Czech Language Processing - PoS Tagging

Jan Hajič & Barbora Hladká

Institute of Formal and Applied Linguistics
Charles University
Malostranské nám. 25
118 00 Prague, CZECH REPUBLIC
[{hajic,hladka}@ufal.ms.mff.cuni.cz]

Abstract

In the specification of the Conference aims, the following keywords appear in the LREC materials: availability of language resources, methods for evaluation of resources, comparing different approaches to a given problem, choosing the best solution etc. To meet these goals, we present here an overview of the state-of-art of Czech part-of-speech (PoS) tagging. We concentrate on the data creation and availability problems, then we discuss the results we obtained when using various methods to tag texts written in a highly inflectional language, and finally we conclude by an outline of future perspectives.[1]

1 Natural Language Processing

One of the meanings of the headword - *process* - speaks about "the analysis (of information) using a computer". That is exactly what we mean by natural language processing (NLP) - an analysis of language information using a computer.

However, the computer alone is not good enough. We need an electronic database covering written and spoken language resources[2].

The starting points for NLP are building a structured corpus and annotating corpus according to the needs of further processing. A corpus is a vast, electronically processed collection of language texts containing a variety of (as much explicit as possible) information the corpus might (implicitly) provide.

If we look at any NLP conference proceedings from 80s and 90s that we can see at the first sight that the vast majority of frequently processed languages are English, French, German, Italian, Spanish. Why they are so few contributions on processing of some typologically different, i.e. Slavic or similar language? There are many reasons for that but the key reason lies in the absence of the main resources of NLP - corpora for these languages.

2 Czech Language Processing

One of the main tasks of the most recent Czech Language Processing (CzLP) project in the Czech Republic, so called "Integrated Project: Czech in the Age of Computers[1]" (started in 1996), is an investigation of present-day Czech based on contemporary methods and techniques for computational linguistics. This task includes i.a. a development of a Part-of-Speech (PoS) tagging system. This is not trivial task in a view of the fact that most of the existing tagging systems have been developed for languages typologically different from Czech.

3 Czech PoS Tagging[3]

3.1 Language Resources

[1] The results described herein have been obtained within various projects sponsored by the Czech Grant Agency (405/96/K214, 405/95/0190), by the Ministry of Education project No.VS96151, by the Charles University Grant Agency project No. 39/94 and by the individual grant of the OSF/HESP No. 195/1995.

[2] In what follows, we will concentrate on the processing of written language.

[3] Is there any relationship between NLP terms **PoS tagging, morphological disambiguation, morphological annotation and morphological analysis**? The *morphological analysis* of a given wordform provides for all possibilities of a *morphological annotation*. For illustration, let's assume wordform "zdi" ('walls'). One of the morphological annotations corresponds to the genitive singular for feminine nouns, other to the dative, vocative and locative singular, or nominative and accusative plural of the same word.. The other corresponds to the imperative of singular of the verb and so on. Each morphological category (case, gender, number,...) may take a set of possible values (gender -- masculine animate, masculine inanimate, neuter, feminine). The morphological annotations of a wordform represent the combinations of morphological categories for the particular part of speech classes. In order to automatically process a morphological analysis it is very useful to mark the values of morphological categories and part of speech classes positively (gender -- masculine animate (M), masculine inanimate (I), neuter (N), feminine (F), nouns (N), verbs (V),....). Afterwards we can rewrite the morphological annotations of wordform "zdi" mentioned above in the following way NFS[2,3,5,6], NFP[1,4], VM. A task - called *morphological disambiguation* or *PoS tagging* - uses the context of the given wordform in the input text to select the correct tag from the list of all possible tags.

For the experiments described herein, we have used two different corpora: one ”old” (texts from the 60s and early 70s), and one ”new” (smaller volume but modern Czech and technically compatible with our new morphological analysis system). Due to the technical incompatibility of these two resources we performed different experiments on them (see sect. 3.2 for the experiments using the ”old” corpus and sect. 3.3 for the description of experiments using the ”new” one.)

Czech Corpus (CC - ”old”) Thanks to the enthusiasm of a group of people from the Institute of Czech Language the main working material - written and spoken Czech Corpus - has been created during the 70s. The quantitative characteristics of present-day Czech were the main motivation for building CC. The corpus includes newspaper, magazine and scientific texts. The quantitative research (Těšitelová et al., 1984) has concentrated among other things on the frequency of part of speech classes, frequency of morphological categories and syntactic phenomena. For these purposes CC was morphologically and syntactically manually tagged. The format of CC is exemplified in Table 1 as the only tagged corpus for a Slavic language then available.

TOKEN	POS TAG	LEMMA	SYNT. TAG	ORDER (POSITION)
v	776	v	911	0012690
šestém	242416	šestý	31+01	0012691
kole	117416	kolo	51+02	0012692
se			+01	0012693
hrála	5266 74	hráti	21	0012694
dvě	410431	dva	31+02	0012695
nejpřitaž-livější	22 4213	přitažlivý nej	31+01	0012696
utkání	150421	utkání	1 _03	0012697
v	776	v		0012698
praze	107316	praha	31_02	0012699

Table 1: The format of CC illustrated by the Czech sentence *V šestém kole se hrála dvě nejpřitažlivější utkání v Praze*. [lit. *In sixth round Refl. played two most-attractive matches in Prague*]

The fact that there exists a morphologically tagged corpus of Czech was an encouraging moment for us and we used it in the very first experiments of PoS tagging of Czech (Hladká, 1994).

Czech National Corpus (CNC - ”new”) The Czech National Corpus is being built up by an concerted effort of a number of institutions, mostly by the Institute of Czech National Corpus. The work on CNC has started at the beginning of the 90s. In the Figure 1, we illustrate the SGML format of the CNC.

```
<p n=2>
<s id="s/inf/j/1993/vesm9301:045-p2s1">
<f cap>Starořecký
<f>bůh
<f cap>Pan
<f>děsil
<f>noční
<f>poutníky
<f>nevázaností
<f>reje
<f>své
<f>družiny
<D>
<d>.
```

Figure 1: SGML format of CNC illustrated by the Czech sentence *Starořecký bůh Pan děsil noční poutníky nevázaností reje své družiny*[lit. ‘Ancient-Greek God Pan was-horrifying night (Acc.) pilgrims (Acc.) by-wild by-rounds of-his company.’]

There is no tagging (nor manual nor automatic) available yet for the current ”official” version of the CNC.

3.2 Experiments on the Czech Corpus (CC)

Conversion As we have mentioned above, CC was originally morphologically hand-tagged, including lemmatisation and syntactic tags. For the purpose of our Czech PoS tagging experiments, we have used only a part of the CC (600K tokens, newspaper texts) and we have disregarded the lemmatization information and the syntactic tags, as we were interested in wordforms and tags only. Tags used in CC were different from our suggested tags especially as for the number of processed morphological categories and the notation. Thus we carried out conversions of the original data into our tag system (Hajič, Hladká, 1997a) which resulted in the so-called Czech Tagged Corpus (CTC). Table 2 exemplifies the converted format of CTC.

TOKEN	POS TAG
v	Rv
šestém	ANS61A
kole	NNS6
se	X
hrála	V3PAMONA
dvě	CNP1
nejpřitažlivější	ANP13A
utkání	NNP1
v	Rv
Praze	NFS6

Table 2: CTC format (example sentence from Table 1)

Training Data Characteristics The following table presents the basic features of CTC. For comparison, the average number of tags per token in English is 3.2 (based on the WSJ data, see Marcus et al., 1993).

tokens	621 015
wordforms (types)	72 445
different tags (tag types)	1 171
average number of tags per token	3.65

Table 3: Basic features of CTC

Methods The first five experiments have been based on probabilistic methods. They have used the basic source channel model technique (Merialdo, 1992). The probabilistic models (HMMs -unigram, bigram, trigram) have been trained on all available Czech tagged data, i.e. on the CTC. As we were interested in the influence of tag system on the performance of the method used, we have also reduced the tag system from 1 171 tags to a less detailed tagset that contains 206 tags (Hladká, Ribarov, 1998). The following table characterizes the probabilistic model and the PoS tagset of the particular experiments.

experiment No.	1	2	3	4	5
N-gram model	uni-	bi-	tri-	bi-	tri-
PoS tagset	1 171	1 171	1 171	206	206

Table 4: Characteristics of the statistical experiments 1-5

The common question of all experiments can be expressed by the sentences *How to improve tagging accuracy or how to manage the tagging accuracy let us say 97%?* To try to apply another approach. We have had opportunity to apply two another approach - rule-based one (Brill, 1992) and Xerox PoS tagging tools (Hajič, Hladká, 1997a). Totally, we have performed two rule-based experiments (No. 6, 7) and one experiment (No. 8) using Xerox tools (see Table 5).

experiment No.	6	7	8
method	Rule-Based	Rule-Based	Xerox
PoS tagset	1 171	206	89
training data	38K	38K	15K

Table 5: Experiments (6, 7, 8) characteristics

Results All results reported here are based on best-only approach using an accuracy criterion (number of correct results divided by number of input words). It should be said that whereas the experiments 1-5 do not use any morphological pre-processing, the experiments 6-7 (Brill's tagger) (kind of) learns the morphology from the training data, and the Xerox tagger differs even more: not only it uses the full morphology available for XLT, but also the training data is different (it has been prepared specifically for training the XLT tagger).

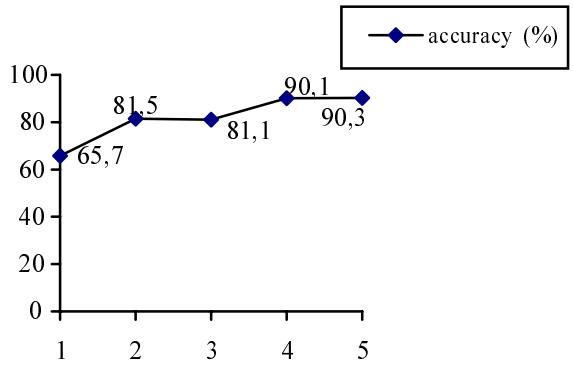

Figure 2: The results of the 'pure' HMM experiments No. 1–5

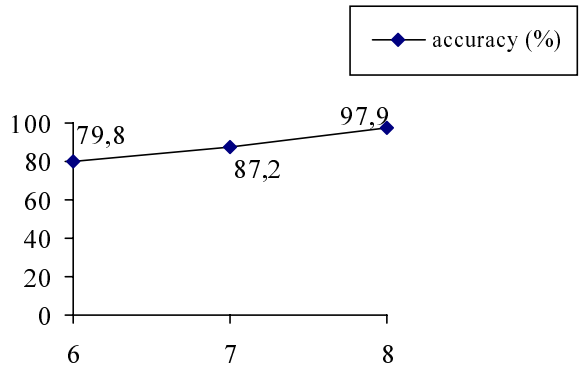

Figure 3: The results of the rule-based and the Xerox Language Tools experiments No. 6-8

3.3 Experiments on CNC

Software Tools for the Data Collection Most of the NLP learning algorithms need very large corpora to get reliable estimates of their parameters. CNC offers new working material for the experiments of Czech PoS tagging**.** At present, CNC covers nearly 70 million tokens. 70 million is so high a number that it is necessary to tag the texts automatically (in order to get its annotated version for, say, lexicographic work). As we have noted earlier, all present tagging experiments need tagged text as training data. That is why a part of CNC must be manually tagged to get the best results. Manual PoS tagging does not mean

that people are pairing the appropriate morphological tags with the wordforms in the text totally manually. That would be a very time consuming, exhausting and never-ending process with respect to the size of the texts that should be manually disambiguated.

In order to make the manual tagging of texts more human-friendly and comfortable a special purpose tool has been developed. The tool was first implemented under the Linux OS and then reimplemented also for the Windows 95 platform. The tools work on texts in the SGML format (see Figure 1) as pre-processed by the morphological analyzer (see Figure 4)[4].

```
<p n=2>
<s id="s/inf/j/1993/vesm9301:045-p2s1">
<f cap>Starořecký<l>starořecký<t>AIS11A
 <t>AIS41A<t>AIS51A<t>AMS11A<t>AMS51A
<f>bůh<l>bůh<t>NMS1A
<f cap>Pan<l>pan<t>NMS1A
<f>děsil<l>děsit_:T<t>VRYSA
<f>noční<l>noční<t>AFP11A<t>AFP41A<t>AFP51A
 <t>AFS11A<t>AFS21A<t>AFS31A<t>AFS41A<t>AFS51A
 <t>AFS61A<t>AFS71A<t>AIP11A<t>AIP41A<t>AIP51A
 <t>AIS11A<t>AIS41A<t>AIS51A<t>AMP11A<t>AMP41A
 <t>AMP51A<t>AMS11A<t>AMS51A<t>ANP11A<t>ANP41A
 <t>ANP51A<t>ANS11A<t>ANS41A<t>ANS51A
<f>poutníky<l>poutník<t>NMP4A<t>NMP7A
<f>nevázaností<l>vázanost_^(např._provazem;
 _odvozeniny_řídké)_(*4t)<t>NFP2N<t>NFS7N
 <l>nevázanost-2^(rozpustilost,_veselí)
 <t>NFP2A<t>NFS7A
<f>reje<l>rej<t>NIP1A<t>NIP4A<t>NIP5A<t>NIS2A
<f>své<l>svůj-1<t>PRSFP1-1<t>PRSFP4-1
 <t>PRSFS2-1<t>PRSFS3-1<t>PRSFS6-1<t>PRSIP1-1
 <t>PRSIP5-1<t>PRSNS1-1<t>PRSNS4-1
 <t>PRSNS5-1<t>PRSYP4-1
 <l>svůj-2_^(být_svůj)<t>A1FP<t>A1IP<t>A1NS
<f>družiny<l>družina
 <t>NFP1A<t>NFP4A<t>NFP5A<t>NFS2A
<D>
<d>.
```

Figure 4: SGML text processed by morphological analyzer exemplified on the Czech sentence from Fig. 1.

The list of ambiguous words found in the input text - `starořecký, noční, poutníky, nevázaností, reje, své, družiny` - (Figure 5 - the left window), the full text context (Figure 5 - the lower right window) and the right upper window devoted to the disambiguation process are the three main parts of the disambiguation tools.

The annotator chooses the correct lemma and then the correct tag.

[4]The morphological tagset contains currently 3111 tags (full description see in Hajič (1998)). The morphological tag system is more detailed than the one used for tagging of CTC.

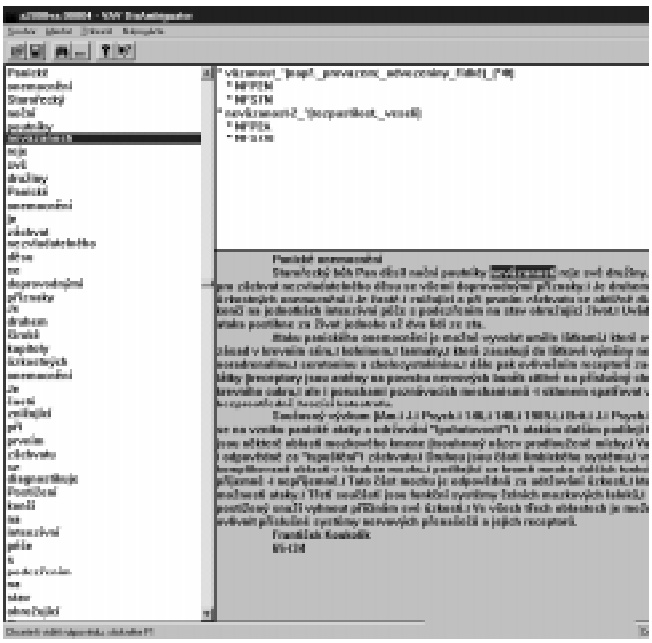

Figure 5: Disambiguation tool (Win95)

Training Data Characteristics Today, our training data consists of about 133K tokens of newspaper and magazine text. For the training process we have separated part of these data (called CNC_t) which covers nearly 125K tokens tagged by 860 different tags.

tokens	124 692
wordforms	29 903
tags	860

Table 6: Basic features of CNC_t.

Methods Since the beginning of tagging experiments it has been clear that including linguistic information into purely statistical approaches should be a step ahead. The term linguistic information means (in our case) linguistic information got from morphological analysis. However, we could not use morphological information in the first experiments based on CTC because the automatic morphological analyzer was not complete and it worked with a different tag system. With the finished automatic morphological analyzer the idea to connect morphology with a probabilistic approach could finally be made.

In the pure probabilistic approach the probability of a word **w** given a tag **t - p(w/t)** - is calculated for **t∈T**, **T** is set of all tags from the training data. When we include the morphological analysis, for given a word **w** we calculate **p(w/t)** or **p(w,t)** as well but **t∈T_w**, **T_w** being the set of all possible tags given by morphological analysis for **w.** Obviously, **$|T_w| < |T|$**.

Results

In order to compare statistical tagging with and without morphological pre-processing we have performed three

experiments (No. 9, 10 and 11). The experiment No. 9 has worked on the part of CTC - 110 874 tokens tagged by 882 different tags.

experiment No.	9	10	11
morphology	-	+	+
N-gram model	bi-	bi-	tri-
tags	882	860	860
training data	110 874	124 692	124 692

Table 7: Experiments (9, 10, 11) characteristics.

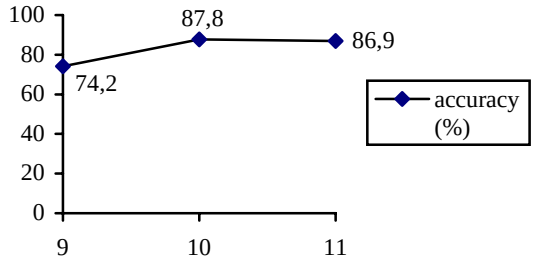

Figure 6: The results of statistical experiments No. 9-11 without and with morphological pre-processing.

4 Result Comparison

In sections 3.2 and 3.3 the experiment results presented in Fig. 2, 3 and 6 without any discussion. We would like to devote the present section to a more detailed discussion on the results.

The sequence of the experiments came into life step by step and each an experiment was based on an idea how to improve the tagging accuracy of the previous experiment. The increasing character of the accuracy curves shows that we have been successful in the selection of the model 'parameters' – *more training data, a less detailed tagset, a different tagging method, including linguistic information.* The choice of such 'parameters' has emerged mostly from the comparison of our different approaches to tagging. The experience from other tagging experiments had a very important influence on our decisions as well.

Let us look at the maximums in the graph Figures 2 & 3: 90.3% is comparable with results for Swedish (Elworthy, 1995), 97.9% is the same as for English (Schiller, 1996). These numbers show that a smaller tagset achieves better tagging performance than the bigger does (as expected) and the statistical approach seems to be a little bit better than a rule-based one, even though the rule-based results has been influenced by the smaller training data size which could not be exceeded because of the time needed for training.

Consequently, the results mean that many sentences will contain at least one error. What is the magic point that we would like to achieve it (we know it is - for various reasons -not 100%)? So what about 98%? Using Xerox tagging tools, the tagging accuracy (97.9%) is becoming closer and closer to 98%. But, the Xerox experiment has been performed upon a smaller tagset containing tags concentrating mostly on PoS classes and, in not all but in many applications, it is too coarse for the subsequent processing of tagged text like automatic syntactic analysis, spelling correction, speech recognition, etc.

The main conclusion, which we drew from the experiments, is the following: tagset should be chosen according to the requirements of a given application rather than to optimise it for the tagger. The more detailed tagset the better - but again, one must primarily consider the application at hand and (if at all possible) to optimize the accuracy/tagset-size ratio.

We can now identify three areas for further research. First, we will add more manually tagged data and possibly convert the "old" CTC into the "new" CNC-compatible SGML format (together with morphology conversion and editing). Second, we will be improving all the Czech taggers, and on finishing the development of a new tagger that uses an exponential probabilistic model based on automatically selected features. This last tagger gave preliminary results which seem to outperform the other taggers (for a detailed account and latest results, see Hajič, Hladká, 1998). Having more data in the 'new' format will allow us to make a 100% fair comparison of all the taggers. The evaluation of the results then reveals whether the taggers differ in where they make their respective errors. We believe that they make substantially different types of errors and thus we plan to develop a model which will combine the results of the morphological stochastic tagger, rule-based tagger and the stochastic exponential tagger with the hope that the final result will be an improvement over all and each of them.

References

Brill E. (1992). A Simple Rule-Based Tagger. In: *Proceedings of The Third Conference on Applied Natural Language Processing.* Trento.

Hajič J (1998). Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová (ed. by E. Hajičová)* (pp. 106-132). Prague: Karolinum

Hajič J. and Hladká B. (1997). Probabilistic and Rule-Based Tagger of an Inflective Language - a Comparison, In: *Proceedings of 5th Conference on Applied Natural Language Processing* (pp. 111-118), ACL, Washington, DC, USA,

Hajič J. and Hladká B. (1998). Error-driven Tagging for a Rich, Structured Tagset, Based on an Exponential Model, accepted for COLIN-ACL'98.

Hladká B. (1994). Programové vybavení pro zpracování velkých českých textových korpusů [Software for Large Czech Corpora Annotation], MSc thesis, MFF UK, Prague, Czech Republic

Hladká B. and Ribarov K. (1998). PoS Tags for Automatic Tagging and Syntactic Structures. In *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová (ed. by E. Hajičová)* (pp. 226-240). Prague: Karolinum.

Marcus M. et al. (1993). Building A Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, 19(2) (pp. 313-330).

Merialdo B. (1992). Tagging Text with a Probabilistic Model. Computational Linguistics 20(2) (pp. 155-171).

Schiller A. (1996). Multilingual Finite-State Noun Phrase Extraction. ECAI´96. Budapest.

Těšitelová M. et al. (1984). Kvantitativní charakteristiky současné češtiny [Quantitative characteristic of the present-day Czech language]. Prague: Academia.