Hunpars: A Rule-based Sentence Parser for Hungarian

Anna Babarczy[1], Bálint Gábor[1], Gábor Hamp[2], András Kárpáti[3], András Rung[4], István Szakadát[2]

[1] Department of Cognitive Science, Budapest University of Technology and Economics, Sztoczek u. 2, H-1111 Budapest, Hungary
babarczy@cogsci.bme.hu, bgabor@cogsci.bme.hu

[2] Department of Sociology and Communication, Budapest University of Technology and Economics, Sztoczek u. 2, H-1111 Budapest, Hungary
hamgp@eik.bme.hu, syi@axelero.hu

[3] T-online Hungary Plc, Neumann J. u. 1/B, H-1117 Budapest, Hungary
karpati.andras@t-online.co.hu

[4] Linguistics Institute, MTA-ELTE, Benczúr Gy. u. 33, H-1068 Budapest, Hungary
runga@artitude.hu

Abstract: The paper describes Hunpars, an experimental application for automatically parsing Hungarian natural language sentences. The parser is purely rule-based and has been developed as the third step in natural language processing, following morphological tagging and part-of-speech disambiguation. The project is experimental in that different processing algorithms may be selected for different sections of the grammar.

Keywords: parsing, Hungarian, grammar, syntactic structure, lexical database

1 Introduction

A syntactic parser may be needed for various applications working with natural language data, such as grammar checking, machine translation, manuscript recognition, automatic summarising, question answering systems, etc. The task of the syntactic parser as a tool is to identify sentence structure, that is, to find and unambiguously show formal or semantic dependencies between the words of a sentence. The aim of our project is to develop an open source application that carries out this task fully automatically for Hungarian natural language sentences.

The project started life in 2003 as one of the modules of a Hungarian question answering system[1]. Its goals were later extended and from a simple question-pattern recognition system, the application has now developed into a flexible syntactic parser.

1.1 Overview of Approaches to Syntactic Parsing

The key notion in syntactic analysis is the dependency relations between words or units larger than words in a sentence. There are two main approaches to expressing these dependencies: one is the lexical approach, where syntactic structure is built in terms of link (or transition) types between individual words in the sentence string (e.g., Karlsson et al 1995). A Hungarian example for a parser of this kind is the GeLexi project, where valid lexical link types are defined by a grammar and individual transitions are listed in an extensive lexicon (Alberti et al. 2003). While this method is both theoretically appealing and remarkably accurate for the range of data the lexicon is worked out for, it has the practical drawback of requiring an unreasonable amount of human effort to expand the range. A less labour-intensive method of realising a lexical dependency grammar is statistical grammar acquisition, where the parser learns the probabilities of link types from a hand-parsed corpus. A learning algorithm for Hungarian is currently being developed by Hócza (2004).

The alternative theoretical approach to parsing is phrase structure-grammars, where words are grouped into a hierarchical structure of phrasal constituents and major dependencies are computed with reference to this hierarchical arrangement (for a review of the method see Appelt and Israel 1997). For instance, the sentence in (1) is assigned the bracketed phrasal structure shown in (2):

1 After the show I was left, quite crestfallen, outside the derelict cinema building.

2 [[after [the show]] [I] [[was left], [[quite] crestfallen], [outside [the [derelict] [cinema] building]]]]

Each constituent (bracketed unit) has a *head,* the word that determines the behaviour of the unit within the constituent one level higher in the hierarchy. The phrase *the show,* for instance is a noun phrase (NP) headed by the noun *show,* while the sequence *after the show* is a prepositional phrase (PP) headed by the preposition *after.* Hungarian parsing initiatives in this framework include special purpose applications developed in line with international systems (Váradi 2003, Váradi & Gábor 2004).

[1] "In the web of words", Budapest University of Technology and Economics and T-online project supported by NKFP and OM.

Phrase-structure grammars can be supplemented by lexical dependency information (e.g., Head-Driven Phrase Structure Grammar, Sag & Wasow 1999). In addition to constituency structure, the broad interpretation of the sentence minimally requires a grammar of modifiers (e.g., that the preposition *after* may introduce a freely occurring time adverbial in sentence 1 above) and information on the subcategorisation frames of predicates, i.e., the specifications of the type of word or constituent that can satisfy the argument requirements of the predicate. The verb *leave*, for instance, subcategorises for a subject NP, an optional object NP and an optional locative PP, while the verb *want* would need either two NPs or a subject NP and an infinitival verb phrase complement. A complex grammar of this kind may be defined by a set of generalised and lexical construction templates (e.g., Kálmán et al. 2003) or by a combination of phrase-structure composition rules and lexical dependency databases. Hunpars uses this latter method with some modifications, as will be described in more detail in the next sections.

2 The Principles of Hunpars

2.1 The Major Components

The major components of Hunpars are the following:

- A phrase-structure grammar
- Lexical databases
- A number of search algorithms.

2.1.1 The Phrase-Structure Grammar

The parser has been developed specifically for Hungarian, with two major characteristics of the language in mind: a rich morphological system and variable word and constituent order. In order to make full use of the rich morphology, the parser uses the Hunmorph morphological analyser (Németh et al. 2004) as a resource – the input to the parser is a tokenised sentence with labelled words. The tags supplied by Hunmorph include complete morphological feature-structures of the individual word tokens of the sentence in addition to part-of-speech information. The word form *dolg-ok-nak* (thing-s-to), for instance, has the analysis: dolog<NOUN<PLUR><CAS<DAT>>>, i.e., the plural of the noun root “dolog” in the dative case. The core of the phrase-structure grammar therefore relies less on linear order, and more on morphological tags. As phrasal bracketing proceeds, bracketed constituents inherit the feature structures of their heads or, in some cases, are assigned a feature structure with features of both the head and other member words or constituents.

If one or more words are morphologically ambiguous, a candidate sentence is created for each combination of morphological tags. A large proportion of ambiguities can be resolved with the help of a statistical disambiguator module. As this module has not yet been fully implemented, however, the parser currently incorporates a pre-processing phase, where lexical ambiguities are reduced based on a few rules. Any remaining sentence candidates are filtered by the phrase-structure grammar: if the algorithms fail to converge, the sentence candidate is rejected. In sentence 3, for example, the word *megint* is ambiguous between a third person singular verb (warns) and an adverb (again):

3 A bíró megint minden játékost a második félidőben.
the referee-NOM warns/again every player-ACC the second halfmatch-in.
The referee warns every player in the second half of the match.

With a nominative (subject) NP, an accusative (object) NP and no other verb candidates in the sentence, only the VERB label for *megint* should satisfy the grammar. In other cases, however, morphological ambiguity results in genuine structural ambiguity. The sentence in (4) has two possible parses due to the ambiguity of the word *ki* (who or out) combined with fact that subjects may be dropped in Hungarian. The two interpretations are shown in (5a) and (b).

4 Ki tud menni a kertbe?
who/out can go-INF the garden-into

5 a) Can he/she go out into the garden?
b) Who can go out into the garden?

In case like this one, both candidate sentences are retained.

2.1.2 The Lexical Databases

In addition to feature structures, the grammar makes use of lexical databases. These include subcategorisation specifications for verbs, adjectives (e.g., *egy fiára büszke anya* – a mother proud of her son, but not **fiának büszke* – proud to her son, **fiával büszke* – proud with her son, etc.) and postpositions. Subcategorised complements of adjectives and postpositions are bracketed with the head predicate, as the linear ordering of these constituents relative to the head is syntactically determined in Hungarian:

6 [egy [[fiára] büszke] anya]
[a mother [proud of [her son]]]

The relative linear position of the complements of verbs, on the other hand, is mostly determined by the semantic/pragmatic information structure of the Hungarian sentence, which can only be predicted from discourse and/or physical context. Thus, all the linear arrangements in (7) below are well-formed Hungarian sentences, with the core meaning “hurry(Cathy, home) & cook(Cathy, dinner)”:

7 a) Kati hazasiet vacsorát főzni.
Cathy-NOM home-hurries dinner-ACC cook-INF
b) Kati siet haza vacsorát főzni.
Cathy-NOM hurries home dinner-ACC cook- INF
c) Kati vacsorát főzni siet haza.
Cathy-NOM dinner-ACC cook- INF hurries home.
d) Hazasiet Kati vacsorát főzni.
home-hurries Cathy-NOM dinner-ACC cook- INF

Not all linear orders are grammatical, however. The sequence in (8), for instance, is not a well-formed Hungarian sentence:

8 *Főzni haza Kati vacsorát siet.
cook- INF home Cathy dinner-ACC hurries

Ordering constraints of this kind are used by the parser to reduce morphological ambiguity, but otherwise play no major role in identifying constituent structure. For this reason, the complements of verbs appear as sister constituents of the head directly under the clause:

9 [[A bíró] [megint] [minden játékost]].
[[The referee-NOM] [warns] [every player-ACC]].

The rich morphology of Hungarian allows grammatical relations (subject, object, etc.) to be determined on the basis of morphological features. For the structure above, for instance, the constituent [a bíró] (the referee) inherits the nominative case of the head noun and is therefore identified as the subject of the clause. The constituent [minden játékost] will be linked to the object function through its inherited accusative case. Since the linking between morphological case and grammatical function is to a large extent lexically determined, the linking rules are given in the subcategorisation databases of verbs. The implementation of this component of Hunpars, however, has only been partially realised to date.

A number of other lexical-type databases are also used as resources by the parser. These include a list of permissible combinations of Hungarian verbs and verbal particles. A verbal particle may be attached to the verb as a prefix or may occur as an orthographically independent word preceding or following the verb in linear order, depending on sentence structure. The list is supplemented by a small set of construction templates, representing structures where the verb and the particle may be separated.

2.1.3 The Search Algorithms

The parser is divided into ordered parsing phases. Each phase consists of a series of rule-matching, where the bracketing of a certain type of constituent is carried out. At each phase, the search algorithms scan the input sentence either from left to right or from right to left until a word is found whose morphological tag matches the features of the head of the current constituent type, as specified by the grammar. When a head has been found, the parser opens a bracket and moves one unit at a time to the left or to the right, adding units that satisfy the grammar of the constituent to the head. Once a bracket has been closed, the algorithm searches for the next head in the clause string. The direction of the search is specified by the grammar and may change at any point. When the parser reaches the end of the clause string, it moves to the next phase. At each parsing phase, the constituents bracketed at the previous phases are regarded as single units with feature structures inherited from their member words. Wherever necessary, the grammar directs the parser to database components for lexical dependencies to be satisfied.

The result of a sentence parse can be graphically represented as a parse-tree with the sentence as the top node and the word senses (i.e., morphologically unambiguous word tokens) as the terminal nodes. Figure 1 below shows the tree diagram of the sentence in (10)

10 Az élelem hiánya űzte a különböző közösségeket újabb és újabb területek felkutatására.
the food lack-of-NOM drove the various communities-ACC newer and newer territories exploration-of
Lack of food drove the various communities to explore ever newer territories.

2.2 The Parsing Phases

The parsing phases of Hunpars are the following:

- Pre-processing: simple morphological ambiguity reduction
- Segmenting the sentence into finite clauses
- Bracketing the verbal complex, identifying the clausal predicate
- Adverb phrases
- Coordination 1
- Quantifier phrases
- Coordination 2
- Adjective phrases
- Coordination 3
- Noun phrases
- Coordination 4
- Postpositional phrases
- Coordination 5

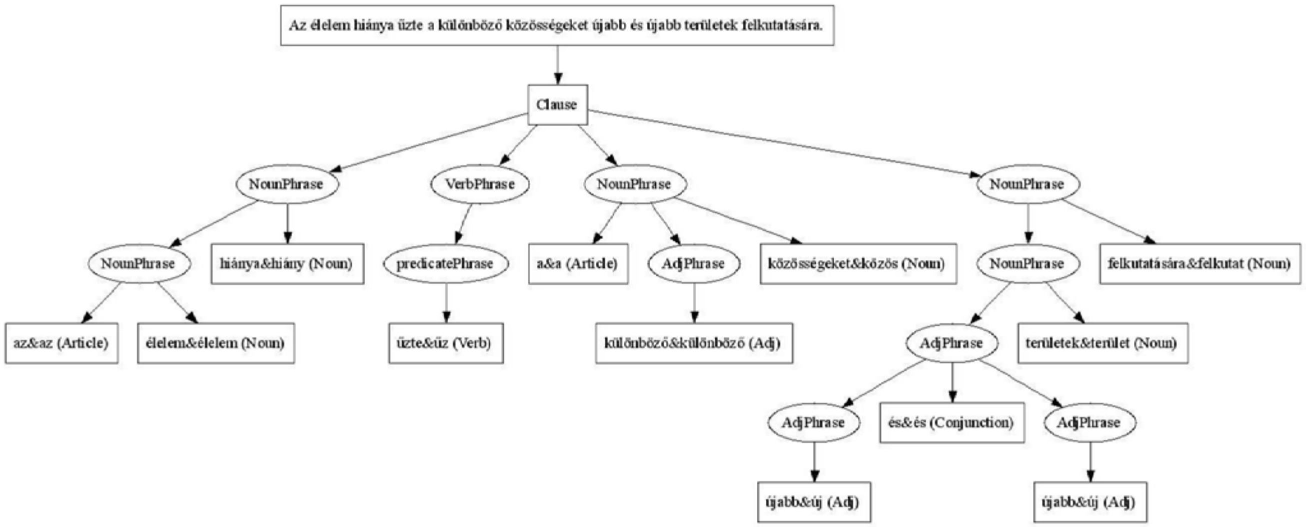

Figure 1

Tree diagram of the sentence

„Az élelem hiánya űzte a különböző közösségeket újabb és újabb területek felkutatására.”

(Lack of food drove the various communities to explore ever newer territories.)

Following the phases of pre-processing the parser slices the sentence at the boundaries of finite clauses. At subsequent phases the clauses are processed separately and are merged back into the sentence again at the end of the parsing process. The first phase of clausal analysis is the bracketing of the verbal complex consisting in the identification of the finite verb, any modal or auxiliary verbs and any verbal particles that may be separated from the verb stem. A particle and a verb will be matched provided that a) the lexical combination occurs in the dictionary of permissible particle-verb sequences and b) the sentence structure fits one of the construction templates with separated particle and verb. Finally, the verbal element with clause level subcategorisation requirements is identified and labelled as the predicate of the clause.

The parsing procedures for the remaining phases follow the basic principles outlined in the previous section. Heads are identified and complements and modifiers are bracketed, building a hierarchically organised phrase structure tree. The phases mainly differ in the feature-sets used by the phrase-structure grammar and in the direction of the search algorithm. After each phase, the coordination algorithm checks the most recently bracketed constituents for syntactic cues to coordination at that level.

3 The Evaluation of Hunpars

3.1 The Beginnings

The first, restricted, version of the parser was completed at the end of 2004. A number of informal tests were carried out at that stage and several shortcomings were identified. Both the grammar and the search algorithms of Hunpars have undergone considerable changes since, partly in response to the development of the morphological tagger.

The major areas of improvement were the following:

- Modification and expansion of lexical databases, adjustment of feature-structure terminology in line with Hunmorph.
- Introduction of the pre-processing phase for lexical disambiguation.
- Implementation of sentence division into finite clauses.
- Introduction of subcategories of parts of speech in the feature structures.
- Improved handling of multiple possessive constructions.
- Redesign of the coordination algorithm.

As well as expanding the data coverage of the parser, the restructuring has also affected previously stable parsing phases. For this reason, an entirely new series of tests have recently been devised.

3.2 Recent Results

As we currently lack a hand-parsed corpus for Hungarian that could be used as a gold standard, the evaluation of the parser cannot presently be automated. Testing was therefore carried out manually. The test corpus consisted of independent Hungarian written sentences selected in a quasi random manner from a variety of sources. The sources included contemporary literature, daily newspapers, popular science texts and texts of legal advice. Sentences including relative clauses were removed from the corpus, as the competence of the parser does not extend to this construction at present. The parser processed the test corpus and for each parse of each input sentence created a tagged text file with the results of the parse. The text files were subsequently processed by the Graphwiz application, which produced graphical parse trees, as shown in Figure 1 above. The trees were manually checked by a linguistically trained tester, who had no knowledge of the details of the grammar or the algorithms implemented for Hunpars.

The quantitative evaluation of the parser was carried out on the basis of the following simple measures:

- The proportion of input sentences that had at least one successful parse.
- The proportion of input sentences that had at least one accurate parse.
- The total number of successful (accurate or inaccurate) parses for the corpus.

In this context successful parse designates any parses where the parsing algorithm converged, i.e., sentence candidates that were not rejected as incorrect by the parser. Accurate parse refers to the smaller set of those successful parses where the bracketing was regarded by the human tester as errorless.

The results of the two most recent quantitative evaluation tests are summarised in Table 1:

Measure	**May 2005**	**Sept 2005**
Number of input sentences	407	309
Total number of successful parses	656	600
Proportion of sentences with at least one successful parse	-	97%
Proportion of sentences with at least one accurate parse	64%	72%
Proportion of input sentences with only erroneous parses due to lexical database gaps or errors (see below)	7%	11%

The error analysis of the results categorised inaccurate parses into the following broad classes:

- Unsuccessful parse due to lexical database error, including words not recognised or miscategorised by the morphological tagger, complements not included in subcategorisation databases and named entities, idioms, etc not listed in any database (11% of input sentences).
- Unsuccessful parse due to grammar error (17% of input sentences).
- Invalid parse due to implementation error. This category covers errors where the output parse fails to conform to the rules of the grammar, presumably due to a hidden incompatibility of parsing algorithms or other programming bug (less than 1% of input sentences).

The results of the error analysis indicate that the performance of the parser could be considerably improved by expanding lexical databases, especially the subcategorisation specifications of postpositions. Further progress could be achieved by improving the coverage of the grammar. A detailed examination of the errors reveals that embedded non-finite (various types of gerundive and participial) clauses pose the most problems. Some examples are shown below with the bracketing showing the target analysis of the erroneously bracketed embedded constituents:

11 A férfi a szóbeszéd szerint [egy [a felesége telefonjában talált] SMS] miatt kezdett gyanakodni házastársára.
The man is said to have become suspicious of his wife because of [a text message [found in his spouse's phone]].

12 A legnagyobb hazai gyorséttermi lánc múlt vasárnaptól [az egyik fizetős wifiszolgáltatóval együttműködve] drótnélküli internettel csalogatja a fizetőképes keresletet.
Since last Sunday, the largest local fast food chain [cooperating with a commercial wifi provider] has been trying to attract solvent demand with wireless internet connection.

13 A vipassana meditáció gyakorlása során a meditáló [[[[a testében megjelenő] pszicho-fizikai jelenségek] természetének] helyes megértésére] törekszik.
In the course of practicing vipassana meditation, the goal to be achieved is [the correct understanding of [the nature of [the psycho-physical phenomena [appearing in the body]]]].

A second problem area is structurally ambiguous sentences where the semantics of the sentence clearly excludes one interpretation for the human processor. Hunpars, however, is not sensitive to semantic information. In the example in (14) the attachment of the locative phrase is incorrect:

14 [Az országos televízió főműsoridőben] legalább húsz perc, országos rádió legalább tizenöt perc önálló hírműsort köteles egybefüggően szolgáltatni.

> [The national TV channel at prime viewing time] must broadcast at least twenty minutes, and the national radio station at least fifteen minutes of uninterrupted news.

The effective filtering of errors of this kind requires a statistical component, which is trained on large pre-tagged corpora. The lack of resources in Hungarian unfortunately does not allow us at present to move in this direction.

References

[1] Alberti G., Kleiber J., Viszket A. (2003): Főnévi GeLExi projekt: GEneratív LEXIkonon alapuló mondatelemzés. [Nominal GeLexi project: parsing based on Generative LEXicon] In: MSZNY 79-84

[2] Appelt, D. E. & Israel D. (1997): ANLP-97 Tutorial: Building information extraction systems. Available as http://www.ai.sri.com/appelt/ie-tutorial

[3] Hócza A. (2004): Teljes mondatszintaxis tanulása és felismerése. [Acquiring and processing a complete sentential syntax] In: MSZNY 127-135

[4] Kálmán L., Balázs L., Erdélyi Szabó M. (2003): Tudásalapú természetesnyelv-feldolgozás. [Knowledge-based natural language processing] In: MSZNY 109-114

[5] Karlsson, F., Voutilainen, A., Heikkila, J., Anttila, A. (1995): *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin

[6] Kis B., Naszódi M., Prószéky G. (2003): Komplex (magyar) szintaktikai elemző rendszer mint beágyazott rendszer. [A complex (Hungarian) syntactic analyses as an embedded system] In: MSZNY 145-152

[7] Németh L., Halácsy P., Kornai A., Trón V. (2004): Nyílt forráskódú morfológiai elemző. [An open-source morphological tagger] In: MSZNY 163-171

[8] Sag, I & Wasow T. (1999): *Syntactic Theory: A Formal Introduction*. CSLI Publications, Stanford

[9] Váradi T. (2003): Főnévi csoportok annotálása a CLaRK rendszerben. [Annotating noun phrases in the CLaRK system] In: MSZNY 65-70

[10] Váradi T., Gábor K. (2004): A magyar Intex fejlesztésről. [On developing a Hungarian Intex] In: MSZNY 3-10