Probabilistic and Rule-Based Tagger of an Inflective Language - a Comparison

Jan Hajič Barbora Hladká
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Malostranské nám. 25
CZ−118 00 Prague 1
e-mail: {hajic,hladka}@ufal.mff.cuni.cz

Abstract

We present results of probabilistic tagging of Czech texts in order to show how these techniques work for one of the highly morphologically ambiguous inflective languages. After description of the tag system used, we show the results of four experiments using a simple probabilistic model to tag Czech texts (unigram, two bigram experiments, and a trigram one). For comparison, we have applied the same code and settings to tag an English text (another four experiments) using the same size of training and test data in the experiments in order to avoid any doubt concerning the validity of the comparison. The experiments use the source channel model and maximum likelihood training on a Czech hand-tagged corpus and on tagged Wall Street Journal (WSJ) from the LDC collection. The experiments show (not surprisingly) that the more training data, the better is the success rate. The results also indicate that for inflective languages with 1000+ tags we have to develop a more sophisticated approach in order to get closer to an acceptable error rate. In order to compare two different approaches to text tagging — statistical and rule-based — we modified Eric Brill's rule-based part of speech tagger and carried out two more experiments on the Czech data, obtaining similar results in terms of the error rate. We have also run three more experiments with greatly reducedtagset to get another comparison based on similar tagset size.

1 INTRODUCTION

Languages with rich inflection like Czech pose a special problem for morphological disambiguation (which is usually called tagging [1]). For example, the ending "-u" is not only highly ambiguous, but at the same time it carries complex information: it corresponds to the genitive, the dative and the locative singular for inanimate nouns, or the dative singular for animate nouns, or the accusative singular for feminine nouns, or the first person singular present tense active participle for certain verbs. There are two different techniques for text tagging: a stochastic technique and a rule-based technique. Each approach has some advantages — for stochastic techniques there exists a good theoretical framework, probabilities provide a straightforward way how to disambiguate tags for each word and probabilities can be acquired automatically from the data; for rule-based techniques the set of meaningful rules is automatically acquired and there exists an easy way how to find and implement improvements of the tagger. Small set of rules can be used, in contrast to the large statistical tables. Given the success of statistical methods in different areas, including text tagging, given the very positive results of English statistical taggers and given the fact that there existed no statistical tagger for any Slavic language we wanted to apply statistical methods even for the Czech language although it exhibits a rich inflection accompanied by a high degree of ambiguity. Originally, we expected that the result would be plain negative, getting no more than about two thirds of the tags correct. However, as we show below, we got better results than we had expected. We used the same statistical approach to tag both the English text and the Czech text. For English, we obtained results comparable with the results presented in (Merialdo, 1992) as well as in (Church, 1992). For Czech, we obtained results which are less satisfactory than those for English. Given the comparability of the accuracy of the rule-based

[1]Procedures of automatic tagging of Czech are or were supported fully or partially by the following grants/projects: Charles University GAUK 39/94, Grant Agency of the Czech Republic GAČR 405/96/K214 and Ministry of Education VS96151

part-of-speech (POS) tagger (Brill, 1992) with the accuracy of the stochastic tagger and given the fact that a rule-based POS tagger has never been used for a Slavic language we have tried to apply rule-based methods even for Czech.

2 STATISTICAL EXPERIMENTS

2.1 CZECH EXPERIMENTS

2.1.1 CZECH TAGSET

Czech experiment is based upon ten basic POS classes and the tags describe the possible combinations of morphological categories for each POS class. In most cases, the first letter of the tag denotes the part−of−speech; the letters and numbers which follow it describe combinations of morphological categories (for a detailed description, see Table 2.1 and Table 2.2).

Morph. Categ.	Cat. Var. (see Tab. 2.2)	Poss. Val.	Description
gender	*g*	M I N F	masc. anim. masc. inanim. neuter feminine
number	*n*	S P	singular plural
tense	*t*	M P F	past present future
mood	*m*	O R	indicative imperative
case	*c*	1 2 3 4 5 6 7	nominative genitive dative accusative vocative locative instrumental
voice	*s*	A P	active voice passive voice
polarity	*a*	N A	negative affirmative
deg. of comp.	*d*	1 2 3	base form comparative superlative
person	*p*	1 2 3	1st 2nd 3rd

Table 2.1

Note especially, that Czech nouns are divided into four classes according to the gender (Sgall, 1967) and into seven classes according to the case.

POS Class	
nouns	N*gnc*
noun, abbreviations	NZ
adjectives	A*gncda*
verbs, infinitives	VT*a*
verbs, transgressives	VW*ntsga*
verbs, common	V*pnstmga*
pronouns, personal	PP*pnc*
pronouns, 3rd person	PP3*gnc*
pronouns, possessive	PR*gncpgn*
"svůj" — "his" refering to subject	PS*gnc*
reflexive particle "se"	PE*c*
pronouns demonstrative	PD*gnca*
adverbs	Od*a*
conjunctions	S
numerals	C*gnc*
prepositions	Rpreposition
interjections	F
particles	K
sentence boundaries	T_SB
punctuation	T_IP
unknown tag	X

Table 2.2

Not all possible combinations of morphological categories are meaningful, however. In addition to these usual tags we have used special tags for sentence boundaries, punctuation and a so called "unknown tag". In the experiments, we used only those tags which occurred at least once in the training corpus. To illustrate the form of the tagged text, we present here the following examples from our training data, with comments: word|tag

do|Rdo	#"to" (prepositions have their own individuals tags)
oddílu|NIS2	#"unit" (noun, masculine inanimate, singular, genitive)
k|Rk	#"for" (preposition)
snídani|NFS3	#"breakfast" (noun, feminine, singular, dative)
použije|V3SAPOMA	#"uses" (verb, 3rd person, singular, active, present, indicative, masc. animate, affirmative)
pro|Rpro	#"for" (preposition)
nás|PP1P4	#"us" (pronoun, personal, 1st person, plural, accusative)

2.1.2 CZECH TRAINING DATA

For training, we used the corpus collected during the 1960's and 1970's in the Institute for Czech Language at the Czechoslovak Academy of Sciences. The corpus was originally hand-tagged, including the lemmatization and syntactic tags. We had to do some cleaning, which means that we have disregarded the lemmatization information and the syntactic tag, as we were interested in words and tags only. Tags used in this corpus were different from our suggested tags: number of morphological categories was higher in the original sample and the notation was also different. Thus we had to carry out conversions of the original data into the format displayed above, which resulted in the so-called Czech "modified" corpus, with the following features:

tokens	621 015
words	72 445
tags	1 171
average number of tags per token	3.65

Table 2.3

We used the complete "modified" corpus (621 015 tokens) in the experiments No. 1, No. 3, No. 4 and a small part of this corpus in the experiment No. 2, as indicated in Table 2.4.

tokens	110 874
words	22 530
tags	882
average number of tags per token	2.36

Table 2.4

2.2 ENGLISH EXPERIMENTS

2.2.1 ENGLISH TAGSET

For the tagging of English texts, we used the Penn Treebank tagset which contains 36 POS tags and 12 other tags (for punctuation and the currency symbol). A detailed description is available in (Santorini, 1990).

2.2.2 ENGLISH TRAINING DATA

For training in the English experiments, we used WSJ (Marcus et al., 1993). We had to change the format of WSJ to prepare it for our tagging software. We used a small (100k tokens) part of WSJ in the experiment No. 6 and the complete corpus (1M tokens) in the experiments No. 5, No. 7 and No. 8. Table 2.5 contains the basic characteristics of the training data.

	Experiment No. 6	Experiments No. 5, No. 7, No. 8
tokens	110 530	1 287 749
words	13 582	51 433
tags	45	45
average number of tags per token	1.72	2.34

Table 2.5

2.3 CZECH VS ENGLISH

Differences between Czech as a morphologically ambiguous inflective language and English as language with poor inflection are also reflected in the number of tag bigrams and tag trigrams. The figures given in Table 2.6 and 2.7 were obtained from the training files.

	Czech corpus		WSJ
x <= 4	24 064	x < = 10	459
4 < x <= 16	5 577	10 < x < = 100	411
16 < x <= 64	2 706	100 < x < = 1000	358
x > 64	1 581	x > 1000	225
bigrams	33 928	bigrams	1 453

Table 2.6 Number of bigrams with frequency x

	Czech corpus		WSJ
x <= 4	155 399	x < = 10	11 810
4 < x <= 16	16 371	10 < x < = 100	4 571
16 <x <= 64	4 380	100 < x < = 1000	1 645
x >64	933	x > 1000	231
trigrams	177 083	trigrams	18 257

Table 2.7 Number of trigrams with frequency x

It is interesting to note the frequencies of the most ambiguous tokens encountered in the whole "modified" corpus and to compare them with the English data. Table 2.8 and Table 2.9 contain the first tokens with the highest number of possible tags in the complete Czech "modified" corpus and in the complete WSJ.

Token	Frequency in train. data	#tags in train. data
jejich	1 087	51
jeho	1 087	46
jehož	163	35
jejichž	150	25
vedoucí	193	22

Table 2.8

In the Czech "modified" corpus, the token "vedoucí" appeared 193 times and was tagged by twenty two different tags: 13 tags for adjective and 9 tags for noun. The token "vedoucí" means either: "leading" (adjective) or "manager" or "boss" (noun). The following columns represent the tags for the token "vedoucí" and their frequencies in the training data; for example "vedoucí" was tagged twice as adjective, feminine, plural, nominative, first degree, affirmative.

2	vedoucí|AFP11A		
4	vedoucí|AFP41A		
6	vedoucí|AFS11A	10	vedoucí|NFS1
11	vedoucí|AFS21A	1	vedoucí|NFS2
1	vedoucí|AFS31A	1	vedoucí|NFS3
4	vedoucí|AFS41A	1	vedoucí|NFS4
5	vedoucí|AFS71A	2	vedoucí|NFS7
2	vedoucí|AIP11A	34	vedoucí|NMP1
11	vedoucí|AMP11A	17	vedoucí|NMP4
3	vedoucí|AMP41A	61	vedoucí|NMS1
12	vedoucí|AMS11A	1	vedoucí|NMS5
2	vedoucí|ANP11A		
2	vedoucí|ANS41A		

Token	Frequency in train. data	#tags in train. data
a	25 791	7
down	1 052	7
put	380	6
set	362	6
that	10 902	6
the	56 265	6

Table 2.9

It is clear from these figures that the two languages in question have quite different properties and that nothing can be said without really going through an experiment.

2.4 THE ALGORITHM

We have used the basic source channel model (described e.g. in (Merialdo, 1992)). The tagging procedure ϕ selects a sequence of tags T for the sentence W: $\phi : W \rightarrow T$. In this case the optimal tagging procedure is

$$\begin{aligned}\phi(W) &= argmax_T Pr(T|W) = \\ &= argmax_T Pr(T|W) * Pr(W) = \\ &= argmax_T Pr(W,T) = \\ &= argmax_T Pr(W|T) * Pr(T).\end{aligned}$$

Our implementation is based on generating the (W,T) pairs by means of a probabilistic model using approximations of probability distributions $Pr(W|T)$ and $Pr(T)$. The $Pr(T)$ is based on tag bigrams and trigrams, and $Pr(W|T)$ is approximated as the product of $Pr(w_i|t_i)$. The parameters have been estimated by the usual maximum likelihood training method, i.e. we approximated them as the relative frequencies found in the training data with smoothing based on estimated unigram probability and uniform distributions.

2.5 THE RESULTS

The results of the Czech experiments are displayed in Tables 2.10.

	No. 1	No. 2	No. 3	No. 4
test data (tokens)	1 294	1 294	1 294	1 294
prob. model	unigram	bigram	bigram	trigram
incorrect tags	444	334	239	244
tagging accuracy	65.70%	74.19%	81.53%	81.14%

Table 2.10

These results show, not surprisingly, of course, that the more data, the better (results experiments of No.2 vs. No.3), but in order to get better results for a trigram tag prediction model, we would need far more data. Clearly, if 88% trigrams occur four times or less, then the statistics is not reliable. The following tables show a detailed analysis of the errors of the trigram experiment.

	A	C	F	K	N	O
A	32	0	0	0	6	3
C	0	4	0	0	1	0
F	0	0	0	0	0	0
K	0	0	0	0	0	0
N	4	0	0	0	64	8
O	0	0	0	0	1	0
P	0	0	0	0	0	3
R	0	0	0	0	1	1
S	0	0	0	0	0	0
V	0	0	0	0	3	8
T	0	0	0	0	1	0
X	0	0	0	0	0	0

Table 2.11a

	P	R	S	V	T	X	
A	2	2	2	2	1	0	50
C	0	0	0	0	0	0	5
F	0	0	0	0	0	0	0
K	0	0	1	0	0	1	2
N	0	4	2	2	5	4	93
O	0	0	0	1	1	0	3
P	19	0	0	0	1	2	23
R	0	0	0	0	0	2	4
S	0	0	0	0	0	2	2
V	0	3	8	28	1	2	53
T	0	0	0	0	0	0	1
X	5	0	1	2	0	0	8

Table 2.11b

The letters in the first column and row denote POS classes, the interpunction (T) and the "unknown tag" (X). The numbers show how many times the tagger assigned an incorrect POS tag to a token in the test file. The total number of errors was 244. Altogether, fifty times the adjectives (A) were tagged incorrectly, nouns (N) 93 times, numbers (C) 5 times and etc. (see the last unmarked column in Table 2.11b); to provide a better insight, we should add that in 32 cases, when the adjective was correctly tagged as an adjective, but the mistakes appeared in the assignment of morphological categories (see Table 2.12), 6 times the adjective was tagged as a noun, twice as a pronoun, 3 times as an adverb and so on (see the second row in Table 2.11a). A detailed look at Table 2.12 reveals that for 32 correctly marked adjectives the mistakes was 17 times in gender, once in number, three times in gender and case simultaneously and so on.

A	g	n	c	g&c	g&n	c&a	g&n&c	g&c&d
32	17	1	6	3	2	1	1	1

Table 2.12

Similar tables can be provided for nouns (Table 2.13), numerals (Table 2.14), pronouns(Table 2.15) and verbs (Table 2.16a, Table 2.16b).

N	g	n	c	g&c	n&c	−> NZ
64	11	5	41	2	4	1

Table 2.13

C	g	c
4	1	3

Table 2.14

P	g	c	g&c	PD −> PP
19	8	7	3	1

Table 2.15

V	p	t	n	s	n&t	p&t	t&a
22	3	6	5	5	1	1	1

Table 2.16a

V	g&a	p&n&t	V −>VT
6	1	1	4

Table 2.16b

The results of our experiments with English are displayed in Table 2.17.

	No. 5	No. 6	No. 7	No. 8
test data (tokens)	1 294	1 294	1 294	1 294
prob. model	unigram	bigram	bigram	trigram
incorrect tags	136	81	41	37
tagging accuracy	89.5%	93.74%	96.83%	97.14%

Table 2.17

To illustrate the results of our tagging experiments, we present here short examples taken from the test data. Cases of incorrect tag assignment are in boldface.

— Czech

<table>
<tr><th>word|hand tag</th><th>exp. No.4</th><th>exp. No.3</th><th>exp. No.2</th><th>exp. No.1[2]</th></tr>
<tr><td>na|Rna</td><td>Rna</td><td>Rna</td><td>Rna</td><td>Rna</td></tr>
<tr><td>půdě|NFS6</td><td>NFS6</td><td>NFS6</td><td>NFS6</td><td>NFS6</td></tr>
<tr><td>vlasti|NFS2</td><td>NFS2</td><td>NFS2</td><td>NFS2</td><td>NFS2</td></tr>
<tr><td>rady|NFS2</td><td>NFS2</td><td>NFS2</td><td>NFS2</td><td>NFS2</td></tr>
<tr><td>žen|NFP2</td><td>NFP2</td><td>NFP2</td><td>NFP2</td><td>NFP2</td></tr>
<tr><td>Gusta|NFS1</td><td>T_SB</td><td>T_SB</td><td>AFP21A</td><td>XX</td></tr>
<tr><td>Fučíková|NFS1</td><td>NFS1</td><td>NFS1</td><td>NFP2</td><td>NFS1</td></tr>
<tr><td>a|SS</td><td>SS</td><td>SS</td><td>SS</td><td>SS</td></tr>
<tr><td>předseda|NMS1</td><td>NMS1</td><td>NMS1</td><td>NMS1</td><td>NMS1</td></tr>
<tr><td>úv|NZ</td><td>NZ</td><td>NZ</td><td>NZ</td><td>NZ</td></tr>
<tr><td>ssm|NZ</td><td>NZ</td><td>NZ</td><td>NZ</td><td>NZ</td></tr>
<tr><td>Juraj|NMS1</td><td>NMS1</td><td>NMS1</td><td>NMS1</td><td>XX</td></tr>
<tr><td>Varholík|NMS1</td><td>NMS1</td><td>NMS1</td><td>NMS1</td><td>NMS1</td></tr>
</table>

[2] We used a special tag XX for unknown words.

— English

word | hand tag	exp. No.8	exp. No.7	exp. No.6	exp. No.5
With|IN	IN	IN	IN	IN
stock|NN	NN	NN	NN	NN
prices|NNS	NNS	NNS	NNS	NNS
hovering|VBG	VBG	VBG	**IN**	VBG
near|IN	IN	IN	**JJ**	IN
record|NN	NN	NN	NN	NN
levels|NNS	NNS	NNS	NNS	NNS
,|,	,	,	,	,
a|DT	DT	DT	DT	DT
number|NN	NN	NN	NN	NN
of|IN	IN	IN	IN	IN
companies|NNS	NNS	NNS	NNS	NNS
have|VBP	VBP	VBP	VBP	VBP
been|VBN	VBN	VBN	VBN	VBN
announcing|VBG	VBG	VBG	**IN**	VBG
stock|NN	NN	NN	NN	NN
splits|NNS	NNS	**VBZ**	**NN**	**VBZ**
.|.	.	.	.	.

2.6 A PROTOTYPE OF RANK XEROX POS TAGGER FOR CZECH

(Schiller, 1996) describes the general architecture of the tool for noun phrase mark-up based on finite-state techniques and statistical part-of-speech disambiguation for seven European languages. For Czech, we created a prototype of the first step of this process — the part-of-speech (POS) tagger — using Rank Xerox tools (Tapanainen, 1995), (Cutting et al., 1992).

2.6.1 POS TAGSET

The first step of POS tagging is obviously adefinition of POS tags obviously. We performed three experiments. These implementations differ in the POS tagset. During the first experiment we designed tagset which contains 47 tags. The POS tagset can be described as follows:

Category	Symbol	Pos. Val.	Description
case	*c*	NOM	nominative
		GEN	genitive
		DAT	dative
		ACC	accusative
		VOC	vocative
		LOC	locative
		INS	instrumental
		INV	for abbreviations
kind of verb	*t*	PAP	past paticiple
		PRI	present participle
		INF	infinitive
		IMP	imperative
		TRA	transgressive

Table 2.18

POS tag	Description
NOUN_*c*	nouns + case
ADJ_*c*	adjectives + case
PRON_*c*	pronouns + case
VERB_*t*	verbs + kind of verb
ADV	adverbs
PROP	proper names
PREP	prepositions
PSE	reflexive "se"
CLIT	clitics
CONJ	conjunctions
INTJ	interjections
PTCL	particles
DATE	dates
CM	comma
PUNCT	interpunction
SENT	sentence bundaries

Table 2.19

Analysing the results of the first implementation declared very high ambiguity between nominative and accusative of nouns, adjectives, pronouns and numerals. That is why we replaced the tags for nominative and accusative of nouns, adjectives, pronouns and numerals by new tags NOUN_NA, ADJ_NA, PRON_NA and NUM_NA (meaning nominative or accusative undestinguished). The rest of the tags stayed unchanged. This led POS tags — 43. In the third experiment we deleted the morphological information for nouns and adjectives all together. This process resulted in the final 34 POS tags.

2.6.2 THE RESULTS

Figures representing the results of all experiments are presented in the following table. We have also included the results of English tagging using the same Xerox tools.

language	tags	ambiguity[3]	tagging accuracy
Czech	47	39%	91.7%
Czech	43	36%	93.0%
Czech	34	14%	96.2%
English	76	36%	97.8%

Table 2.20

The results show that the more radical reduction of Czech tags (from 1171 to 34) the higher accuracy of the results and the more comparable are the Czech and English results. However, the difference in the error rate is still more than visible — here we can speculate that the reason is taht Czech is "free" word order language, whereas English is not.

3 A RULE-BASED EXPERIMENT FOR CZECH

A simple rule-based part of speech (RBPOS) tagger is introduced in (Brill, 1992). The accuracy of this tagger for English is comparable to a stochastic English POS tagger. From our point of view, it is very interesting to compare the results of Czech stochastic POS (SPOS) tagger and a modified RBPOS tagger for Czech.

3.1 TRAINING DATA

We used the same corpus used in the case of the SPOS tagger for Czech. RBPOS requires different input format; we thus converted the whole corpus into this format, preserving the original contents.

3.2 LEARNING

It is an obvious fact that the Czech tagset is totally different from the English tagset. Therefore, we had to modify the method for the initial guess. For Czech the algorithm is: "If the word is W_SB (sentence boundary) assign the tag T_SB, otherwise assign the tag NNS1."

3.2.1 LEARNING RULES TO PREDICT THE MOST LIKELY TAG FOR UNKNOWN WORDS

The first stage of training is learning rules to predict the most likely tag for unknown words. These rules operate on word types; for example, if a word ends by "dý", it is probably a masculine adjective. To compare the influence of the size of the training files on the accuracy of the tagger we performed two subexperiments[4]:

	No. 1	No. 2
TAGGED-CORPUS (tokens)	37 971	9 576
TAGGED-CORPUS (words)	15 297	5 031
TAGGED-CORPUS (tags)	738	495
UNTAGGED-CORPUS (tokens)	621 015	621 015
UNTAGGED-CORPUS (words)	72 445	72 445
LEXRULEOUTFILE (rules)	101	75

Table 3.1

We present here an example of rules taken from LEXRULEOUTFILE from the exp. No. 1:

u hassuf 1 NIS2	# change the tag to NIS2 if the suffix is "u"
y hassuf 1 NFS2	# change the tag to NFS2 if the suffix is "y"
ho hassuf 2 AIS21A	# change the tag to AIS21A if the suffix is "ho"
ch hassuf 3 NFP6	# change the tag to NFP6 if the suffix is "ch"
nej addpref 3 O2A	# change the tag to O2A if adding the prefix "nej" results in a word

3.2.2 LEARNING CONTEXTUAL CUES

The second stage of training is learning rules to improve tagging accuracy based on contextual cues. These rules operate on individual word tokens.

[3] The percentage of ambiguous word forms in the test file.

[4] We use the same names of files and variables as Eric Brill in the rule-based POS tagger's documentation. TAGGED-CORPUS — manually tagged training corpus, UNTAGGED-CORPUS — collection of all untagged texts, LEXRULEOUTFILE — the list of transformations to determine the most likely tag for unknown words, TAGGED-CORPUS-2 — manually tagged training corpus, TAGGED-CORPUS-ENTIRE — Czech "modified" corpus (the entire manually tagged corpus), CONTEXT-RULEFILE — the list of transformations to improve accuracy based on contextual cues.

	No. 1	No. 2
TAGGED-CORPUS-2 (tokens)	37 892	9 989
TAGGED-CORPUS-2 (words)	12 676	4 635
TAGGED-CORPUS-2 (tags)	717	479
TAGGED-ENTIRE-CORPUS (tokens)	621 015	621 015
TAGGED-ENTIRE-CORPUS (words)	72 445	72 445
TAGGED-ENTIRE-CORPUS (tags)	1 171	1 171
CONTEXT-RULEFILE (rules)	487	61

Table 3.2

We present here an example of the rules taken from CONTEXT−RULEFILE from the exp. No. 1:

AFP21A AIP21A	# change the tag AFP21A to AIP21A
NEXT1OR2TAG NIP2	if the following tag is NIP2
NIS2 NIS6	# change the tag NIS2 to NIS6
PREV1OR2OR3TAG Rv	if the preceding tag is Rv
NIS1 NIS4	# change the tag NIS1 to NIS4
PREV1OR2TAG Rna	if the preceding tag is Rna

3.2.3 RESULTS

The tagger was tested on the same test file as for the statistical experiments. We obtained the following results:

	No. 1	No. 2
TEST-FILE	1 294	1 294
errors	262	294
tagging accuracy	79.75%	77.28%

Table 3. 3

4 CONCLUSION

The results, though they might seem negative compared to English, are still better than our original expectations. Before trying some completely different approach, we would like to improve the current simple approach by some other simple measures: adding a morphological analyzer (Hajič, 1994) as a front-end to the tagger (serving as a "supplier" of possible tags, instead of just taking all tags occurring in the training data for a given token), simplifying the tagset, adding more data. However, the desired positive effect of some of these measures is not guaranteed: for example, the average number of tags per token will increase after a morphological analyser is added. Success should be guaranteed, however, by certain tagset reductions, as the original tagset (even after the reductions mentioned above) is still too detailed. This is especially true when comparing it to English, where some tags represent, in fact, a set of tags to be discriminated later (if ever). For example, the tag VB used in the WSJ corpus actually means "one of the (five different) tags for 1st person sg., 2nd person sg., 1st person pl., etc.". First, we will reduce the tagset to correspond to our morphological analyzer which already uses the reduced one. Then, the tagset will be reduced even further, but nevertheless, not as much as we did for the Xerox-tools-based exapriment, because that tagset is too "rough" for many applications, even though the results are good.

Another possibility of an improvement is to add more data; also, the use of trigrams rather than smoothing may lead to better results. We will also may add contemporary newspaper texts to our training data in order to account for recent language development. Hedging against failure of all these simple improvements, we are also working on a different model using independent predictions for certain grammatical categories (and the lemma itself), but the final shape of the model has not yet been determined. This would mean to introduce constraints on possible combinations of morphological categories and take them into account when "assembling" the final tag.

References

Eric Brill. 1992. A Simple Rule-Based Part of Speech Tagger. In: *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy.

Eric Brill. 1993. A Corpus Based Approach To Language Learning. *PhD Dissertation*, Department of Computer and Information Science, University of Pennsylvania.

Eric Brill. 1994. Some Advances in Transformation–Based Part of Speech Tagging. In: *Proceedings of the Twelfth National Conference on Artificial Intelligence.*

Jan Hajič. 1994. Unification Morphology Grammar. *PhD Dissertation*, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.

Kenneth W. Church. 1992. Current Practice In Part Of Speech Tagging And Suggestions For The Future. For Henry Kučera, *Studies in*

Slavic Philology and Computational Linguistics, Michigan Slavic Publications, Ann Arbor.

Doug Cutting, Julian Kupiec, Jan Pedersen and Penelope Sibun 1992. A Practical Part-of-Speech Tagger. In: *Proceedings of the Third Conference on Applied Natural Language Processing* , Trento, Italy.

Mitchell P. Marcus, Beatrice Santorini, and Mary-Ann Marcinkiewicz 1993. Building A Large Annotated Corpus Of English: The Penn Treebank. *Computational Linguistics*, 19(2):313—330.

Bernard Merialdo. 1992. Tagging Text With A Probabilistic Model. *Computational Linguistics*, 20(2):155—171

Beatrice Santorini. 1990. Part Of Speech Tagging Guidelines For The Penn Treebank Project. *Technical report MS-CIS-90-47*, Department of Computer and Information Science, University of Pennsylvania.

Anne Schiller. 1996. Multilingual Finite-State Noun Phrase Extraction. *ECAI'96*, Budapest, Hungary.

Petr Sgall. 1967. The Generative Description of a Language and the Czech Declension (In Czech). *Studie a práce lingvistické, 6*. Prague.

Pasi Tapanainen. 1995. RXRC Finite-State Compiler. *Technical Report MLTT-20*, Rank Xerox Research Center, Meylen, France.