Morpho-syntactic ambiguity and tagset design for Hungarian*

Tamás Váradi, Csaba Oravecz
Linguistics Institute, Hungarian Academy of Sciences
`{varadi,oravecz}@nytud.hu`

Abstract

The paper reports on work in progress to develop a tag set for Hungarian. The rich morphological structure of the language makes tagging feasible only after a full-scale morphological analysis, which results in a magnitude of patterns that do not easily translate into a corpus tag set of manageable size. The paper analyses the extent and types of morpho-syntactic ambiguity found in a 21m word sample of the Hungarian National Corpus as a preparatory stage in establishing a practical tag set.

1 Introduction

Hungarian morphology is reputed to be notoriously difficult to handle computationally. One recent estimate (Tihanyi, 1996), which only considered the combinatorial possibilities of inflection and derivation, produced the figure of slightly more than 4 million forms for a single verb. Even without compounding and only considering a medium sized stock of lemmas (e.g. 50,000 nouns, 9,400 verbs) the total possible combination of different forms would result in close to 20 billion word forms. This clearly rules out the possibility of processing word forms by lexical lookup from tables. It also poses a challenge for corpus annotation in that it is difficult to establish a tag set that does justice to the rich morpho-syntactic information encoded within the words and at the same time remains computationally tractable. As was pointed out by Elworthy (1995), prior morphological analysis of the corpus may take over a large part of the job of the tagger.

*The research reported in this paper was supported by grant number T026091 of the Országos Tudományos Kutatási Alap.

Before embarking on the delicate task of tag set design it was important to find out what amount and kind of ambiguities one can expect to find in Hungarian. In section 2 we describe the data and the tools used in the morphological analysis of the HNC. Section 3 presents the MSD tag set, i.e. the morpho-syntactic categories derived from the morphological analysis. Section 4 will present an analysis of ambiguity found in the data, and in section 5 we explore the issue of what effect modifying the tag set will have on ambiguity classes. In section 6 we describe the type and amount of disambiguation that seems possible at the lexical level. Some preliminary conclusions will be attempted in Section 7.

2 The morphological analysis

As language resource for our analysis we used roughly half (21m words) of the current stock of the Hungarian National Corpus, which at its present, initial phase comprises newspaper material and parliamentary transcripts. First, a word form frequency list was compiled, which served as input for the morphological analysis. Table 1 presents some basic statistics on the range of word form variation found in the corpus.

Entries	Word forms	Lemmas
21285868	763291	214068[1]

Table 1: The distribution of word forms

The word form list was processed with HUMOR, the morphological analysis tool developed by Morphologic Kft. (Prószéky and Tihanyi, 1996). The main statistical figures of the results

[1]The number of lemmas were calculated on the assumption that alternatives in ambiguous cases were evenly distributed. This is obviously false but the correct figure could only be arrived after the corpus has been completely disambiguated

Units	Unique	Ambiguous	Unknown	Total
Word forms	493182 (64.6%)	112971 (14.8%)	157138 (20.6%)	763291
Tokens	14768298 (69.4%)	5360903 (25.2%)	1156667 (5.4%)	21285868

Table 2: Summary figures of the morphological analysis

are displayed in Table 2. Provided that the morphological analysis is correct, the remaining ambiguity amounts to 25.2% of the tokens and 14.8% of the word forms. The high ratio of unknown words both in terms of lexical and token frequency stresses the importance of integrating a guesser, which was not available at this stage. Enhancing the lexicon with abbreviations also promises to drastically reduce token frequency of unknown cases as a great deal of the high frequency unknown words were abbreviations in film listings.

3 The morpho-syntactic description (MSD)

The output of HUMOR was not suitable to be applied directly in the tagset for two reasons: a) it was not designed to return a POS tag for each analysis of a given word form and b) it returns several analyses at varying levels of specificity. Figure 1 shows the analysis of `lehetőségekben` 'within possibilities'. As regards point a) note that the leftmost item in each line is tagged with a POS label but this POS may change as suffixes are added to the stem. In the first line we find that the noun stem `lehetőség` features in the lexicon as a unit and in this particular case the two suffixes `PL` and `INE` did not modify the POS status of the resulting word form. However, in the following line the derivation suffix `COL` does turn the adjective stem into a noun but this fact remains implicit in the analysis. Point b) is illustrated by lines 2-4, which unfold a derivational tree at successively finer levels. The multitude of analyses in themselves do not create any ambiguity as in this particular example they all amount to the same reading as a noun. They are mentioned here merely to illustrate the need to interpret HUMOR's output to make the data tractable.

For the time being we have decided to eliminate all the derivational details about the internal structure of the rightmost stem but in the future we intend to explore how this rich morphological detail could be preserved in the notation, or at least recovered if necessary.

The analysis provided by HUMOR was processed by a program[2] that established the lemma,

```
lehetőség[FN]+ek[PL]+ben[INE]
lehető[MN]+ség[COL]+ek[PL]+ben[INE]
lehet[IGE]+ő[MIF]+ség[COL]+ek[PL]+ben[INE]
lesz[IGE]=le+het[HAT]+ő[MIF]+ség[COL]
          +ek[PL]+ben[INE]
```

legend:

FN = N	MIF = Present participle
MN = Adj	COL = Adj →N deriv. suffix
IGE = Verb	INE= inessive case
HAT = modal	PL = plural

Figure 1: A sample output of HUMOR

the POS and inflectional structure. Our example was thus transformed into the following form:
`lehetőségekben    lehetőség\[N][PL][INE]`
This format was further converted into a single string representation in the manner of the encoding scheme employed in the MULTEXT-EAST project[3] The intent at this stage was merely to preserve in a concise and consistent notation all the information provided by HUMOR that was relevant for tagging. Table 3 displays the features encoded in the morpho-syntactic descriptions for the major POS categories. One of the major aspects in which the current scheme differs from the one used in the MULTEXT-EAST project lies in the inclusion of the feature "stem category". This is devised to preserve the derivational history of the lemma as well as to indicate the syntactic behaviour of the word as a head category. This scheme allows to treat various kinds of pronouns according to the major POS category they may fulfil so that a nominal pronoun like *rajta* 'on it' is encoded as a N with stem category P. The example above is accordingly recoded as `lehetőségekben\[NP3NO]` (i.e. third person plural noun of noun stem class in 'other' case).

As the size of the full MSD tag set was thought to be prohibitively difficult to handle, different levels of granularity in the tag set were explored. As a result, two reduced MSD tag sets were compiled. One collapsed all noun cases except nominative, accusative and dative into a single 'other

[2]We are indebted to Attila Novák for the software tools mentioned in this section as well as for the design of the MSD scheme displayed in Table 3.

[3]See `http://nl.ijs.si/ME` for the final report.

POS	Num	Pers	Stem [NAR] Mood/Tense [V]	Case [N] Def [V]	Owner's Num	Owner's Pers	Total
N	2	3	6	21	2	3	4536
A	2		2		2	3	24*
R	2	3	2				12*
V	2	3	5	3			90*
Invariant minor categories: Q, D, PRE, RP, C, Int, Y							7
							4670

N = Noun A = Adjective R = Adverb V = Verb
Q = Numeral D = Article PRE = Verbal prefix RP = Postposition
C = conjunction Y = Abbreviation Int = Interjection
Def = Agreement in definiteness with object (def, indef, 2nd person)
Owner's Num = sing. or plural owner Owner's Pers = person marker of owner
* = upper bound, not all combinations are possible
[NAR][V][N] = POS categories to which the attribute apply

Table 3: The MSD scheme

case' category. This produced a maximum of 996 elements. A further reduction to 342 tags was achieved by 1) retaining only the stem category P for nouns, adjectives and adverbs (to accommodate pronouns), and 2) collapsing the tense/mode category for verbs so that it only indicates infinitive forms. The three experimental MSD tag sets will be referred to as F (full), M (medium) and R (reduced) respectively.

It is interesting to find in Table 4 how the larger the tag set is the smaller the ratio of actually occurring tags to the potential maximum number becomes.

F (4670)	M (996)	R (342)
1133 (24.3%)	505 (50.7 %)	332 (97%)

Table 4: Distribution of actual tag frequency relative to max. number of tags in different tag sets

4 Analysis of ambiguity

A ranked ordering of the cumulative frequency of the ambiguous word forms reveals a radically different pattern than that reported for French (Chanod and Tapanainen, 1995), where it was found that "the 16 most frequent ambiguous words account for 50% of all ambiguity". In Hungarian, ambiguity is more evenly dispersed: the top 16 ambiguous forms account for one third of the ambiguities and it takes a little over the hundred most frequent ambiguous cases to cover half the total ambiguity.

As an overall measure of the distribution of ambiguity per tokens, Table 5 displays corresponding figures for our corpus of the index proposed in Tufiş (1998). Compared to Romanian, another highly inflected language, ambiguity in Hungarian is more evenly distributed. Another important finding that emerges from the identical value of the ambiguity measure AW between the F and M tagsets is that the finer resolution tagset does not increase the average number of alternatives to ambiguous words. In other words, the same amount of ambiguity can be tackled with the M tagset, which is one fourth the size of the other.

Tagset	TW	KW	AW
F	1.30	1.32	2.21
M	1.30	1.32	2.21
R	1.27	1.28	2.15

Table 5: Different measures of text ambiguity at various tagset levels

TW = number of tags / number of word tokens
KW = number of tags / number of word tokens – unknown cases
AW = number of tags assigned to ambiguous cases / number of ambiguous tokens

5 Ambiguity classes and tag-set size

Applying the three tag sets, lists of ambiguity classes were drawn up using the thousand most frequent ambiguous word forms (they covered roughly eighty per cent of all ambiguities). Table 6 shows that while the size of the tag set is reduced to its one fourth if case is eliminated, there is no corresponding drop in the number of ambiguity classes, much less in the coverage of the tokens involved. The R tagset brings about a sizeable drop in ambiguity types without a proportional decrease in the tokens.

	F (4670)	M (996)	R (342)
amb. classes	284	236	167
tokens	3968341	3967256	3808927

Table 6: Number of ambiguity classes and their coverage across the different tag sets

One important consideration in reducing MSD tagsets by merging categories is the extent to which the eliminated distinctions can be restored. Table 7 displays the number of ambiguity classes where the MSD could not be restored on the basis of the POS category alone. The figures are fairly high even with the reduced tagset, which suggests that the tiered tagging approach suggested in Tufiş (op.cit) should make recourse to finer distinctions than POS membership, including perhaps those features already collapsed in the R tagset.

The merging of stem category features for nouns appears to be justified, many of the N..N/N..V distinctions proved spurious and originated from redundant overlap in the lexicon.

On the other hand, the overall merging of the Mood/Tense category does not seem to be judicious in that they may serve to identify homonymous imperative/subjunctive and indicative forms of some verbs.

6 Disambiguation at the lexical level

A close analysis of the particular ambiguity classes is beyond the scope of the present paper but a few general remarks suggest themselves.

Even a cursory look at the data showed that there were a great number of ambiguity types which can be disambiguated at the lexical level. Part of such cases obviously originate from inappropriate analyses by the morphological analyser, which can be corrected. An even greater part of the spurious ambiguities involve cases where one or several of the forms involved are extremely unlikely (such as pseudo-compounds) or forms where the legal combination of morphemes results in a word form that not only does not exist but can even be ruled out. An example for the latter is the homograph `levél`, the perfectly common reading of which is the noun meaning 'letter'. In addition, however, it is also spuriously analyzed as a PREVERB+VERB combination `le+vél` 'down' + 'consider,think', which lacks any possible semantic reading. Obviously, we do not pretend to have a recipe for dealing with compounding but many of such impossible or extremely improbable formations could be filtered out even if at the risk of misidentifying a negligible number of cases.

Unfortunately, Hungarian does not seem to bear out the observation made for Romanian in Tufiş (op.cit.) that inflected words tend to be unambiguous. One source of widespread ambiguity involves suffix homonymy where one inflectional ending or a string of suffixes correspond to an individual lemma e.g. *-ét* 'possessive suffix+accusative' vs. 'foodstuff' etc. Many of such cases are the single representative of fairly large ambiguity classes.

As a preliminary indication of the scope for lexical disambiguation in our data consider Table 8 showing the number of ambiguity classes that can be partially or fully disambiguated. Partial disambiguation means reducing the number of alternatives in the ambiguity class.

7 Conclusions

The method of studying the extent and types of ambiguity on word form lists derived from an extensive corpus has provided useful orientation about the rough strategy to follow in tagset design. The pattern of distribution of ambiguity is fairly even and widespread. It is not to be expected that tackling a handful of ubiquitous cases will spectacularly reduce the total ambiguity. The high ratio of unidentified tokens clearly points to the vital need of a guesser for any system to yield acceptable results.

In varying the size of the tag set, a major consideration is the complex case system, which can produce a four-fold increase in the number of tags. Case does matter in Hungarian in that, for example, there is agreement in case within a noun phrase between the head and demonstrative pronouns. However, as the figures in Table 6 indicate, case itself is a negligible source of ambiguity. Eliminating case categories hardly changes the number of ambiguous tokens whereas it results in a more than fourfold decrease in the tagset, improving the feasibility of applying statistical tagging methods.

The finding that barely half of the ambiguity classes allow the recovery of the original MSD on the basis the POS category stresses the importance of the finer distinctions within the inflectional endings that play an equal role in ambiguity resolution. This suggests that overall merging of categories prove too crude a tactic and each fea-

	F (4670)	M (996)	R (342)
amb. classes	148 (52.1%)	125 (52.9%)	69 (41.3%)
tokens	927474 (23.3%)	926389 (23.3%)	546280 (14.3%)

Table 7: Number of non-recoverable MSD ambiguity classes and their coverage

	Total	Reduced	Eliminated
amb. classes	284 (100%)	84 (29.6%)	41 (14.4%)
tokens	3968341 (100%)	890,700 (22.4%)	647,800 (16.3%)

Table 8: Number of ambiguity classes partially or fully disambiguated (F tagset)

ture should be evaluated on its merit. The pernicious homonymy of some suffix combinations with individual stems calls for lexical weighting to be applied.

The method used in our enquiry derives its strength from its context independence. Ambiguity patterns of the corpus were studied at the level of word form types. However, this is also its weakness in that categories that suggest themselves for merger may prove valuable clues in resolving ambiguity in the immediate environment. Unfortunately, at the current stage or our work we have no tagging results against which to evaluate and balance our findings. The next phase of the work must evaluate the choice of tags against both the paradigmatic and the syntagmatic behaviour of word classes.

References

J.-P. Chanod and P. Tapanainen. Tagging French – comparing a statistical and a constraint-based method. *cmp-lg/9503003*, 1995.

D. Elworthy. Tagset Design and Inflected Languages. In *Proceedings of the ACL-SIGDAT Workshop*, Dublin, 1995. (also available as cmp-lg/9504002).

G. Prószéky and L. Tihanyi. Humor – a Morphological System for Corpus Analysis. In *Proceedings of the first TELRI Seminar in Tihany*, pages 149–158, Budapest, 1996.

L. Tihanyi. *MULTEXT-EAST Delivearable D1.2. Application to Hungarian. Appendix 2*, chapter Number of Hungarian Word Forms. , 1996.

D. Tufiş. Tagging Romanian Texts: a Case Study for qtag, a Language Independent Probabilistic Tagger. In A. Rubio, N. Callardo, R. Castro, and A. Tejada, editors, *First International Conference on Language Resources and Evaluation*, pages 589–596, Granada, Spain, 1998. ELRA.