Szeged Treebank 2.0

A Hungarian natural language database with detailed syntactic analysis

Introduction

In Hungarian, like in many other languages, the role of morphemes and syntagmas in sentences and their relation to one another is of key importance. Syntactic analysis and annotation, that is the marking of different syntactic units (e.g. nouns or adjectival phrases, postpositional structures, verbs and their arguments). A treebank representation that describes the syntactic structure of sentences already exists for most Western European languages and a number of Middle and Eastern European languages, so it is time to create a precisely analyzed Hungarian treebank as well.

We relied on known sources and already existing theories when forming the treebank. After studying and comparing them, our linguistic experts developed a consistent syntactic system of rules. The defined syntactic units were marked by an automatic pre-annotating unit on texts of the Szeged Corpus 2.0, then linguistic experts checked and corrected the marked structures. Szeged Treebank 2.0 is based on the first version of the treebank, so it contains its NP and CP annotations.

The database formed I this way forms a reliable basis for the development of different computer applications. The determination of marked syntagmas and their relationship helps further linguistic processing, among others the semantic analysis of texts. We marked syntactic structures on 82 000 sentences (1.2 million word entries + 250 thousand punctuation marks) of the Szeged Corpus 2. 0 file. Treebank files are stored in XML-format, their inner structure is described by TEI P4 DTD (Document Type Definition) scheme.

Texts of Szeged Treebank 2.0

Text files of Szeged Treebank 1.0 correspond to Szeged Corpus 2.0. Texts from six different topics were selected, with each topic containing roughly 200.000 words. The topics are as follows:

- Fiction
- Short essays of 14 to 16 year-old students
- Newspaper articles
- Texts related to computer science
- Legal texts
- Short business and economic news

Further information on the types and sizes of texts is available in the description of Szeged Corpus 2.0.

The formation process of Szeged Treebank 2.0

1. Preliminaries

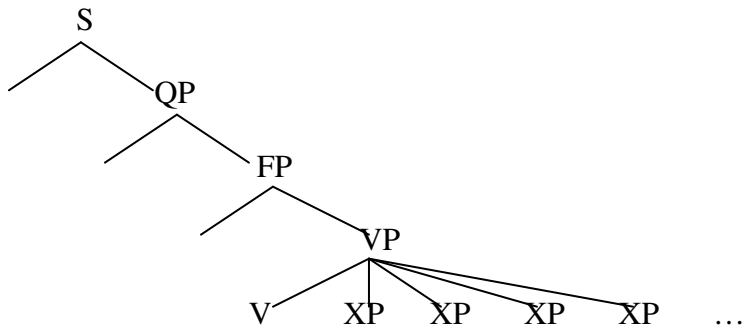
Szeged Treebank 2.0 is based on Szeged Corpus 2.0, which describes the sentences in the following way. Each sentence is surrounded by <s> and </s> XML tags. Within the tags, the entire text of the sentence can be found first, then the words and punctuation marks of the sentence is listed among <w> and </w> and <c> and </c> tags, respectively. Within <w> and </w> XML tags, the actual word form can be found first, which is followed by all possible POS-tags of the word together with their stems. Each <w> tag, that is, word form contains all possible morpho-syntactic (MSD) codes of the word form (together with stems) between <anav> tags. The MSD code selected from possible codes on the basis of the context is always given in <ana> tags together with the stem.
For the partial syntactic annotation of the texts we used internationally accepted
We used the internationally accepted NP (noun phrase) and CP (clausal phrase) tags to label the texts with syntactic tags. It was evident to build the coding of syntactic structure on the basis of strings of <w> and <c> tags, considering them as terminal symbols. The string consisting of <w> and <c> tags within an <s> tag is between <CP> and </CP> tags separated from the text of the sentence. Coordinated and subordinated clauses also received a CP label. Within <CP> tags created in this way, noun phrases had to be determined between <NP> and </NP> tags. During the annotation process, annotators of the group used not only <CP> and <NP> but <XP> tags as well, which were to separate parts of the text not being in close connection with the main body of the text (interpolations between dashes or parentheses, resolution of an abbreviation given in parentheses etc.).This NP, CP and XP tag file was at hand at the beginning of syntactic analysis. For further information on the marking of clauses and noun phrases see the description of the 1.0 version of Szeged Treebank.

2. Linguistic considerations

There has been considerable effort made in the research on Hungarian sentence structure for the formation of a consistent syntactic rule system, however it is still not available in an implementable form. For this reason, taking already existing results and experiences into consideration, we tried to construct such a criteria system for the annotation, which is best adaptable to principles of computer processing.

2.1. Theory

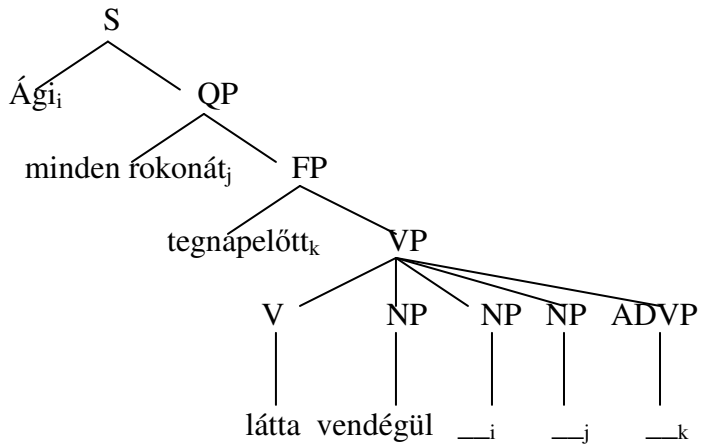
The broad theory chosen is Hungarian generative syntax. The output of syntactic analysis is a (or more) syntactic tree. As a first step, rewriting rules and the lexicon create the initial or deep structure of the sentence. From this, the final or so-called surface structure is realizable through transformations (movements, deletions). In the deep structure of the Hungarian sentence the verb precedes its complements, which follow in an optional order.



Denotations:

S:	sentence
QP:	quantifier position
FP:	focus position
VP:	verb phrase
V:	verb
XP:	optional complement phrase:
	NP (noun phrase)
	ADVP (adverbial phrase)
	PP (postpositional phrase)

Free, that is leafless, branches are left blank in the deep structure, they are so-called functional positions, and components can be moved to these places from behind the verb. The components moved leave so-called traces in their original places. Traces are indexed with the moved components, so movements need not be indicated with arrows. The syntactic tree – neither its deep nor its surface structure – does not show which of the verb arguments are complements. This information can be found in the lexicon. Let us see an example, in which some components are moved from behind the verb:



(Ági hosted each of her relatives the day after tomorrow.)

The analysis can be carried on beyond sentence level, too – the inner structure of sentence-level components is also revealable. This, however, does not influence sentence level analysis, it does not overwrite it.

2.2. Practical realization

The linguists working on the building of Szeged Treebank 2.0 decided not to represent empty categories. Empty categories can be the traces left by moved components, or phonologically not empty pronouns (pro, PRO, or phonologically empty deictic words). Thus, theory is not curtailed since traces are recoverable: components preceding the verb leave their traces behind the verb (in an arbitrary order). Phonologically empty pronouns can also be generated from personal suffixes.

Another important difference from theory is the omission of the representation of functional positions (projections). In some cases their saturation can be concluded from the position of the verb and the verb modifier (verbal prefix or singular common noun without an article), in

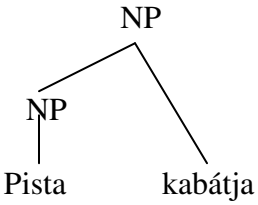
other cases it depends on the prosodic features of live speech, which are not coded in the written texts of the corpus.

Noun phrases are handled uniformly; definite (DP), indefinite (NUMP), and predicative (predNP) noun phrases are not distinguished.

Syntactic trees in the treebank do not appear as trees, but they are realized with labelled parenthesizing for technical reasons and the sake of simplicity only. Labelled parenthesizing was realized in the widely used XML format. Labelled parenthesizing and tree structure are equivalent with each other:

<NP>
    <NP>
        Pista
    </NP>
    kabátja
</NP>

≡



(Pista's coat)

Inventory and short overview of the use of syntactic labels used in the corpus

ADJP: boundary of attributive adjectives

ADVP: boundary of adverbial phrases; adverbial adjectives (gyorsan [quickly], kétségtelenül [undoubtedly]), postpositional personal pronouns (e.g. mögötte [behind him/her], utánatok [after you]), and tokens not belonging to any other category (szervusz [hi], igen [yes])

c: punctuation mark

C0: conjunction

CP: boundary of clauses; also the realisation of the starting symbol of theory, S in the corpus; in the case of subordinate sentences it is the deictic word that is represented as the verb complement, or the CP in case it is missing. (For further information on the marking of CPs see the description of the 1.0 version of Szeged Treebank.)

INF_: boundary of the infinitive and its complement list

- INF0: boundary of the infinitive
- CHILDREN: boundary of the complement list
- NODE: label of the attributes of a given complement

NEG: negative particle

NP: boundary of noun phrases; we considered only movable, noun-headed sentences as noun phrases; predicative (non-attributive) adjectives and inflected personal pronouns (nekem [for me], tőlünk [from us]) are also marked as noun phrases. (For further information on the marking of NPs see the description of the 1.0 version of Szeged Treebank.)

PA_: boundary of the adverbial participle and its complement list

- PA0: boundary of adverbial participle
- CHILDREN: boundary of the complement list

- NODE: label of the attributes of a given complement

PP: boundary of postpositional structures

PREVERB: verbal prefix

V_: boundary of the verb and its complement list

- V0: boundary of the verb; it contains past tense conditional composite verb form in one
- CHILDREN: boundary of the complement list
- NODE: label of the attributes of a given complement

XP: an interjected (not organic) part of the sentence, e.g. interjection in parentheses, between dashes.

List of attributes:

id: automatically generated identifier of uppermost level components

preverb_ref: attribute and value of verb-like labels (V0, INF0, PA0) is the identifier of inseparable verbal prefixes

preverb_body: attribute and value of verb-like labels (V0, INF0, PA0) is the lower-case form of inseparable verbal prefixes

idref: attribute and value of NODE is the identifier of the complement

type: attribute and value of NODE is the type of the complement label

role: attribute and value of NODE is the morpho-syntactic and semantic role of the complement. The *role* type attribute can take the values represented in the chart below:

Case, description	MSD	Inflection, example	role
nominative	n	Ø	NOM
accusative	a	-t	ACC
genitive	g	Ø, -nak/-nek	GEN
dative	d	-nak/-nek	DAT
instrumental	i	-val/-vel	INS
illative	x	-ba/-be	ILL
inessive	2	-ban/-ben	INE
elative	e	-ból/-ből	ELA
allative	t	-hoz/-hez/-höz	ALL
adessive	3	-nál/-nél	ADE
ablative	b	-tól/-től	ABL
sublative	s	-ra/-re	SUB
superessive	p	-n/-on/-en/-ön	SUP
delative	h	-ról/-ről	DEL
terminative	9	-ig	TER
essive	w	-ul/-ül	ESS
(essive-)formal	f	-ként, -képp(en)	FOR
temporalis	m	-kor	TEM
causalis	c	-ért	CAU
sociative	q	-stul/-stül	SOC

factive	y	-vá/-vé	FAC
distributive	u	-nként	DIS
locativus	l	-tt	LOC
place: point of location		there; under the tree	LOCY
place: endpoint		there; under the tree	TO
place: starting point		from there; from under the tree	FROM
time: point of location		today; during the meeting	TLOCY
time: endpoint		next year; till then	TTO
time: starting point		from that time	TFROM
predikatív nominal			PRED
question word		whether	QUE
result		infinitive	GOAL
other		ill, because of him	MODE

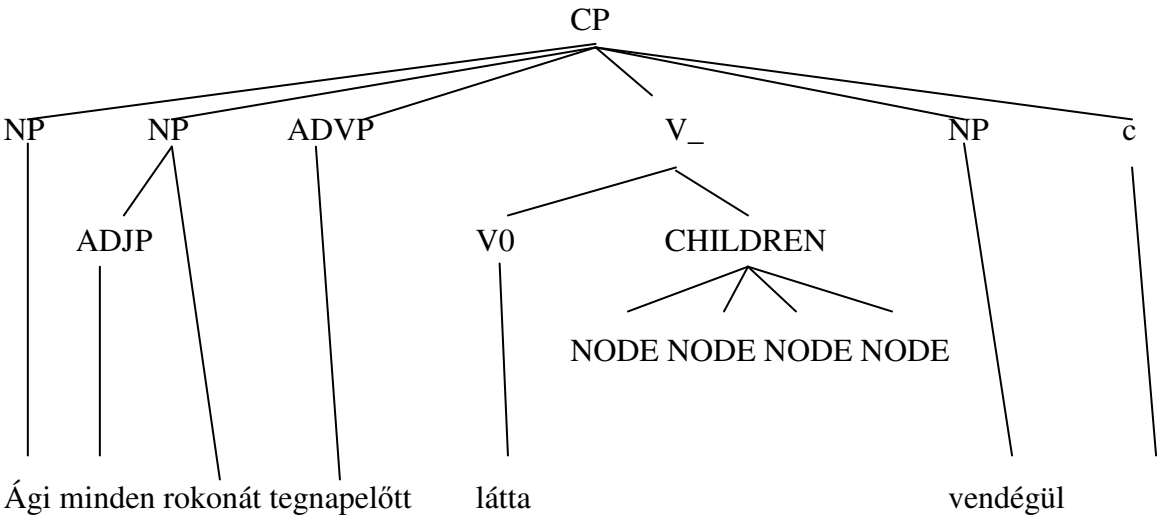
Table 1.: Possible values of the role attribute

Thus, the analysis of the sentence above in the treebank looks as follows:

```
<CP id="fajl.1.1">
        <NP id="fajl.1.2">
            Ági
        </NP>
        <NP id="fajl.1.3">
            <ADJP>
                minden
            </ADJP>
            rokonát
        </NP>
        <ADVP id="fajl.1.4">
            tegnapelőtt
        </ADVP>
        <V_ id="fajl.1.5">
            <V0>
                látta
            </V0>
            <CHILDREN>
                <NODE idref="fajl.1.2" type="NP" role="NOM">
                </NODE>
                <NODE idref="fajl.1.3" type="NP" role="ACC">
                </NODE>
                <NODE idref="fajl.1.4" type="ADVP" role="TLOCY">
                </NODE>
                <NODE idref="fajl.1.6" type="NP" role="ESS">
                </NODE>
            </CHILDREN>
        </V_>
        <NP id="fajl.1.6">
```

```
            vendégül
        </NP>
        <c>
            .
        </c>
</CP>
```

The equivalent tree is:



2.3. Further possible developments

- Indication of the components of nonverbal components
- Indexing members of possessive structures with the inflection -nak/-nek together
- Indexing subordinate clauses and their representative deictic words together
- Classification of NPs (DP, NUMP, predNP)

3. Text processing

3.1. Preliminary analysis of syntactic structures

Since NPs and CPs were already marked in texts deriving from the 1.0 version of the Szeged Treebank, only a preliminary annotation of the remaining syntactic structures was necessary. The syntactic role of words could mostly be given with the help of its MSD code, that is, its morpho-syntactic traits. The use of regular rules defined by experts was not necessary here. For the automatic annotation of syntactic units, we used a self-developed program. Naturally, the program was not expected to work with a hundred percent precision in the definition of the structures, so the control and correction of experts could not be omitted in this phase, either.

3.2. The process of manual annotation

The next step of the process was the control and correction of automatically developed syntactic annotation. To simplify the task, we adapted the earlier developed program – for the annotation of clauses and noun phrases – to the purpose. The window, in which annotators

could not only see the XML structure of sentences but also their syntactic tree structures, can be seen in figure 1 below.

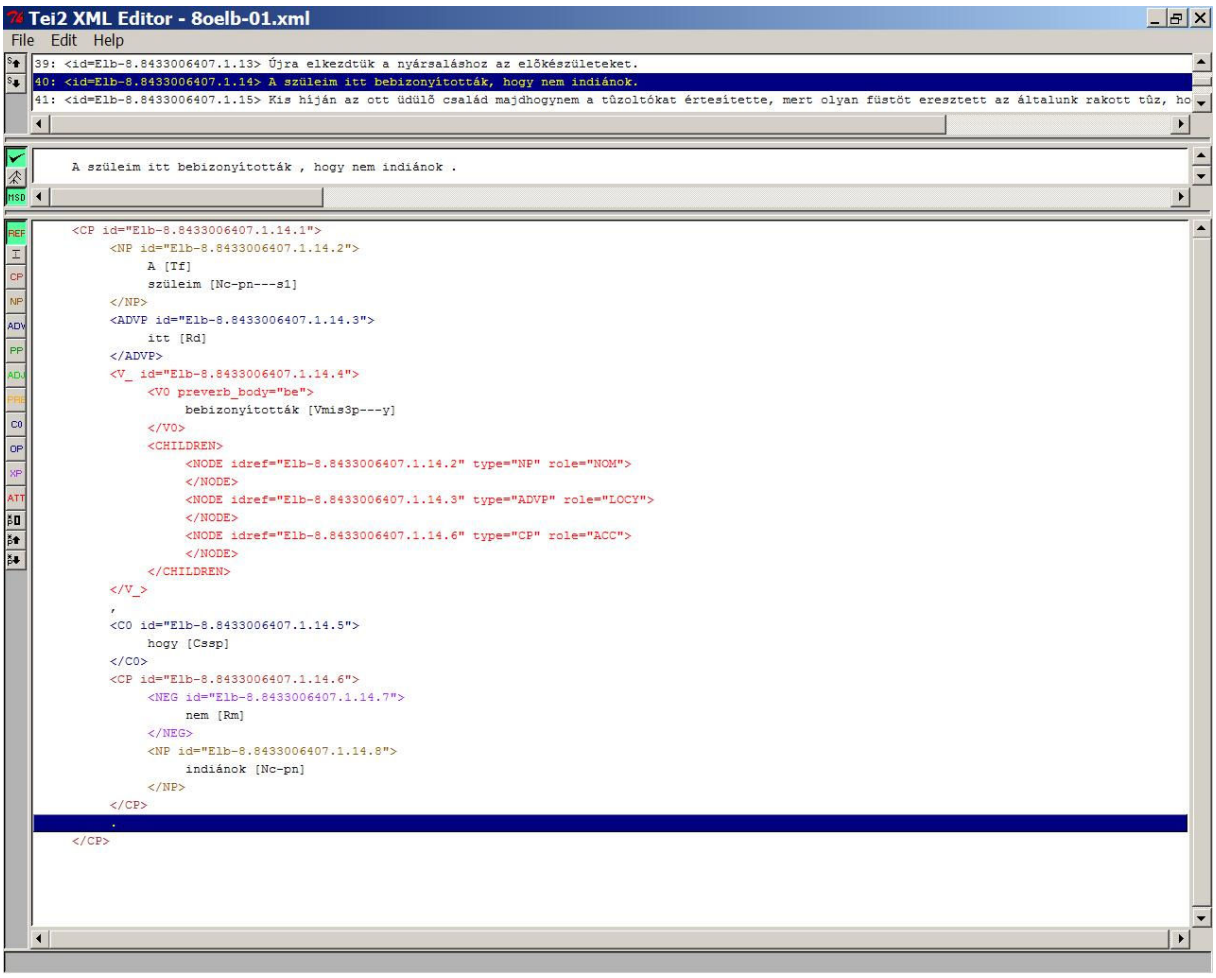


Figure 1. The editing window of the XML Editor

4. Treebank data in numbers

We did statistic measuring on the whole treebank file in order to be able to examine the distribution of different features of the entire syntax tree. The following two tables summarize these results according to topics.

4.1. The depth of the entire syntax tree

The depth of the entire syntax tree is the length of the longest way from the leaf to the root (symbol S), that is, the number of tree levels. The table below refers to the whole syntax tree of all the sentences in the treebank summarized according to topics. The columns of the table comprise the frequency of trees of certain depth. Depth data are represented per level up to a depth of 5 levels, from 6 level on they appear contracted into increasingly larger groups. Distribution is greatest in the case of 4 level, complete syntax trees.

Syntax tree depth								
	1	**2**	**3**	**4**	**5**	**6-7**	**8-10**	**11-20**

Short essays	141	2922	7898	8388	3942	1380	62	0
Legal texts	2	110	687	1554	2127	3346	1337	115
Newspaper articles	29	577	1466	2469	2545	2567	534	24
Business news	0	75	864	2396	2844	2933	455	10
Fiction	493	4649	5230	4170	2373	1495	152	2
Computer technology	9	541	1133	2413	2654	2638	373	7
All	**674**	**8874**	**17278**	**21390**	**16485**	**14359**	**2913**	**158**

Table 2. Distribution of the entire syntax tree depth in treebank sentences

4.2. The width of the entire syntax tree

The width of the whole syntax tree is practically equivalent with the length of the sentences, that is, how many words and punctuation marks there are in the sentence. The table below refers to the whole syntax tree of all the sentences in the treebank summarized according to topics. The columns of the table comprise the frequency of trees of certain width. Width data are represented separately up to 5 words, from 6 words on they appear contracted into increasingly larger groups. Distribution according to width spreads more than that of depth. A syntax tree width of 21-50 is the most common, however a width of above 50 also occurs in every topic.

Syntax tree width										
	1	**2**	**3**	**4**	**5**	**6-7**	**8-10**	**11-20**	**21-50**	**50-**
Short essays	25	126	319	578	1109	2811	4738	11309	3667	51
Legal texts	20	56	60	72	48	147	429	2640	5153	653
Newspaper articles	1	83	97	120	156	438	1000	3693	4401	222
Business news	1	0	2	11	158	114	502	3741	5006	42
Fiction	15	434	1099	1336	1397	2691	3095	5487	2864	146
Computer technology	104	142	108	80	130	266	681	3643	4430	184
All	**166**	**841**	**1685**	**2197**	**2998**	**6467**	**10445**	**30513**	**25521**	**1298**

Table 3. Distribution of the entire syntax tree width in treebank sentences

Creators of Szeged Treebank 2.0

Consortium partners:

- University of Szeged, Department of Informatics, HLT Group
- MorphoLogic Ltd. Budapest
- Research Institute for Linguistics at the Hungarian Academy of Sciences, Department of Corpus Linguistics

Project Leaders

János Csirik	University of Szeged	e-mail: csirik@inf.u-szeged.hu
Tibor Gyimóthy	University of Szeged	e-mail: gyimothy@inf.u-szeged.hu
Gábor Prószéky	MorphoLogic Ltd.	e-mail: proszeky@morphologic.hu

Balázs Kis	MorphoLogic Ltd.	e-mail: kis@morphologic.hu
Tamás Váradi	Research Institute for Linguistics	e-mail: varadi@nytud.hu

Further project members from the University of Szeged

Zoltán Alexin	University of Szeged	e-mail: alexin@inf.u-szeged.hu
Dóra Csendes	University of Szeged	e-mail: dcsendes@inf.u-szeged.hu
Richárd Farkas	University of Szeged	e-mail: rfarkas@inf.u-szeged.hu
András Hócza	University of Szeged	e-mail: hocza@inf.u-szeged.hu
András Kocsor	University of Szeged	e-mail: kocsor@inf.u-szeged.hu
Kornél Kovács	University of Szeged	e-mail: kkornel@inf.u-szeged.hu
György Szarvas	University of Szeged	e-mail: szarvas@inf.u-szeged.hu

Annotators (students of Hungarian linguistics at the University of Szeged)

Anikó Formanek, Kinga Konczer, Ildikó Korpa, Éva Nagy, Krisztián Pálmai, Ágnes Szabó, Bernadett Szőke, Csilla Tóth, Veronika Vincze

Programmers (students of informatics at the University of Szeged)

András Appelshoffer, Tibor Bakota, Csongor Barta, Szabolcs Iván, András Miháczi, Miklós Rácz, György Soponyai