Manually Annotated Hungarian Corpus

Zoltán Alexin
Department of Informatics
University of Szeged
alexin@inf.u-szeged.hu

János Csirik
Department of Informatics
University of Szeged
csirik@inf.u-szeged.hu

Tibor Gyimóthy
Research Group on Artifical
Intelligence at University of Szeged
gyimothy@inf.u-szeged.hu

Károly Bibok
Slavic Institute
University of Szeged
kbibok@lit.u-szeged.hu

Csaba Hatvani
Department of Informatics
University of Szeged
hacso@inf.u-szeged.hu

Gábor Prószéky
MorphoLogic
Budapest
proszeky@morphologic.hu

László Tihanyi
MorphoLogic
Budapest
tihanyi@morphologic.hu

Abstract

Current paper presents the results of a two-year project during which a consortium of the University of Szeged and the MorphoLogic Ltd. Budapest developed a morpho-syntactically parsed and annotated (disambiguated) corpus for Hungarian. For morpho-syntactic encoding, the Hungarian version of MSD (Morpho-Syntactic Description) has been used. The corpus contains texts of five different topic areas: schoolchildren's compositions, fiction, computer-related texts, news, and legal texts. During annotation, linguists have checked the morpho-syntactic parsing of each word. Finding part-of-speech tagging (disambiguation) rules by machine learning algorithms was also studied by the researchers of the consortium. Due to the fact that the size of the corpus reaches up to 1 million text words without punctuation characters, it may serve as a reference source for numerous future research applications. The corpus can be obtained freely via Internet for research and educational purposes.

1 Introduction

The beginning of the work dates back to 1998 when the authors started a research project on the application of ILP (Inductive Logic Programming) learning methods for part-of-speech tagging. This research was done within the framework of a European ESPRIT project (LTR 20237, "ILP2"), where first studies were based on the so-called TELRI corpus (Erjavec et al., 1998). Since the corpus annotation had several deficiencies and its size proved to be small for further research, a national project has been organized with the main goal to create a suitably large training corpus for machine learning applications, primarily for POS (Part-of-speech) tagging.

POS tagging plays a central role in NLP (natural language processing). Hungarian words – similarly to other languages – may have more than one part-of-speech labels (e.g. the word *ég* may be a noun or a verb).[1] In many natural language processing software systems, including web-based dictionaries and optical character recognition programs, determining the part-of-

[1] *Ég* is an ambiguous word in Hungarian, it corresponds either to *sky* (noun) or *to burn* (verb) in English.

speech tag of a particular word in a given context is significant. Syntactic and semantic parsing of natural language sentences are greatly influenced by adequate part-of-speech tagging. In their preliminary studies, the consortium members found that ambiguous words are very frequent in Hungarian language. Hence, developing an annotation (disambiguation) technology proved to be a real necessity.

When choosing the form of representation of the corpus it was taken into consideration that it should comply with international standards. Therefore, the tag encoding system of the annotated Hungarian corpus was based on a technology (MSD) that has already been applied to other – mainly European – languages.

2 Preliminaries

Collecting special text corpora in Hungary has already begun in the eighties. These texts have been thematically grouped, but were not analyzed morpho-syntactically. The development of the morphological parser Humor (High-speed Unification Morphology) began in the early nineties (Prószéky and Kis, 1999). In the framework of the Copernicus project 106 "MULTEXT-EAST" between 1995 and 1997, the participants created an augmented morpho-syntactic coding scheme, called MSD (Erjavec et al., 1997) to be applicable to Central and East European languages. To demonstrate the behavior of this coding technique, a parallel annotated corpus was developed based on Orwell's novel, "1984". Part-of-speech tagging of this corpus was completed manually by linguists. It is widely known as TELRI corpus and published on a CD-ROM (Erjavec et al., 1998). For automatic generation of morpho-syntactic labels for the Hungarian part of the TELRI corpus the above-mentioned Humor system was used.

Unfortunately, the Hungarian part of the TELRI corpus did not implement the whole encoding scheme; more precisely, it did not classify the pronouns, numerals, adverbs, and conjunctions. For example, all pronouns got the same [P] tag, without any attributes encoded. Other attempts for making a Hungarian annotated corpus was not known before the presently described project was started in 2000. A comparison of the manually annotated Hungarian corpus and the Hungarian part of the TELRI corpus can be seen in Table 1.

Manually annotated Hungarian corpus	**TELRI corpus**
Size: 1 million text words (excluding punctuation characters)	Size: 100 000 tokens (including punctuation characters)
Specially selected texts	Single novel (special literary language)
XML technology	SGML technology
Full MSD encoding	Partial implementation of the Hungarian MSD

Table 1. Comparing the main features of the manually annotated Hungarian corpus and the TELRI corpus

Using the TELRI corpus Horváth, Alexin, Gyimóthy, and Wrobel, (1999) investigated the applicability of several machine learning algorithms for learning part-of-speech tagging rules for Hungarian. The manually annotated Hungarian corpus can significantly enlarge the learning database for applying similar methods.

In section 3 the main feature of the corpus is presented, in section 4 more statistical data and two connecting projects is presented. Section 5 summarizes the main achievements of the work.

3 Manually Annotated Hungarian Corpus

Participants of the project aimed not only to increase the amount of corpus text up to 1 million text words, but to improve the quality of the annotation as well. By quality we mean both full conformity to the MSD coding scheme and accurate manual morpho-syntactic parsing and tagging.

The parts of the 1-million-word corpus were selected and put together by the project partners. Naturally, a corpus of this size could not cover the whole written language, but the consortium tried to mainly include most recent texts, well representing the major types available through the Internet, including the special language used by the youth – the primary users of the Internet. Based on this idea, the consortium decided to gather texts belonging to five different topic areas listed below. Parts of the corpus belonging to each topic area contains roughly 200 000 words respectively.

- **Schoolchildren's compositions.** This material was collected from pupils of the age 16 (grade 10). They were asked to write two one-page-long compositions with the titles *The most interesting day of my life* and *Why do/don't I like school?* This type of text caused lot of headaches for the consortium, because it contained many misspelled, mistyped or incorrectly written words – a phenomenon that occurs frequently in Internet texts as well.
- **Fiction.** Three novels were included in the corpus, one of which was the Hungarian translation of Orwell's *1984* and two more Hungarian novels. The first has been completely re-parsed and re-annotated.
- **Computer-related texts.** Some issues of *ComputerWorld Számítástechnika* magazine and three chapters from a book about Windows 2000 were selected.
- **News.** One complete issue each from 1999 of four well-known Hungarian newspapers (*Magyar Hírlap, Népszabadság, Népszava and HVG*)
- **Legal texts.** Two complete Acts (*Act on economic companies, Act on authors' rights*) were included in the corpus.

The developed corpus is available in XML[2] format. The inner structure of the files is described in TEIXLITE DTD (Document Type Definition).[3] This "light" version of the TEI XML DTD is widely used for corpus representations.

The text of the corpus has been divided into divisions between <div> and </div> tags), where one division comprised a single composition, a newspaper article, etc.; paragraphs (marked by <p> and </p> tags); and sentences (between <s> and </s> tags). Each structural element is uniquely identified by an *id* attribute. Text words are marked by <w> and </w>, punctuation characters marked by <c> and </c> tags. Some statistical data can be seen in Table 2.

The next step of processing was morpho-syntactic parsing. Preliminary steps were executed by a segmenting tool and the HuMor morpho-syntactic parser. A lexicon has been built that contained all of the 163 000 different word-forms and a 15 000-word-long list of named entities, mainly proper nouns occurring in the corpus. Either since HuMor could not produce some of the attributes needed for MSD encoding or because the results of this automatic tool were sometimes incorrect, linguists had to manually check the lexicon and create a relatively large list of exceptions. Most of this work was based on the Hungarian Explanatory Dictionary (Juhász, Szőke, Nagy, and Kovalovszky, 1972), however annotators had to rely on their intuition in a large number of neologies. Finally, the whole text was re-parsed using the created exception dictionary.

Tags	Number of tags
<div>	3365
<p>	17 144
<s>	68 932
<w> words	1 009 024
<c> punctuations	203 005

Table 2. Data exhibiting the size of the manually annotated Hungarian corpus

To make the manual annotation easier, a software tool was written. Annotators worked on 400-500-sentence-long pieces of the corpus. Senior linguists and computer programs checked the quality of their work. Producing a POS tagger prototype was among the final goals of the project.

4 Discussion

The development of the first version of the corpus was finished in summer of 2002. Since then two major projects have been started using the described corpus. Each of which aims to add new features to the existing material.

The goal of the first project is to create an information extraction system from short business news. To accomplish this, participants augment the manually annotated Hungarian corpus with a 200 000-word-long part containing short business news. Moreover, a newer version of the corpus is created containing partial syntactic parsing namely, hierarchic NP annotations. During this project, the participants extensively use tools for determining syntax rules and machine learning techniques. The goal of the second project is to

[2] http://www.xml.org
[3] The TEI consortium http://www.tei-c.org is an international organization that elaborates guidelines for computer text representations.

create a complete treebank for Hungarian by the end of 2004.

The distribution of words' main categories occurring in the manually annotated Hungarian corpus is shown in Table 3.

Category	Number of words	
	count	%
Adjectives	130727	10.79%
Conjunctions	86531	7.14%
Interjection	1856	0.15%
Numerals	29802	2.46%
Nouns	281811	23.25%
Pronouns	60833	5.02%
Adverbs	116410	9.60%
Suffixes	16096	1.33%
Articles	129680	10.70%
Verbs	141231	11.65%
Unknown	9605	0.79%
Abbreviation	1370	0.11%
Mistyped	3071	0.25%
All text words	1009023	83.25%
Punctuations	203006	16.75%
All tokens	1212029	100.00%

Table 3. Number of words by main categories of their part-of-speech tags in the manually annotated Hungarian corpus

5 Conclusion

During 2000-2002 the consortium developed the manually annotated Hungarian corpus as well as a part-of-speech tagging method (prototype system and technology), possessing the following characteristics:

- establishment of a medium-size (1 million-word-long) manually annotated Hungarian learning corpus;
- efficient disambiguation of texts belonging to different domains;
- development of an adaptable system that is able to keep track of the changes in the (spoken or written) language;
- a technology applicable to other European languages;
- internationally accepted part-of-speech (morpho-syntactic) classes, augmented with the special attributes of Hungarian language necessary partly because of the highly inflectional character of Hungarian language.

From the scientific point of view, future forthcoming papers dealing with applications, accuracy, combinations, and limits of existing learning algorithms can be of international scientific interest.

6 Internet Availability

The corpus presented in the current paper can be obtained through the following URL address: http://www.inf.u-szeged.hu/lll/szegedcorpus.html. Downloading the corpus via Internet requires preliminary registration. The size of the corpus is 161 MB or 15 MB with WinZip compression.

Acknowledgement

The project was partially supported by the Hungarian Ministry of Education (grant: IKTA 27/2000). The authors also would like to thank researchers of the Research Institute for Linguistics at the Hungarian Academy of Sciences for their kind help and advice.

References

Tomaž Erjavec, and M. Monachini, editors, 1997. *Specification and Notation for Lexicon Encoding*, Copernicus project 106 "MULTEXT-EAST", Work Package WP1 – Task 1.1 Deliverable D1.1F.

Tomaž Erjavec, A. Lawson, and L. Romary, editors, 1998. *TELRI: East meets West — A Compendium of Multilingual Resources* http://www.ids-mannheim.de/telri/cdrom.html

Tamás Horváth, Z. Alexin, T. Gyimóthy, and S. Wrobel, 1999. *Application of Different Learning Methods to Hungarian Part-of-speech Tagging*, in Proceedings of 9th International Workshop on Inductive Logic Programming (ILP99) Bled, Slovenia, in the LNAI series Vol **1634** p. 128–139, Springer Verlag http://www.cs.bris.ac.uk/~ilp99/

József Juhász, I. Szőke, G. Nagy O., and M. Kovalovszky editors, 1972. *Magyar Értelmező Kéziszótár* (Hungarian Explanatory Dictionary) Akadémiai Kiadó, Budapest, Hungary

Gábor Prószéky, and Balázs Kis, 1999. A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* 261–268. College Park, Maryland, USA