Tagging Portuguese with a Spanish Tagger Using Cognates

Jirka Hana
Department of Linguistics
The Ohio State University
hana.1@osu.edu

Anna Feldman
Department of Linguistics
The Ohio State University
afeldman@ling.osu.edu

Chris Brew
Department of Linguistics
The Ohio State University
cbrew@acm.org

Luiz Amaral
Department of Spanish and Portuguese
The Ohio State University
amaral.1@osu.edu

Abstract

We describe a knowledge and resource light system for an automatic morphological analysis and tagging of Brazilian Portuguese.[1] We avoid the use of labor intensive resources; particularly, large annotated corpora and lexicons. Instead, we use (i) an annotated corpus of Peninsular Spanish, a language related to Portuguese, (ii) an unannotated corpus of Portuguese, (iii) a description of Portuguese morphology on the level of a basic grammar book. We extend the similar work that we have done (Hana et al., 2004; Feldman et al., 2006) by proposing an alternative algorithm for cognate transfer that effectively projects the Spanish emission probabilities into Portuguese. Our experiments use minimal new human effort and show 21% error reduction over even emissions on a fine-grained tagset.

1 Introduction

Part of speech (POS) tagging is an important step in natural language processing. Corpora that have been POS-tagged are very useful both for linguistic research, e.g. finding instances or frequencies of particular constructions (Meurers, 2004) and for further computational processing, such as syntactic parsing, speech recognition, stemming, word-sense disambiguation. *Morphological tagging* is the process of assigning POS, case, number, gender and other morphological information to each word in a corpus. 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.

Applications of taggers include syntactic parsing, stemming, text-to-speech synthesis, word-sense disambiguation, information extraction. For some of these getting all the tags right is inessential, e.g. the input to noun phrase chunking does not necessarily require high accuracy fine-grained tag resolution.

Cross-language information transfer is not new; however, most of the existing work relies on parallel corpora (e.g. Hwa et al., 2004; Yarowsky and Ngai, 2001) which are difficult to find, especially for lesser studied languages. 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 the results of the experiments done on Brazilian Portuguese and Peninsular Spanish, however, our system is not tied to these particular languages. The method is easily portable to other (inflected) languages. Our method assumes that an annotated corpus exists for the source language (here, Spanish) and that a text book with basic linguistic facts about the source language is available (here, Portuguese). We want to test the generality and specificity of the method. Can the systematic commonalities and differences between two genetically related languages be exploited for cross-language applications? Is the processing of Portuguese via Spanish different from the processing of Russian via Czech (Hana et al., 2004; Feldman et al., 2006)?

[1] We thank the anonymous reviewers for their constructive comments on an earlier version of the paper.

	Spanish	Portuguese
1. sg.	canto	canto
2. sg.	cantas	cantas
3. sg.	canta	canta
1. pl.	catamos	cantamos
2. pl.	cantais	cantais
3. pl.	cantan	cantam

Table 1: Verb conjugation present indicative: *-ar* regular verb: *cantar* ‘to sing’

2 Brazilian Portuguese (BP) vs. Peninsular Spanish (PS)

Portuguese and Spanish are both Romance languages from the Iberian Peninsula, and share many morpho-syntactic characteristics. Both languages have a similar verb system with three main conjugations (*-ar*, *-er*, *-ir*), nouns and adjectives may vary in number and gender, and adverbs are invariable. Both are pro-drop languages, they have a similar pronominal system, and certain phenomena, such as clitic climbing, are prevalent in both languages. They also allow rather free constituent order; and in both cases there is considerable debate in the literature about the appropriate characterization of their predominant word order (the candidates being SVO and VSO).

Sometimes the languages exhibit near-complete parallelism in their morphological patterns, as shown in Table 1.
The languages are also similar in their lexicon and syntactic word order:

(1) Os estudantes já comparam os
Los estudiantes ya compraron los
The students already bought the
livros. [BP]
libros. [PS]
books
‘The students have already bought the books.’

One of the main differences is the fact that Brazilian Portuguese (BP) accepts object dropping, while Peninsular Spanish (PS) doesn’t. In addition, subjects in BP tend to be overt while in PS they tend to be omitted.

(2) a. A: O que você fez com o livro? [BP]
What you did with the book?
A: ‘What did you do with the book?’
B: *Eu* dei para Maria.
I gave to Mary
B: ‘I gave it to Mary.’

b. A: ¿Qué hiciste con el libro? [PS]
What did with the book?
A: ‘What did you do with the book?’
B: Se *lo* di a María.
Her.dat it.acc gave to Mary.
B: ‘I gave it to Mary.’

Notice also that in the Spanish example (2b) the dative pronoun *se* ‘her’ is obligatory even when the prepositional phrase *a María* ‘to Mary’ is present.

3 Resources

3.1 Tagset

For both Spanish and Portuguese, we used positional tagsets developed on the basis of Spanish CLiC-TALP tagset (Torruella, 2002). Every tag is a string of 11 symbols each corresponding to one morphological category. For example, the Portuguese word *partires* ‘you leave’ is assigned the tag `VM0S---2PI-`, because it is a verb (`V`), main (`M`), gender is not applicable to this verb form (`0`), singular (`S`), case, possesor’s number and form are not applicable to this category(`-`), 2nd person (`2`), present (`P`), indicative (`I`) and participle type is not applicable (`-`).

A comparison of the two tagsets is in Table 2.[2] When possible the Spanish and Portuguese 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. In addition, we added a tag for “treatment” Portuguese pronouns.

The Spanish tagset has 282 tags, while that for Portuguese has 259 tags.

3.2 Training corpora

Spanish training corpus. The Spanish corpus we use for training the transition probabilities as well as for obtaining Spanish-Portuguese cognate pairs is a fragment (106,124 tokens, 18,629 types) of the Spanish section of CLiC-TALP (Torruella,

[2] 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)

No.	Description	No. of values	
		Sp	Po
1	POS	14	11
2	SubPOS – detailed POS	30	29
3	Gender	6	6
4	Number	5	5
5	Case	6	6
6	Possessor’s Number	4	4
7	Form	3	3
8	Person	5	5
9	Tense	7	9
10	Mood	8	9
11	Participle	3	3

Table 2: Overview and comparison of the tagsets

2002). CLiC-TALP is a balanced corpus, containing texts of various genres and styles. We automatically translated the CLiC-TALP tagset into our system (see Sect. 3.1) for easier detailed evaluation and for comparison with our previous work that used a similar approach for tagging (Hana et al., 2004; Feldman et al., 2006).

Raw Portuguese corpus. For automatic lexicon acquisition, we use NILC corpus,[3] containing 1.2M tokens.

3.3 Evaluation corpus

For evaluation purposes, we selected and manually annotated a small portion (1,800 tokens) of NILC corpus.

4 Morphological Analysis

Our morphological analyzer (Hana, 2005) is an open and modular system. It allows us to combine modules with different levels of manual input – from a module using a small manually provided lexicon, through a module using a large lexicon automatically acquired from a raw corpus, to a guesser using a list of paradigms, as the only resource provided manually. The general strategy is to run modules that make fewer errors and less overgenerate before modules that make more errors and overgenerate more. This, for example, means that modules with manually created resources are used before modules with resources automatically acquired. In the experiments below, we used the following modules – lookup in a list of (mainly) closed-class words, a paradigm-based guesser and an automatically acquired lexicon.

4.1 Portuguese closed class words

We created a list of the most common prepositions, conjunctions, and pronouns, and a number of the most common irregular verbs. The list contains about 460 items and it required about 6 hours of work. 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. The reason why the creation of such a list took 6 hours is that the words were annotated with detailed morphological tags used by our system.

4.2 Portuguese paradigms

We also created a list of morphological paradigms. Our database contains 38 paradigms. We just encoded basic facts about the Portuguese morphology from a standard grammar textbook (Cunha and Cintra, 2001). The paradigms include all three regular verb conjugations (*-ar*, *-er*, *-ir*), the most common adjective and nouns paradigms and a rule for adverbs of manner that end with *-mente* (analogous to the English *-ly*). We ignore majority of exceptions. The creation of the paradigms took about 8 h of work.

4.3 Lexicon Acquisition

The morphological analyzer supports a module or modules employing a lexicon containing information about lemmas, stems and paradigms. There is always the possibility to provide this information manually. That, however, is very costly. Instead, we created such a lexicon automatically.

Usually, automatically acquired lexicons and similar systems are used as a backup for large high-precision high-cost manually created lexicons (e.g. Mikheev, 1997; Hlaváčová, 2001). Such systems extrapolate the information about the words known by the lexicon (e.g. distributional properties of endings) to unknown words. Since our approach is resource light, we do not have any such large lexicon to extrapolate from.

The general idea of our system is very simple. The paradigm-based Guesser, provides all the possible analyses of a word consistent with Portuguese paradigms. Obviously, this approach mas-

[3] 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.

sively overgenerates. Part of the ambiguity is usually real but most of it is spurious. We use a large corpus to weed the spurious analyses out of the real ones. In such corpus, open-class lemmas are likely to occur in more than one form. Therefore, if a lemma+paradigm candidate suggested by the Guesser occurs in other forms in other parts of the corpus, it increases the likelihood that the candidate is real and vice versa. If we encounter the word *cantamos* 'we sing' in a Portuguese corpus, using the information about the paradigms we can analyze it in two ways, either as being a noun in the plural with the ending *-s*, or as being a verb in the 1st person plural with the ending *-amos*. Based on this single form we cannot say more. However if we also encounter the forms *canto*, *canta*, *cantam* the verb analysis becomes much more probable; and therefore, it will be chosen for the lexicon. If the only forms that we encounter in our Portuguese corpus were *cantamos* and (the non-existing) *cantamo* (such as the existing word *ramo* and *ramos*) then we would analyze it as a noun and not as a verb.

With such an approach, and assuming that the corpus contains the forms of the verb *matar* 'to kill', *mato*$_{1sg}$ *matas*$_{2sg}$, *mata*$_{3sg}$, etc., we would not discover that there is also a noun *mata* 'forest' with a plural form *matas* – the set of the 2 noun forms is a proper subset of the verb forms. A simple solution is to consider not the number of form types covered in a corpus, but the coverage of the possible forms of the particular paradigm. However this brings other problems (e.g. it penalizes paradigms with large number of forms, paradigms with some obsolete forms, etc.). We combine both of these measures in Hana (2005).

Lexicon Acquisition consists of three steps:

1. A large raw corpus is analyzed with a lexicon-less MA (an MA using a list of mainly closed-class words and a paradigm based guesser);

2. All possible hypothetical lexical entries over these analyses are created.

3. Hypothetical entries are filtered with aim to discard as many nonexisting entries as possible, without discarding real entries.

Obviously, morphological analysis based on such a lexicon still overgenerates, but it overgenerates much less than if based on the endings alone.

Lexicon	no	yes
recall	**99.0**	98.1
avg ambig (tag/word)	4.3	**3.5**
Tagging (cognates) – accuracy	79.1	82.1

Table 3: Evaluation of Morphological analysis

Consider for example, the form *funções* 'functions' of the feminine noun *função*. The analyzer without a lexicon provides 11 analyses (6 lemmas, each with 1 to 3 tags); only one of them is correct. In contrast, the analyzer with an automatically acquired lexicon provides only two analyses: the correct one (noun fem. pl.) and an incorrect one (noun masc. pl., note that POS and number are still correct). Of course, not all cases are so persuasive.

The evaluation of the system is in Table 3. The 98.1% recall is equivalent to the upper bound for the task. It is calculated assuming an oracle-Portuguese tagger that is always able to select the correct POS tag if it is in the set of options given by the morphological analyzer. Notice also that for the tagging accuracy, the drop of recall is less important than the drop of ambiguity.

5 Tagging

We used the TnT tagger (Brants, 2000), an implementation of the Viterbi algorithm for second-order Markov model. In the traditional approach, we would train the tagger's transitional and emission probabilities on a large annotated corpus of Portuguese. However, our resource-light approach means that such corpus is not available to us and we need to use different ways to obtain this information.

We assume that syntactic properties of Spanish and Portuguese are similar enough to be able to use the transitional probabilities trained on Spanish (after a simple tagset mapping).

The situation with the lexical properties as captured by emission probabilities is more complex. Below we present three different ways how to obtains emissions, assuming:

1. they are the same: we use the Spanish emissions directly (§5.1).

2. they are different: we ignore the Spanish emissions and instead uniformly distribute

the results of our morphological analyzer. (§5.2)

3. they are similar: we map the Spanish emissions onto the result of morphological analysis using automatically acquired cognates. (§5.3)

5.1 Tagging – Baseline

Our lowerbound measurement consists of training the TnT tagger on the Spanish corpus and applying this model directly to Portuguese.[4] The overall performance of such a tagger is 56.8% (see the the *min* column in Table 4). That means that half of the information needed for tagging of Portuguese is already provided by the Spanish model. This tagger has seen no Portuguese whatsoever, and is still much better than nothing.

5.2 Tagging – Approximating Emissions I

The opposite extreme to the baseline, is to assume that Spanish emissions are useless for tagging Portuguese. Instead we use the morphological analyzer to limit the number of possibilities, treating them all equally – The emission probabilities would then form a uniform distribution of the tags given by the analyzer. The results are summarized in Table 4 (the *e-even* column) – accuracy 77.2% on full tags, or 47% relative error reduction against the baseline.

5.3 Tagging – Approximating Emissions II

Although it is true that forms and distributions of Portuguese and Spanish 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.

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 to accurate Spanish-Portuguese translations or even a sentence-aligned corpus. All of these are resources that we could not expect to obtain in a resource poor setting. In the absence of this knowledge, we automatically identify cognates, using the edit distance measure (normalized by word length).

Unlike in the standard edit distance, the cost of operations is dependent on the 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, e.g. replacing a *q* with *c* is less costly than replacing *m* with, say *s*. 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 Spanish-Portuguese cognate pairs, we can use these to map the emission probabilities acquired on Spanish corpus to Portuguese.

[4] Before training, we translated the Spanish tagset into the Portuguese one.

Let's assume Spanish word w_s and Portuguese word w_p are cognates. Let T_s denote the tags that w_s occurs within the Spanish 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_p denote tags assigned to the Portuguese word w_p by our morphological analyzer, and the $p_p(t)$ is the even emission probability: $p_p(t) = \frac{1}{|T_p|}$. Then we can assign the new emission probability $p'_p(t)$ to every tag $t \in T_p$ in the following way (followed by normalization):

$$p'_p(t) = \frac{p_s(t) + p_p(t)}{2} \qquad (1)$$

Results. This method provides the best results. The full-tag accuracy is 82.1%, compared to 56.9% for baseline (58% error rate reduction) and 77.2% for even-emissions (21% reduction). The accuracy for POS is 87.6%. Detailed results are in column *e-cognates* of Table 4.

	min	e-even	e-cognates
Tag:	56.9	77.2	82.1
POS:	65.3	84.2	87.6
SubPOS:	61.7	83.3	86.9
gender:	70.4	87.3	90.2
number:	78.3	95.3	96.0
case:	93.8	96.8	97.2
possessor's num:	85.4	96.7	97.0
form:	92.9	99.2	99.2
person:	74.5	91.2	92.7
tense:	90.7	95.1	96.1
mood:	91.5	95.0	96.0
participle:	99.9	100.0	100.0

Table 4: Tagging Brazilian Portuguese

6 Evaluation & Comparison

The best way to evaluate our results would be to compare it against the TnT tagger used the usual way – trained on Portuguese and applied on Portuguese. We do not have access to a large Portuguese corpus annotated with detailed tags. However, we believe that Spanish and Portuguese are similar enough (see Sect. 2) to justify our assumption that the TnT tagger would be equally successful (or unsuccessful) on them. The accuracy of TnT trained on 90K tokens of the CLiC-TALP corpus is 94.2% (tested on 16K tokens). The accuracy of our best tagger is 82.1%. Thus the error-rate is more than 3 times bigger (17.9% vs. 5.4%).

Branco and Silva (2003) report 97.2% tagging accuracy on 23K testing corpus. This is clearly better than our results, on the other hand they needed a large Portuguese corpus of 207K tokens. The details of the tagset used in the experiments are not provided, so precise comparison with our results is difficult.

7 Related work

Previous research in resource-light language learning has defined *resource-light* in different ways. Some have assumed only partially tagged training corpora (Merialdo, 1994); some have begun with small tagged seed wordlists (Cucerzan and Yarowsky, 1999) for named-entity tagging, while others have exploited the automatic transfer of an already existing annotated resource in a different genres or a different language (e.g. cross-language projection of morphological and syntactic information in (Yarowsky et al., 2001; Yarowsky and Ngai, 2001), requiring no direct supervision in the target language).

Ngai and Yarowsky (2000) observe that the total weighted human and resource costs is the most practical measure of the degree of supervision. Cucerzan and Yarowsky (2002) observe that another useful measure of minimal supervision is the additional cost of obtaining a desired functionality from existing commonly available knowledge sources. They note that for a remarkably wide range of languages, there exist a plenty of reference grammar books and dictionaries which is an invaluable linguistic resource.

7.1 Resource-light approaches to Romance languages

Cucerzan and Yarowsky (2002) present a method for bootstrapping a fine-grained, broad coverage POS tagger in a new language using only one person-day of data acquisition effort. Similarly to us, they use a basic library reference grammar book, and access to an existing monolingual text corpus in the language, but they also use a medium-sized bilingual dictionary.

In our work, we use a paradigm-based morphology, including only the basic paradigms from a standard grammar textbook. Cucerzan and Yarowsky (2002) create a dictionary of regular inflectional affix changes and their associated POS and on the basis of it, generate hypothesized inflected forms following the regular paradigms.

Clearly, these hypothesized forms are inaccurate and overgenerated. Therefore, the authors perform a probabilistic match from all lexical tokens actually observed in a monolingual corpus and the hypothesized forms. They combine these two models, a model created on the basis of dictionary information and the one produced by the morphological analysis. This approach relies heavily on two assumptions: (i) words of the same POS tend to have similar tag sequence behavior; and (ii) there are sufficient instances of each POS tag labeled by either the morphology models or closed-class entries. For richly inflectional languages, however, there is no guarantee that the latter assumption would always hold.

The accuracy of their model is comparable to ours. On a fine-grained (up to 5-feature) POS space, they achieve 86.5% for Spanish and 75.5% for Romanian. With a tagset of a similar size (11 features) we obtain the accuracy of 82.1% for Portuguese.

Carreras et al. (2003) present work on developing low-cost Named Entity recognizers (NER) for a language with no available annotated resources, using as a starting point existing resources for a similar language. They devise and evaluate several strategies to build a Catalan NER system using only annotated Spanish data and unlabeled Catalan text, and compare their approach with a classical bootstrapping setting where a small initial corpus in the target language is hand tagged. It turns out that the hand translation of a Spanish model is better than a model directly learned from a small hand annotated training corpus of Catalan. The best result is achieved using cross-linguistic features. Solorio and López (2005) follow their approach; however, they apply the NER system for Spanish directly to Portuguese and train a classifier using the output and the real classes.

7.2 Cognates

Mann and Yarowsky (2001) present a method for inducing translation lexicons based on trasduction modules of cognate pairs via bridge languages. Bilingual lexicons within language families are induced using probabilistic string edit distance models. Translation lexicons for abitrary distant language pairs are then generated by a combination of these intra-family translation models and one or more cross-family online dictionaries. Similarly to Mann and Yarowsky (2001), we show that languages are often close enough to others within their language family so that cognate pairs between the two are common, and significant portions of the translation lexicon can be induced with high accuracy where no bilingual dictionary or parallel corpora may exist.

8 Conclusion

We have shown that a tagging system with a small amount of manually created resources can be successful. We have previously shown that this approach can work for Czech and Russian (Hana et al., 2004; Feldman et al., 2006). Here we have shown its applicability to a new language pair. This can be done in a fraction of the time needed for systems with extensive manually created resources: days instead of years. Three resources are required: (i) a reference grammar (for information about paradigms and closed class words); (ii) a large amount of text (for learning a lexicon; e.g. newspapers from the internet); (iii) a limited access to a native speaker — reference grammars are often too vague and a quick glance at results can provide feedback leading to a significant increase of accuracy; however both of these require only limited linguistic knowledge.

In this paper we proposed an algorithm for cognate transfer that effectively projects the source language emission probabilities into the target language. Our experiments use minimal new human effort and show 21% error reduction over even emissions on a fine-grained tagset.

In the near future, we plan to compare the effectiveness (time and price) of our approach with that of the standard resource-intensive approach to annotating a medium-size corpus (on a corpus of around 100K tokens). A resource-intensive system will be more accurate in the labels which it offers to the annotator, so annotator can work faster (there are fewer choices to make, fewer keystrokes required). On the other hand, creation of the infrastructure for such a system is very time consuming and may not be justified by the intended application.

The experiments that we are running right now are supposed to answer the question of whether training the system on a small corpus of a closely related language is better than training on a larger corpus of a less related language. Some preliminary results (Feldman et al., 2006) suggest that using cross-linguistic features leads to higher precision, especially for the source languages which have target-like properties complementary to each other.

9 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 Spanish part of the CLiC-TALP corpus.

References

Branco, A. and J. Silva (2003). Portuguese-specific Issues in the Rapid Development of State-of-the-art Taggers. In *Workshop on Tagging and Shallow Processing of Portuguese: TASHA'2000*.

Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of ANLP-NAACL*, pp. 224–231.

Carreras, X., L. Màrquez, and L. Padró (2003). Named Entity Recognition for Catalan Using Only Spanish Resources and Unlabelled Data. In *Proceedings of EACL-2003*.

Cucerzan, S. and D. Yarowsky (1999). Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In *Proceedings of the 1999 Joint SIGDAT Conference on EMNLP and VLC*, pp. 90–99.

Cucerzan, S. and D. Yarowsky (2002). Bootstrapping a Multilingual Part-of-speech Tagger in One Person-day. In *Proceedings of CoNLL 2002*, pp. 132–138.

Cunha, C. and L. F. L. Cintra (2001). *Nova Gramática do Português Contemporâneo*. Rio de Janeiro, Brazil: Nova Fronteira.

Feldman, A., J. Hana, and C. Brew (2006). Experiments in Morphological Annotation Transfer. In *Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing)*.

Hana, J. (2005). Knowledge and labor light morphological analysis. Unpublished manuscript.

Hana, J., A. Feldman, and C. Brew (2004). A Resource-light Approach to Russian Morphology: Tagging Russian using Czech resources. In *Proceedings of EMNLP 2004*, Barcelona, Spain.

Hlaváčová, J. (2001). Morphological Guesser or Czech Words. In V. Matoušek (Ed.), *Text, Speech and Dialogue*, Lecture Notes in Computer Science, pp. 70–75. Berlin: Springer-Verlag.

Hwa, R., P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak (2004). Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural Language Engineering 1*(1), 1–15.

Mann, G. S. and D. Yarowsky (2001). Multipath Translation Lexicon via Bridge Languages. In *Proceedings of NAACL 2001*.

Merialdo, B. (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics 20*(2), 155–172.

Meurers, D. (2004). On the Use of Electronic Corpora for Theoretical Linguistics. Case Studies from the Syntax of German. *Lingua*.

Mikheev, A. (1997). Automatic Rule Induction for Unknown Word Guessing. *Computational Linguistics 23*(3), 405–423.

Ngai, G. and D. Yarowsky (2000). Rule Writing or Annotation: Cost-efficient Resource Usage for Base Noun Phrase Chunking. In *Proceedings of the 38th Meeting of ACL*, pp. 117–125.

Solorio, T. and A. L. López (2005). Learning named entity recognition in Portuguese from Spanish. In *Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing)*.

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

Yarowsky, D. and G. Ngai (2001). Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora. In *Proceedings of NAACL-2001*, pp. 200–207.

Yarowsky, D., G. Ngai, and R. Wicentowski (2001). Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of HLT 2001, First International Conference on Human Language Technology Research*.

Yarowsky, D. and R. Wicentowski (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Meeting of the Association for Computational Linguistics*, pp. 207–216.