A Resource-light Approach to Russian Morphology: Tagging Russian using Czech resources

Jiri Hana and **Anna Feldman** and **Chris Brew**
Department of Linguistics
Ohio State University
Columbus, OH 43210

Abstract

In this paper, we describe a resource-light system for the automatic morphological analysis and tagging of Russian. We eschew the use of extensive resources (particularly, large annotated corpora and lexicons), exploiting instead (i) pre-existing annotated corpora of Czech; (ii) an unannotated corpus of Russian. We show that our approach has benefits, and present what we believe to be one of the first full evaluations of a Russian tagger in the openly available literature.

1 Introduction

Morphological processing and part-of-speech tagging are essential for many NLP tasks, including machine translation, information retrieval and parsing. In this paper, we describe a resource-light approach to the tagging of Russian. Because Russian is a highly inflected language with a high degree of morpheme homonymy (cf. Table 1[1]) the tags involved are more numerous and elaborate than those typically used for English. This complicates the tagging task, although as has been previously noted (Elworthy, 1995), the increased complexity of the tags does not necessarily translate into a more demanding tagging task. Because no large annotated corpora of Russian are available to us, we instead chose to use an annotated corpus of Czech. Czech is sufficiently similar to Russian that it is reasonable to suppose that information about Czech will be relevant in some way to the tagging of Russian.

The languages share many linguistic properties (free word order and a rich morphology which plays a considerable role in determining agreement and argument relationships). We created a morphological analyzer for Russian, combined the results with information derived from Czech and used the TnT (Brants, 2000) tagger in a number of different ways, including a a committee-based approach, which turned out to give the best results. To evaluate the results, we morphologically annotated (by hand) a small corpus of Russian: part of the translation of Orwell's "1984" from the MULTEXT-EAST project (Véronis, 1996).

krasiv-a	beautiful (short adjective, feminine)
muž-a	husband (noun, masc., sing., genitive)
	husband (noun, masc., sing., accusative)
okn-a	window (noun, neuter, sing., genitive)
	window (noun, neuter, pl., nominative)
	window (noun, neuter, pl., accusative)
knig-a	book (noun, fem., sing., nominative)
dom-a	house (noun, masc., sing., genitive)
	house (noun, masc., pl., nominative)
	house (noun, masc., pl., accusative)
skazal-a	say (verb, fem., sing., past tense)
dv-a	two (numeral, masc., nominative)

Table 1: Homonymy of the *a* ending

2 Why TnT?

Readers may wonder why we chose to use TnT, which was not designed for Slavic languages. The short answer is that it is convenient and successful, but the following two sections address the issue in rather more detail.

2.1 The encoding of lexical information in TnT

TnT records some lexical information in the emission probabilities of its second order Markov Model. Since Russian and Czech do not use the same words we cannot use this information (at least not directly) to tag Russian. Given this, the move from Czech to Russian involves a loss of detailed lexical information. Therefore we implemented a morphological analyzer for Russian, the output of which we use to provide surrogate emission probabilities for the TnT tagger (Brants, 2000). The details are described below in section 4.2.

[1] All Russian examples in this paper are transcribed in the Roman alphabet. Our system is able to analyze Russian texts in both Cyrillic and various transcriptions.

2.2 The modelling of word order in TnT

Both Russian and Czech have relatively free word order, so it may seem an odd choice to use a Markov model (MM) tagger. Why should second order MM be able to capture useful facts about such languages? Firstly, even if a language has the potential for free word order, it may still turn out that there are recurring patterns in the progressions of parts-of-speech attested in a training corpus. Secondly, n-gram models including MM have indeed been shown to be successful for various Slavic languages, e.g., Czech (Hajič et al., 2001) or Slovene (Džeroski et al., 2000); although not as much as for English. This shows that the transitional information captured by the second-order MM from a Czech or Slovene corpus is useful for Czech or Slovene.[2] The present paper shows that transitional information acquired from Czech is also useful for Russian.

3 Russian versus Czech

A deep comparative analysis of Czech and Russian is far beyond the scope of this paper. However, we would like to mention just a number of the most important facts. Both languages are Slavic (Czech is West Slavonic, Russian is East Slavonic). Both have extensive morphology whose role is important in determining the grammatical functions of phrases. In both languages, the main verb agrees in person and number with subject; adjectives agree in gender, number and case with nouns. Both languages are free constituent order languages. The word order in a sentence is determined mainly by discourse. It turns out that the word order in Czech and Russian is very similar. For instance, old information mostly precedes new information. The "neutral" order in the two languages is Subject-Verb-Object. Here is a parallel Czech-Russian example from our development corpus:

(1) a. [Czech]
Byl jasný,
was$_{Masc.Past}$ bright$_{Masc.Sg.Nom}$
studený dubnový
cold$_{Masc.Sg.Nom}$ April$_{Masc.Sg.Nom}$
den i hodiny
day$_{Masc.Sg.Nom}$ and clocks$_{Fem.Pl.Nom}$
odbíjely třináctou.
stroke$_{Fem.Pl.Past}$ thirteenth$_{Fem.Sg.Acc}$

b. [Russian]
Byl jasnyj,
was$_{Masc.Past}$ bright$_{Masc.Sg.Nom}$
xolodnyj aprel'skij
cold$_{Masc.Sg.Nom}$ April$_{Masc.Sg.Nom}$
den' i časy
day$_{Masc.Sg.Nom}$ and clocks$_{Pl.Nom}$
probili trinadtsat'.
stroke$_{Pl.Past}$ thirteen$_{Acc}$
'It was a bright cold day in April, and the clocks were striking thirteen.' [from Orwell's '1984']

Of course, not all utterances are so similar. Section 5.4 briefly mentions how to improve the utility of the corpus by eradicating some of the systematic differences.

4 Realization

4.1 The tag system

We adopted the Czech tag system (Hajič, 2000) for Russian. Every tag is represented as a string of 15 symbols each corresponding to one morphological category. For example, the word *vidjela* is assigned the tag VpFS---XR-AA---, because it is a verb (V), past participle (p), feminine (F), singular (S), does not distinguish case (-), possessive gender (-), possessive number (-), can be any person (X), is past tense (R), is not gradable (-), affirmative (A), active voice (A), and does not have any stylistic variants (the final hyphen).

No.	Description	Abbr.	No. of values	
			Cz	Ru
1	POS	P	12	12
2	SubPOS – detailed POS	S	75	32
3	Gender	g	11	5
4	Number	n	6	4
5	Case	c	9	8
6	Possessor's Gender	G	5	4
7	Possessor's Number	N	3	3
8	Person	p	5	5
9	Tense	t	5	5
10	Degree of comparison	d	4	4
11	Negation	a	3	3
12	Voice	v	3	3
13	Unused		1	1
14	Unused		1	1
15	Variant, Style	V	10	2

Table 2: Overview and comparison of the tagsets

The tagset used for Czech (4290+ tags) is larger than the tagset we use for Russian (about 900 tags). There is a good theoretical reason for this choice

[2] Respectively, and if the techniques in the present paper generalize, probably also irrespectively.

– Russian morphological categories usually have fewer values (e.g., 6 cases in Russian vs. 7 in Czech; Czech often has formal and colloquial variants of the same morpheme); but there is also an immediate practical reason – the Czech tag system is very elaborate and specifically devised to serve multiple needs, while our tagset is designed solely to capture the core of Russian morphology, as we need it for our primary purpose of demonstrating the portability and feasibility of our technique. Still, our tagset is much larger than the Penn Treebank tagset, which uses only 36 non-punctuation tags (Marcus et al., 1993).

4.2 Morphological analysis

In this section we describe our approach to a resource-light encoding of salient facts about the Russian lexicon. Our techniques are not as radical as previously explored unsupervised methods (Goldsmith, 2001; Yarowsky and Wicentowski, 2000), but are designed to be feasible for languages for which serious morphological expertise is unavailable to us. We use a paradigm-based morphology that avoids the need to explicitly create a large lexicon. The price that we pay for this is overgeneration. Most of these analyses look very implausible to a Russian speaker, but significantly increasing the precision would be at the cost of greater development time than our resource-light approach is able to commit. We wish our work to be portable at least to other Slavic languages, for which we assume that elaborate morphological analyzers will not be available. We do use two simple pre-processing methods to decrease the ambiguity of the results handed to the tagger – longest ending filtering and an automatically acquired lexicon of stems. These were easy to implement and surprisingly effective.

Our analyzer captures just a few textbook facts about the Russian morphology (Wade, 1992), excluding the majority of exceptions and including information about 4 declension classes of nouns, 3 conjugation classes of verbs. In total our database contains 80 paradigms. A paradigm is a set of endings and POS tags that can go with a particular set of stems. Thus, for example, the paradigm in Table 3 is a set of inflections that go with the masculine stems ending on the "hard" consonants, e.g., *slon* 'elephant', *stol* 'table'.

Unlike the traditional notions of *stem* and *ending*, for us a stem is the part of the word that does not change within its paradigm, and the ending is the part of the word that follows such a stem. For example, the forms of the verb *moč'* 'can.INF': *mogu* '1sg', *možeš'* '2sg', *možet* '3sg', etc. are analyzed as

0	NNMS1----------	y	NNMP1----------
a	NNMS2----------	ov	NNMP2----------
u	NNMS3----------	am	NNMP3----------
a	NNMS4----------	ov	NNMP4----------
u	NNMS4---------1		
e	NNMS6----------	ax	NNMP6----------
u	NNMS6---------1		
om	NNMS7----------	ami	NNMP7----------

Table 3: A paradigm for "hard" consonant masculine nouns

the stem *mo* followed by the endings *gu*, *žeš'*, *žet*. A more linguistically oriented analysis would involve the endings *u*, *eš'*, *et* and phonological alternations in the stem. All stem internal variations are treated as suppletion.[3]

Unlike the morphological analyzers that exist for Russian (Segalovich and Titov, 2000; Segalovich, 2003; Segalovich and Maslov, 1989; Kovalev, 2002; Mikheev and Liubushkina, 1995; Yablonsky, 1999; Segalovich, 2003; Kovalev, 2002, among others) (Segalovich, 2003; Kovalev, 2002; Mikheev and Liubushkina, 1995; Yablonsky, 1999, among others), our analyzer does not rely on a substantial manually created lexicon. This is in keeping with our aim of being resource-light. When analyzing a word, the system first checks a list of monomorphemic closed-class words and then segments the word into all possible prefix-stem-ending triples.[4] The result has quite good coverage (95.4%), but the average ambiguity is very high (10.9 tags/token), and even higher for open class words. We therefore have two strategies for reducing ambiguity.

4.2.1 Longest ending filtering (LEF)

The first approach to ambiguity reduction is based on a simple heuristic – the correct ending is usually one of the longest candidate endings. In English, it would mean that if a word is analyzed either as having a zero ending or an *-ing* ending, we would consider only the latter; obviously, in the vast majority of cases that would be the correct analysis. In addition, we specify that a few long but very rare endings should not be included in the maximum length calculation (e.g., 2nd person pl. imperative).

[3] We do in fact have a very similar analysis, the analyzer's run-time representation of the paradigms is automatically produced from a more compact and linguistically attractive specification of the paradigms. It is possible to specify the basic paradigms and then specify the subparadigms, exceptions and paradigms involving phonological changes by referring to them.

[4] Currently, we consider only two inflectional prefixes – negative *ne* and superlative *nai*.

4.2.2 Deriving a lexicon

The second approach uses a large raw corpus[5] to generate an open class lexicon of possible stems with their paradigms. In this paper, we can only sketch the method, for more details see (Hana and Feldman, to appear). It is based on the idea that open-class lemmata are likely to occur in more than one form. First, we run the morphological analyzer on the text (without any filtering), then we add to the lexicon those entries that occurred with at least a certain number of distinct forms and cover the highest number of forms. If we encounter the word *talking*, using the information about paradigms, we can assume that it is either the *-ing* form of the lemma *talk* or that it is a monomorphemic word (such as *sibling*). Based on this single form we cannot really say more. However, if we also encounter the forms *talk*, *talks* and *talked*, the former analysis seems more probable; and therefore, it seems reasonable to include the lemma *talk* as a verb into the lexicon. If we encountered also *talkings*, *talkinged* and *talkinging*, we would include both lemmata *talk* and *talking* as verbs.

Obviously, morphological analysis based on such a lexicon overgenerates, but it overgenerates much less than if based on the endings alone. For example, for the word form *partii* of the lemma *partija* 'party', our analysis gives 8 possibilities – the 5 correct ones (noun fem sg gen/dat/loc sg and pl nom/acc) and 3 incorrect ones (noun masc sg loc, pl nom, and noun neut pl acc; note that only gender is incorrect). Analysis based on endings alone would allow 20 possibilities – 15 of them incorrect (including adjectives and an imperative).

4.3 Tagging

We use the TnT tagger (Brants, 2000), an implementation of the Viterbi algorithm for second order Markov models. We train the transition probabilities on Czech (1.5M tokens of the Prague Dependency Treebank (Bémová et al., 1999)). We obtain surrogate emission probabilities by running our morphological analyzer, then assuming a uniform distribution over the resulting emissions.

5 Experiments

5.1 Corpora

For evaluation purposes, we selected and morphologically annotated (by hand) a small portion from the Russian translation of Orwell's '1984'. This corpus contains 4011 tokens and 1858 types. For development, we used another part of '1984'. Since we want to work with minimal language resources, the development corpus is intentionally small – 1788 tokens. We used it to test our hypotheses and tune the parameters of our tools.

In the following sections, we discuss our experiments and report the results. Note that we do not report the results for tag position 13 and 14, since these positions are unused; and therefore, always trivially correct.

5.2 Morphological analysis

As can be seen from Table 4, morphological analysis without any filters gives good recall (although on a non-fiction text it would probably be lower), but also very high average ambiguity. Both filters (the longest-ending filter and automatically acquired lexicon) reduce the ambiguity significantly; the former producing a considerable drop of recall, the latter retaining high recall. However, we do best if we first attempt lexical lookup, then apply LEF to the words not found. This keeps recall reasonably high at the same time as decreasing ambiguity. As expected, performance increases with the size of the unannotated Russian corpus used to generate the lexicon. All subsequent experimental results were obtained using this best filter combination, i.e., the combination of the lexicon based on the 1Mword corpus and LEF.

LEF	no	no	no	yes	yes	yes
Lexicon based on	0	100K	1M	0	100K	1M
recall	**95.4**	94	93.1	84.4	88.3	90.4
avg ambig (tag/word)	10.9	7.0	4.7	4.1	3.5	**3.1**
Tagging – accuracy	50.7	62.1	67.5	62.1	66.8	69.4

Table 4: Morph. analysis with various parameters

5.3 Tagging

Table 7 summarizes the results of our taggers on test data. Our baseline is produced by the morphological analyzer without any filters followed by a tagger randomly selecting a tag among the tags offered by the morphological analyzer. The direct-full tag column shows the result of the TNT tagger with transition probabilities obtained directly from the Czech corpus and the emission symbols based on the morphological analyzer with the best filters.

To further improve the results, we used two techniques: (i) we modified the training corpus to remove some systematic differences between Czech

[5] We used The Uppsala Russian Corpus (1M tokens), which is freely available from Uppsala University at `http://www.slaviska.uu.se/ryska/corpus.html`.

and Russian (5.4); (ii) we trained batteries of taggers on subtags to address the data sparsity problem (5.5 and 5.6).

5.4 Russification

We experimented with "russified" models. We trained the TnT tagger on the Czech corpus with modifications that made the structure of training data look more like Russian. For example, plural adjectives and participles in Russian, unlike Czech, do not distinguish gender.

(2) a. Nadaní muži soutežili.
Gifted$_{masc.pl}$ men competed$_{masc.pl}$
'Gifted sportsmen were competing.' [Cz]

b. Nadané ženy soutežily.
Gifted$_{fem.pl}$ women competed$_{fem.pl}$
'Gifted women were competing.' [Cz]

c. Nadaná děvčata soutěžila.
Gifted$_{neut.pl}$ girls$_{neut}$ competing$_{neut.pl}$
'Gifted girls were competing.' [Cz]

d. Talantlivye mužčiny/ženščiny
Gifted$_{pl}$ men/women
sorevnovalis'.
competed$_{pl}$
'Gifted men/women were competing.' [Ru]

Negation in Czech is in the majority of cases is expressed by the prefix *ne-*, whereas in Russian it is very common to see a separate particle (*ne*) instead:

(3) a. Nic **neřekl**.
nothing not-said
'He didn't say anything.' [Cz]

b. On ničego **ne skazal**.
he nothing not said
'He didn't say anything.' [Ru]

In addition, reflexive verbs in Czech are formed by a verb followed by a reflexive clitic, whereas in Russian, the reflexivization is the affixation process:

(4) a. Filip **se** ještě neholí.
Filip REFL-CL still not-shaves
'Filip doesn't shave yet.' [Cz]

b. Filip esče ne breet+**sja**.
Filip still not shaves+REFL.SUFFIX
'Filip doesn't shave yet.' [Ru]

Even though auxiliaries and the copula are the forms of the same verb *byt'* 'to be', both in Russian and in Czech, the use of this verb is different in the two languages. For example, Russian does not use an auxiliary to form past tense:

(5) a. Já **jsem** psal.
I aux$_{1sg}$ wrote
'I was writing/I wrote.' [Cz]

b. Ja pisal.
I wrote
'I was writing/I wrote.' [Ru]

It also does not use the present tense copula, except for emphasis; but it uses forms of the verb *byt'* in some other constructions like past passive.

We implemented a number of simple "russifications". The combination of random omission of the verb *byt'*, omission of the reflexive clitics, and negation transformation gave us the best results on the development corpus. Their combination improves the overall result from 68.0% to 69.4%. We admit we expected a larger improvement.

5.5 Sub-taggers

One of the problems when tagging with a large tagset is data sparsity; with 1000 tags there are 1000^3 potential trigrams. It is very unlikely that a naturally occurring corpus will contain all the acceptable tag combinations with sufficient frequency to reliably distinguish them from the unacceptable combinations. However, not all morphological attributes are useful for predicting the attributes of the succeeding word (e.g., tense is not really useful for case). We therefore tried to train the tagger on individual components of the full tag, in the hope that each sub-tagger would be able to learn what it needs for prediction. This move has the additional benefit of making the tag set of each such tagger smaller and reducing data sparsity. We focused on the first 5 positions – POS (P), SubPOS (S), gender (g), number (n), case (c) and person (p). The selection of the slots is based on our linguistic intuition – for example it is reasonable to assume that the information about part-of-speech and the agreement features (gnc) of previous words should help in prediction of the same slots of the current word; or information about part-of-speech, case and person should assist in determining person. On the other hand, the combination of tense and case is *prima facie* unlikely to be much use for prediction. Indeed, most of our expectations have been met. The performance of some of the models on the development corpus is summarized in Table 5. The bold numbers indicate that the tagger outperforms the full-tag

tagger. As can be seen, the taggers trained on individual positions are worse than the full-tag tagger on these positions. This proves that a smaller tagset does not necessarily imply that tagging is easier – see (Elworthy, 1995) for more discussion of this interesting relation. Similarly, there is no improvement from the combination of unrelated slots – case and tense (ct) or gender and negation (ga). However, the combinations of (detailed) part-of-speech with various agreement features (e.g., Snc) outperform the full-tag tagger on at least some of the slots.

	full-tag	P	S	g	n	c
1 (P)	89.0	87.2	–	–	–	–
2 (S)	86.6	–	84.5	–	–	–
3 (g)	81.4	–	–	78.8	–	–
4 (n)	92.4	–	–	–	91.2	–
5 (c)	80.9	–	–	–	–	78.4

	full-tag	Pc	gc	ga	nc	cp	ct
1 (P)	89.0	87.5	–	–	–	–	–
2 (S)	86.6	–	–	–	–	–	–
3 (g)	81.4	–	80.4	78.7	–	–	–
4 (n)	92.4	–	–	–	91.8	–	–
5 (c)	80.9	80.6	**81.1**	–	**81.5**	79.3	79.5
8 (p)	98.3	–	–	–	–	96.9	–
9 (t)	97.0	–	–	–	–	–	96.1
11 (a)	97.0	–	–	95.4	–	–	–

	full-tag	Pgc	Pnc	Sgc	Snc	Sgnc
1 (P)	89.0	87.9	87.5	–	–	–
2 (S)	86.6	–	–	86.1	86.4	**87.1**
3 (g)	81.4	80.3	–	81.4	–	**82.7**
4 (n)	92.4	–	92.4	–	**93.0**	**92.8**
5 (c)	80.9	**81.8**	**81.4**	80.9	**82.9**	**82.3**

Table 5: Performance of the TnT tagger trained on various subtags (development data)

5.6 Combining Sub-taggers

We now need to put the sub-tags back together to produce estimates of the correct full tags. We cannot simply combine the values offered by the best taggers for each slot, because that could yield illegal tags (e.g., nouns in past tense). Instead we select the best tag from those offered by our morphological analyzer using the following formula:

(6) $\mathsf{bestTag} = \mathsf{argmax}_{t \in T_{MA}} \mathsf{val}(t)$

T_{MA} – the set of tags offered by MA
$\mathsf{val}(t) = \sum_{k=0}^{14} N_k(t)/N_k$
$N_k(t)$ – # of taggers voting for k-th slot of t
N_k – the total # of taggers on slot k

That means, that the best tag is the tag that received the highest average percentage of votes for each of its slots. If we cared about certain slots more than about others we could weight the slots in the val function.

We ran several experiments, the results of three of them are summarized in Table 6. All of them work better than the full-tag tagger. One ('all') uses all available subtaggers, other ('best 1') uses the best tagger for each slot (therefore voting in Formula 6 reduces to finding a closest legal tag). The best result is obtained by the third tagger ('best 3') which uses the three best taggers for each of the *Pgcp* slots and the best tagger for the rest. We selected this tagger to tag the test corpus, for which the results are summarized in Table 7.

	full-tag	all	best 1	best 3
overall	69.5	70.3	70.7	71.1
1 (P)	89.0	88.9	89.1	89.2
2 (S)	86.6	86.5	86.9	86.9
3 (g)	81.4	81.8	83.0	83.2
4 (n)	92.4	92.6	93.1	93.2
5 (c)	80.9	82.1	83.0	83.2
6 (G)	98.5	98.5	98.7	98.7
7 (N)	99.6	99.7	99.8	99.8
8 (p)	98.3	98.2	98.4	98.3
9 (t)	97.0	97.0	97.0	97.0
10 (G)	96.0	96.0	96.0	96.0
11 (a)	97.0	97.0	96.9	97.0
12 (v)	97.4	97.3	97.5	97.4
15 (V)	99.1	99.1	99.0	99.0

Table 6: Combining sub-taggers (development data)

Tagger	Baseline random	Direct full-tag	Russified full-tag	Russified voting
Accuracy				
Tags	33.6	69.4	72.6	73.5
1 (POS)	63.2	88.5	90.1	90.4
2 (SubPOS)	57.0	86.8	88.1	88.6
3 (Gender)	59.2	82.5	84.5	85.0
4 (Number)	75.9	91.2	92.6	93.4
5 (Case)	47.3	80.4	84.1	85.3
6 (PossGen)	83.4	98.4	98.8	99.0
7 (PossNr)	99.6	99.6	99.6	99.8
8 (Person)	97.1	99.3	98.9	98.9
9 (Tense)	86.6	96.5	97.6	97.6
10 (Grade)	90.1	95.9	96.6	96.6
11 (Neg)	81.4	95.3	95.5	95.5
12 (Voice)	86.4	97.2	97.9	97.9
15 (Variant)	97.0	99.1	99.5	99.5

Table 7: Tagging with various parameters (test data)

Russian	Gloss	Correct	Xerox	Ours
Člen	member	noun_nom		_gen
partii	party	noun_gen	_obl	
po	prep	prep_obl	_acc	
vozmožnosti	possibility	noun_obl	_acc	
staralsja	tried	vfin		
nje	not	ptcl		
govorit'	to-speak	vinf		
ni	nor	ptcl		
o	about	prep_obl		
Bratstvje	Brotherhood	noun_obl		
,		cm		
ni	nor	ptcl		
o	about	prep_obl		
knigje	book	noun_obl		
Errors			3	1

'Neither the Brotherhood nor the book was a subject that any ordinary Party member would mention if there was a way of avoiding it.' [Orwell: '1984']

Table 8: Tagging with Xerox & our tagger

5.7 Comparison with Xerox tagger

A tagger for Russian is part of the Xerox language tools. We could not perform a detailed evaluation since the tool is not freely available. We used the online demo version of Xerox's Disambiguator[6] to tag a few sentences and compared the results with the results of our tagger. The Xerox tagset is much smaller than ours, it uses 63 tags, collapsing some cases, not distinguishing gender, number, person, tense etc. (However, it uses different tags for different punctuation, while we have one tag for all punctuation). For the comparison, we translated our tagset to theirs. On 201 tokens of the testing corpus, the Xerox tagger achieved an accuracy of 82%, while our tagger obtained 88%; i.e., a 33% reduction in error rate. A sample analysis is in Table 8.

5.8 Comparison with Czech taggers

The numbers we obtain are significantly worse than the numbers reported for Czech (Hajič et al., 2001) (95.16% accuracy); however, they use an extensive manually created morphological lexicon (200K+ entries) which gives 100.0% recall on their testing data. Moreover, they train and test their taggers on the same language.

6 Ongoing Research

We are currently working on improving both the morphological analysis and tagging. We would like to improve the recall of filters following morphological analysis, e.g., using n maximal values instead of 1, using some basic knowledge of derivational morphology, etc. We are incorporating phonological conditions on stems into the guesser module as well as trying to deal with different morphological phenomena specific to Russian, e.g., verb reflexivization. However, we try to stay language independent (at least within Slavic languages) as much as possible and limit the language dependent components to a minimum.

Currently, we are working on more sophisticated russifications that would be still easily portable to other languages. For example, instead of omitting auxiliaries randomly, we want to use the syntactic information present in Prague Dependency Treebank to omit only the 'right' ones.

If possible, we would like to avoid entirely throwing away the Czech emission probabilities, because our intuition tells us that there are useful lexical similarities between Russian and Czech, and that some suitable process of cognate detection will allow us to transfer information from the Czech to the Russian emission probabilities. Just as a knowledge of English words is sometimes helpful (modulo sound changes) when reading German, a knowledge of the Czech lexicon should be helpful (modulo character set issues) when reading Russian. We are seeking the right way to operationalize this intuition in our system, bearing in mind that we want a sufficiently general algorithm to make the method portable to other languages, for which we assume we have neither the time nor the expertise to undertake knowledge-intensive work. A potentially suitable cognate algorithm is described by (Kondrak, 2001).

Finally, we would like to extend our work to Slavic languages for which there are even fewer available resources than Russian, such as Belarusian, since this was the original motivation for undertaking the work in the first place.

Acknowledgements

We thank Erhard Hinrichs and Eric Fosler-Lussier for giving us feedback on previous versions of the paper and providing useful suggestions for subtaggers and voting; Jan Hajič for the help with the Czech tag system and the morphological analyzer; to the Clippers discussion group for allowing us to interview ourselves in front of them, and for ensuing discussion, and to two anonymous EMNLP reviewers for extremely constructive feedback.

[6] http://www.xrce.xerox.com/competencies/content-analysis/demos/russian

References

Alena Bémová, Jan Hajič, Barbora Hladká, and Jarmila Panevová. 1999. Morphological and syntactic tagging of the prague dependency treebank. In *Proceedings of ATALA Workshop*, pages 21–29. Paris, France.

Thorsten Brants. 2000. TnT - A Statistical Part-of-Speech Tagger. In *Proceedings of ANLP-NAACL*, pages 224–231.

Sašo Džeroski, Tomaž Erjavec, and Jakub Zavrel. 2000. Morphosyntactic Tagging of Slovene:Evaluating Taggers and Tagsets. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 1099–1104.

David Elworthy. 1995. Tagset design and inflected languages. In *EACL SIGDAT workshop "From Texts to Tags: Issues in Multilingual Language Analysis"*, pages 1–10, Dublin, April.

John Goldsmith. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198.

Jan Hajič, Pavel Krbec, Pavel Květoň, Karel Oliva, and Vladimír Petkevič. 2001. Serial Combination of Rules and Statistics: A Case Study in Czech Tagging. In *Proceedings of ACL Conference*, Toulouse, France.

Jan Hajič. 2000. Morphological Tagging: Data vs. Dictionaries. In *Proceedings of ANLP-NAACL Conference*, pages 94–101, Seattle, Washington, USA.

Jiri Hana and Anna Feldman. to appear. Portable Language Technology: The case of Czech and Russian. In *Proceedings from the Midwest Computational Linguistics Colloquium, June 25-26, 2004*, Bloomington, Indiana.

Greg Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*, pages 103–110, June.

Andrey Kovalev. 2002. A Probabilistic Morphological Analyzer for Russian and Ukranian. http://linguist.nm.ru/stemka/stemka.html.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Andrei Mikheev and Liubov Liubushkina. 1995. Russian Morphology: An Engineering Approach. *Natural Language Engineering*, 3(1):235–260.

Ilya Segalovich and Michail Maslov. 1989. Dictionary-based Russian morphological analysis and synthesis with generation of morphological models of unknown words (in Russian). http://company.yandex.ru/articles/article1.html.

Ilya Segalovich and Vitaly Titov. 2000. Automatic morphological annotation MYSTEM. http://corpora.narod.ru/article.html.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. http://company.yandex.ru/articles/iseg-las-vegas.html.

Jean Véronis. 1996. MULTEXT-EAST (Copernicus 106). http://www.lpl.univaix.fr/projects/multext-east.

Terence Wade. 1992. *A Comprehensive Russian Grammar*. Blackwell. 582 pp.

Serge A. Yablonsky. 1999. Russian Morphological Analysis. In *Proceedings VEXTAL*.

David Yarowsky and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Meeting of the Association for Computational Linguistics*, pages 207–216.