SCIENTIFIC PAPER UDC 004.9:811.163.41’374

Annotating the Corpus of Contemporary Serbian[1]

Miloš Utvić [2]
Faculty of Philology,
University of Belgrade,
Department of Library and Information Science

Abstract:

This article describes stages in annotation of the 113 million Corpus of Contemporary Serbian (preparation and implementation). There are several levels of annotation which have been conducted. Corresponding bibliographical information is attached to each corpus text. Part-of-speech (PoS) tagset is prepared, based on the electronic morphological dictionary of Serbian, as well as dictionary of possible annotations adapted for TreeTagger, the PoS tagging system. The Corpus of Contemporary Serbian has been automatically, morphosyntactically annotated with TreeTagger software, i.e. information about part of speech and lemma has been attached to each corpus word form. TreeTagger used manually tagged one million word corpus INTERA as a training set. Ten-fold cross-validation is used for evaluation of applied annotation procedure.

Keywords: annotation, corpus, tagger, TreeTagger

---

1 This paper presents results of research that have been achieved during 2011 supported by the Serbian Ministry of Education and Science under the grant 178006 (Serbian Language and its Resources) and by project CESAR, as a part of a wider network of excellence called META-NET, funded by the European Union.
2 misko@matf.bg.ac.rs

1. Corpus Annotation - Pro et Contra

Corpus, in its broadest sense, is usually defined as a collection of texts. Although corpus linguists distinguish between pre-electronic and electronic corpora, i.e. distinguish between collections of machine-readable texts and other (non-machine-readable) collections of texts, contemporary corpora are almost always electronic. Therefore, in the rest of this article when we refer to a corpus we mean electronic corpus, unless noted otherwise. Corpus linguists produce different definitions of (electronic) corpus, but they mostly agree that corpus is "a collection of machine-readable authentic texts (...) that is sampled to be representative of a particular natural language or language variety" (Xiao, 2010).

Corpus word form is a sequence of (corpus) text characters between two consecutive delimiters. The set of delimiters can be defined in many ways, usually as a set of non-alphanumeric characters. Both corpus words and single delimiters are called tokens. Considering given definition, in the sentence *If a=0, then end of a proof*. corpus words are *If, a, 0, then, end, of, a, proof*, and delimiters are space, comma, full stop and = character, and all of them together are the tokens of the given sentence.

Annotation of corpus is a procedure by which extra information is added to different parts of corpus (texts, logical units inside texts, tokens).

Extra information can be added at several levels:

(1) corresponding bibliographical reference (information about text source), information about text size (number of tokens and types), information about creation and updates of electronic version of a text (creation date, original text version, procedure used for transformation to machine-readable form, persons responsible for electronic version creation, error correction etc.) can be added to a corpus text;

(2) logical structure of text can be encoded (chapters, headings, paragraphs, sentences);
(3) the following information can be added to each corpus word
    (I) part-of-speech (noun, adjective, verb, etc.),
    (II) lemma (nominative singular for a noun, infinitive for a verb etc.),
    (III) values of inflectional categories (gender, number, case, tense, aspect etc.), i.e. inflectional stem and affixes;
    (IV) derivational stem, prefixes, infixes and suffixes;
    (V) pronunciation (accent);
    (VI) syllable boundaries;
(4) a tag for corresponding meaning can be attached to each token;
(5) mutual information can be added to a sequence of one or more corpus words concerning their role in the sentence (subject, predicate, object, verb complement) or that sequence can be encoded as phrase (noun phrase, adjective phrase etc.).
(6) coreference relationships (anaphoric and cataphoric reference) between corpus words can be encoded at the level of discourse;
(7) speech act can be annotated for
    (I) pragmatic and
    (II) stylistic information.

Although most of authors use the term *corpus annotation* for all mentioned information addings to a corpus, some authors (Xiao, 2010) mean by that term only an adding of linguistic information (lemma, part-of-speech etc.), while for other (non-linguistic) information (text bibliographical data, original text formatting, etc.) the term *corpus markup* is used. Also, specified addings (1)-(7) have their own names: *structural annotation/markup* (2), *Part-of- Speech tagging*

or PoS tagging (3I), lemmatization (3II), grammatical annotation (3III), semantic annotation (4), parsing (5), coreference annotation (6), pragmatic annotation (7I), stylistic annotation (7II). Types of annotation (1), (2) and (3I)-(3III) are are already widespread, the type (5) is getting close, while the other types of annotation are still rarely present.

The main goal of corpus annotation is making the corpus search more efficient. In case where the information about a lemma is present, corpus user can find, for example, all word forms of a given noun just specifying its canonical form – nominative singular. Morphosyntactic description (MSD) of token enables precise specification of syntactic constructions that corpus user wants to find (e.g. all adjective phrases of the form "adverb followed by adjective", like *very quick*). In case where a lexicographer is interested in some specific meaning of the polysemous lexeme (e.g. Serbian noun *čas* with meanings *clock* and *lesson*) or lexeme which has the same spelling as some other lexeme(s) (e.g. Serbian word form *kosi* can be a word form of a verb *kositi* "mow" or word form of noun *kosa* "hair"), and corpus is annotated with meanings, lexicographer will easily filter corpus search results, i.e. eliminate the "noise" produced by homography and polisemy.

Since corpus search results are only language samples extracted from a wider context, data about original text version are necessary in order to compensate information lost with wider context.

Corpus annotation is done as part of the text preprocessing phase. During text preprocessing, elimination of non-textual elements (figures, tables etc.) requires annotation text explaining the location and type of omitted text units. Similiarly, it's necessary to annotate paralinguistic features (laughter, pause, intonation, voice modulation, etc.) in transcribed spoken texts. Also, the only way to include comments of text editors inside edited corpus text is to place them as a part of corpus annotation (Xiao, 2010).

Corpus annotation simplifies statistical analysis of corpus, i.e. it can be used to automatically determine the distribution of linguistic features. One example is creation of frequency lists. *Frequency list* is a list of corpus words and their number of occurrences. In case when there is no information about lemma in corpus, then there is no possibility to count frequencies of lemmas automatically (e.g. frequency of Serbian noun *kuća* "house"), but the number of occurrences is counted separately for each word form of lemma (*kuća, kuće, kući* etc.). If there is no information about part of speech, it's impossible to determine frequency of nouns and other parts of speech without manual counting. Also, without information about meaning of corpus word one cannot calculate frequencies of all possible meanings for particular corpus words. Same goes for other unannotated linguistic information.

Besides mentioned advantages, there are annotation misadvantages which cause that certain corpus linguists oppose to the idea. This is one of the key differences between corpus-based linguists and corpus-driven linguists. The first ones assume that the role of corpus is to test, correct and complement existing theories, as well as to find examples which confirm those theories. This assumption is the main reason why they insist on detailed corpus annotation. Corpus-driven linguists argue that an approach to corpus should be without any preconceived theories in order to postulate linguistic categories based only on data itself. Therefore, latter approach finds corpus annotation unnecessary, because annotation is actually one particular corpus analysis (done by annotators), so if one analyses annotated corpus one will only get repeated results of someone else's previous corpus analysis (Lindquist, 2009, стр. 45).

Implementation of corpus annotation itself is more important thing than this theoretical discussion. Namely, annotation often requires hard work and significant human resources, not only numerous, but also familiar with disciplines of linguistics (morphology, syntax, semantics etc.). Although there are software tools which are capable of annotating corpus with precision of 95-97% (Brants, 2005), many of them require previously prepared manually annotated corpus which would be used as the training set for annotation of arbitrary text.

Two most prevalent difficulties following the procedure of automatic annotation are disambiguation and annotation of "unknown" words. The first difficulty emerges when there are two or more mutually exclusive information which can be associated with a token (e.g. information about lemma and part-of-speech for Serbian token more can be noun *more* "sea" and verb *moriti* "torment"). Human disambiguates using several mechanisms based on different sources of information (context of a sentence or a wider logical unit of text, text domain, common knowledge, etc.). Software tools for automatical annotation have access to a limited amount of information in regard to a human, and they are still not capable to link information and draw conclusions from them with efficiency which is close enough to human.

The second difficulty is represented by "unknown words", tokens which annotation tool hasn't "met" during its training, i.e. it has no available information about them. Tipical example of "unknown" words are hapax legomena (transliteration of Greek απαξ λεγομενον), words that one coined (using some known derivational mechanisms) and used once.

Considering that annotation results are usually input data for further analysis (e.g. syntactic analyzer uses results of morphological annotation), "unknown" words cannot be ignored, but annotation program has to use some heuristics which will attach necessary information even to them. Also, in case when annotation is just one of the first steps of processing, disambiguation can be postponed until further steps, especially if further steps have access to additional knowledge which will make a decision of choosing the "right" annotation easier. Then, instead of a single annotation, a set of all possible or several most probable annotations are attached to a token, and disambiguation is postponed (Guengoer, 2010).

In this article, we shall mainly focus on the part-of-speech tagging and lemmatization where each word form is attached to exactly one lemma and one part-of-speech.

2. Corpus Annotation Standards

There is no corpus annotation standard which is unanimously accepted and applied. We shall consider some most commonly accepted and widely adopted „unofficial standards".

2.1. TEI

A need for standardization of encoding and annotation of machine-readable text has been recognized since the beginning of text processing by computer. The first attempt to establish such a standard which was widely accepted by the users and practically applied, are the Text Encoding Initiative Guidelines (TEI Guidelines for short). The first proposal (TEI P1) emerged in 1990., and the current fifth proposal (TEI P5) – in 2007. (TEI, 2009). TEI comes from academic community and at first it was maintained by ACH (Association of Computer in the Humanities), ALLC (Association of Literary and Linguistic Computing) and ACL (Association for Computational

Linguistics), and during 1999/2000. special non-profit consortium (TEI Consortium) was established in order to develop, maintain and promote TEI.

TEI Guidelines specify the set of tags which can be embedded in electronic representation of text in order to markup the text structure and other features of interest (text bibliographical information, information about linguistic elements of text, etc.). The first versions of TEI used SGML as the markup language, while versions published after 2002. (TEI P4 and later) are defined using XML.

TEI Guidelines try to deal with annotation of as many kinds of electronic text. That is the reason why the Guidelines are very extensive (more than one and a half thousand pages of TEI P5 proposal are used to explain the usage of almost five hundred tags), but also too general for specific tasks (e.g. corpus linguistics). Because of this generality, annotation proposed by TEI is organized like hierarchy of modules with inheritance of elements and attributes, which enables users to adapt annotation to their needs by modification – adding, deleting, renaming etc. – of elements and attributes names, updating element content model or changing values of attributes.

Some of participants in creation of TEI annotation were also involved in annotation of British National Corpus (BNC), and that is surely the most significant example of usage of this annotation in corpus linguistics.

However, generality and extensiveness of TEI annotation affected many corpus researchers to decide not to use TEI annotation. Moreover, compliance with TEI Guidelines doesn't mean that annotation is applied consistently, especially if Guidelines offer different ways to annotate the same phenomena.

As an attempt to bring TEI Guidelines closer to users, and to expand their usage, a standardized subset of most frequent or most important TEI elements has been extracted under name TEI-Lite („lite version of TEI").

2.2. CES/XCES

One of TEI markup language derivatives, adapted to the needs of corpus linguistics, is Corpus Encoding Standard (CES). This TEI-compliant standard was published in 1996. by Expert Advisory Groups on Language Engineering Standards (EAGLES), as a part of their own guidelines. EAGLES was formed by the European Community (EC) in 1993. in order to develop standards based on existing practice of encoding and annotation for the official languages in EU, which would be used in future projects supported by the EU.

As well as TEI, CES was defined as SGML application at first (Ide, 1998), but after the development of XML technologies. XCES emerged (Ide et al, 2000), as an XML version of CES (definition).

CES and XCES use the subset of TEI elements whose meaning is more precisely described, and whose content model is reduced. Special attention was paid to the morphosyntactic annotation.

Important characteristic of XCES standard is the possibility to keep text and its annotation in separate files (stand-off annotation) and to link them with pointers (XPointer). Separation enables that multiple different annotations in separate files, which don't have to be used in the same time, can be attached to the same text. Thus, time complexity of natural language processing (NLP) applications is reduced by excluding layers of annotation which are of no interest at a given time. Pointers are especially significant for creation of parallel corpora because they are used to link corresponding

translation units of source and target texts. Annotation layering solves the problem of different overlapping annotations which is typical when text and its annotation are kept in the same file.

2.3. MULTEXT-East

On of the first applications of CES is the annotation of multilingual corpus JOC (Official Journal of European Community), which was created during the series of projects known as *Multilingual Tools and Corpora or MULTEXT* (Ide and Veronis, 1994). The goals of MULTEXT projects were to develop standards and specifications for the encoding and processing of linguistic corpora, and to develop tools and resources which would apply these standards. MULTEXT resources comprise texts in five Western European languages (English, French, Spanish, Italian and German). Texts are aligned at the sentence level, and corpus words are part-of-speech tagged.

Extension of resources, tools, metodologies and experiences of MULTEXT projects to Central and Eastern European languages became the goal of MULTEXT-East project (Erjavec, 2010). Although project officialy ran from 1995. to 1997, resulting resources, specifications and tools were published several times until today, in 1998, 2002, 2004. and 2010, each time corrected and extended with resources of additional languages, including Serbian (Krstev et al, 2004).

Morphosyntactic annotation applied in MULTEXT-East project (Table 1) is a positional annotation, i.e. each position in description represents one attribute. Values of attributes are marked with letters and digits, and a special character (-) is used to indicate the abscence of attribute for a given token. The same mark can be used at several positions, whereby the position determines the meaning of the mark. For example, marks in the description Afpmsnn mean that this is an adjective (A), qualificative adjective (f), positive adjective (p), masculine adjective (m), singular (s), nominative case (n), indefinite (n).

токен	морфосинт. опис
Bio	Vmps-sman-n---p
je	Va-p3s-an-y---p
vedar	Afpmsnn
i	C-s
hladan	Afpmsnn
aprilski	Aopmpn
dan	Ncmsn—n

Table 1 MULTEXT-East (morphosyntactic annotation applied to the beginning of the first sentence in the Serbian version of Orwell's novel 1984)

The specification of MULTEXT-East annotation is one of the candidates for annotation of corpus comprising Serbian texts. Unfortunately, this proposed standard also has its deficiencies. The principles taken into consideration when particular attributes and values are included in the Multext-East morphosyntactic descriptions (MSD) are not always clear and consistent. Also, important information, like those that determine the complex agreement conditions in Serbian, cannot be expressed within the Multext-East MSD (Vitas et al, 2007).

3. Corpus of Contemporary Serbian

Corpus of Contemporary Serbian (SrpKor) is available online since 2002, at http://www.korpus.matf.bg.ac.yu at first, and now at http://www.korpus.matf.bg.ac.rs. The first version of SrpKor represented the 22 million words collection of unannotated texts, without information about source texts (Krstev and Vitas, 2005). SrpKor has gradually changed its appearance during the first decade of its existence. First of all, corpus was supplemented with source information, and users were enabled to have insight to bibliographical references of texts from which corpus search extracted the concordances.

Necessity of extending SrpKor dictates its further development into two directions. One direction is to gradually expand corpus in a way that keeps the balance between particular functional styles and registers represented in corpus texts. The second direction is to create a large opportunistic corpus of Serbian of at least 100 million words. Both directions have a goal to produce not only bibliographical information about corpus texts, but also part-of-speech tagging and lemmatization as additional annotation.

New 113 million words version of SrKor was produced in July 2011. Bibliographical information is provided for all corpus texts. Besides, texts are differentiated with respect to functional styles (literature, scientific, publicistic, administrative and remainder). More information about this corpus and how to access it can be found at http://www.meta-net.eu/meta-share.

In the rest of this article we'll present the recent results of SrpKor annotation.

4. Part-of-Speech Tagging and Lemmatization

As already mentioned in section 1, part-of-speech tagging (PoS tagging) is a process of assigning a part-of-speech or other syntactic class marker to each corpus token. This kind of annotation is also applied to punctuation, which is usually tagged with a mutual marker or special tags can be used for those punctuation marks which are of interest, depending on the corpus purpose.

There are certain requirements in order to begin with PoS tagging:

- Precisely defined set of markers or tags which will be attached to single tokens (tagset).
- Choice of the software (PoS-tagger) which will be used for the automatic annotation.
- Preparation of auxiliary resources required by PoS tagger, mostly manually annotated corpus which is used for training the tagger, and optional lexicon containing all possible tags for a particular word form.

4.1. Tagset

Instead of inventing a tagset from scratch, we decided to adapt existing morphosyntactic descriptions, which can be found in electronic morphological dictionary of Serbian (Krstev and Vitas, 2005). The format applied in the dictionary is known as DELA and it was originally presented for French in (Courtois and Silberztein, 1990). LADL/DELA format (Table 2) enables, not only information about word form itself, but also a description of lemma, part-of-speech, inflection categories (gender, case, number etc.). It also includes syntactic and semantic markers which can be used to retrieve information about name entities, pronunciation (dialect), derivation type (derivation of diminutive, possessive/relational adjective, gender motion), semantic roles (agent, instrument, etc.).

During the selection of tagset it is necessary to make balance between size of the tagset, i.e. extent of information provided by tags, on the one hand, and influence of the ambiguity to the annotation precision, on the other hand. The richer tagset provides more information, but it makes the task of precise tagging heavier, and vice versa.

Example of one entry in morphological electronic dictionary of Serbian		
korisnikovog,korisnikov.A+Hum+Pos+Der:adms4v		
Explanation of morphosyntactic codes		
korisnik*ovog*	word form (element of the dictionary)	
korisnik*ov*	lemma (canonical form of the word form)	
A	part-of-speech (**adjective**)	
+Hum+Pos+Der	syntactic and semantic markers	
	+Hum	**Human**
	+Pos	**possessive adjective**
	+Der	**derived form**
:adms4v	inflection categories	
	a	**positive** (degree)
	d	**definite** (definiteness)
	m	**masculine** (gender)
	s	**singular** (number)
	4	**accusative** (case)
	v	**animate**/ non-animate

Table 2 LADL/DELA format

Considering that one of the main tagging objectives is obtaining the highest possible precision, the basic set of only 16 tags is used, which include Serbian parts-of-speech, as well as some specific tokens which require special treatment (Roman numerals, abbreviations, prefixes, suffixes):

1. N (Noun)
2. A (Adjective)
3. V (Verb)
4. PRO (Pronoun)
5. NUM (Numeral)
6. PREP (Preposition)
7. CONJ (Conjunction)
8. INT (Interjection)
9. PAR (Particle)
10. ADV (Adverb)
11. PREF (Prefix)
12. ABB (Abbreviation)
13. RN (Roman numeral)
14. PUNCT (Punctuation)
15. SENT (Sentence end marker)
16. ? (A tag for everything else: foreign words in text, suffixes like *ux* in 1990-*ux* "nineties" etc.).

4.2. Tagging Tools

The way to adapt morphosyntactic descriptions in electronic dictionary of Serbian depends on a chosen PoS tagger. Considering the results presented in (Popović, 2010) and functionalities described in (Paumier, 2008), we decided to test three tools: Unitex (Paumier, 2008), TnT (Brants, 2000) and TreeTagger (Schmid, 1994).

Unitex is a corpus processing system based on technology of finite state automata and recursive transition networks. It uses lexical resources (electronic dictionaries and grammars) to process texts and create corpus. Unitex corpus can be searched not only by using powerful regular expressions (which include all morphosyntactic categories available in dictionaries), but also by applying complex graphs which can describe both morphological and syntax phenomena.

Since Unitex dictionaries use LADL/DELA format, Unitex was a natural first choice for a tagger. One of the most important results of processing corpus with Unitex is a text automaton. The text automaton consists of segments (usually sentences) and represents all possible lexical interpretations of the words in each segment. Since an equivalent graph exists for each automaton, every path in a sentence graph (from the beginning to the end of a sentence, or generally a segment) describes one possible tagging of a sentence. The essential problem in Unitex is handling ambiguity. Beside manual elimination of offered „false" suggestions, a formalism called

ELAG (Laporte, 1998) is one available solution. Basically, ELAG uses manually created tagging rules (again, implemented as automata) which resolve ambiguity by specifying an allowed or forbidden context of the corpus word which has been tagged in a certain way.

The important advantage of Unitex is that tagging process attaches to a token only those lexical interpretations which exist in a dictionary. However, it requires a great effort to create all the necessary ELAG grammars in order to achieve at least the same precision as the statistical taggers like TnT and TreeTagger. Also, ELAG grammars for Serbian are still at early stage of development, so testing Unitex as a PoS tagger is left for the future work.

Evaluation described in (Popović, 2008) and based on 10-fold cross-validation shows that TnT and TreeTagger give similar results on words included in the training set (TnT: 93.86%, TreeTagger: 91.78%), but TnT is better in tagging "unknown" (or better said, unrecognized) words (TnT: 58.36%, TreeTagger: 36.71%). The used tagset consists of 908 tags – morphosyntactic descriptions for Serbian defined in MULTEXT-East project specifications (http://nl.ijs.si/me/V3/msd/html/). The size of tested corpus is approximately 105K corpus words (18K word types, i.e. different corpus words, and 7.6K lemmas).

Considering obtained results for the tagging of "unknown" words and the need that annotated corpus contains information about lemma, further choice has been narrowed to the TreeTagger.

4.3. Auxiliary Resources

Both statistical taggers, TnT and TreeTagger, require a training set, while TreeTagger also requires so-called full lexicon.

Corpus INTERA was used as the training set. This corpus was named after the project (Gavrilidou, 2006) which produced SELFEH (*Serbian-English Law Finance Education and Health*), parallel English-Serbian corpus of texts pertaining to finance, health, law and education (http://www.korpus.matf.bg.ac.rs/prezentacija/selfeh.html). SELFEH was developed by Human Language Technology group at the Faculty of Mathematics, University of Belgrade, and it comprises 150 TMX documents. Serbian version of this corpus (INTERA) was adapted to format required by TreeTagger, and it contains information about part-of-speech and lemma for each of 1,100,281 tokens. Tagset has 16 tags listed in section 4.1. INTERA has 907,633 corpus words and 55,488 corpus types.

Number of corpus words tagged as '?' ("unknown words") is 4404, which is less than 0.5% of total number of words in the training set.

The full lexicon for TreeTagger is a text file whose lines contains, as columns, one word form and its possible tags. Columns are separated by tabs. Every possible tag is an ordered pair (part-of-speech, lemma) whose components are separated by space Table 3 Excerpt containing two tokens (corpus words) from the full lexicon.

token	MSD.	MSD	MSD.	MSD.
bacili	N bacili	V baciti		
vrelo	Nvrelo	V vreti	Avrelo	ADV vrelo

Table 3 Excerpt containing two tokens (corpus words) from the full lexicon

On of the more difficult problems during TreeTagger training was to create a full lexicon from existing morphological electronic dictionary in LADL/DELA format. We need to point here that TreeTagger isn't a "true" lemmatizer, but its work is reduced to choosing the most likely part-of-speech tag, afterwards tagger simply concatenates lemma from full lexicon, which corresponds

to the chosen part-of-speech. Hence, word forms with the same part-of-speech, but different lemma cannot coexist in the full lexicon (Table 4).

Token	Word type lema	Word type lema	Word type lema
kapi	*N kap*	*N kapa*	*N kapo*
Brankom	*N Branko*	*N Branka*	
donesen	*V doneti*	*V donijeti*	

Table 4 Prohibited entries in the full lexicon

This limitation had a great impact on distinguishing the homographs whose lemmas represent different dialects or similar personal names with different gender. Also, homographs whose lemmas are homographs too, cannot be distinguished. Instead of defining new tags which would be clones of the existing ones, the decision was made to eliminate all (token, part-of-speech) duplicates. The used criteria were to leave Ekavian pronunciation and masculine proper names, while all homographs whose lemmas are also homographs have a unique common representation in the full lexicon.

Ambiguity still exists in full lexicon after elimination of (*token, part-of-speech*) duplicates (Table 5). The most common case of ambiguity is possibility to tag token as an adjective or as a verb (84.77%).

5. Evaluation

Evaluation of applied annotation is based on 10-fold cross-validation of the training set. During test unannotated version of training set is partitioned to 10 complementary subsets, tagger uses nine subsets for training and then annotates the remaining tenth of corpus. Choice of the tenth to be annotated unambiguously determines the remaining nine tenths of corpus which are to be used as the training set. In this way, tagger can be trained on nine tenths of corpus and automatically annotate remaining tenth for each partition, 10 times, and each time results of manual and automatic annotation can be compared. Results of comparison are arrays, each having 10 elements (for each tenth of corpus), which represent, respectively, the size of the tenth (total number of tokens), and number of tenth of the tokens which were annotated manualy and automatically in the same way. All tokens whose manualy produced tags differ from their automatically produced tags, and they are not present in the full lexicon, are treated as unknown words.

relative frequency	tag sequences		
84,77%	A	V	
7,02%	N	V	
2,96%	A	N	
2,75%	A	ADV	
0,32%	?	N	
0,31%	A	N	V
0,29%	A	ADV	V
0,23%	ADV	N	

Table 5 Ambiguity distribution in the full lexicon

Based on the obtained results, we calculated precision of each single annotation, ratio of unknown words per total number of mismatches between manual and automatic annotation, and afterwards the minimum, maximum and average of these values, as well as the variance and standard deviation (Table 6).

	precision	„unknown" words
min.	95,38%	7,68%
max.	96,98%	16,40%
average	96,57%	11,93%
standard deviation	0,43%	2,20%

Table 6 Results of the evaluation

Precision of particular annotation (rj) was calculated as the fraction of the total number of tokens that were annotated the same way manually and automatically and the total number of tokens in the tenth which was annotated. Average precision and variance var were calculated using following formulas

$$\bar{r} = \frac{1}{\mathbb{0}} \sum_{j=1}^{\mathbb{0}} r_j$$

$$var = \frac{1}{\mathbb{0}} \sum_{j=1}^{\mathbb{0}} \left(\bar{r} - r_j\right)^2$$

6. Conclusion

Evaluation results show that precision of tagging is within the limits of the accuracy of tagging systems with similar tagset size. Training corpus INTERA was used for annotation of new version of SrpKor which will, beside bibliographical information about texts, include information about part-of-speech and lemma for each token. New version of SrpKor will also use a new user interface that will be presented during 2011/12 as one of the results within the CESAR project (http://www.meta-net.eu/projects/cesar/), which is a part of the wider network of projects called META-NET.

References

Brants, Thorsten. 2000. TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000, Seattle, WA.

Brants, Thorsten. 2005. Part-of-speech tagging. In The Encyclopedia of Language and Linguistics. (Second ed.), Volume 1-14, Ed. Brown, K., 221-230. Oxford: Elsevier.
Courtois, Blandine and Max Silberztein. 1990. Dictionnaires électroniques du français. Paris: Larousse.
Erjavec, Tomaž. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: Proceedings of the LREC 2010, Malta, 19-21 May 2010.

Gavrilidou, Maria, Penny Labropoulou, Stelios Piperidis, Voula Giouli, Nicoletta Calzolari, Monica Monachini, Claudia Soria, and Khalid Choukri. 2006. Language Resources Production Models: the case of the INTERA multilingual corpus and terminology, In: Proceedings of the Fifth International Conference on Language Resources and Evaluation-LREC2006, 24-26 May 2006, Genoa, Italy, 609 – 614.

Guengoer, T. 2010. Part-of-speech tagging. In, Handbook of Natural Language Processing (Second ed.), Eds. Nitin Indurkhya and Fred J. Damerau, Machine Learning and Pattern Recognition, Chapter 10, 205-235. Boca Raton, London, New York: Chapman & Hall/CRC, Taylor & Francis Group.

Ide, Nancy. 1998. Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. In First International Conference on Language Resources and Evaluation, LREC'98, Granada, 463-470. ELRA.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary. 2000. XCES: An XML-based Encoding Standard for Linguistic Corpora. In Second International Conference on Language Resources and Evaluation, LREC'00.

Ide, Nancy and Jean Véronis. 1994. Multext (Multilingual Tools and Corpora). In Proceedings of the 15th International Conference on Computational Linguistics, Kyoto, 90-96. ACL.

Krstev, Cvetana and Duško Vitas. 2005. Corpus and Lexicon - Mutual Incompleteness. In Proceedings of the Corpus Linguistics Conference, Eds. Pernilla Danielsson, P., Wagenmakers, M., 14-17 July 2005, Birmingham, http://www.corpus.bham.ac.uk/PCLC/.

Krstev, Cvetana., Duško Vitas, and Tomaž Erjavec. 2004. MULTEXT-East Resources for Serbian. In Zbornik 7. mednarodne multikonference „Informacijska druzba IS 2004" Jezikovne tehnologije, Eds. Tomaž Erjavec and Jerneja Z. Gros, 9-15 Oktober 2004, Ljubljana, Slovenija. Institut „Jozef Stefan".

Laporte, Eric and Anne Monceaux, A. 1998. Elimination of lexical ambiguities by grammars: The ELAG system. Lingvisticæ Investigationes, 22:341–367. Amsterdam-Philadelphia: John Benjamins Publishing Company.

Lindquist, Hans. 2009. Corpus Linguistics and the Description of English. Edinburgh University Press.

Paumier, Sébastien 2008. Unitex 2.1 User Manual.

http://www-igm.univmlv.fr/unitex/UnitexManual2.1.pdf.

Popović, Zoran. 2010. Tggers Applied on Texts in Serbian. Infotheca 11(2):21a–38a.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, UK. http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.

TEI Consortium (Ed.). 2009. TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Vitas, Duško, Cvetana Krstev and Svetla Koeva. 2007. Towards a Complex Model for Morpho-Syntactic Annotation. Proceedings of the Workshop on a Common Natural Language Processing Paradigm for Balkan Languages, Eds. E. Paskaleva and M. Slavcheva, 26 September 2007, Borovets, Bulgaria, 65-71.

Vitas, Duško, Cvetana Krstev, Ivan Obradović, Ljubomir Popović and Gordana Pavlović-Lažetić. 2003. Processing Serbian Written Texts: An Overview of Resources and Basic Tools. Workshop on Balkan Language Resources and Tools, 21 Novembar 2003, Thessaloniki, Greece, Eds. Piperidis, S., Karkaletsis, V., 97-104.

Xiao, Richard. 2010. Corpus Creation. In, Handbook of Natural Language Processing, Eds. N. Indurkhya and F. Damerau, Machine Learning & Pattern Recognition Series, Chapter 7, 147-165. CRC Press, Taylor and Francis Group.