# Unsupervised adaptation of supervised part-of-speech taggers for closely related languages

**Yves Scherrer**
LATL-CUI
University of Geneva
Route de Drize 7, 1227 Carouge, Switzerland
yves.scherrer@unige.ch

## Abstract

When developing NLP tools for low-resource languages, one is often confronted with the lack of annotated data. We propose to circumvent this bottleneck by training a supervised HMM tagger on a closely related language for which annotated data are available, and translating the words in the tagger parameter files into the low-resource language. The translation dictionaries are created with unsupervised lexicon induction techniques that rely only on raw textual data. We obtain a tagging accuracy of up to 89.08% using a Spanish tagger adapted to Catalan, which is 30.66% above the performance of an unadapted Spanish tagger, and 8.88% below the performance of a supervised tagger trained on annotated Catalan data. Furthermore, we evaluate our model on several Romance, Germanic and Slavic languages and obtain tagging accuracies of up to 92%.

## 1 Introduction

Recently, a lot of research has dealt with the task of creating part-of-speech taggers for languages which lack manually annotated training corpora. This is usually done through some type of annotation projection from a language for which a tagger or an annotated corpus exists (henceforth called RL for *resourced language*) towards another language that lacks such data (NRL for *non-resourced language*). One possibility is to use word-aligned parallel corpora and transfer the tags from the RL to the NRL along alignment links. Another possibility is to adapt the parameters of the RL tagger using bilingual dictionaries or manually built transformation rules.

In this paper, we argue that neither parallel corpora nor hand-written resources are required if the RL and the NRL are closely related. We propose a generic method for tagger adaptation that relies on three assumptions which generally hold for closely related language varieties. First, we assume that the two languages share a lot of cognates, i.e., word pairs that are formally similar and that are translations of each other. Second, we suppose that the word order of both languages is similar. Third, we assume that the set of POS tags is identical. Under these assumptions, we can avoid the requirements of parallel data and of manual annotation.

Following Feldman et al. (2006), the reasoning behind our method is that a Hidden Markov Model (HMM) tagger trained in a supervised way on RL data can be adapted to the NRL by translating the RL words in its parameter files to the NRL. This requires a bilingual dictionary between RL words and NRL words. In this paper, we create different HMM taggers using the bilingual dictionaries obtained with the unsupervised lexicon induction methods presented in our earlier work (Scherrer and Sagot, 2014).

The paper is organized as follows. In Section 2, we present related work on tagger adaptation and lexicon induction. In Section 3, we review Hidden Markov Models and their relevance for tagging and for our method of tagger adaptation. Section 4 presents a set of different taggers in some detail and evaluates them on Catalan, using Spanish as RL. In Section 5, we demonstrate the validity of the proposed approach by performing small-scale evaluations on a number of Romance, Germanic and Slavic languages: we transfer part-of-speech tags from Spanish to Aragonese, from Czech to Slovak and Sorbian, from Standard German to Dutch and Palatine German. We conclude in Section 6.





*Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 30–38,
Dublin, Ireland, August 23 2014.

## 2 Related work

The task of creating part-of-speech taggers (and other NLP tools) for new languages without resorting to manually annotated corpora has inspired a lot of recent research. The most popular line of work, initiated by Yarowsky et al. (2001), draws on parallel corpora. They tag the source side of a parallel corpus with an existing tagger, and then project the tags along the word alignment links onto the target side of the parallel corpus. A new tagger is then trained on the target side, using aggressive smoothing to reduce the noise caused by alignment errors.

In a similar setting, Das and Petrov (2011) use a more sophisticated graph-based projection algorithm with label propagation to obtain high-precision tags for the target words. Follow-up work by Li et al. (2012) uses tag dictionaries extracted from Wiktionary instead of parallel corpora, and Täckström et al. (2013) attempt to combine these two data sources: the Wiktionary data provides constraints on word *types*, whereas the parallel data is used to filter these constraints on the *token* level, depending on the context of a given word occurrence. Duong et al. (2013) show that the original approach of Das and Petrov (2011) can be simplified by focusing on high-confidence alignment links, thus achieving equivalent performance without resorting to graph-based projection. The research based on parallel corpora does not assume any particular etymological relationship between the two languages, but Duong et al. (2013) note that their approach works best when the source and target languages are closely related.

Other approaches explicity model the case of two closely related languages, such as Feldman et al. (2006). They train a tagger on the source language with standard tools and resources, and then adapt the parameter files of that tagger to the target language using a hand-written morphological analyzer and a list of cognate word pairs. Bernhard and Ligozat (2013) use a similar approach to adapt a German tagger to Alsatian; they show that manually annotating a small list of closed-class words leads to considerable gains in tagging accuracy. In a slightly different setting, Garrette and Baldridge (2013) show that taggers for low-resource languages can be built from scratch with only two hours of manual annotation work.

Even though recent work on closely related and low-resource languages presupposes manually annotated data to some extent, we believe that it is possible to create a tagger for such languages fully automatically. We adopt the general model proposed by Feldman et al. (2006), but use automatically induced bilingual dictionaries to translate the source language words in the tagger parameter files. The bilingual dictionaries are obtained with our unsupervised lexicon induction pipeline (Scherrer and Sagot, 2013; Scherrer and Sagot, 2014). This pipeline is inspired by early work by Koehn and Knight (2002), who propose various methods for inferring translation lexicons using monolingual data.

Our lexicon induction pipeline is composed of three main steps. First, a list of formally similar word pairs (cognate pairs) is extracted from monolingual corpora using the BI-SIM score (Kondrak and Dorr, 2004). Second, regularities occurring in these word pairs are learned by training and applying a character-level statistical machine translation (CSMT) system (Vilar et al., 2007; Tiedemann, 2009). Third, cross-lingual contextual similarity measures are used to induce additional word pairs. The main idea is to extract word n-grams from comparable corpora of both languages and induce word pairs that co-occur in the context of already known word pairs (Fung, 1998; Rapp, 1999; Fišer and Ljubešić, 2011). In our pipeline, the already known word pairs are those induced with CSMT.

In this paper, we extend our previous work (Scherrer and Sagot, 2014) in two aspects. First, we use a more powerful HMM tagging model instead of the simple unigram tagger that insufficiently accounts for the ambiguity in language. Second, we assess the impact of each lexicon induction step separately rather than merely evaluating the final result of the pipeline.

## 3 HMM tagging

Hidden Markov Models (HMMs) are a simple yet powerful formal device frequently used for part-of-speech tagging. A HMM describes a process that generates a joint sequence of tags and words by decomposing the problem into so-called transitions and emissions. Transitions represent the probabilities of a tag given the preceding tag(s), and emissions represent the probabilities of a word given the tag assigned to it (Jurafsky and Martin, 2009).

The main advantage of HMM taggers for our work lies in the independence assumption between transitions and emissions: crucially, the emission probability of a word only depends on its tag; it does not depend on previous words or on previous tags. Assuming, as stated in the introduction, that the word order is similar and the tag sets identical between the RL and the NRL, we argue that the transition probabilities estimated on RL data are also valid for NRL. Only the emission probabilities have to be adapted since RL words are formally different from NRL words.

Following earlier work (Feldman et al., 2006; Duong et al., 2013), we use the TnT tagger (Brants, 2000), an implementation of a trigram HMM tagger that includes smoothing and handling of unknown words. In contrast to other implementations that use inaccessible binary files, TnT stores the estimated parameters in easily modifiable plain text files.

### 3.1 Adapting emission counts

The goal of this work is to adapt an existing RL HMM tagger for a closely related NRL by replacing the RL words in the emission parameters by the corresponding NRL words. Let us explain this process with an example, using Spanish as RL and Catalan as NRL.

The TnT tagger creates an emission parameter file that contains, for each word, the tags and their frequencies observed in the training corpus. For example, a tagger trained on Spanish data may contain the following lines (word on the left, tag in the middle, frequency on the right):

(1)

| | | |
|---|---|---|
| intelectual | AQ | 11 |
| intelectual | NC | 3 |
| intelectuales | AQ | 3 |
| intelectuales | NC | 7 |

Furthermore, suppose that we have a dictionary that associates Catalan words (left) with Spanish words (center), where the weight (right) indicates the ambiguity level of the Catalan word, which is simply defined as the inverse of the number of its Spanish translations:

(2)

| | | |
|---|---|---|
| intel·lectual | intelectual | 0.5 |
| intel·lectual | intelectuales | 0.5 |
| intel·lectuals | intelectuales | 1 |

A new Catalan emission file is then created by taking, for each Catalan word, the union of the tags of its Spanish translations and by multiplying the tag weights with the dictionary weights. This yields the following entries:

(3)

| | | |
|---|---|---|
| intel·lectual | AQ | $(0.5 \cdot 11) + (0.5 \cdot 3) = 7$ |
| intel·lectual | NC | $(0.5 \cdot 3) + (0.5 \cdot 7) = 5$ |
| intel·lectuals | AQ | $1 \cdot 3 = 3$ |
| intel·lectuals | NC | $1 \cdot 7 = 7$ |

Or more formally: for each dictionary triple $\langle w_{RL}, w_{NRL}, f_d \rangle$ and each emission triple $\langle w_{RL}, t, f_e \rangle$ with matching $w_{RL}$, add the new emission triple $\langle w_{NRL}, t, f_d \cdot f_e \rangle$. Merge emission triples with identical $w_{NRL}$ and $t$ and sum their weights.

Finally, RL words occurring in the emission file that have not been translated to NRL (because no appropriate word pair existed in the dictionary) are copied without modification to the new emission file. In particular, this allows us to cover punctuation signs and numbers as well as named entities (which are mostly spelled identically in both languages).

## 4 Tagger adaptation for Catalan

In this section, we present seven taggers for Catalan. Three of them (Sections 4.2 to 4.4) are supervised taggers and serve as baseline taggers and as upper bounds. The four remaining taggers (Sections 4.6 to 4.9) are taggers created by adaptation from a Spanish tagger, using the method presented in Section 3.1;

they differ in the lexicons used to translate the emission counts. These four taggers represent the main contribution of this paper. We start by listing the data used in our experiments.

### 4.1 Data

Most taggers presented below are initially trained on a part-of-speech annotated corpus of Spanish. We use the Spanish part of the AnCora treebank (Taulé et al., 2008), which contains about 500 000 words.

The AnCora morphosyntactic annotation includes the main category (e.g. noun), the subcategory (e.g. proper noun), and several morphological categories (e.g., gender, number, person, tense, mode), yielding about 280 distinct labels. Since we are mainly interested in part-of-speech information, we simplified these labels by taking into account the two first characters of each label, corresponding to the main category and the subcategory. This simplified tagset contains 42 distinct labels, which is still considerably more than the 12 tags of Petrov et al. (2012) commonly used in comparable settings.

All taggers need to be evaluated on a Catalan gold standard that shares the same tagset as Spanish. For this purpose, we use the Catalan part of AnCora, which also contains about 500 000 words. We simplified the tags in the same way as above. The Catalan part of AnCora is also used to train the supervised models presented in Sections 4.3 and 4.4.

Finally, the lexicon induction algorithms require data on their own, which we present here for completeness. As in Scherrer and Sagot (2013), we use Wikipedia dumps consisting of 140M words for Catalan and 430M words for Spanish.[1]

### 4.2 Baseline: a Spanish tagger

Since Spanish and Catalan are closely related languages, one could presume that a lot of words are identical, and that a tagger trained on Spanish data would yield acceptable performance on Catalan test data without modifications. In order to test this hypothesis, we trained a TnT tagger on Spanish AnCora and tested it on Catalan AnCora. We obtained a tagging accuracy of 58.42% only, which suggests that this approach is clearly insufficient. (The results of all experiments are summed up in Table 1.) For comparison, Feldman et al. (2006) obtain 64.5% accuracy on the same languages with a smaller training corpus (100k instead of 500k words), but also with a smaller tagset (14 instead of 42).

We view this model as a baseline that we expect to beat with the adaptation methods.

### 4.3 Upper bound 1: a supervised Catalan tagger

The upper bound of the Catalan tagging experiments is represented by a tagger created under ideal data conditions: a tagger trained in a supervised way on an annotated Catalan corpus. We train a TnT tagger on Catalan AnCora and test it on the same corpus, using 10-fold cross-validation to avoid having the same sentences in the training and the test set. This yields an averaged accuracy value of 97.96%.

For comparison, Feldman et al. (2006) obtain 97.5% accuracy on their dataset. More recently, Petrov et al. (2012) report an accuracy of 98.5% by training on the CESS-ECE corpus, but do not mention the tagging algorithm used. In any case, our result obtained with TnT can be considered close to state-of-the-art performance on Catalan.

### 4.4 Upper bound 2: a tagger with Spanish transition counts and Catalan emission counts

We introduce a second upper bound that shares the assumption of structural similarity underlying the adaptation-based models. Concretely, we combine the transition probabilities from the baseline Spanish tagger (Section 4.2) with the emission probabilities of the supervised Catalan tagger (Section 4.3). The resulting tagger is evaluated again on Catalan AnCora using 10-fold cross-validation. We get an accuracy value of 97.66%, or just 0.3% absolute below the supervised tagger of Section 4.3.[2] This suggests that the transition probabilities are indeed very similar between the two languages, and that they can safely be kept constant in the adaptation-based models presented below.

[1] This is not exactly a realistic setting for the intended use for low-resource languages. However, Section 5 will illustrate the performance of the proposed models on smaller data sets. Note also that the lexicon induction methods do not require the two corpora to be of similar size.

[2] This difference is significant: $\chi^2(1; N = 1\,064\,002) = 109.9747799; p < 0.01$.

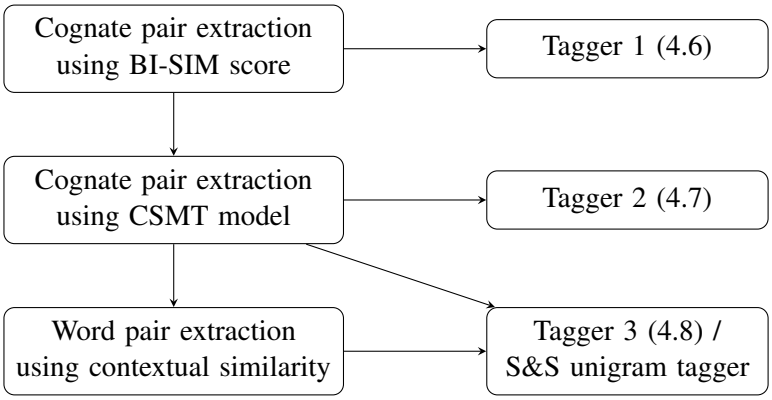


Figure 1: Flowchart of the lexicon induction pipeline and of the resulting taggers.

### 4.5 Lexicon induction methods for adaptation-based taggers

The adaptation-based taggers presented in Sections 4.6 to 4.8 differ in the bilingual dictionaries used to adapt the emission counts. These dictionaries have been created using the pipeline of Scherrer and Sagot (2014), which we summarize in this section (see Figure 1).

The pipeline starts with a cognate pair extraction step that uses the BI-SIM score to identify likely cognate pairs. The result of this step is used as training data for the second step, in which a CSMT model is trained to identify likely cognate pairs even more reliably. The result of the second step is in turn used as seed data for the third step, in which additional word pairs are extracted on the basis of contextual similarity. Scherrer and Sagot (2014) create a single unigram tagger (abreviated S&S in Figure 1) with the union of the word pairs obtained in the second and third steps (plus additional clues like word identity and suffix analysis, which are not required here).

The three steps are evaluated separately: Tagger 1 relies on the lexicon induced in the first step; Tagger 2 relies on the lexicon induced in the second step; Tagger 3 relies on the union of the lexicons induced in the second and third steps.

### 4.6 Tagger 1: cognate pairs induced with BI-SIM score

As first step of the lexicon induction pipeline, word lists are extracted from both Wikipedia corpora, and short words (words with less than 5 characters) as well as rare words (words accounting for the lowest 10% of the frequency distribution) are removed. Then, the BI-SIM score is computed between each Catalan word $w_{\text{ca}}$ and each Spanish word $w_{\text{es}}$. For each $w_{\text{ca}}$, we keep the $\langle w_{\text{ca}}, w_{\text{es}} \rangle$ pair(s) that maximize(s) the BI-SIM value, provided it is above the empirically chosen threshold of 0.8. When a $w_{\text{ca}}$ is associated with several $w_{\text{es}}$, we keep all of them. This creates a list of cognate pairs, albeit a rather noisy one since it does not take into account regular correspondences between languages, but merely counts letter bigram differences.

Tagger 1, the first adaptation-based tagger, is created by replacing the Spanish emission counts with their Catalan equivalents using the list of cognate pairs. Tagger 1 yields an accuracy of 68.32%, which is a full 10% higher than the baseline. This improvement is surprisingly high, as the cognate list is not only noisy, but also incomplete: only 17.91% of the words in the emission file could be translated with it.

### 4.7 Tagger 2: cognate pairs induced with CSMT

In this model, the Spanish emission counts are replaced using the list of cognate pairs obtained in the second step of the lexicon induction pipeline.

We train a CSMT system on the list of potential cognate pairs of the first step. We then apply this system to translate each Catalan word again into Spanish. We assume that the CSMT system learns useful generalizations about the relationship between Catalan and Spanish words, which the generic BI-SIM measure was not able to make. Moreover, the CSMT system is able to translate Catalan words even

| | Baseline | Tagger 1 | Tagger 2 | Tagger 3 | Tagger 4 | Upper bound 2 | Upper bound 1 |
|---|---|---|---|---|---|---|---|
| Tagging accuracy | 58.42% | 68.32% | 72.32% | 88.72% | 89.08% | 97.66% | 97.96% |
| Translated words | | 17.91% | 64.03% | 65.62% | | | |

Table 1: Results of the Catalan tagging experiments. The first line reports tagging accuracies of the different taggers. The second line shows – where applicable – how many words of the emission files could be translated.

if their Spanish translations have not been seen, on the basis of the character correspondences observed in other words.

This new dictionary allowed us to translate 64.03% of the words in the emission file. In consequence, the resulting tagger shows improved performance compared with Tagger 1: its accuracy lies at 72.32%, suggesting that the CSMT system yields a dictionary that is at the same time more precise and more complete than the one obtained with BI-SIM in the previous step.

### 4.8 Tagger 3: word pairs induced with CSMT and context similarity

In previous work (Scherrer and Sagot, 2014), we have argued that lexicon induction methods based on formal similarity alone are not sufficient, for the following reasons: (1) even in closely related languages, not all word pairs are cognates; (2) high-frequency words are often related through irregular phonetic correspondences; (3) pairs of short words may just be too hard to predict on the basis of formal criteria alone; (4) formal similarity methods are prone to inducing false friends, i.e., words that are formally similar but are not translations of each other. For these types of words, we have proposed a different approach that relies on contextual similarity.

We extract 3-gram and 4-gram contexts from both languages and form context pairs whenever the first and the last word pairs figure in the dictionary obtained with CSMT, allowing the word pair(s) in the center to be newly inferred. Several filters are added in order to remove noise.

In order to create Tagger 3, we merge the dictionary induced with CSMT and the dictionary induced with context similarity, giving preference to the latter. Again, the emission parameters of the baseline Spanish tagger are adapted using this dictionary. 65.62% of the words in the emission file could be translated, i.e. only 1.59% more than for Tagger 2. Nevertheless, the accuracy of Tagger 3 (88.72%) lies about 18% absolute above Tagger 2. This large gain in accuracy is due to the fact that context similarity mostly adds high-frequency words, which are few but crucial to obtain satisfactory tagging performance.

One goal of these experiments was to show whether the improved handling of ambiguity provided by HMMs in comparison with the unigram model used by Scherrer and Sagot (2013) is reflected in better overall tagging performance. This goal has been reached: the unigram model of Scherrer and Sagot (2013) shows a tagging accuracy of 85.1%, which is 3% absolute below Tagger 3, the most directly comparable HMM-based tagger.[3]

### 4.9 Tagger 4: re-estimate transition probabilities

In this last model, we challenge the initial assumption that the Spanish transition probabilities are "good enough" for tagging Catalan. Concretely, we use Tagger 3 to tag the entire Catalan Wikipedia corpus (the one also used for the lexicon induction tasks) and then train Tagger 4 in a supervised way on this data. The idea behind this additional step is that the transition (and emission) counts estimated on the large Catalan corpus are more reliable than those obtained by direct tagger adaptation.

Tagger 4 yields an accuracy value of 89.08%, outperforming Tagger 3 by only 0.36%.[4] This difference is consistent with the one observed between Upper Bound 1 and Upper Bound 2, suggesting once more

[3]The Catalan results reported in Scherrer and Sagot (2014) are based on a different test set, which is why we rather refer to the directly comparable Scherrer and Sagot (2013) results in this section.

[4]This difference is significant: $\chi^2(1; N = 1\,064\,002) = 35.84835013; p < 0.01$.

that transition counts only marginally influence the tagging performance if the former are estimated on a language that is structurally similar.

## 5 Multilingual experiments

In addition to the Spanish–Catalan experiment, we have induced taggers for several closely related languages from Romance, Germanic and Slavic language families and tested them on the multilingual data set used by Scherrer and Sagot (2014). Although the results of these additional experiments are less reliable than the Spanish–Catalan data due to the small test corpus sizes, they allow us to generalize our findings to other languages and language families. The experiments are set up as follows:

- The **Aragonese** taggers were adapted from a **Spanish** tagger trained on AnCora. They are tested on a Wikipedia excerpt of 100 sentences that was manually annotated with the simplified AnCora labels of Section 4.1. The Wikipedia corpora used for lexicon induction contained 5.4M words for Aragonese, and 431M words for Spanish.

- The **Dutch** and **Palatine German** taggers were adapted from a **Standard German** tagger trained on the TIGER treebank (900 000 tokens; 55 tags; Brants et al. (2002)). The gold standard corpora are Wikipedia excerpts of 100 sentences each, manually annotated with TIGER labels. The Wikipedia corpora used for lexicon induction contained 0.5M words for Dutch, 0.3M words for Palatine German, and 612M words for Standard German.

- The **Upper Sorbian**, **Slovak** and **Polish** taggers were adapted from a **Czech** Tagger trained on the Prague Dependency Treebank 2.5 (2M tokens; 57 simplified tags).[5] The gold standard corpora are Wikipedia excerpts of 30 sentences each, manually annotated with simplified PDT labels. The Wikipedia corpora used for lexicon induction contained 0.9M words for Upper Sorbian, 30M words for Slovak, 206M words for Polish, and 85M words for Czech.

The tagging accuracies are reported in the left part of Table 2. The accuracy values vary widely across languages, with baseline performances ranging from 24% to 81%. This variation essentially reflects the linguistic distance between the RL and the NRL: German and Dutch seem to be particularly distant, while Czech and Slovak are particularly closely related. In contrast, the overall tendency of the tagging models is the same for all languages: there are consistent gradual improvements from the baseline tagger to Tagger 3. These findings are in line with the Catalan experiments. The differences between Tagger 3 and Tagger 4 are not significant for any language, whereas the Catalan experiment showed a slight but significant improvement. Finally, Taggers 3 and 4 slightly outperform the unigram tagger of Scherrer and Sagot (2014) (S&S in Table 2) on most languages, although the difference is less marked than for Catalan.

The right half of Table 2 shows what percentage of the emission files could be translated at each step, analogously to the figures reported for Catalan in Table 1. The variation observed here mainly depends on the language proximity and on the size of the corpora used for lexicon induction.

Globally, the Germanic languages obtain the lowest accuracy scores. This is due to a combination of factors. First, as stated above, the baseline performance is already lower than in the other language families, which essentially results from a lower number of identical NRL–RL word pairs than in other language families. Second, the lexicon induction corpora are much smaller than for the other language families.[6] Third, Germanic languages tend to have longer words due to compounding, so that the BI-SIM threshold is more difficult to satisfy. The combination of the second and third factors lead to poor performance of the first lexicon induction step: less than 4% of the German words could be translated

[5] Similarly to AnCora, the morphosyntactic labels of the PDT consist of 15 positions that encode the main morphosyntactic category, the subcategory as well as various morphological categories. We simplify the tagset analogously to AnCora, keeping only the main category and the subcategory, which leads to 57 distinct labels.
The PDT is available at `http://ufal.mff.cuni.cz/pdt2.5/`.

[6] As in our earlier work, we used all of the Palatine German Wikipedia, whereas we reduced the Dutch Wikipedia corpus on purpose to better simulate the low-resource scenario.

| Language | Tagging accuracy | | | | | | Translated words | | |
|---|---|---|---|---|---|---|---|---|---|
| | Baseline | T1 | T2 | T3 | T4 | S&S | T1 | T2 | T3 |
| Aragonese | 72% | 74% | 74% | 87% | 87% | 85% | 16.11% | 42.65% | 43.23% |
| Dutch | 24% | 30% | 39% | 60% | 62% | 59% | 3.69% | 6.73% | 6.79% |
| Palatine German | 50% | 54% | 57% | 70% | 70% | 65% | 3.86% | 5.52% | 5.58% |
| Upper Sorbian | 70% | 72% | 77% | 84% | 84% | 84% | 5.70% | 11.60% | 11.69% |
| Slovak | 81% | 85% | 88% | 93% | 93% | 92% | 29.39% | 52.40% | 54.41% |
| Polish | 66% | 69% | 72% | 78% | 79% | 78% | 8.50% | 42.27% | 42.73% |

Table 2: Results of the multilingual tagging experiments. The left half of the table reports tagging accuracies and compares them with the results reported by Scherrer and Sagot (2014) (S&S column). The right half of the table shows how many words of the emission files could be translated.

when building Tagger 1. This obviously reduces the potential for accuracy gains in Tagger 1, but it also hampers the training of the CSMT system at the origin of Tagger 2. However, one should note that good tagging results can be achieved even with relatively low translation coverage, as shown by the Upper Sorbian experiment.

## 6 Conclusion

One goal of the experiments presented here was to validate the pipeline proposed earlier in Scherrer and Sagot (2014). By showing that there are gradual improvements from the baseline tagger to Tagger 3 on a large number of languages, we demonstrate that the overall approach of inducing word pairs in subsequent steps is sound, and that the order of these steps is reasonably chosen. Furthermore, we find that re-estimating the tagger parameters on a large monolingual corpus (Tagger 4) does not improve its performance substantially, as we have predicted in Section 4.4 on the basis of supervised Catalan taggers.

A second goal of these experiments was to show that the HMM taggers offer improved handling of ambiguity compared with the unigram tagger of Scherrer and Sagot (2014). We have indeed noted an accuracy gain of 3% on the Catalan data, and the multilingual data set shows similar (yet less marked) tendencies.

However, the Catalan experiments show that there still is a gap of about 10% absolute accuracy between the adaptation taggers and fully supervised taggers. We see two main reasons for this gap. First, the completely unsupervised lexicon induction algorithms obviously produce a number of erroneous word pairs, which may then result in erroneous tagging. Second, the lexicon induction algorithms currently do not allow a given NRL word to relate to two different RL words. As a result, the taggers are not able to model tagging ambiguities arising from translation ambiguities. Better ambiguity handling, for instance on the basis of token-level constraints as suggested by Täckström et al. (2013), could thus further improve tagging accuracy.

Finally, discriminative models using Maximum Entropy or Perceptron training have largely superseded HMMs for part-of-speech tagging in the last few years.[7] Such models take into account a larger set of features such as word suffixes, word structure (presence of punctuation signs, numerals, etc.) and external lexicon information. Further research will be needed to investigate how our adaptation methods can be applied to feature-based tagging models.

## Acknowledgements

The author would like to thank Benoît Sagot for his collaboration on earlier versions of this work. This work was partially funded by the Labex EFL (ANR/CGI), Strand 6, operation LR2.2.

[7]For an overview on recent English taggers, see for example `http://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art)`.

## References


Delphine Bernhard and Anne-Laure Ligozat. 2013. Hassle-free POS-tagging for the Alsatian dialects. In Marcos Zampieri and Sascha Diwersy, editors, *Non-Standard Data Sources in Corpus Based-Research*, volume 5 of *ZSM Studien*, pages 85–92. Shaker.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Treebank. In *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT 2002)*, pages 24–41.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of ANLP 2000*, pages 224–231.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL-HLT 2011*, pages 600–609.

Long Duong, Paul Cook, Steven Bird, and Pavel Pecina. 2013. Simpler unsupervised POS tagging with bilingual projections. In *Proceedings of ACL 2013*, pages 634–639.

Anna Feldman, Jirka Hana, and Chris Brew. 2006. A cross-language approach to rapid creation of new morpho-syntactically annotated resources. In *Proceedings of LREC 2006*, pages 549–554.

Darja Fišer and Nikola Ljubešić. 2011. Bilingual lexicon extraction from comparable corpora for closely related languages. In *Proceedings of RANLP 2011*, pages 125–131.

Pascale Fung. 1998. A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. *Machine Translation and the Information Soup*, pages 1–17.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of NAACL-HLT 2013*, pages 138–147.

Daniel Jurafsky and James H. Martin. 2009. *Speech and language processing*. Pearson, 2nd edition.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition (SIGLEX 2002)*, pages 9–16.

Grzegorz Kondrak and Bonnie Dorr. 2004. Identification of confusable drug names: A new approach and evaluation methodology. In *Proceedings of COLING 2004*, pages 952–958.

Shen Li, João Graça, and Ben Taskar. 2012. Wiki-ly supervised part-of-speech tagging. In *Proceedings of EMNLP-CoNLL 2012*, pages 1389–1398.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC 2012*, pages 2089–2096.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of ACL 1999*, pages 519–526.

Yves Scherrer and Benoît Sagot. 2013. Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources. In *Proceedings of the RANLP 2013 Workshop on Adaptation of language resources and tools for closely related languages and language variants*.

Yves Scherrer and Benoît Sagot. 2014. A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages. In *Proceedings of LREC 2014*, pages 502–508.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of LREC 2008*, pages 96–101.

Jörg Tiedemann. 2009. Character-based PSMT for closely related languages. In *Proceedings of EAMT 2009*, pages 12–19.

David Vilar, Jan-Thorsten Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of WMT 2007*, pages 33–39.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT 2001*.