# Principled Hidden Tagset Design for Tiered Tagging of Hungarian

**Dan Tufiş***, **Péter Dienes**†,
**Csaba Oravecz**†⋆, **Tamás Váradi**†

*Romanian Academy (RACAI)
13, '13 septembrie', 74311, Bucharest 5, Romania
tufis@valhalla.racai.ro

†Research Institute for Linguistics
Hungarian Academy of Sciences, Budapest
{dienes,oravecz,varadi}@nytud.hu

**Abstract**

For highly inflectional languages, the number of morpho-syntactic descriptions (MSD), required to descriptionally cover the content of a word-form lexicon, tends to rise quite rapidly, approaching a thousand or even more set of distinct codes. For the purpose of automatic disambiguation of arbitrary written texts, using such large tagsets would raise very many problems, starting from implementation issues of a tagger to work with such a large tagsets to the more theory-based difficulty of sparseness of training data. Tiered tagging is one way to alleviate this problem by reformulating it in the following way: starting from a large set of MSDs, design a reduced tagset, Ctag-set, manageable for the current tagging technology. We describe the details of the reduced tagset design for Hungarian, where the MSD-set cardinality is several thousand. This means that designing a manageable C-tagset calls for severe reduction in the number of the MSD features, a process that requires careful evaluation of the features.

## 1. Introduction

The combinatorial possibilities of inflection and derivation in Hungarian morphology (for an estimate see (Tihanyi, 1996)) pose a challenge for corpus annotation in that it is difficult to establish a set of morphosyntactic descriptions that does justice to the rich morpho-syntactic information encoded within the words and at the same time remains computationally tractable. Tiered tagging (Tufiş, 1998) is one way to alleviate this problem by reformulating it in the following way: starting from a large set of MSDs, design a reduced tagset, Ctag-set, manageable for the current tagging technology. The Ctag-set is used as a hidden tagset for the proper tagging of a text. This text, tagged in terms of the Ctag-set, is subject to a procedure aiming at recovering all (or most of) the information left out from the Ctag-set with respect to the MSD-set. In other words, each Ctag assigned to an item in the tagged text, is replaced with an appropriate and more informative descriptor, namely a MSD.

In section 2. we will give an overview of the general principles one can follow in the design process. Section 3. presents the data analysis mostly along the lines described in (Váradi and Oravecz, 1999), but with much larger data sets and further investigations than those presented there. Section 4. will describe the process of reducing the MSD set into a Ctag set of manageable size. In section 5. we show some preliminary results on tagging accuracy and error analysis comparing the performance of the tagging process with a verbose tagset and that of the tiered tagging with a more constrained tagset. Conclusions and suggestions for further work will follow in section 6.

⋆The author was supported by the Research Support Scheme of the Open Society Support Foundation, grant No.: 320/1998

## 2. General requirements for tiered tagging

The design process of a reduced tagset has to consider two fundamental requirements: to identify and leave out the features/values in the MSDs which do not provide relevant clues for the contextual disambiguation, and to make it possible to recover as accurately and fast as possible the information eliminated in the previous phase.

Fortunately, these two objectives, although not very simple to reach, are feasible and rewarding. The process is a trial-and-error one and relies both on human introspection and evidence provided by the data analysis. One possible approach would be to use an information loss-less algorithm to convert the MSD-set into a Ctag-set. Such an algorithm might reduce the size of the tagset with 10-20%, which is too little for a large initial tagset. However, modifying such an algorithm to allow for limited ambiguity (that is losing a limited amount of information), could result in a drastic reduction of the Ctag-set, up to a cardinality which is within the restrictions imposed by the available training data and computing power.

The remaining problem is deciding what kind of ambiguities to accept in the output of such a generalization algorithm, so that by using a subsequent process we will be able to resolve them. In our approach, the reduced tagset is designed as a subsuming one for the MSD-set and as such once a Ctag was assigned to a lexical item in the tagged text, the recovery process has to identify the relevant MSD, out of the set of the MSDs that are subsumed by the Ctag in case. The recovering process could be lexicon driven (the lexicon would be encoded in terms of the large MSD-set) and can be conceived of as the intersection between the set of MSDs subsumed by a Ctag assigned to a wordform *w*, and the set of MSDs for *w* as provided by the lexicon (Tufiş, 2000). This model can be compiled as a database, so that the recovery process could be a simple look-up in this database.

Actually, for Hungarian the construction of such a system is a bit cumbersome. The huge number of possible wordforms in Hungarian rules out the possibility of lexical lookup from precompiled tables for unconstrained corpora and makes the use of a morphological analyzer necessary at least in the preparation phase of the corpus for the tagging process. The output of the morphological analysis is then converted into the MSD encoding, and, in principle, a specific lexicon could be constructed containing the lexical items with their corresponding MSDs for the corpus to be tagged. This lexicon can then be used in the recovery process for the lexical items in the tagged corpus but obviously will not suffice for other corpora. Thus, one can either construct specific lexicons for each chunk of corpora to be tagged and use them in the MSD lookup, or resort to the morphological analyzer in the recovery process as well to provide the set of possible MSDs for the lexical items "on-line". This whole issue basically boils down to an efficiency problem and needs further investigation (besides a fast morphological analyzer).

The items that this recovering process makes ambiguous are more often than not the difficult cases for statistical disambiguation methods. Therefore, the tiered-tagging approach might use for such cases a rule-based disambiguation phase as well.

## 3. Data analysis

### 3.1. The morphological analysis and morphosyntactic descriptions (MSD)

The language resource of our analysis consisted of the whole current stock of the Hungarian National Corpus (approximating 80m words) compiled into a word frequency list as input to the morphological analysis. Table 1 presents some basic statistics on the range of word form variation found in the corpus.

| Entries | Word forms | Lemmas |
|---|---|---|
| 74,063,211 | 1,728,771 | 429,612[1] |

Table 1: The distribution of word forms

The word form list was processed with HUMOR, the morphological analyzer developed originally for Hungarian (Prószéky and Tihanyi, 1996). The main statistical figures of the results are displayed in Table 2. Provided that the morphological analysis is correct, the remaining ambiguity amounts to 27.7% of the tokens and 13.2% of the word forms, indicating that ambiguous items tend to appear in the upper regions of the word frequency list.

The output notation of the morphological analyzer was not suitable to be applied directly as a MSD set for two reasons: a) it was not designed to return a POS tag and a lemma for each analysis of a given word form and b) it returns several analyses at varying levels of specificity. For illustration purposes an example is repeated here from (Váradi and Oravecz, 1999): 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 derivational 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 inflectional suffixes `PL` and `INE` obviously 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 the analyzer's output to make the data tractable.

```
1. lehetőség[FN]+ek[PL]+ben[INE]
2. lehető[MN]+ség[COL]+ek[PL]+ben[INE]
3. lehet[IGE]+ő[MIF]+ség[COL]+ek[PL]
   +ben[INE]
4. 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 the morphological analyzer

To construct an initial MSD notation we eliminated all derivational details about the internal structure of the rightmost POS category[2]. Only the lemma, the POS category and the inflectional structure is preserved. So the above example is transformed into the following form:
`lehetőségekben   lehetőség\[N][PL][INE]`
This format represents roughly the same information as and can in principle be mapped into the EAGLES compliant encoding scheme developed in Multext–East (Erjavec and Monachini, 1997). However, the presence or lack of some of the distinctions in one representation with respect to the other does not make a fully automatic mapping from one format to the other possible, so for the time being the above format is used as an internal MSD notation as output from the morphological analysis.

Still, to establish the possibility of referring to positional attributes and their values in MSD representations, which facilitates the identification of reducible features for the corpus tagset, the MSD scheme, as an initial step in tagset creation, is converted into an attribute/value single

[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 at after the corpus has been completely disambiguated.

[2]The possibility of preserving this information in a concise way is currently under investigation. However, this needs a reformulation of the notation of the morphological analyzer into a labelled bracketing-like representation.

| Units | Unique | Ambiguous | Unknown | Total |
|---|---|---|---|---|
| Word forms | 1,048,263 (60.6%) | 228,105 (13.2%) | 452,403 (26.2%) | 1,728,771 |
| Tokens | 50,437,483 (68.1%) | 20,542,442 (27.7%) | 3,083,286 (4.2%) | 74,063,211 |

Table 2: Summary figures of the morphological analysis

string representation. The intent at this stage is merely to preserve in a concise and consistent notation all the information provided by the MSD that is relevant for tagging. Table 3 displays the features encoded in this initial Ctag scheme (F set) 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 for instance various kinds of pronouns according to the major POS category they may fulfill 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\[NP3N2]` (i.e. third person plural noun of noun stem class in '2' (=inessive) case).

### 3.2. Analysis of ambiguity at the MSD level

A ranked ordering of the cumulative frequency of the ambiguous word forms reveals a fairly even pattern of the coverage of top ambiguity classes in terms of tokens: it takes a little over the hundred most frequent ambiguous cases to cover half of the total ambiguity. As an overall measure of the distribution of ambiguity per tokens, Table 4 displays corresponding figures for our corpus of the index proposed in (Tufiş, 1998). The comparison of these values to those of different tagset schemes in section 4.2. will provide some insight into the recoverability of information from and coverage of tagsets over ambiguity classes.

| | TW | KW | AW |
|---|---|---|---|
| MSD | 1.334 | 1.349 | 2.230 |

Table 4: Different measures of text ambiguity at the MSD level

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

## 4. The design of Ctag sets

### 4.1. The reduction of the initial tagset

As the cardinality of the full initial tagset is too high to be handled by current tagging methods, especially by statistical taggers, different levels of granularity in the tagset have been explored. This section addresses the problems of possible reductions of the initial tagset. There are three important principles we have to consider during this process:

(i) when merging MSDs into the the same Ctag, we have to retain the *recoverability* of the original MSD of each word;

(ii) we should not lose any information giving contextual clues for the disambiguation of other words;

(iii) ambiguity classes should be merged when contextual information is not enough to disambiguate.

Merging MSD tags in the light of these principles is an empirical issue: checking the fulfillment of the principles involves either the investigation of the ambiguity classes occurring in the corpus (principle (i)) or the comparison of tagging results obtained by using the merged/non-merged tagsets (principles (ii) and (iii)).

As an initial attempt to design the reduced corpus tagset, we made use of the algorithm proposed by (Tufiş, 2000) to remove features from the full tags that can be recovered from the intersection of the set of MSDs for a lexical item with the set of MSDs the proposed Ctag for this item subsumes. Basically, the algorithm removes an attribute from a tag if this attribute is recoverable, i.e. the deletion of the attribute does not merge two tags in an ambiguity class. Applied to a 74 million word lexicon, the algorithm yielded the results in Table 5.

| Head | Pos. | Meaning |
|---|---|---|
| `A` | 2 | number of the Adjective (always S) |
| `A` | 3 | not in use |
| `R` | 3 | not in use |
| `R` | 4 | marking verbal root in Adverbs |
| `N` | 6 | number of the possessor of the Noun |
| `N` | 3 | person of the Noun |
| `V` | 2 | number of the Verb |

Table 5: Recoverable automatic reductions

Although the deletion of the proposed features results in a recoverable tagset, the resulting reduction in the size of the tagset is significant but not satisfactory. The first three items in Table 5 do not involve any reduction in the tagset since they are the same in all tags of the given category (`S`, _ and _, respectively). Deleting the verbal root in adverbs brings about a most minimal decrease (1). In the final two rows, the deletion of number of the verb or the person of a noun is very problematic, since we lose an important clue for contextual disambiguation (in Hungarian the verb and the subject must agree in number and person). Thus, the only relevant feature that can be deleted, retaining recoverability, is the number of the possessor of a noun. After such a deletion, the cardinality of the tagset is still too high to be convenient for current tagging methods: 1265 different tags would remain.

| POS | Num | Pers | Stem [NAR] Mood/Tense [V] | Case [N] Def [V] | Owner's Num | Owner's Pers | Total |
|---|---|---|---|---|---|---|---|
| N | 2 [PS] | 3 [123] | 5 [QAVNP] | 21 | 2 [PS] | 3 [123] | 2058* |
| A | | | 2 [AV] | | | | 2* |
| R | | | 2 [RV] | | | | 2* |
| V | 2 [PS] | 3 [123] | 5 [PRCSI] | 3 [ID2] | | | 79* |
| Invariant minor categories: Q, D, PRE, RP, C, Int, Y | | | | | | | 7 |
| | | | | | | | 2148 |

| | | | |
|---|---|---|---|
| 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
* = not all combinations are possible, so not a simple product
[NAR][V][N] = POS categories to which the attribute apply

Table 3: The initial Ctag scheme (F set)

These results might be attributed to two main reasons. First, each feature within a tag seems to be relevant, that is, the tagset is very compact. However, this insight is not supported by the fact that within the 74 million corpus only 1105 tags occur out of 2148 theoretical possibilities (see Table 6 for details on tag statistics). The second reason for such unsatisfactory results lies within the algorithm itself. This algorithm can only remove certain attributes from the full description, instead of applying a merger of some of the distinctions in attribute values. This difference is crucial for the interpretation of the results. Consider, for example, case marking on nouns. Hungarian has 21 cases, which are represented by the fifth attribute of the noun tag. The algorithm does not remove case marking, since ambiguity due to case (eg. `NS3NN`∼`NS3NA`) actually occurs (eg. *párt* 'party+NOM' or 'couple+ACC'). However, there is no ambiguity class where, all other attributes being equal, the word can be analyzed as eg. being either in superessive or in dative case. This means that the cases dative and superessive can be merged. This option is not available for the algorithm proposed by (Tufiş, 2000). Hence, we have also made recourse to non-algorithmic methods, relying on linguistic intuition.

Starting from the F tagset, we made experiments with three reduced tagsets to be described presently. The *medium* (M) tagset has been derived by merging non-nominantive and non-accusative cases of nouns (represented by the letter `O` standing for "other"). Hence, this tagset distinguishes 3 cases only: `N`, `A` and `O`, which yields in the tagset cardinality of 384 tags, out of which 309 actually occurs in the corpus. Further reductions were made in order to arrive at the so called *optimal* (O) tagset. The merged attributes are the following:

**Possessor on Nouns** The distinction between S1, S2, P1, P2 and P3 possessors is recoverable. Furthermore, they do not bear any information giving clue to the disambiguation of any words.

**Objects on Verbs** The full tagset marks three kind of objects in the case of verbal tags: `I` – no object or indefinite object, `D` – definite object and `2` – incorporated second person object. The classes `I` and `2`, however, can be merged, since this distinction is recoverable and they behave in the same syntactic way.

These reductions result in a tagset of 240 tags. Finally, we blurred stem-category distinctions in nouns and adjectives (i.e. the fourth attribute was removed), which gave a *reduced* (R) tagset with 119 tags. However, in this step we strongly violated the recoverability principle. The rationale behind this move was principle (iii): it was thought that this information could not be extracted from the context and the disambiguation of other tokens could not make use of this information either. The findings with tagging test in section 5., however, seem to question this assumption. Table 6 summarizes the results.

| Tags | Number of tags | | Saturation |
|---|---|---|---|
| | Theoretical | Actual | |
| F | 2148[3] | 1105 | 51.4% |
| M | 384 | 309 | 80.4% |
| O | 240 | 205 | 85.4% |
| R | 119 | 96 | 80.6% |

Table 6: Corpus-tagsets

Note that the decrease in the saturation value of the reduced tagset indicates non-recoverability: this move basically affected tags that were used frequently within the lexicon.

### 4.2. Ambiguity classes and tagset size

Applying the four tagsets, lists of ambiguity classes were drawn up using all ambiguous word forms from the corpus. Table 7 presents the measures of text ambiguity at the four levels. The MSD values are repeated here for convenience of comparison.

[3]The number of MSDs actually occurring is 5261 out of a possible value of around 10,000.

| | TW | KW | AW |
|---|---|---|---|
| MSD | 1.334 | 1.349 | 2.230 |
| F | 1.330 | 1.345 | 2.224 |
| M | 1.330 | 1.3445 | 2.223 |
| O | 1.330 | 1.3445 | 2.223 |
| R | 1.291 | 1.303 | 2.164 |

Table 7: Measures of text ambiguity

| Ctag set | Error rate | |
|---|---|---|
| | MULTEXT-ISSCO | TnT |
| F | 6.04% | 3.92% |
| M | 5.65% | 3.34% |
| O | 5.45% | 2.75% |
| R | 6.00% | 2.51% |

Table 10: Error rate with the tagsets

An important finding that emerges from the identical value of the ambiguity measure AW between the F and O tagsets is that the finer resolution tagset does not significantly increase the average number of alternatives to ambiguous words. In other words, the same amount of ambiguity can be tackled with the O tagset, which is almost one tenth the size of the other.

Table 8 also gives some supporting evidence that the O tagset complies with the requirement in principle (i), inasmuch as while the size of the tagset is significantly reduced, there is no corresponding drop in the coverage of the tokens involved. The R tagset brings about a sizeable drop in the ambiguity types but there is some decrease in the tokens as well compared to the difference found between the F, M and O tagsets. The huge drop in the number of the classes between the MSD and F notation seems to justify the need for the latter as the initial notation for further reductions: many of the spurious ambiguities present between MSDs can be resolved by preserving only features relevant for the tagging process if only limited or no irrecoverable lexical information is lost.

The distribution of intracategorial ambiguities across the different tagsets proposed in Table 9 underlines the importance of the evaluation of each of the features distinguishing such classes and suggests that if losing a limited amount of information is allowed significant decrease can be achieved in the number of ambiguity classes, again without a corresponding loss in token coverage.

## 5. Evaluation of tagsets

The three reduced Ctag sets proposed have been subject to practical evaluation in actual tagging experiments. We have made use of two HMM taggers: Thorsten Brants' 3-gram TnT tagger (Brants, 1998) and the MULTEXT-ISSCO bigram tagger (Gilbert and Amstrong, 1995) used in the Multext-East project (Erjavec and Ide, 1998). The training corpus consisted of two register-diverse corpora: the first three quarters of Orwell's 1984 and newspaper text, adding up to 87969 tokens altogether. The test corpus included the rest of the Orwell and newspaper texts, 21267 tokens in total. The MULTEXT-ISSCO tagger was trained with the Baum-Welch algorithm. The TnT tagger has the problem of learning possible ambiguity classes and words from the training corpus only. To remedy this situation, after the training phase, we enriched the generated lexicon file with further ambiguities and added words from the test corpus with their ambiguity classes. The results of the tagging are summarized in Table 10.

To some extent, the tagset cardinality correlates with the test results, except for the R tagset with the MULTEXT tagger. The increase of error rate in this case might be attributed the lack of contextual information which could have been provided by features already missing from the R tagset. That there is practically no information loss if switching from the M to the O set only the latter being more compact is nicely justified by the improved results with both taggers. Obviously, only much more extensive testing could provide reliable justification, however, these preliminary experiments can also indicate whether feature reduction/merger applied so far are on the right track.

## 6. 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 finding that a large percentage of ambiguous tokens belong to intracategorial ambiguity classes stresses the importance of the finer distinctions within the inflectional endings that play an equal role in ambiguity resolution. This suggests that overall merging or deletion of features might prove too crude a tactic and each feature should be evaluated on its merit. In varying the size of the tagset one can gain important insight by looking at the distribution of ambiguity classes and their coverage over tokens in the corpus independent of any contextual information. However, it is the actual evaluation in tagging experiments that plays the role if the resulting tagset is to comply with principle (ii).

## 7. References

Brants, Thorsten, 1998. *TnT – A Statistical Part-of-Speech Tagger, Instalation and User Guide*. University of Saarland.

Erjavec, Tomaž and Nancy Ide, 1998. The MULTEXT-EAST corpus. In Antonio Rubio, Natividad Gallardo, Rosa Castro, and Antonio Tejada (eds.), *First International Conference on Language Resources and Evaluation, LREC'98*. Granada: ELRA.

Erjavec, Tomaz and M. Monachini, 1997. Specifications and notation for lexicon encoding. COP Project 106 Multext-East, Deliverable D1.1 F (Final Report).

Gilbert, R. and S. Amstrong, 1995. Tagging tool. MULTEXT Deliverable 2.4.1.

Prószéky, Gábor and László Tihanyi, 1996. Humor – a Morphological System for Corpus Analysis. In *Proceedings of the first TELRI Seminar in Tihany*. Budapest.

| | MSD | F | M | R | O |
|---|---|---|---|---|---|
| amb. classes | 7205 | 3123 | 1542 | 1370 | 590 |
| tokens | 20155486 | 19998444 | 19994060 | 19993592 | 18526417 |

Table 8: Number of ambiguity classes and their coverage across the different tag sets

| | F | M | O | R |
|---|---|---|---|---|
| amb. classes | 2218 (71%) | 759 (49.2%) | 602 (44%) | 120 (20%) |
| tokens | 6355199 (31.7%) | 6336289 (31.7%) | 6335331 (31.7%) | 3770062 (20.3%) |

Table 9: Intracategorial ambiguities and their coverage


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

Tufiş, Dan, 1998. Tiered tagging. Technical Report 32, RACAI.

Tufiş, Dan, 2000. Using a large set of eagles-compliant morpho-syntactic descriptors as a tagset for probabilistic tagging. In *Proceedings of Second International Conference on Language Resources and Evaluation*. Athens. This volume.

Tufiş, Dan, 1998. Tagging Romanian Texts: a Case Study for qtag, a Language Independent Probabilistic Tagger. In Antonio Rubio, Natividad Callardo, Rosa Castro, and Antonio Tejada (eds.), *First International Conference on Language Resources and Evaluation*. Granada, Spain: ELRA.

Váradi, Tamás and Csaba Oravecz, 1999. Morpho-syntactic ambiguity and tagset design for Hungarian. In *Proceedings of the EACL LINC Workshop on Annotated Corpora*. Bergen, Norway.