can productively feature as second compound members. In the experiments described in section 5 below, we did not include such stems in the guesser. This results in the program not being able to correctly identify the lemma of such words in some cases. The performance of the guesser could to some extent be improved by adding them to the database.

We made a particularly restrictive assumption about verbs. The set of root verbs in Hungarian is closed. Any new verb stem must have a clearly identifiable ending or it must end in a genuine verb forming derivational suffix. The partial analyser proposes verbal analyses only in cases when the stem has one of these endings. This quite drastically reduces the number of hypothetical analyses produced. The assumption behind these decisions was that the regular morphological analyser can (or should be able to) handle all closed class items.

When creating the guesser database, we decided not to distinguish noun and adjective as distinct stem categories. The reason for this was that the same set of inflectional suffixes can be attached to noun and adjective (and also numeral) stems in Hungarian and thus these categories are in practice impossible to distinguish solely on morphological grounds. Only a few derivational suffixes are specific to adjectives and numerals, respectively. Numerals form a closed class, so we did not include numeral endings. Two of the three suffixes which are normally attached only to adjectives are in most cases homophonous to various inflectional endings. In the cases where the ending is an unambiguous adjective suffix, the category can correctly be recognised by the guesser itself, these cases are quite rare, though. Otherwise, when the guesser produces an analysis containing a noun stem, this can in fact also be an adjectival form. The noun tag attached to the stem by the guesser can later be overridden if the word proves to be an adjective. We hoped to be able to detect differences in the distribution of suffixes between nouns and adjectives and use this information to identify and correct the category.

Another category which the guesser itself does not distinguish from nouns is that of words the form of which is invariable (particles, parts of names etc.). When the program analyses a word as the nominative case of a noun (the nominative form never bears an overt suffix in Hungarian and there are no restrictions on the form of noun stems, thus such an analysis is in fact always returned), the word may also be regarded as invariable.

Figure 3 presents the output of the symbolic guesser module. The format of the output is identical to that of the lemmatiser.

```
guesser>Torgyán
Torgyán[FN][NOM]
Torgyá[FN][SUP]
Torgya[FN][SUP]
Torgy[FN][PSe3][SUP]
```

Figure 3: Sample output of the symbolic guesser module

4 Data

For a comprehensive test of the performance of the morphological analyser and the guesser module, as well as for the statistical information on word form and suffix frequencies, the whole stock of the Hungarian National Corpus (Váradi, 2002) was used as a language resource after some initial filtering[2]. The texts were tokenised using a simple customised tokeniser with the aim of letting the analyser and the guesser try to cope with a wide scale of different tokens, using minimal preprocessing (e.g. named entities were not handled by a separate processing step because morphological information is necessary for an efficient named entity recogniser). This "let's first of all see what morphological information can be available for a token" language processing approach is motivated by the fact that various types of tokens, like abbreviations, proper names, addresses, titles etc. are often suffixed in Hungarian, and even separate modules custom tailored to handle a particular class of them would all need to have access to morphological information.

The corpus was compiled into a word frequency list as input to the morphological analyser/lemmatiser (MA). Table 1 presents the main statistical figures of the result of the morphological analysis. The unknown tokens were then processed by the guesser, which produced a list of possible analyses for each word form the MA left unrecognised. Altogether the guesser assigned 2,360,845 morphosyntactic tags to the 995,396 word forms, which gives an average of 2.37 tags per word. This value is significantly higher then that of the MA (where 3,065,988 tags were assigned to the 2,222,280 forms, averaging 1.38 tag/word). There are two reasons for this difference: on the one hand a number of constraints originally present in the MA are relaxed in the guesser to make analyses of foreign word forms with regular Hungarian inflections possible, and on the other hand the lemmatiser often conflates different analyses which the guesser does not. The main reason for the guesser to eagerly look for (productive) derivational suffixes is that since the more suffixes are analysed the more possible stem tokens can be identified, this approach produces much more solid grounding to the statistical methods we use to evaluate and rank the proposed analyses. Note that if there were not so strict restrictions on verbal analyses and the guesser would also analyse words as adjectives and invariable forms, the average number of analyses per word would be above 5.

Units	Analysed by the MA	Unknown to the MA	Total
Word form types	2,222,280 (69.06%)	995,396 (30.94%)	3,217,676
Word form tokens	125,319,357 (95.50%)	5,907,372 (4.50%)	131,226,729

Table 1: Summary figures of the morphological analysis

5 The experiments and evaluation of the guesser

The hypothesis space generated by the symbolic guesser module needs to be further pruned to exclude improbable analyses. A possible source of information relevant for decreasing the number of analyses is the statistics concerning word form and suffix sequence distribution gathered from the corpus. Based on this kind of data a number of models were used in the experiments to test the performance of the guesser.

5.1 Experiments

In model 1a the selection of a particular analysis for a word form was driven by the corpus frequency of the form the guesser proposed as a stem in the given analysis. That is, the analysis whose stem appeared the highest number of times as an independent token in the corpus was chosen as the preferred reading of the word form. All frequency data was calculated with tokens normalised to lower case. Figure 4 illustrates the output of the guesser weighted by the corpus frequency of occurrence.

[2] Special tokens normally considered as belonging to the domain of tokenisers/segmenters, such as mathematical expressions etc. were discarded. Numbers containing characters other that digits, however, were retained, since inflections are often attached to these, which must be identified by the morphological analyser.

```
19957  Torgyán      Torgyán[FN][NOM] (19957)
                    Torgy[FN][PSe3][SUP] (0)
                    Torgyá[FN][SUP] (0)
                    Torgya[FN][SUP] (0)
1635   mindenképp   minden[FN][_KEPP] (175547)
                    mindenképp[FN][NOM] (1635)
598    Monde        Mond[FN][PSe3][NOM] (6792)
                    Monde[FN][NOM] (598)
```

Figure 4: Output of the guesser in model 1a: form frequency measure

Model 1b was a slightly modified version of the previous model. This model included a filter which removed certain analyses before ranking the rest using the same measure (plain stem form token frequency) as in model 1a. The filter worked as follows: if the MA did manage to assign analyses to the guessed stem, but none of these analyses was compatible with the proposed stem category (e.g. the stem had an analysis as a verb form and the proposed stem was a noun), the analysis was discarded.

Figure 5 illustrates the filtered and ranked output of the guesser. The word *mond* 'say' is a very common Hungarian verb and the analysis containing it was thus filtered out.

```
19957  Torgyán      Torgyán[FN][NOM] (19957)
                    Torgy[FN][PSe3][SUP] (0)
                    Torgyá[FN][SUP] (0)
                    Torgya[FN][SUP] (0)
1635   mindenképp   minden[FN][_KEPP] (175547)
                    mindenképp[FN][NOM] (1635)
598    Monde        Monde[FN][NOM] (598)
```

Figure 5: Output of the guesser in model 1b: filtered version of 1a

In model 2, the analyses produced by the MA were consulted like in model 1b. But in addition to filtering out incompatible analyses, the stem category tag for compatible analyses was changed to that proposed by the MA, and the measure used for these modified analyses was not the plain stem form frequency, but the frequency of all analyses produced by the MA containing the proposed stem. For stems left unanalysed by the MA, word form frequency is used instead of stem frequency as in the previous models. Note that the covert assumption behind this algorithm is that the reason for the lack of analysis by the MA is often not that the stem is missing from the database of the MA but is caused by either a paradigm error in the MA or some orthographical deviation at the stem-suffix boundary (e.g. a hyphen was used to attach a suffix when it should not have according to the rules of orthography or vice versa).

Figure 6 illustrates the output for model 2. *Mond* is filtered out and the tag and frequency of *minden* is changed.

```
19957  Torgyán      Torgyán[FN][NOM] (19957)
                    Torgy[FN][PSe3][SUP] (0)
                    Torgyá[FN][SUP] (0)
                    Torgya[FN][SUP] (0)
1635   mindenképp   minden[FN|NM][_KEPP] (216310)
                    mindenképp[FN][NOM] (1635)
598    Monde        Monde[FN][NOM] (598)
```

Figure 6: Output of the guesser in model 2: hybrid frequency measure

In model 3, we tried to devise and use a measure of the degree to which a proposed stem behaves like genuine stems of the same category, i.e. a measure of nounness, adjectiveness, verbness etc. was calculated. The distribution of all suffix sequences attached to stems of the same category in all word forms which could be analysed by the MA was calculated and stored for all stem categories. We counted all occurrences of tag sequences beginning with the tag of the stem category and normalised these by dividing them with the number of occurrences of the stem category. We thus obtained a normalised histogram (H_C) for each stem category (C). The same procedure was repeated for all analyses proposed by the symbolic guesser. Then the histogram for each proposed stem form plus category ($H(s_C)$) was compared to the global histogram ($H(C')$) of all categories (C') compatible to the stem category, and the absolute difference ($AD_C(s_C)$) of the two histograms was calculated. The absolute difference of the two histograms is a number between 0 and 2 ($AD_C(s_C) \in [0,2]$); 0 if the two are identical and 2 if they do not have any suffix sequence in common. The value that was assigned to the hypothetical stem form plus category as a measure of similarity to genuine stems of category C' was $SM_C(s_C) = (2-AD_C(s_C))/2$. The overall measure for an analysis containing a stem of category C' was the number of occurrences of the stem with category C ($F(s_C)$) multiplied by $SM_C(s_C)$: $OM_C(s_C) = SM_C(s_C) \cdot F(s_C)$.

Unfortunately, this measure did not turn out to be a good indicator. First of all it tended to very strongly overemphasise spurious analyses of words as invariable (being compatible with analyses as a noun in nominative case), since for these the similarity measure was often 1. This effect could possibly have been counterbalanced by defining the overall measure as $OM_C(s_C) = F(s_C)\cdot(1+\lambda\cdot SM_C(s_C))$ with a convenient λ. But the measure also turned out always to prefer adjectival analyses over nouns, because the global distribution of nouns calculated from analyses by the MA as described above contained much less nominative case forms than the analyses produced by the symbolic guesser. On the other hand, adjectives do occur much more often without a suffix than nouns, because it is nouns that normally appear at the end of noun phrases and thus get the inflectional endings. The source for this difference between the global noun histogram and the ones for guessed words might be the difference between the way the lemmatiser and the guesser work: the lemmatiser does not normally return an analysis as a noun in nominative case if the word ends in a derivational suffix, while the guesser does. We are going to repeat this experiment after modifying the lemmatiser so that it will work more like the guesser does. For the time being, we filtered out analyses where the stem category was identified as invariable or adjective.

Figure 7 illustrates the output for model 3.

```
19957  Torgyán      Torgyán[FN][NOM] (10100)
                    Torgyá[FN][SUP] (462)
                    Torgya[FN][SUP] (462)
                    Torgy[FN][PSe3][SUP] (218)
1635   mindenképp   mindenképp[FN][NOM] (679)
                    minden[FN][_KEPP] (6)
598    Monde        Monde[FN][NOM] (338)
                    Mond[FN][PSe3][NOM] (58)
```

Figure 7: Output of the guesser in model 3: suffix distribution similarity measure

5.2 Evaluation

In the lack of information on any standard methodology with respect to the evaluation of the performance of unknown word guessers, we have chosen the following scenario. In the corpus frequency list of unknown forms, we stipulated a threshold for the minimal number of occurrence (10) to eliminate large numbers of thrash tokens to which no meaningful analysis could be assigned (except for "residual" or "miscellaneous") — these are unavoidable in large corpora — and from the remaining set we randomly selected 100 tokens from 10 frequency domains evenly divided along the range from the lowest to highest ranked token. This procedure resulted in a 1000 word list, on which the guessing process was run and evaluated in terms of general precision or performance. For the present experiments, a simplified evaluation was carried out, with each model producing only one preferred analysis for a form, so no separate recall and precision values could be calculated[3].

In order to be able to estimate how much the statistical data gained from the corpus improves the performance of the symbolic guesser, two simple baseline models were evaluated along with the models described in section 5.1. Model 0a was a random choice with even distribution over the proposed analyses for each word form, while model 0b always selected the *noun in nominative case* analysis.

The analyses selected by each model were hand checked. The results are reported in Table 2.

	Model		Performance	
			Types	Tokens
no corpus data used	0a	random choice	69.76%	53.39%
	0b	noun in nominative case	78.09%	88.72%
corpus statistics used	1a	stem form frequency	84.18%	91.89%
	1b	stem form frequency with filtering	84.61%	92.73%
	2	hybrid frequency	84.61%	92.69%
	3	histogram comparison	84.29%	91.85%

[3] We plan to implement an extended framework, where POS tagging methods are also applied to disambiguate the output of the guesser module. In that framework, the model will be allowed to come up with more than one tag for a form. Recall would be calculated as the number of all possible correct analyses proposed by the guesser divided by the number of all correct analyses possible for the word forms, while precision as the number of all possible correct analyses proposed by the guesser divided by the number of all analyses proposed by the guesser.

Table 2: Guesser performance using different statistical models

Since nouns in nominative case are very frequent among unknown words, even the simple 0b baseline model performed quite well. The same fact (along with a similar tendency for verbs to appear most often in present tense indicative third person singular) allows the 1a model with the rudimentary statistics to achieve a good result. The simple stem filtering used in model 1b leads to a further increase in performance. We were disappointed by the performance achieved by the most complicated model 3, the performance of which did not surpass that of model 1a. Therefore, further investigation is needed into the question how a workable similarity measure of stem category can be defined.

The results are a bit difficult to compare to the outcome of other similar efforts. Alegria et al. (2002) report results (around 93% precision) in the wider context of POS disambiguation, where only one reading is considered possible for a word form, while Chanod and Tapanainen (1995) use a fairly similar evaluation method to ours and achieve 85% precision, albeit with a POS tagset of low cardinality (the tagset we use is of a cardinality of several thousand).

6 Conclusion and further work

We have discussed the development of a morphological guesser combining symbolical and statistical methods for an agglutinative language with fairly complex morphology and hope to have shown that the performance of a fundamentally symbolic tool can be effectively aided by statistical information gathered from a large corpus. We evaluated the results from the perspective of a subsequent POS disambiguation task. The same methods could also be applied to the task of identifying candidates for inclusion in the lexicon of the morphological analyser to improve its performance, but then the global set of stems proposed by the symbolic guesser must be ranked. That task would also require that idiosyncratic lexical properties of stems other than category be inferred as well. Some manual checking would also be needed to remove orthographically ill-formed word forms. This method can be used not only to enhance the stem lexicon but also to discover and correct possible paradigm errors in the morphological description built into the analyser. Model 2 seems especially suited to this purpose. The whole process can be iterated to diminish the need for human intervention.

In order to further facilitate selection of a correct analysis from the hypothesis space, part of speech tagging methods (Oravecz and Dienes, 2002) can be applied to utilise the contextual information found around the hypothesised forms in the corpus, making the system especially well suited and robust for morphological processing of unconstrained Hungarian language data.

7 References

Alegria I, Aranzabe M, Ezeiza A, Ezeiza N, Urizar R 2001. Using finite state technology in natural language processing of Basque. In *Proceedings of the Conference on Implementations and Applications of Automata*, pp 2–12, Pretoria.

Alegria I, Aranzabe M, Ezeiza A, Ezeiza N, Urizar R 2002. Robustness and customisation in an analyser/lemmatiser for Basque. In *Proceedings of the LREC-2002 Workshop on Customizing Knowledge in NLP Applications*, Las Palmas.

Chanod J–P, Tapanainen P 1995. Creating a tagset, lexicon and guesser for a French tagger. In Tzoukermann E, Armstrong S (eds), *From Texts to Tags: Issues in Multilingual Language Analysis: Proceedings of the ACL SIGDAT Workshop*, pp 58–64, Geneva.

Daciuk J 2000. Finite state tools for natural language processing. In *Proceedings of the COLING 2000 workshop Using Toolsets and Architectures to Build NLP Systems*, pp 34–37, Luxembourg, Luxembourg.

Hockett C F 1954. Two models of grammatical description. *Word*, 10(2): 210–234.

Karttunen L 2000. Applications of finite-state transducers in natural language processing. In *Proceedings of CIAA-2000*, Lecture Notes in Computer Science. Springer Verlag.

Oravecz Cs, Dienes P 2002. Efficient stochastic part of speech tagging for Hungarian. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pp 710–717, Las Palmas.

Prószéky G, Kis B 1999. Morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp 261–268, College Park, Maryland, USA.

Váradi T 2002. The Hungarian National Corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pp 385–389, Las Palmas.