The Linguistics Journal

Tagset Design, Inflected Languages, and N-gram Tagging

Anna Feldman
***Montclair State University*, USA**

Bio Data:

Anna Feldman is an assistant professor of linguistics and computer science. Her interests are corpus linguistics and computational linguistics.

Abstract

This paper explores the relationship between the tagset design and linguistic properties of inflected languages for the task of morphosyntactic tagging. Some information theoretic measures and statistics on these languages are reported which show, unsurprisingly, that the tagsets for morphologically rich languages are larger than tagsets for English and the average tag/token ambiguity is higher. The surprising outcome of the experiments is that for Catalan, Czech, Polish, Portuguese, and Russian – which are considered to be "word order" free languages (to various degrees) – the knowledge about the preceding tag reduces the uncertainty about the tag in question if the *detailed* tagset is used, but when the tagset is *reduced* to the size of the English tagset (eliminating the detailed information), the two adjacent tags are relatively independent of each other. The experiments provide additional support to Elworthy (1995)'s results.

Moreover, even though the word order of richly inflected languages is considered to be relatively free, such languages seem to behave like English with respect to context, and therefore, it is concluded that n-gram tagging techniques are well

justified for such languages. Experiments with cross-lingual projection of morphosyntax described in Hana et al. (2004); Feldman et al. (2006b,a); Hana et al. (2006) provide additional empirical evidence for this claim.

A disclaimer: Due to the difficulty in obtaining tagged training data for these experiments, the corpora used here are relatively small. Hana et al. (2004); Feldman et al. (2006b,a); Hana et al. (2006) describe tagging experiments with the languages discussed in the present paper and prove that the predictions made here are correct. Further similar investigations on larger datasets should verify the claims made in this paper.

Keywords: morphosyntax, tagset, accession rate, average ambiguity, inflected languages.

1 Introduction

In this paper, we examine a number of properties of Slavic and Romance languages quantitatively. The results of the experiments discussed below provide a strong motivation for using *n*-gram tagging techniques for richly inflected (fusional) languages. Before we turn to the actual experiments, a description of the tagsets used in this work is provided.

From the practical point of view, the results of the experiments are important for deciding what algorithm to use for POS-tagging of highly inflected languages. Hidden Markov Models (HMMs) are independent of the language to which they are applied. Typically, training will make use of a manually tagged corpus, or an untagged corpus with some initial bootstrapping probabilities (e.g. Cutting et al. (1992)).

Another important question is how much training data is needed for avoiding the data sparsity problem maximally. The experiments that measure the accession rate try to answer this question.

The next question, both theoretical and practical, is what tagset design is suitable for languages like the ones explored in these experiments. As Elworthy

(1995) mentions in his paper, there are two criteria to consider: 1) The tagset must be capable of making the linguistic distinctions required in the output corpora (the external criterion); 2) Make the tagging as effective as possible (the internal criterion). The problem of tagset design is particularly important for highly inflected languages, such as Russian, Czech, Polish, Portuguese, Spanish, etc. The question is whether all syntactic variations, realized in these languages by means of morphological affixes, should be represented in the tagset.

The experiments reported in this paper look at the coverage of the tagset for different text sizes as well as the text coverage by a small number of high frequency tags. The results of these experiments suggest that the languages under consideration are more data-sparsity prone compared to English and will require a large training corpus. The next, naturally arising question is how much training data is necessary. The experiments described in section 6 are devoted to this question.

2 Related Work

Good tagset design is particularly important for highly inflected languages. If all of the syntactic variations that are realized in the inflectional system were represented in the tagset, there would be a huge number of tags, and it would be practically impossible to implement or train a tagger.

As has been mentioned above, Elworthy (1995) distinguishes external and internal criteria for tagset design. The external criterion is that the tagset must be capable of making the linguistic (for example, syntactic or morphological) distinctions required in the output corpora. The internal criterion on tagsets is the design criterion of making the tagging as effective as possible.

Elworthy (1995) designs an experiment to explore the relationship between tagging accuracy and the nature of the tagset, using corpora in English, French, and Swedish. The experiment addresses the internal design criterion. The aim of the experiment is to determine, crudely, whether a bigger tagset is better than a

smaller one, or whether external criteria requiring human intervention should be used to choose the best tagset.

It turns out that a larger tagset generally gives higher accuracy for Swedish, French, and English for texts with no unknown words (with notable exceptions in French, where gender marking was the key factor). For the test corpora that includes "unknown" words — words not seen during training and for which the (HMM) tagger hypothesizes all open-class tags — the results are slightly different. For the three test languages, the accuracy improves on the known words, but for unknown words, smaller tagsets give higher accuracy (again, for French, gender marking is the key factor). What seems to come out of these results is that there is not a consistent relationship between the size of the tagset and the tagging accuracy. Elworthy's general conclusion is that the external criterion should be the one to dominate tagset design.

Elworthy (1995) suggests that what is important is to choose the tagset required for the application, rather than to optimize it for the tagger. An additional comment that can be made here is that a large tagset could be always reduced to a smaller and less-detailed one if the application demands it.

3 Tag system

Various tag systems used for Slavic and Romance languages have been proposed. Here we discuss only the tagsets used in our experiments.

3.1 *Slavic tagsets*

The experiments with Slavic languages described in this paper deal with Russian, Polish, and Czech (see Tables 1 and 2).

The Czech tagset used for the experiments is an unmodified version of Hajič (2004)'s tag system. It contains 4290+ tags. This tag system is *positional*, which means that a tag is a string of 15 positions and each slot corresponds to one morphological category.

Since there is no available corpus of Russian annotated with detailed morphological information, the Russian tagset has been developed from scratch, based on the Hajič system. The tagset is very similar to the one used for Czech. However, it is smaller – it has only 1,000 tags. The reasons for this are both theoretical and practical. From the linguistic point of view, Russian does not make as many distinctions as Czech (e.g. no dual number, no auxiliary or pronominal clitics, no distinction between inanimate and animate masculine gender etc.). From a practical point of view, unlike the Czech system which was developed during several years and involved detailed analysis of the language and some theoretical assumptions, the Russian tagset is not as fine-grained as that for Czech because not as much time was spent on its development. Therefore, some fine-grained distinctions were omitted deliberately. This includes various types of numerals (e.g. multiplicative, definite, and indefinite numerals). Numerals often behave either as nouns or adjectives from both the morphological and syntactic points of view and, therefore, they are difficult to capture without a predefined lexicon.

Table 1 makes a comparison of the Czech and Russian tagsets. Consider the gender values, for example. Czech has 11 values for this attribute: M (masculine animate), F (feminine), N (neuter), X (any), ' -' (N/A), H (feminine or neuter), I (masculine inanimate), Q (feminine singular or neuter plural), T (masculine inanimate or feminine plural), Y (masculine animate or inanimate), and Z (not feminine). Russian does not include the ambiguous cases as separate attribute values. Thus, it distinguishes only between M (masculine), F (feminine), N (neuter), X (any), and ' -' (N/A). The number of case values differs as well because unlike Czech, Russian does not have the vocative case.

The original Polish tagset translated into the current system is taken from the IPI PAN corpus (Przepiórkowski (2004)). This corpus is morphosyntactically annotated, but the structure of its morphosyntactic tags is different from the tagset used for Czech and Russian. The inventory of grammatical categories used in the IPI PAN corpus is different from the Czech tagset. For example, some Polish pronouns are tagged as adjectives because they have adjectival inflections, whereas

No.	Description	Abbr.	No. of values		
			Czech	Russian	Polish
1	POS	P	12	12	12
2	SubPOS – detailed POS	S	75	42	20
3	Gender	g	11	5	5
4	Number	n	6	4	5
5	Case	c	9	8	9
6	Possessor's Gender	G	5	4	2
7	Possessor's Number	N	3	3	2
8	Person	p	5	5	5
9	Tense	t	5	5	5
10	Degree of comparison	d	4	4	4
11	Negation	a	3	3	3
12	Voice	v	3	3	3
13	Unused		1	1	1
14	Unused		1	1	1
15	Variant, Style	V	10	2	1

Table 1: Overview and comparison of the Slavic tagsets

the Czech system makes more fine-grained distinctions. The map between the original Polish tags and their multiple translations was randomly selected. Traditional categories which are represented only partially in the IPI PAN tagset include tense, mood, and voice. Table 1 summarizes the number of values for each

Language	Tagset size
Czech	4290+
Polish	800+
Russian	900+

Table 2: Size of Slavic tagsets

attribute of the Polish tag.

3.2 *Romance tagsets*

Experiments were also conducted with Spanish, Portuguese, and Catalan (see Tables 3 and 4).

The original CLiC-TALP tagset Civit (2000) developed for Spanish has been translated into the current system. The new tag is positional with 11 slots, each responsible for a particular morphological category. The original tagset is also positional, but unlike the new system, the attributes of the subsequent position depend on the first (POS). Here, the attribute positions are fixed for all categories, but the attribute values depend on POS and subPOS. To illustrate, for the CLiC-TALP tag *DA0CS0*, the values of the positions indicate it is 1) a determiner, 2) an article, 3) not a personal, 4) of indeterminate gender, 5) singular, and 6) not possessive; whereas the tag *VMIF1P0* stands for 1) verb, 2) main, 3) indicative, 4) future, 5) 1st person, 6) plural, and 7) undefined gender. So, CLiC-TALP tags can be of different lengths and each position can be occupied by a different attribute depending on the POS category. In the new system, all tags are of the same length, and the position and interpretation of an attribute does not depend on the value of any other attribute. The boolean value '0' ('undetermined for this particular form') from '-' ('inapplicable for this category') is distinguished. Thus, those same example tags would be translated into *DACS-0-0---* and *VM0P---1FI-* in this system, where each position, no matter what POS that is, stands for 1) POS, 2) detailed POS, 3) gender, 4) number, 5) case, 6) possessor's number, 7) form, 8) person, 9) tense, 10) mood, and 11) participle. This system makes comparison and evaluation easy. Table 3 provides an overview of the tagset.

The CLiC-TALP corpus has a portion of Catalan annotated with a similar tagset to that described above (for Spanish). Similar translations were implemented for Catalan as for Spanish, and the results are summarized in Table 3.

For the experiments with Portuguese, no accessible corpus annotated with detailed morphological information was available. Therefore, a completely new

No.	Description	Abbr.	No. of values		
			Spanish	Portuguese	Catalan
1	POS	p	14	14	14
2	SubPOS – detailed POS	s	29	30	29
3	Gender	g	6	6	6
4	Number	n	5	5	5
5	Case	c	6	6	6
6	Possessor's Number	N	4	4	4
7	Form	f	3	3	3
8	Person	P	5	5	5
9	Tense	t	7	8	7
10	Mood	m	7	7	7
11	Participle	R	3	3	3

Table 3: Overview and comparison of the Romance tagsets

tagset was created for specifying the paradigms and annotating a test corpus. The tagset for Portuguese is very similar to the tagsets described for Catalan and Spanish. From Table 3, one can see that the Spanish, Portuguese, and Catalan tagsets in the majority of cases use the same values. However, some differences are unavoidable. For instance, the pluperfect is a compound verb tense in Spanish, but a separate word that needs a tag of its own in Portuguese. Notice there are 6 possible values for the gender position in all the tagsets. These correspond to 'M' (masculine), 'F' (feminine), 'N' (neuter, for certain pronouns), 'C' (common, either M or F), '0' (unspecified for this form within the category), and '-' (the category does not distinguish gender).

4 Corpora

Several corpora were used for the experiments with the Slavic and Romance languages.

Language	Tagset size
Spanish	280+
Portuguese	280+
Catalan	280+

Table 4: Size of Romance tagsets

4.1 *Slavic corpora*

The experiments with the Slavic languages are based on several resources. The size of the corpus for each language was 2K tokens.

The Czech corpus used is the Prague Dependency Treebank (Bémová et al. (1999)). The corpus is a collection of newspaper articles.

The Polish corpus used in the experiments is 2K tokens of the IPI PAN Polish corpus Przepiórkowski (2004), translated into the current system as described above.

For Russian, we manually annotated 1,788 word tokens of the Russian translation of Orwell's *1984* taken from Multext-East (Erjavec (2004)).

4.2 *Romance corpora*

The Spanish corpus is 2K tokens of the Spanish section of CLiC-TALP. The CLiC-TALP tagset was automatically translated into the current system for easier detailed evaluation and comparison.

For Portuguese, we used 2K of the PALAVRAS corpus Bick (2000), but the original tags were substituted with the positional system described above.

For Catalan, we used 2K tokens of the translated CLiC-TALP corpus.

5 Tagset size, tagset coverage

First, the coverage of the tagset for different text sizes is measured as well as the text coverage by a small number of high frequency tags.

Tables 5 and 6 provide information about the size of the tagsets and the corpora for the Slavic and the Romance languages used in these experiments, as well as for English.

	Ca	Cz	Pol	Por	Ru	Sp	En
Distinct tags in corpus	92	221	173	74	186	112	38
Tagset size	289	4,290+	800+	259	900+	282	45
Tokens	2,000	2,000	2,000	1,915	1,788	2,000	2,000
Types	736	1,102	1,119	629	856	830	836
Distinct bigrams	535	1,068	943	484	830	674	363
Distinct trigrams	1,180	1,723	1,674	1,140	1,435	1,435	1,037
$H(X)$	5.004	6.160	5.661	4.690	5.623	5.191	4.305
$I(X;Y)$	2.074	2.852	2.194	1.576	2.361	1.937	1.206
$I(Y;X)$	2.075	2.849	2.200	1.574	2.359	1.934	1.204
Average tag/token ambiguity	1.109	1.165	1.219	1.229	1.159	1.124	1.072
Average tag/token, context w = -1	1.024	1.022	1.044	1.052	1.031	1.024	1.013
Average tag/token, context w = -2	1.009	1.002	1.006	0.017	1.006	1.005	1.006
Average tag/token, context w = +1	1.035	1.010	1.020	1.050	1.015	1.028	1.011
Average tag/token, context w = +2	1.007	1.001	1.001	1.009	1.006	1.008	1.004

Table 5: The corpus and detailed tagset size, n-gram counts, entropy (H), mutual information (I), and average tag/token ambiguity: Slavic, Romance, English.

The former table provides this data for the full tagset, and the latter for the reduced tagset, where the reduced set is limited only to POS+SubPOS (i.e. is comparable to English).

The potential sparsity problem can be seen by comparing the number of distinct tags that appear in a 2K-token corpus to the number of tags in the whole tagset. Figures 1, 2, 3, and 4 illustrate the coverage of the tagset in a 2K-token corpus for each language. The graphs in Figures 1 and 2 depict the number of distinct tags learned using the detailed and reduced tagsets, respectively, as the

	Ca	Cz	Pol	Por	Ru	Sp	En
Distinct tags in corpus	27	39	20	20	36	28	38
Tagset size	30	75	29	29	42	30	45
Tokens	2,000	2,000	2,000	1,915	1,788	2,000	2,000
Types	736	1,102	1,119	629	856	830	836
Distinct bigrams	213	279	155	175	263	235	363
Distinct trigrams	634	750	588	561	758	760	1,037
$H(X)$	3.629	3.462	3.124	3.369	3.681	3.768	4.305
$I(X;Y)$	1.103	0.606	0.322	0.737	0.700	0.917	1.206
$I(Y;X)$	1.105	0.603	0.319	0.734	0.699	0.915	1.204
Average tag/token ambiguity	1.097	1.036	1.057	1.210	1.032	1.105	1.072
Average tag/token, context w = -1	1.031	1.009	1.020	1.064	1.007	1.023	1.013
Average tag/token, context w = -2	1.013	1.003	1.012	1.030	1.005	1.009	1.006
Average tag/token, context w = +1	1.046	1.006	1.017	1.084	1.003	1.045	1.011
Average tag/token, context w = +2	1.026	1.002	1.002	1.024	1.001	1.018	1.004

Table 6: The corpus and reduced tagset size, n-gram counts, entropy (H), mutual information (I), and average tag/token ambiguity: Slavic, Romance, English.

corpus size grows. Figure 1 shows that in the case of the large tagset (especially for the Slavic languages, whose tagsets are the most detailed), the number of the new tags continues to grow with the size of the corpus, whereas in the case of the reduced tagset, shown in Figure 2, after processing the first 1K word tokens, new tags are not discovered anymore. The figures in 3 and 4 support the same observation — the percentage of the tagset covered by the corpus stops growing for English after the first 1K word tokens are processed. More than 80% of the whole tagset is discovered at that point. For the other languages, the discovery of new tags does not proceed as fast. For instance, after the first 2K tokens of the text are processed, more than 90% of the Czech tagset are still unknown.

From the tables and the graphs presented thus far, it is evident that languages such as Czech, Russian, and Polish will require a larger training corpus in order to learn the information about the occurrences of a significant subset of all possible

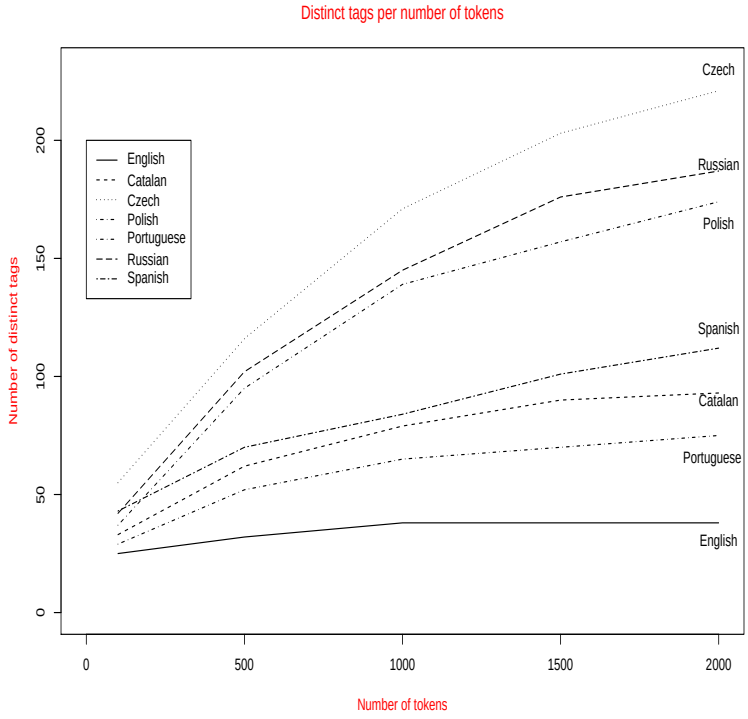

Figure 1: The number of distinct tags plotted against the number of tokens for the detailed tagset.

tags. To a lesser extent than the Slavic languages, Spanish, Portuguese, and Catalan also show the same pattern. They are also more data-sparsity prone compared to English. More training data will be needed for these languages as well.

6 How much training data is necessary?

The next questions to ask are whether it is indeed necessary to see all the possible tags and how much data can be covered just by a set of the most frequent tags. To explore these issues, the first five most frequent tags for each language were selected and the percentage of the corpus which would be covered by such a set

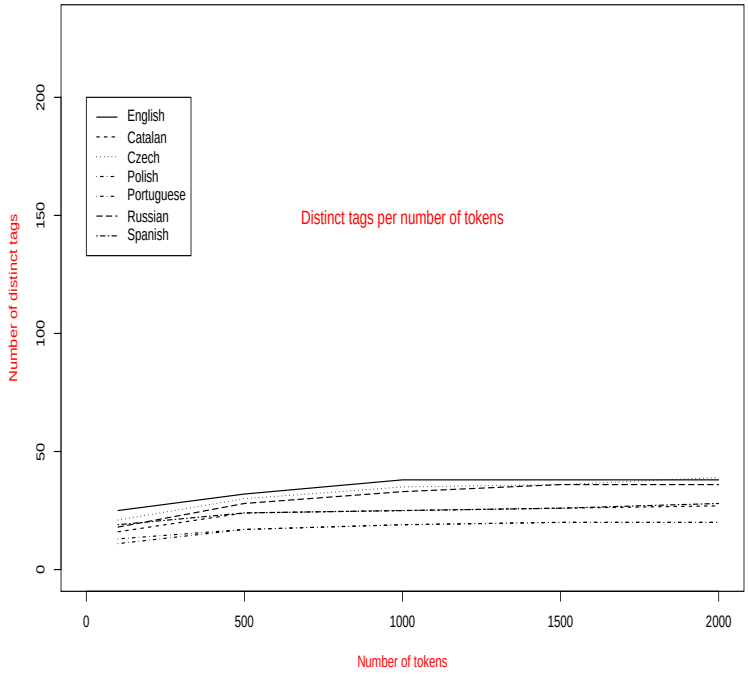

Figure 2: The number of distinct tags plotted against the number of tokens for the reduced tagset.

was calculated. Figures 5 and 6 illustrate the results. The results for the detailed tagsets (Figure 5) are comparable — a range of 30-50% coverage of the corpus seems to be constant across languages and independent of the size of the corpus (e.g. compare the results for 500 tokens, 1,000 tokens etc.). For Czech, for instance, only 30% of the corpus is covered by the five most frequent tags. For the reduced tagset, the coverage is better, as much as 80%, but generally, the graph shows that the increase in the text size does not affect the text coverage of the five most frequent tags.

Entropy $H(Y)$ of the tagsets was also measured using the formula in (1),

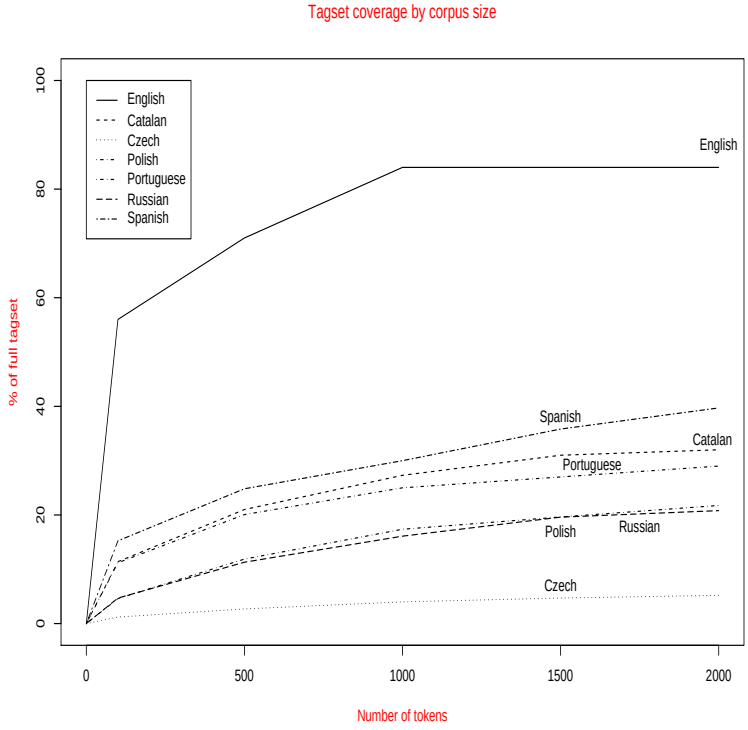

Figure 3: The percentage of the tagset covered by the number of tokens for the detailed tagset.

where Y denotes a random variable over $Tagset$ and $y \in Tagset$.

(1) $$H(Y) = \sum_{y \in Y} p(y) log \frac{1}{p(y)}$$

Intuitively, entropy is a measure of the size of the 'search space' consisting of the possible tags and their associated probabilities. The higher the entropy, the larger the 'search space'. Tables 5 and 6 give the results of the entropy calculations for each tagset and language. These entropy scores provide an additional piece of evidence that Czech, Polish, and Russian, followed by Catalan, Spanish, and Portuguese, are the most challenging languages for tagging (if we use detailed

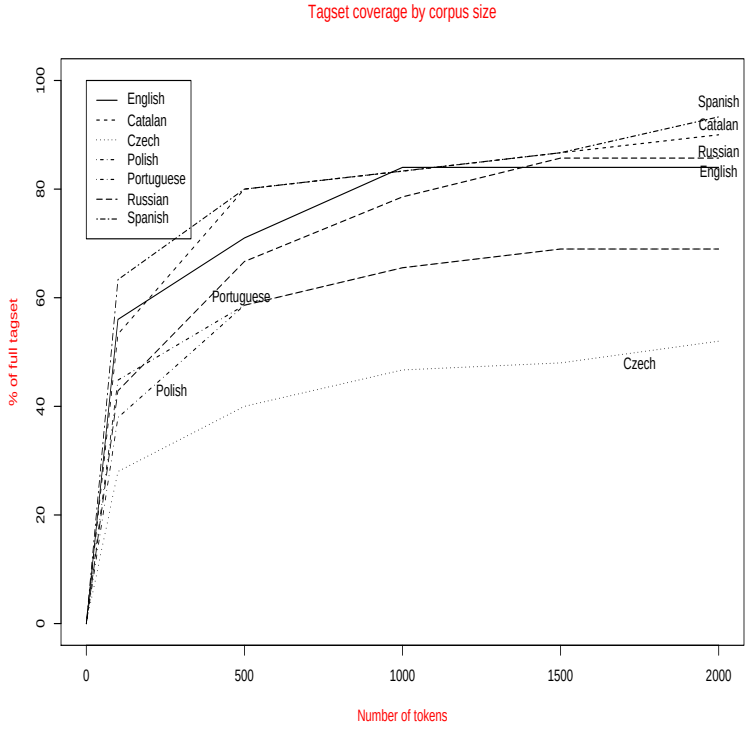

Figure 4: The percentage of the tagset covered by the number of tokens for the reduced tagset.

tagsets).

The discussion above suggests that a large tagset creates a larger 'search space'. In addition, the figures show that even though the tagsets for morphologically rich languages are larger than the English tagset, the percentage of the corpus covered by the five most frequent tags is only slightly higher for English (see Figure 5). To investigate this further, the *accession rate* for new tags, i.e. the rate at which new tags are discovered as more text is processed was examined (see e.g. Krotov et al. (1999) for further explanation of accession rates). One might expect that as more text is processed, the number of new tags added per text will be smaller. The

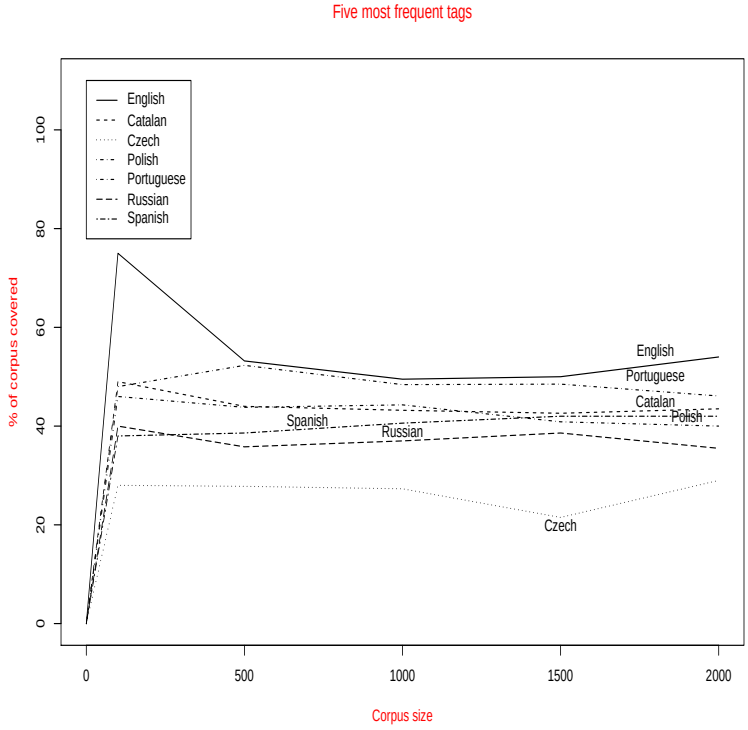

Figure 5: The percentage of the corpus covered by the 5 most frequent tags for the detailed tagset.

accession rate is measured for both the detailed and reduced tagsets. The results, plotted in Figure 7 and Figure 8, show that tag accession drops significantly after the first 100-200 tokens of the text are processed, but then proceeds at a relatively constant rate throughout processing of the remaining 2K tokens corpus. Given that the accession rate is measured on rather small corpora, strong claims cannot be made as to whether the accession rate becomes constant after processing the first 1,500 tokens of text (for the detailed tagset) or the first 800 tokens (for the reduced tagset). Clearly, it slows down significantly, which means that the discovery of new tags does not grow with the size of the corpus. This fact suggests

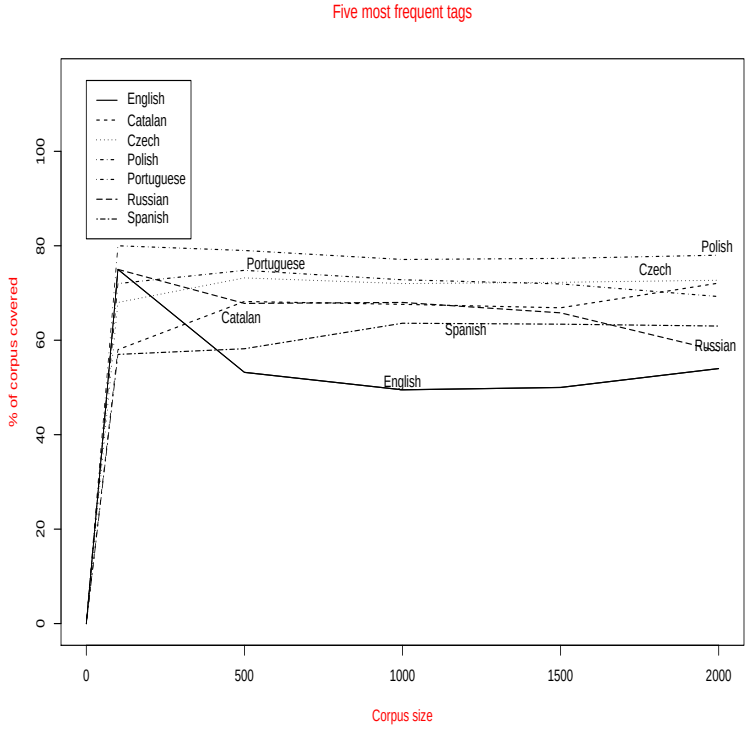

Figure 6: The percentage of the corpus covered by the 5 most frequent tags for the reduced tagset.

that even though a large tagset requires more training data, it is unclear how much data is actually needed to discover the full tagset.

7 Data sparsity, context, and tagset size

How much context contributes to reducing the 'search space' and how much the uncertainty about tag_y is reduced due to knowing about the preceding tag tag_x was also measured. For that, the mutual information $I(X;Y)$ is calculated as in (2), where X denotes a random variable over the set of tags that occur in the first position of all tag bigrams in a corpus, and Y denotes a random variable over the

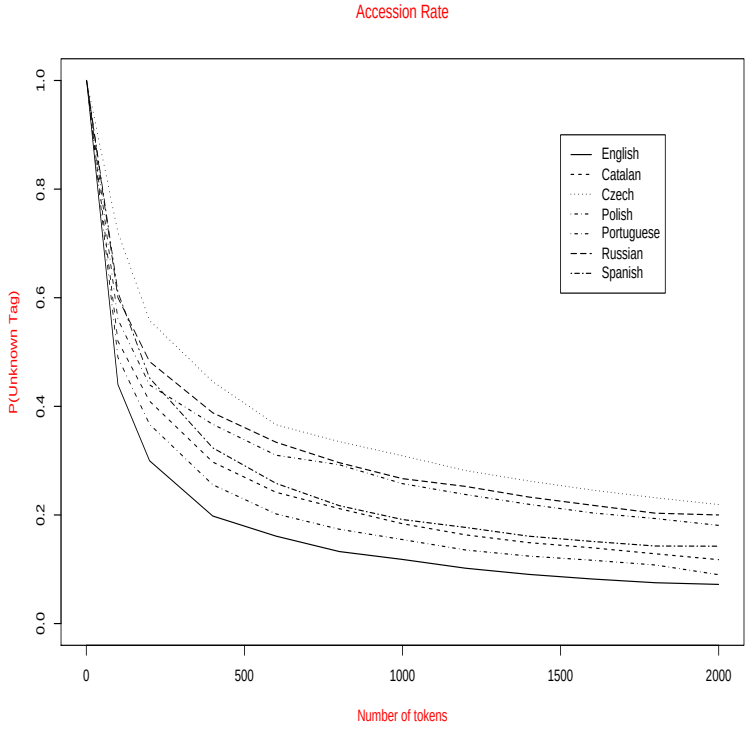

Figure 7: Accession rate for the detailed tagset.

set of tags that occur in the second position of all tag bigrams in the same corpus. The results of the mutual information calculations are summarized in Tables 5 and 6.

$$(2) \qquad I(X;Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}.$$

The higher the $I(X;Y)$ score is, the more dependent the current tag is on the previous tag. Comparing the mutual information scores for the detailed and reduced tagset for the inflected languages clearly shows that the dependence is greater in the case of the detailed tagset. This means that by reducing the tagset for inflected languages, important information is lost about agreement features (gender, num-

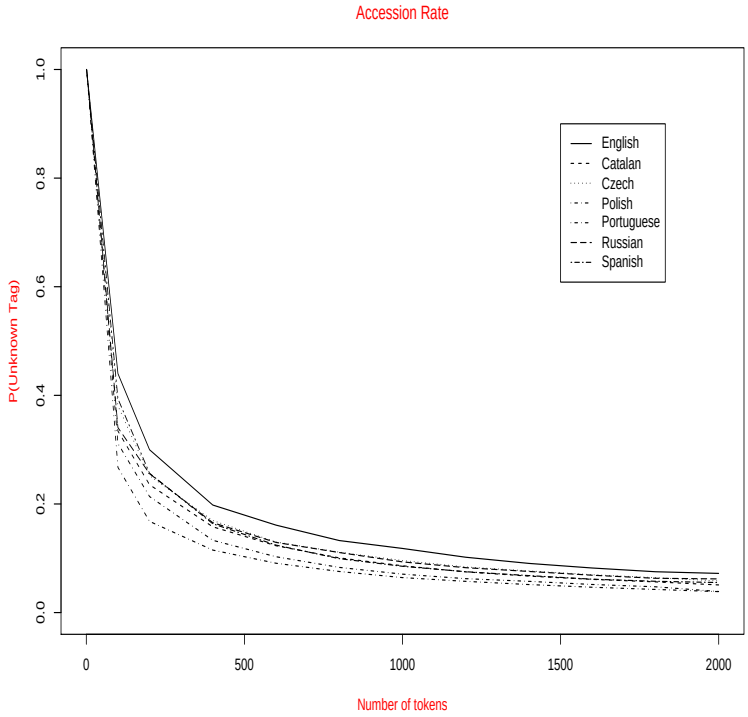

Figure 8: Accession rate for the reduced tagset.

ber, case, etc.), which might, in turn, bring about the reduction in overall tagging accuracy (as is indeed reported in Elworthy (1995)). Among these five inflected languages used in the experiments, Portuguese is the one for which the knowledge about the preceding tag helps the least. This fact suggests that a tagging approach which relies on the preceding context will be less efficient for Portuguese than for the rest of the languages. For the reduced tagsets, the greatest dependency between two tags is for English, a morphologically poor language. For the other five languages, a preceding tag is not as helpful for predicting the current tag when the reduced tagset is used.

8 Summary

This paper investigated some properties of Slavic and Romance languages and their tagsets. These properties were compared with those of English, a well-known and studied case. The comparison suggests that the data sparsity problem for the languages with a large tagset seems to be real. This is observable in the relationship between the corpus size and the number of new tags discovered. This is an expected observation. The surprising outcome of the experiments is that for Catalan, Czech, Polish, Portuguese, and Russian, the knowledge about the preceding tag_{n-1} reduces the uncertainty about tag_n if the detailed tagset is used. Recall that the detailed tagset contains the information about case, gender, number, and other important agreement features. But when the tagset is reduced to the size of the English tagset (eliminating the detailed information), the mutual information score drops significantly for the inflected languages. Compared to the English case, it seems that the two reduced adjacent tags for the Slavic and Romance languages are relatively independent of each other. This fact suggests that using a detailed tagset for languages such as Czech or Portuguese is beneficial and that reduction in the tagset will not necessary lead to better tagging results. In addition, even though the inflected languages are considered to be relatively word-order free, the adjacent information seems to be helpful for reducing the tag/token ambiguity. This another interesting result of the investigation supports the existence of a relatively fixed order of syntactic constituents in so-called "free word order" languages and provides an additional argument in favor of using the n-gram techniques for tagging these languages.

Here, we will not repeat experiments described in Hana et al. (2004); Feldman et al. (2006b,a); Hana et al. (2006) that deal with tagging the languages discussed in the present paper and with cross-language annotation transfer. The tagging experiments used the TnT (a tri-gram) tagger (Brants (2000)) and the results reported in that work prove that the predictions made in the current paper are correct.

Acknowledgements

I thank Sandra Maria Aluísio, Gemma Boleda, Toni Badia, Lukasz Debowski, Maria das Graças Volpe Nunes, Jan Hajic, Ricardo Hasegawa, Vicente López, Maria das Graças Volpe Nunes, Lluís Padró, Carlos Rodríguez Penagos, Adam Przepiórkowski, and Martí Quixal for the help with the corpora. I also thank Chris Brew and Jirka Hana for the invaluable help and suggestions.

References

Bémová, A., J. Hajič, B. Hladká, and J. Panevová (1999). Morphological and syntactic tagging of the Prague Dependency Treebank. In *Proceedings of Association pour le Traitement Automatique des Langues (ATALA) Workshop*, pp. 21–29. Paris, France.

Bick, E. (2000). *The parsing system PALAVRAS: automatic grammatical analysis of Portuguese in a Constraint-Grammar Framework*. Ph. D. thesis, University of Aarhus, DK.

Brants, T. (2000). TnT - A statistical part-of-speech tagger. In *Proceedings of 6th Applied Natural Language Processing Conference and North American chapter of the Association for Computational Linguistics annual meeting (ANLP-NAACL)*, pp. 224–231.

Civit, M. (2000). Guía para la anotación morfológica del corpus CLiC-TALP (Versión 3). Technical Report WP-00/06, X-Tract Working Paper. Centre de Llenguatge i Computaci (CLiC), Barcelona, Catalunya.

Cutting, D., J. Kupiec, J. Pedersen, and P. Sibun (1992). A Practical Part-of-speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing (ANLP)*, pp. 133–140. Association for Computational Linguistics.

Elworthy, D. (1995). Tagset design and inflected languages. In *7th Conference of the European Chapter of the Association for Computational Linguistics (EACL), From Texts to Tags: Issues in Multilingual Language Analysis SIGDAT Workshop*, Dublin, pp. 1–10.

Erjavec, T. (2004). MULTEXT-EAST version 3: multilingual morphosyntactic specifications, lexicons and corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Paris, France, pp. 1535–1538.

Feldman, A., J. Hana, and C. Brew (2006a). A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

Feldman, A., J. Hana, and C. Brew (2006b). Experiments in cross-language morphological annotation transfer. In *Proceedings of Computational Linguistics and Intelligent Text Processing, CICLing*, Lecture Notes in Computer Science, Mexico City, Mexico, pp. 41–50. Springer-Verlag.

Hajič, J. (2004). *Disambiguation of rich inflection: computational morphology of Czech*. Prague, Czech Republic: Charles University Press.

Hana, J., A. Feldman, L. Amaral, and C. Brew (2006). Tagging Portuguese with a Spanish tagger using cognates. In *Proceedings of the Workshop on Cross-language Knowledge Induction hosted in conjunction with the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy, pp. 33–40.

Hana, J., A. Feldman, and C. Brew (2004). A Resource-light approach to Russian morphology: tagging Russian using Czech resources. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, Barcelona, Spain, pp. 222–229.

Krotov, A., M. Hepple, R. Gaizauskas, and Y. Wilks (1999). Evaluating two methods for treebank grammar compaction. *Natural Language Engineering 5*(4), 377–394.

Przepiórkowski, A. (2004). *The IPI PAN corpus: preliminary version*. Warsaw, Poland: Institute of Computer Science (IPI), Polish Academy of Science (PAN).