Shallow parsing of Hungarian business news

Tamás Váradi
Research Institute for Linguistics
Hungarian Academy of Sciences

varadi@nytud.hu

Abstract

The present paper reports on an attempt to annotate noun phrases in Hungarian using cascaded regular grammars. Hungarian presents several difficulties to shallow parsing such as discourse oriented constituent order as well as left-branching recursive possessive and participle structure inside noun phrases. The approach uses cascaded regular grammars and was developed with the CLaRK system. The NP grammar was tested on a morphologically tagged and disambiguated corpus of 928 sentences representing a sample of highly sophisticated written style of journalism. The results are encouraging even on this challenging text type.

1 Introduction

The present paper reports on ongoing work to develop a shallow parser for Hungarian. So far, there is no syntactic parser developed for Hungarian. There are two systems of morphosyntactic taggers developed at Szeged University (Gyimothy et al. 2000) and the Hungarian Academy of Sciences (Dienes Oravecz 2002) both using HUMOR, the morphological analyzer developed by Morphologic Kft. (Prószéky – Tihanyi 1996). These three centres are currently partners in two research projects aiming to develop an information extraction system of short business news and to build a treebank of Hungarian, which give the background to the work reported here.

The paper is structured as follows. First, a brief overview of the peculiarities of Hungarian NPs is given, focusing on the practical problems they present for parsers. This will be followed by the description of the data (section 3) and the basic concept and the annotation used in the CLaRK system (section 4). Section 5 discusses the design principles of the NP grammar, which is presented in some detail in section 6. The paper ends with a discussion of the results and suggestions for future work.

2 Problems of Hungarian Noun Phrase identification[1]

Hungarian is customarily defined as a free word order language. Its basic sentence structure is designed on principles of topic-comment structure in the first place, which obey subtle rules of discourse perspective. Hungarian being a pro-drop language, the subject is often missing and prototypical sentences open with a topic instead. The topic position is optionally followed by the focus in the slot immediately preceding the verb. After the finite verb one finds an array of neutral elements in mostly free order (É. Kiss 1994). The automatic identification of the topic-comment structure is impossible to carry out on written language alone as the phenomenon is inherently intertwined with stress and other prosodic phenomena.

Fortunately, word order within Noun Phrases follows strict rules. NPs are introduced by determiners, which may only be preceded by demonstratives or dative possessors. Determiners can be followed by nu-

[1] This section incorporates contributions to an earlier version of this article by Tibor Szécsényi, which is gratefully acknowledged.

merals, which are followed by adjectives modifying the phrase-final noun head (Kenesei–Vago–Fenyvesi 1998) .

Determiners

Determiners cannot be relied on as anchor points to indicate the boundaries of noun phrases because they can be omitted, particularly when the NP is generic and functions as a kind of verb premodifier. Thus along with [1a] one also finds [1b].

[1a] [Péter] olvas[egy könyv-et].
Peter-NOM read a book-ACC
'Peter is reading a book.'

[1b] [Péter][könyvet] olvas.
Peter-NOM book-ACC read.
'Peter is reading a book.'

Participle structures

In Hungarian there are two participles in active use, which are similar to the present and past participles in English. They are peculiar, however, in that they act as premodifiers and they can freely take any number of their own arguments in front of the NP head. In effect, this means that a participle premodifier can bring in the whole array of verb arguments and, even adjuncts, that would otherwise occur in a full-blown clause. To determine the leftmost boundary of such phrases is often a vexing problem especially because the determiner belonging to the head is often deleted, hence it cannot be used as a safe indicator of phrase boundary. The situation can be compounded by the fact that NPs containing participle premodifiers can be embedded recursively in each other. Such constructions are quite typical of careful, written Hungarian. As an example which is far from unusual consider:

[2] [[[az ország 70 százalékát] [ellenőrzése] alatt tartó Unita] ellen harcoló [angolai kormány]]
the country-NOM 70 percent-POSS-ACC control-POSS-NOM under holding Unita against fight-PrPART Angolan government
'The Angolan government fighting against Unita, which is holding 70 percent of the country under control'

Note how the corresponding English structure contains a mirror image, as it were, and involves a string of post-modifier structures.

Missing noun heads

It is not only the determiner that can be omitted from the NP, in certain cases the noun itself may be absent. If the noun head is identifiable from the context it can be elided. In cases like this the case marking suffix appears on the last constituent of the remaining NP:

[3] [Péter] [a sárga könyv-et] olvas-sa, [Mari] pedig [a piros-at].
Peter-NOM the yellow(-NOM) book-ACC read-3SG Mary-NOM and the red-ACC
Peter read the yellow book and Mary read the red one.

This phenomenon means that practically everything that can serve as premodifier can also function as NP head. The only exception may be the articles. Unfortunately, the definite article is homophonous with the demonstrative pronoun, the indefinite article with the ordinal numeral meaning 'one' both of which can easily take the role of an inflected NP head.

3 Description of the data

The input to the grammar is text that is morphosyntactically annotated using a scaled-down version of the annotation scheme developed for the Hungarian National Corpus (Váradi 2002). For each token (tagged as <w>) its lemma, morphosyntactic description (msd) and a simplified corpus tag (ctag) is given as attributes of <w>.

The quality of parser output depends to a great deal on the granularity of the morphosyntactic annotation and the precision of the disambiguation. As reported in Dienes and Oravecz 2002, the HNC tagging system basically uses the tagset of the HUMOR tool after applying some post-processing (such as stemming and some streamlining of the codes) on its output. The rate of precision of the disambiguation is reported to reach 98 percent.

The texts were taken from a quality weekly economics journal, *Heti Világgazdaság*, which is the Hungarian equivalent of the British magazine *The Economist* both as regards its prestige and its elaborate language style. The choice of this source was required by the information extraction project that focuses on short business news but at the same time it serves as the toughest possible test for the robustness of the NP grammar. This aspect of the data, though not easily amenable to quantitative measures, should be borne in mind when evaluating the precision/recall figures.

At first, the regular grammar operates on a sequence of *msd* tags of words within the scope of a sentence. Figure 1 shows a sample of the annotated input data. The effect of a rule is typically to add an XML annotation over the chunk of data that it covered. The next set of rules operated on the output of the previous one, each adding further and further layers of XML annotation in the process.

```
<?xml version="1.0" encoding="windows-1252"?>
<!-- This Document is created with the Clark System! http://www.bultree-
bank.org -->

<text>
  <div>
<div>
   <p>
    <s>
     <w lemma="új" msd="A.FOK.NOM" ctag="AS_Ac">Újabb</w>
     <w lemma="cég" msd="N.NOM" ctag="NS3NN">cég</w>
     <w lemma="kerül" msd="V.Me3" ctag="VS3PI">került</w>
     <w lemma="csőd" msd="N.ILL" ctag="NS3NX">csődbe</w>
                                    </s>
   </p>
```

Figure 1 A sample of the annotated input data in XML

Figure 1 A sample of the annotated corpus

4 The development environment

The grammar was developed with the help of the CLaRK corpus development system, which is a graphical environment using a full blown XML processor and a finite state engine. The tool offers a whole number of useful facilities for grammar development, described in more detail in Simov 2001, 2002 and elsewhere.

Here, we can only highlight some of the most important features of the system. Each rule of the grammar is defined in terms of a regular expression defining the target pattern together with left- and right-context and the XML markup that will be applied to the expression in case of a hit. For each rule, an Xpath expression is used to define the scope of the rule (using the *apply to* windows). and what part of an XML node will be fed to the regular expression processor (the so called *element value*). For example, most of the time the rules were applied to the msd attribute of the words (or, indeed, at a later stage the msd of constituents). The element value therefore was set to `attribute::msd`. As a unique extension, the system offers the possibility to assign a (user definable) tokenizer to each rule separately, making it possible, for example, to refer to capitalized words or hyphenated words, abbreviations etc. The functionality of XML processor is further enhanced by the use of a facility (called constraints) to manipulate the data in ways that are not easily done otherwise (e.g. automatically copying data, checking the XML tree in terms of the content of their nodes, filling in attributes interactively from a set of dynamically defined alternatives etc. We used the constraints facility to percolate head features upwards the syntactic/XML tree.

5 Design principles of the NP grammar

The difficulties that Hungarian sentence structure presents to parsing are somewhat eased by the extremely rich morphological system, which encodes a lot of syntactic information that serves as useful clues for establishing constituent structure at the sentence level. Many constituents serving as adjuncts to clauses are formed by (extensions of) inflected nouns, which explains the ubiquity of NPs in Hungarian sentences.

As the internal structure of Hungarian NPs is left-branching, the NP head was used as a anchor point of the regular expression for the base NP. In defining the leftmost boundaries, the principle of longest match was used even in the compilation of the noun extensions. The base NP was allowed to contain a premodifier N provided it was in the nominative case. This constraint was imposed on other premodifiers capable of taking case (adjectives and pronouns). Such a rule is a slight oversimplification of the linguistic facts because demonstrative pronouns can, in fact, occur inside NPs in non-nominative case provided they agree with the head. However, case agreement was not implemented in the current version of the grammar.

The possibility for Noun heads to be omitted and their role to be filled by the premodifier standnig closest to them, calls for some heuristics in order to prevent overgeneralization. Accordingly, we adopted a 'depth first' strategy in the definition of NPs. First base NPs with noun heads were assembled and they were extended as far as it was possible. In the process, all premodifiers were associated with heads, thus making them unavailable as possible heads. Next, the premodifiers left out of NPs by earlier rules are identified as heads of noun-less NPs. Even here, it was necessary to introduce two stages in the assignment of heads, first allowing all premodifiers except participles to be heads (rule NP2) and participles are assigned head status only after testing if the new NPs formed with rule NP2 can form larger participle-NPs.

6 The NP grammar

The NP grammar is implemented as a series of cascaded regular grammars (Abney ??), twelve in its current version, applied cyclically fourteen times altogether. The various levels form two main stages: the level of base NP, which is built out of DetP and (simple and coordinated) AdJP chunks and includes a noun or a NamedEntity.

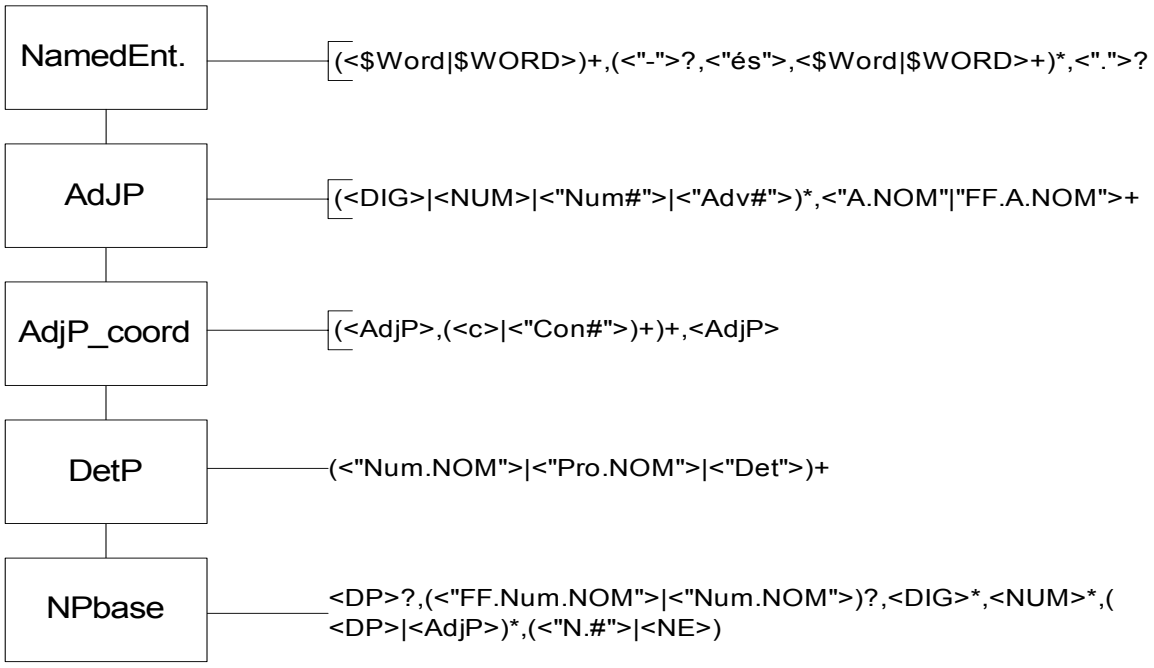

Figure 2. Overview of the grammars producing basic NPs

At present the NamedEntity rule only covers a very elementary approximations of proper names using only the token type $Word or $WORD standing for non-sentence-initial words that begin with an upper-

case letter or words in uppercase. This is admittedly just a placeholder for an in-depth treatment of named entities, which we aim to carry out in the future.

The process of producing basic NPs and the regular expressions used in the rules are illustrated in Figure 2. As an introduction to the notation, concatenation of tokens is indicated with commas, "<AdjP>" refers to a node of the label *AdjP,* <"Adv#"> is a regular expression evaluated on whatever the Element Value setting passes to the Regular Expression evaluator, typically the *msd* attribute of the node in our case. "#" is the wild card character in CLaRK standing for 0 or more accurrence of any character. Concatenation of the tokens in the patterns is signalled with a comma.

The next stage consisted in building maximal extension NPs by combining possessive NPs, coordinated NPs and NPs linked with participles. Possessive and participle-NPs can be chained together, hence they are applied again whenever a larger NP is formed. For reasons detailed in 2) the structure of participle-NPs are obviously the most complex.

After the application of each NP extension rule, the msd attribute of the head was copied onto the node. This was done with the help of the Constraints facility within CLaRK. Unfortunately, while grammars can be grouped together and run at once, constraints cannot be similarly automated, which means the cascade has to be run manually.

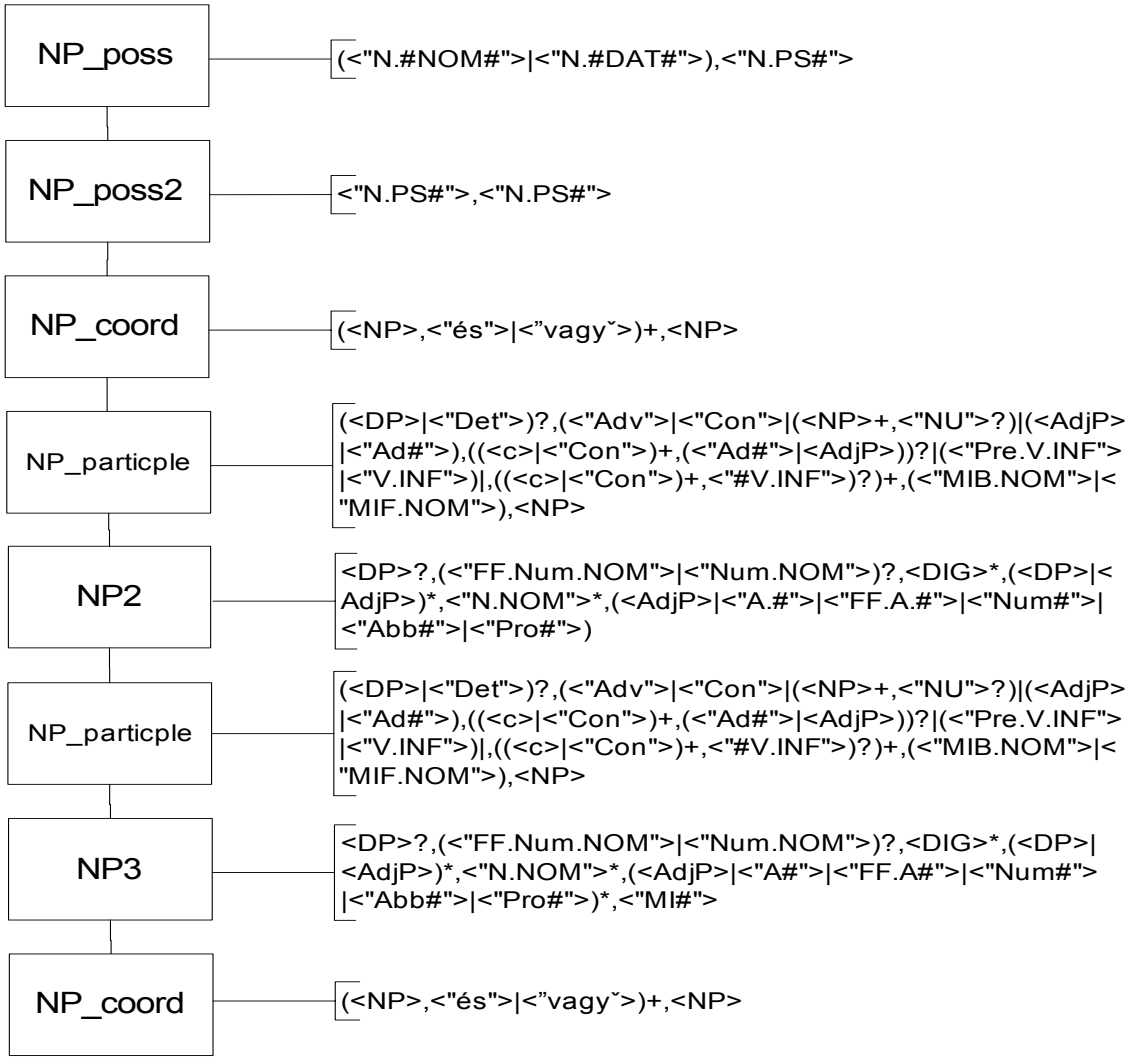

Figure 3. Overview of the grammars producing embedded NPs

7 Results

The grammar was developed on a file of 928 sentences containing 23991 tokens. The NP grammar recognized a total of 10500 NPs, 6807 of which were maximal extensions. At the current stage of our work, we are mostly concerned with the NPs at the sentence level. Therefore the parser will be evaluated in terms of its performance in handling NPs of maximal extension, ignoring their internal structure.

In order to better evaluate the performance of the system, a dual measurement was applied. One reports precision and recall values for the standard structural bracketing, the other focusses on the number of tokens that are analyzed as being under a top level NP chunk. The latter measure is motivated by the fact that the present system is designed as a prepocessor for manual annotation and even if a top level NP is not identified correctly, detected parts of it still serve as valuable information facilitating the annotators' work.

A 100 sentence text chunk containing 2537 tokens served as the test set. A gold standard was produced from this by manual annotation and proof-reading, resulting in 488 top level NPs. Then the test set was processed by the system and evaluation measures were calculated in the standard ways:

(i) structural measures:

- precision: number of common NPs in gold standard and test set / number of NPs in the test set

- recall: number of common NPs in gold standard and test set / number of NPs in gold standard

(ii) per-token measures:

- precision: number of common tokens under (common?) top level NPs in gold standard and test set / number of tokens under top level NPs in test set

- recall: number of common tokens under (common?) top level NPs in gold standard and test set / number of tokens under top level NPs in gold standard

FB1 scores were calculated as usual:

$$FB1=2*prec * recall/(prec + recall)$$

The results for the structural measures are displayed in Table 1, the per-token measures are shown in Table 2.

Number of NPs in gold standard:	488
Number of NPs in test set:	611
Number of correct NPs	323
NP precision:	52.87%
NP recall:	66.17%
Per-word FB1 score:	58.78%

Table 1 Structural measures

Number of tokens in NPs in gold standard:	1660
Number of tokens in NPs in test set:	1577
Number of tokens in NPs in common:	1511
Per-word precision:	95.81%
Per-word recall:	91.02%
Per-word FB1 score:	93.36%

Table 2 Per-token measures

8 Conclusions

Perusing the outcome of the grammar, we could not help finding the results far more encouraging than the numerical indices may otherwise warrant. In order to do justice to the results, the figures dislayed in the tables should be evaluated with the following points kept in mind. 1) The data the grammar is tested on is, without exaggeration, a sample of the most sophisticated written journalism in Hungarian. Even discounting flourishes of style so pervasive in this particular journal, one should note the high number of figures, dates, appellations, parenthetic and appositive materials etc. Thus, the grammar can be realistically expected to stand the test of the most robust applications with at least these figures. To put it in another way, far better results can reasonably be expected on on average run-of the mill text type. 2) Hungarian sentences abound in NPs, owing to the fact that a large number of adjuncts are expressed with inflected nouns. To illustrate, 69% of the total of 2200 words were inside top level NPs. Therefore, once the NPs are tackled, the major part of the parsing of sentences is already done.

There are, obviously, shortcomings in the grammars as well, some were known, some were highlighted during a closer inspection of the results[2]. As mentioned above, the current implementation of the grammar cannot stipulate agreement, which would be important in a few cases even within NPs despite the fact that in general Hungarian does not usually display modifier-head agreement in NPs. In some instances, letting adjectives assume the role of NP heads led to false results.

Inevitably, in a handful of cases, the wrong analysis could be tracked down to the wrong *msd* value. An example would be the occasional fault in disambiguating the past participle/past tense ambiguities. We also found examples where the NP grammar rules served to spot spurious part of speech analyses. The word *szörnyen* monster-adessive 'terribly' is again one of the ubiquitous inflected nouns that function as an adverb. Recall that the NP1 rules allowed no nouns next to the head unless they were in the nominative. However, as a booster expression it can clearly occur inside NPs as in *Szörnyen drága kabát* 'terribly expensive coat'. The fact that our grammar rejects this expression as a single NP indicates that the word *szörnyen* is incorrectly tagged as a noun instead of an adverb. Such cases, and also multi word expressions of a similar type (i.e. spurious nouns while actually adverbs) should be systematically addressed along other named entities. This point leads to the conclusion that a proper syntactic analysis of Hungarian NPs should be done with close integration of some degree of semantic markup.

References

Abney S 1996 Partial Parsing via Finite-State Cascades In *Proceedings of the ESSLLI'96 Robust Parsing Workshop,* pp 1 – 8

Kenesei I, Vago R M, Fenyvesi A 1998 *Hungarian*. Descriptive Grammars. London, Routledge.

É. Kiss K 1994 Sentence structure and word order. In: Kiefer–É. Kiss (eds): *The Syntactic Structure of Hungarian.* San Diego, Academic Press. 1–90.

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*, Las Palmas, pp 710—717

Prószéky G, Tihanyi L 1996 "Humor -- a Morphological System for Corpus Analysis." *Proceedings of the first TELRI Seminar in Tihany.* Budapest, pp 149-58.

Simov K 2001 CLaRK – an XML-based System for Corpora Development in *Proceedings of the Corpus Linguistics 2001 Conference,* Lancaster, pp 553-560.

Simov K et al. 2002 CLaRK System: Construction of Treebanks in *The First Workshop on Treebanks and Linguistics* Theories Sopozol: LML CLPP Bulgarian Academy of Sciences 183-199.

Szabolcsi A 1994 The noun phrase. . In: Kiefer–É. Kiss (eds): *The Syntactic Structure of Hungarian*. San Diego, Academic Press. 179-274.

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

[2] We are indebted to Kata Gábor, Erzsébet Dancsecs and Andrea Littomericzky for their help in checking the results.