A Cross-language Approach to Rapid Creation of New Morpho-syntactically Annotated Resources

Anna Feldman, Jirka Hana, Chris Brew

Department of Linguistics
The Ohio State University
USA
{afeldman,hana,cbrew}@ling.osu.edu

Abstract

We take a novel approach to rapid, low-cost development of morpho-syntactically annotated resources without using parallel corpora or bilingual lexicons. The overall research question is how to exploit language resources and properties to facilitate and automate the creation of morphologically annotated corpora for new languages. This portability issue is especially relevant to minority languages, for which such resources are likely to remain unavailable in the foreseeable future. We compare the performance of our system on languages that belong to different language families (Romance vs. Slavic), as well as different language pairs within the same language family (Portuguese via Spanish vs. Catalan via Spanish). We show that across language families, the most difficult category is the category of nominals (the noun homonymy is challenging for morphological analysis and the order variation of adjectives within a sentence makes it challenging to create a realiable model), whereas different language families present different challenges with respect to their morpho-syntactic descriptions: for the Slavic languages, case is the most challenging category; for the Romance languages, gender is more challenging than case. In addition, we present an alternative evaluation metric for our system, where we measure how much human labor will be needed to convert the result of our tagging to a high precision annotated resource.

1. Introduction

Morpho-syntactically annotated corpora are crucial for many language processing tasks. Applications include syntactic parsing, stemming, text-to-speech synthesis, word-sense disambiguation, information extraction. Despite the importance of morphological tagging, there are many languages that lack annotated resources of this kind, mainly due to the lack of training corpora which are usually required for applying standard statistical taggers.

In this paper, we describe a cross-language method that requires neither training data of the target language nor bilingual lexicons or parallel corpora. We report results of the experiments done on Slavic (Czech and Russian) and Romance (Spanish, Portuguese, Catalan) languages. The overall research question is how to exploit language resources and properties to facilitate and automate the creation of morphologically annotated corpora for new languages. This portability issue is especially relevant to minority languages, for which such resources are likely to remain unavailable in the foreseeable future. From the theoretical point of view, we want to understand and isolate general properties of languages that seem to make a difference in the cross-language transfer approach. We compare the performance of our system on languages that belong to different language families (Romance vs. Slavic), as well as different language pairs within the same language family (Portuguese via Spanish vs. Catalan via Spanish). We show that across language families, the most challenging category is the category of nominals, whereas morpho-syntactic difficulties vary depending on a language family: for Slavic languages, case is the most challenging category; for Romance languages, gender is more difficult than case. In addition, we present an alternative evaluation metric for our system, where we measure how much human labor will be needed to convert the result of our tagging to a high precision annotated resource.

2. Our Approach

The details of our method are described in (Hana et al., 2004; Hana et al., 2006; Feldman et al., 2006). In a nutshell, we train a second-order Markov model tagger (TnT, (Brants, 2000)) on a related source language, apply a resouce-light morphological analyzer (Hana, 2005) to the target language, and then combine the two sources of information in various ways to create a tagger for the target language.

3. Resources

In our work we do not rely on training data for the target languages; instead we approximate the target language model by a model trained on a related language. We use Czech for processing Russian, and Spanish for Portuguese and Catalan. The following sections describe the resources.

3.1. Experiments with Slavic languages

3.1.1. Corpora

For the experiments described below we use 630K tokens of the morphologicallay annotated Prague Dependency Treebank (Bémová et al., 1999). For development purposes, we selected and morphologically annotated (by hand) a small portion from the Russian translation of Orwell's *1984*. This corpus contains 1858 tokens (856 types).

We also acquire a lexicon of Russian automatically, as described in (Hana et al., 2004; Hana et al., 2006; Feldman et al., 2006). For that we use a large raw corpus, the Uppsala Russian Corpus (1M tokens), which is freely availabe from Uppsala University: `www.slaviska.uu.se/ryska/corpus.html`.

3.1.2. Knowledge Encoding

Our morphological analyzer captures just a few textbook facts about the Russian morphology, excluding the majority of exceptions and including information about basic declension and conjugation classes of nouns and verbs, respectively. In total, our database contains 80 paradigms. We use (Wade, 1992) for encoding this information. Based on this reference grammar text, we also created a list of closed class words, which contains about 800 items. In general, the closed class words can be derived either from a reference grammar book, or can be elicited from a native speaker. This does not require native-speaker expertise or intensive linguistic training.

3.1.3. Tagset

We adopted the Czech tag system (Hajič, 2000) for Russian and Polish. Every tag is represented as a string of 15 symbols each corresponding to one morphological category (Hana et al., 2004). A comparison of the tagsets is given in Table 1. 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 – 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 to capture only the core of Russian morphology, as we need it for our primary purpose of demonstrating portability and feasibility of our technique.

3.2. Experiments with Romance languages

3.2.1. Corpora

The Spanish corpus we use for training the transition probabilities as well as for obtaining Spanish-Portuguese or Spanish-Catalan cognate pairs is a fragment (106,124 tokens, 18,629 types) of the Spanish section of CLiC-TALP (Torruella, 2002). CLiC-TALP is a balanced corpus, containing texts of various genres and styles. We automatically translated the CLiC-TALP tagset into our system for easier detailed evaluation and comparison.

For automatic Portuguese lexicon acquisition, we use the NILC corpus, [1] containing 1.2M tokens. For automatic Catalan lexicon acquisition, we use a raw corpus (63M tokens) obtained by collecting “El Periodico” newspaper texts avilable at `www.elperiodico.es`[2].

We also have a development corpus for Catalan. We translated the Catalan system into ours and used 2K tokens for tuning parameters of our system.

3.2.2. Knowledge encoding

For creating a list of morphological paradigms for Portuguese, we used (Perinin, 2002)’s reference grammar book. For the Catalan paradigms we use (Wheeler et al., 1999). Our Portuguese database contains 38 paradigms, whereas the Catalan morphology contains 30 paradigms.

No. Slavic	No. Romance	Description	No. of values				
			Cz	Ru	Sp	Po	Ca
1	1	POS	12	12	14	11	11
2	2	SubPOS	75	32	30	29	30
3	3	Gen	11	5	6	6	6
4	4	Num	6	4	5	5	5
5	5	Case	9	8	6	6	6
6		Poss’s Gen	5	4			
7	6	Poss’s Num	3	3	4	4	4
	7	Form			3	3	3
8	8	Pers	5	5	5	5	5
9	9	Tense	5	5	7	9	7
10		Deg of Comprs	4	4			
	10	Mood			8	9	7
11		Neg	3	3			
	11	Prtcpl			3	3	3
12		Voice	3	3			
13		Unused	1	1			
14		Unused	1	1			
15		Variant	10	2			

Table 1: Overview and comparison of the tagsets

We also made a list of closed class words: 450 for Portuguese, and 500 for Catalan. These mainly contain prepositions, conjunctions, some pronouns, and adverbs.

We should mention that the paradigms for Portuguese were created by a native speaker, whereas the paradigms for Catalan were encoded by a linguist who had no training in Romance languages.

3.2.3. Tagset

For Spanish, Portuguese, and Catalan, we use positional tagsets developed on the basis of the Spanish CLiC-TALP tagset (Torruella, 2002). Every tag is a string of 11 symbols each corresponding to one morphological category. When possible, the Spanish, Portuguese and Catalan tagsets 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. The Spanish tagset has 282 tags; the Portuguese tagset contains 259 tags; and Catalan has 289[3].

4. Languages

A deep contrastive analysis of all the languages used in our experiments is far beyond the scope of this paper. However, we would like to mention just a number of the most important facts.

4.1. Romance vs. Slavic

In our work we use languages from the Slavic family (Russian and Czech), and languages from the Romance fam-

[1] Núcleo Interdisciplinar de Lingüística Computacional; available at `http://nilc.icmc.sc.usp.br/nilc/`, we used the version with POS tags assigned by PALAVRAS. We ignored the POS tags.

[2] Note that this newspaper is published in Spanish and Catalan, and the Catalan version is obtained via a Machine Translation system plus post edition and correction. Thus, the Catalan version might appear more Spanish-like.

[3] Notice that we have 6 possible values for the gender position: M (masc.), F (fem.), N (neutr., for certain pronouns), C (common, either M or F), 0 (unspecified for this form within the category), - (the category does not distinguish gender)

ily (Spanish, Portuguese, and Catalan). We use Czech as a source language for tagging Russian; in the experiments with the Romance languages, Spanish is the source language. Unlike Slavic languages, which have rich inflectional morphology and are constituent order free, Romance languages have lost the declension system of Classical Latin, and as a result have a relatively rigid sentence structure (still not as rigit as English) and make extensive use of prepositions.

Slavic and Romance languages have some properties in common. Nouns, verbs, adjectives, and adverbs are the major classes, each with a specific set of possible syntactic roles. Languages from both families have a complex system of word inflections to indicate syntactic relationships between words. The basic clause structure, both in Romance and in Slavic, consists of a verb and one or more noun arguments.

However, there are many differences between these language families. Romance languages have only two grammatical genders (masculine and feminine), whereas Slavic has three (masculine, feminine, and neuter). Adjectives usually follow the nouns they modify, whereas in Slavic, adjectives usually precede nouns. Romance languages have definite and indefinite grammatical articles, whereas Slavic languages mark (in)definiteness in other ways (e.g. word order).

4.2. Russian and Czech

Czech and Russian belong to different branche of the Slavic family (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.

Russian and Czech, however, differ in a number of properties. To mention a few, plural adjectives and participles in Russian, unlike Czech, do not distinguish gender. Verb negation in Czech in the majority of cases is expressed by prefixation, whereas in Russian it is very common to see a separate negative particle instead. In addition, reflexive verbs in Czech are formed by a verb followed by a reflexive clitic, whereas in Russian, the reflexivization is the affixiation process. Russian, unlike Czech, does not use an auxiliary to form past tense.

4.3. Spanish, Portuguese, and Catalan

Portuguese, Spanish, and Catalan belong to the Romance branch of the Indo-European language family. Galician, Spanish, and Ladino are the closest relatives of Portuguese among the Romance languages. Their speakers generally claim that the languages are mutually intelligible to some extent: while that may be in part a consequence of the extensive cultural ties between the Iberian countries, which inevitably lead to unconscious learning. It is certainly true that a speaker of any of the three languages can learn to read any of the other two just by practicing, without formal study of their grammar. Bilingualism is quite common in the border regions.

It is also claimed that a Portuguese speaker can understand Spanish better than the other way around. This alleged asymmetry could be due to to the general reduction of unstressed vowels in Portuguese, compared to Spanish. Portuguese differs from Spanish in orthography, and even more in phonology, grammar and vocabulary.

Catalan, Portuguese, and Spanish share a number of properties in common. They all have present, past perfect and past imperfect. For each tense there are six distinct inflections encoding each of the three persons and two numbers. There are two copula verbs from Latin *esse* and *stare*. In orthography, the letter *k* is rarely used in these languages – mostly for unassimilated foreign words and names. The plural formation is similar across these languages – by adding the suffix *s*.

Historically, Vulgar Latin split first into Catalan and Iberian Romance, which in turn, was divided up into Castillian (e.g. Spanish) and Gallo-Portuguese (e.g. Portuguese). As a result of this historical development, Catalan is farer from Spanish and Portuguese in its many linguistic properties. To name a few, Some Romance languages have lost the final usntressed vowels from the Latin roots, while others still retain them. Portuguese and Spanish have final vowels retained, while Catalan retains them only in feminine gender. There are also obvious lexical differences between Spanish-Portuguese and Spanish-Catalan pairs. For instance, the word for *nothing* in Portuguese and Spanish is *nada*, whereas in Catalan it is *res* (similar to the French *rien*).

5. Expectations

Many factors should be taken into account when estimating how good the performance of our system will be on a chosen language pair. These include the language properties in general (e.g. word order, morphological complexity), as well as the relationship between the source and the target language (e.g. how close they are in their word order and lexicon) and whether the source language makes fewer/more morpho-syntactic distinctions (either in the language itself or in the tagset). To go from a detailed tagset to a less detailed tagset is obviosly easier than other way around.

It is hard to assess qualitatively what language pair has the best chance. We have described the properties of the languages, the tagsets and the resources. For Russian, the paradigms were created by a native speaker, so were the Portuguese paradigms, whereas the Catalan paradigms were encoded by a person who did not know Catalan (or any related language). Czech and Russian are not mutually intelligible, whereas Portuguese and Spanish are claimed to be so. Both Czech and Russian have a large tagset, but Czech has more detailed morpho-syntactic descriptions. In the case of Portuguese, Spanish and Catalan, the source and the target morpho-syntactic descriptions are comparable. At the same time, the Romance languages use a much smaller tagset than the Slavic languages.

The quality of the tagging is obviously dependent on the quality of morphological analysis. The quality of the morphological analysis depends, on the paradigms and the ac-

Language	Ru	Po	Ca
recall	90.4	98.1	94.8
avg ambig (tag/word)	3.1	3.5	3.9

Table 2: Evaluation of Morphological analysis

quired lexicon, which is in turn dependent on the quality and the size of data for lexicon acquisition.

The comparison of the recall and the ambiguity of the morphological analyses is given in Table 2. We calculate the recall by running our morphological analyzer and assuming an oracle tagger which picks the right tag out of the possible set of tags suggested by the morphological analyzer. This means that the upperbound performance of our system is 90.4% for Russian; 98.1% for Portuguese and 94.8% for Catalan.

In addition, we measure how close the language pairs are. We train TnT on the source language and apply the resulting model directly to the target language (see Table 3) . The size of the training corpora for each language is approximately 100K tokens.

Source	Sp	Sp	Ru
Target	Po	Ca	Cz
Full Tag:	56.9	36.5	45.6
POS:	65.3	64.5	63.8
SubPOS:	61.7	42.8	59.9
gender:	70.4	75.0	63.9
number:	78.3	85.5	73.2
case:	93.8	94.6	62.8
person:	74.5	77.5	89.4
tense:	90.7	82.2	88.4

Table 3: Lowerbound: Source models directly applied to target languages

From Table 3, it is evident that the Spanish-Portuguese pair is the closest. Portuguese and Spanish share more than 50% linguistic properties, whereas the next pair that shares many linguistic properties is Russian-Czech. The Spanish-Catalan pair is the most distant one. So, we expect that the tagging result on Portuguese will be the best, and we realize that tagging Catalan is the most challenging task. The evaluation reveals that the gender slot is challenging across all languages, and case is a difficult category for Russian.

6. Experiments

6.1. Basic approach

Our basic approach consists of training transitions on the source language, running the resource-light morphological analyzer (Hana, 2005) on the target language and using its output for creating evenly distributed emissions. The results of the tagging are summarized in Table 4 (where the *emiss* column says *e* (=even)). Tables 5, 6, and 7, report the tagging resuls on nouns, verbs, and adjectives, separately.

6.2. Cognates

Although it is true that forms and distributions of the target and the source language words are not the same, they are also not completely unrelated. As any Spanish speaker would agree, the knowledge of Spanish words *is* useful when trying to understand a text in Portuguese. The same is true for the other language pairs.

Many of the corresponding Portuguese and Spanish words are cognates, i.e. historically they descend from the same ancestor root or they are mere translations. We assume two things: (i) cognate pairs have usually similar morphological and distributional properties, (ii) cognate words are similar in form.

Obviously both of these assumptions are approximations:

1. Cognates could have departed in their meanings, and thus probably also have different distributions. For example, Spanish *embarazada* ‘pregnant’ vs. Portuguese *embaraçada* ‘embarrassed’.

2. Cognates could have departed in their morphological properties. For example, Spanish *cerca* ‘near’.adverb vs. Portuguese *cerca* ‘fence’.noun (from Latin *circa*, *circus* ‘circle’).

3. There are false cognates – unrelated, but similar or even identical words. For example, Spanish *salada* ‘salty’.adj vs. Portuguese *salada* ‘salad’.noun, Spanish *doce* ‘twelve’.numeral vs. Portuguese *doce* ‘candy’.noun

Nevertheless, we believe that these examples are true exceptions from the rule and that in majority of cases, the cognates would look and behave similarly. The borrowings, counter-borrowings and parallel developments of the various Romance languages have of course been extensively studied, and we have no space for a detailed discussion.

Identifying cognates For the present work, however, we do not assume access to philological erudition, or accurate target-source translations or even a sentence-aligned corpus. All of these are resources that we could not expect to he arguments. Similarly as (Yarowsky and Wicentowski, 2000), we assume that, in any language, vowels are more mutable in inflection than consonants, thus for example replacing *a* for *i* is cheaper that replacing *s* by *r*. In addition, costs are refined based on some well known and common phonetic-orthographic regularities in language pairs. However, we do not want to do a detailed contrastive morpho-phonological analysis, since we want our system to be portable to other languages. So, some facts from a simple grammar reference book should be enough.

Using cognates. Having a list of Source-Target cognate pairs, we can use these to map the emission probabilities acquired on the source corpus to the target language.

Let’s assume Source word w_s and Target word w_r are cognates. Let T_s denote the tags that w_s occurs within the Source corpus, and let $p_s(t)$ be the emission probability of a tag t ($t \notin T_s \Rightarrow p_s(t) = 0$). Let T_r denote tags assigned to the Target word w_r by our morphological analyzer, and the $p_r(t)$ is the even emission probability: $p_r(t) = \frac{1}{|T_r|}$.

Then we can assign the new emission probability $p'_r(t)$ to every tag $t \in T_r$ in the following way (followed by normalization):

$$p'_r(t) = \frac{p_s(t) + p_r(t)}{2} \tag{1}$$

7. Evaluation

We report the results of the following experiments:

1. Lowerbounds: TnT trained on the source language and applied directly to the target language

2. TnT trained on Catalan and applied to Catalan (for comparison of the performance of the monolingual model vs. the cross-lingual approach)

3. Transitions trained on the source language; target language emissions obtained by running our morphological analyzer and assuming the uniform distribution

4. Transitions trained on the source language; target language emissions, enhanced by cognates (as described in section 6.2.)

7.1. Resources

For testing the performance of our system we use the following corpora:

1. *Russian*: 4K tokens of Orwell's *1984*, annotated by hand.

2. *Portuguese*: 1.8K tokens of NILC, annotated by hand.

3. *Catalan*: 20K tokens of CLiC-TALP, translated into our tag system.

7.2. Performance

We summarize the performance of our system on the test corpora overall, across all categories, as well as report detailed evaluations on three major parts of speech: nouns, verbs, and adjectives. In the gold standard Catalan corpus, some compound words were tagged as a unit (e.g. *Centre_Excursionista_de_Banyoles* is tagged with one tag corresponding to proper names). We therefore do not have reliable gold standard tags for the indidual components of these compound words, so these words were excluded from the evaluation.

Table 4 shows that the lowerbounds of the Catalan, Portuguese, and Russian. Examining these values, we conclude that the closest language pair is Portuguese-Spanish, whereas the most distant one is Catalan-Spanish. This supports our linguistic intuitions. Interestingly enough, when applying the basic (even emissions) model to Portuguese, we have 47% error reduction rate comparing to the Spanish model applied directly to Portuguese, which suggests that morphological analysis is an important step in the cross-lingual tagging process. For Catalan and Russian, the error reduction rate is even more significant. In Tables 5, 6, 7, we report how the basic approach affects each lexical category individually. Notice that for the Romance languages, verbs are as challenging as nominals, whereas the Slavic languages seem to have a more straightforward verb morphology. The most difficult category for Slavic languages is adjectives. The reason is that adjectives seem to have a larger variation in their order in a sentence.

Target	Ca				Po			Ru		
trans	Ca	Sp	Sp	Sp	Sp	Sp	Sp	Cz	Cz	Cz
emiss	Ca	Sp	e$_{Ca}$	cog	Sp	e$_{Po}$	cog	Cz	e$_{Ru}$	cog
Full Tag:	96.0	36.5	70.7	75.2	56.9	77.2	82.1	45.6	78.6	80.4
POS:	97.5	64.5	80.2	83.3	65.3	84.2	87.6	63.8	92.7	92.3
SubPOS:	96.8	42.8	77.9	80.9	61.7	83.5	87.0	59.9	90.9	90.7
Gen:	98.1	75.0	81.9	85.3	70.4	87.3	90.2	63.9	91.1	92.5
Num:	98.9	85.5	89.7	90.2	78.3	95.3	96.0	73.2	94.0	94.8
Case:	99.3	94.6	97.8	97.8	93.8	96.8	97.2	62.8	87.6	88.1
Pers:	98.5	77.5	87.1	89.0	74.5	91.2	92.7	89.4	98.9	99.0
Tense:	99.4	82.2	90.7	92.6	90.7	95.1	96.1	88.4	98.8	98.7

Table 4: Accuracy: all categories

Target	Ca				Po			Ru		
trans	Ca	Sp	Sp	Sp	Sp	Sp	Sp	Cz	Cz	Cz
emiss	Ca	Sp	e$_{Ca}$	cog	Sp	e$_{Po}$	cog	Cz	e$_{Ru}$	cog
Full Tag:	94.8	43.0	40.5	51.3	65.2	60.8	70.7	30.3	65.8	69.4
POS:	97.1	69.8	53.3	63.3	80.6	75.3	81.9	77.7	94.5	95.0
SubPOS:	96.7	58.6	50.6	60.1	76.1	75.1	81.6	77.7	94.5	95.0
Gen:	96.4	62.1	59.2	67.3	74.2	72.2	78.1	51.0	83.5	86.8
Num:	99.0	89.1	79.8	81.0	87.8	97.5	97.7	72.5	90.1	91.2
Case:	100.0	99.8	100.0	100.0	100.0	100.0	99.8	46.0	76.9	78.5

Table 5: Accuracy: Nouns

Target	Ca				Po			Ru		
trans	Ca	Sp	Sp	Sp	Sp	Sp	Sp	Cz	Cz	Cz
emiss	Ca	Sp	e$_{Ca}$	cog	Sp	e$_{Po}$	cog	Cz	e$_{Ru}$	cog
Full Tag:	95.8	31.5	67.2	71.4	33.2	78.4	82.4	47.2	87.6	91.0
POS:	97.6	67.8	82.9	85.4	48.7	91.5	93.0	68.2	98.3	98.3
SubPOS:	97.2	44.3	81.0	83.6	37.7	90.5	92.0	51.9	95.7	96.6
Gen:	98.8	80.8	85.4	86.6	54.3	92.0	93.5	64.8	94.0	97.4
Num:	99.6	87.1	87.3	88.6	66.8	95.5	96.0	68.7	92.7	96.1
Pers:	96.7	66.6	71.1	75.2	38.7	81.9	85.9	81.1	95.7	96.6
Tense:	97.0	54.8	75.4	79.1	35.7	80.9	84.9	63.5	94.0	93.1

Table 6: Accuracy: Verbs

7.3. An alternative evaluation

Our goal is to provide methods for the rapid development of annotated resources. Clearly, given the present level of precision, we cannot be sure that the resources that we create will be usable without modification. This modification will require human intervention, but it is not immediately obvious how costly this intervention will be. As an *ad hoc* measure of the cost, we provide figures on the number of changes that would be required to transform the tagger's output into the desired gold standard tags. Table 8 gives the total number of atomic feature changes that are necessary to recreate the gold standard.

Target	Ca				Po			Ru		
trans	Ca	Sp	Sp	Sp	Sp	Sp	Sp	Cz	Cz	Cz
emiss	Ca	Sp	e$_{Ca}$	cog	Sp	e$_{Po}$	cog	Cz	e$_{Ru}$	cog
Full Tag:	89.4	35.8	24.0	47.9	60.3	58.5	68.3	11.9	53.0	55.6
POS:	90.8	49.5	68.4	76.5	71.5	68.3	76.0	26.5	80.8	80.8
SubPOS:	90.6	49.1	66.7	74.6	71.5	68.3	76.0	26.5	71.5	72.2
Gen:	96.9	63.8	50.8	71.5	87.2	80.3	88.0	50.3	89.4	89.4
Num:	98.7	88.9	92.5	93.7	94.4	94.5	95.6	84.1	93.4	94.0
Case:	100.0	99.1	99.2	99.2	98.9	96.2	95.1	41.1	75.5	76.8

Table 7: Accuracy: Adjectives

Model	Ca-even	Ca-cog	Ru-even	Ru-cog	Po-even	Po-cog
Changes	104504	103713	985	935	1605	1282

Table 8: Number of feature changes needed to recreate gold standard

7.4. Discussion

We have shown that potentially useful results are obtainable from an approach to bilingual lexicon creation that does not rely on parallel corpora or bilingual lexicons. Simple use of cognates is advantageous. Unsurprisingly, the approach works best for language pairs that are very closely related, such as Spanish and Portugese, and rather less well for less related languages, such as Catalan and Spanish. We also provide some quantitative basis for the widely shared anecdotal impression that gender is difficult. We also show that the case category is challenging for the Slavic languages. Further work could include an attempt to quantify the extent to which gender and case difficulties are due to pure lexical idiosyncrasy and the extent to which there are systematic differences which could reasonably be explained to a second-language learner.

8. Acknowledgments

We would like to thank Maria das Graças Volpe Nunes, Sandra Maria Aluísio, and Ricardo Hasegawa for giving us access to the NILC corpus annotated with PALAVRAS and to Carlos Rodríguez Penagos for letting us use the CLiC-TALP corpus.

9. 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.

Anna Feldman, Jirka Hana, and Chris Brew. 2006. Experiments in Cross-Language Morphological Annotation Transfer. In *Proceedings of Computational LInguistics and Intelligent Text Processing, CICLing*, Lecture Notes in Computer Science, pages 41–50. Springer-Verlag.

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

Jirka Hana, Anna Feldman, and Chris Brew. 2004. A Resource-light Approach to Russian Morphology: Tagging Russian Using Czech Resources. In *Proceedings of EMNLP (Empirical Methods for Natural Language Processing)*, pages 222–229.

Jirka Hana, Anna Feldman, Chris Brew, and Luiz Amaral. 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-2006)*.

Jirka Hana. 2005. Knowledge and labor light morphological analysis. Unpublished manuscript.

Mário A. Perinin. 2002. *Modern Portuguese: A Reference Grammar*. Yale University Press.

M. Torruella. 2002. Guía para la anotación morfológica del corpus CLiC-TALP (Versión 3). Technical Report WP-00/06, X-Tract Working Paper.

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

Max W. Wheeler, Alan Yates, and Nicolau Dols. 1999. *Catalan: A Comprehensive Grammar*. Routlege.

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.