Cross-lingual and supervised models for morphosyntactic annotation: a comparison on Romanian

Lauriane Aufrant[1,2], **Guillaume Wisniewski**[1], **François Yvon**[1]
[1]LIMSI, CNRS, Univ. Paris-Sud
Université Paris-Saclay, F-91405 Orsay
[2]DGA, 60 boulevard du Général Martial Valin, F-75509 Paris
{lauriane.aufrant, guillaume.wisniewski,francois.yvon}@limsi.fr

Abstract

Because of the small size of Romanian corpora, the performance of a PoS tagger or a dependency parser trained with the standard supervised methods fall far short from the performance achieved in most languages. That is why, we apply state-of-the-art methods for cross-lingual transfer on Romanian tagging and parsing, from English and several Romance languages. We compare the performance with monolingual systems trained with sets of different sizes and establish that training on a few sentences in target language yields better results than transferring from large datasets in other languages.

Keywords: Cross-Lingual transfer, Part-of-Speech tagging, Dependency parsing, Romanian

1. Introduction

While most Romance languages are well studied in the Natural Language Processing field and have large sets of annotated data, Romanian still stays behind. As a result, most of the tools generally used in the pre-processing steps of NLP pipelines such as lemmatizer, PoS-tagger or dependency parser are not available for Romanian and, when they exist, their performance often fall far short of the performance achieved, for instance, on French or English (Straka et al., 2015).

Romanian is therefore a prime candidate for applying transfer methods (Pan and Yang, 2010). Many works have considered the task of transferring PoS-tagger or dependency parser from a resource-rich to a resource-poor language (McDonald et al., 2011; Täckström et al., 2013; Lacroix et al., 2016). However the proposed approaches have generally been evaluated only on resource-rich languages for which the annotated data required to evaluate the transfer models were readily available. In this work, we intend to compare state-of-the-art cross-lingual methods for PoS tagging and dependency parsing on an actual resource-poor language, Romanian, that, in addition, present several linguistic challenges. Indeed, even if Romanian is a Romance language, it has rare properties among this family: it is a morphologically rich language which kept a case system inherited from Latin and uses more clitics than languages from Romance languages. Moreover, the Romanian went through Slavic influences, noticeable on 15 to 20% of its vocabulary (Haspelmath and Tadmor, 2009; Roegiest, 2006).

Our contribution is twofold:

- we implement two state-of-the-art transfer methods for PoS-tagging (Täckström et al., 2013) and dependency (McDonald et al., 2011) parsing for Romanian;
- we evaluate thoroughly the interest of cross-lingual methods in a realistic use-case of transferring syntactic models from a resource-rich into a resource-poor language.

Our work contrasts with previous studies of the two considered transfer methods, as they are evaluated on an actual low-resourced language instead of the usual well-resourced ones. While the data amount in Romanian is gradually increasing, and should make it possible to train reasonably competitive systems with monolingual annotated data, we restrict our study to linguistic information available with simple means and without language-specific tuning, using models in other languages, automatic word alignments and crawling of crowd-sourced dictionaries. Our purpose is also to comparatively quantify the benefits of manually annotating new resources.

The rest of this paper is organized as follows: after a brief overview of available resources in Romanian (§2.), we present experiments on tagging and dependency parsing (§3. and §4.). Finally, in the light of these results, we reassess the interest of cross-lingual methods (§5.) for under-resourced languages.

All tools and resources used in this work are available at `https://perso.limsi.fr/aufrant/`.

2. Resources for Romanian

Over the years, several corpora have been collected for Romanian. We are particularly interested in parallel corpora that will allow us to transfer annotations and corpora annotated with PoS and dependencies that can be used to evaluate cross-lingual taggers and parsers. In this section, we will quickly describe existing corpora.

Most work on cross-lingual transfer rely on the Europarl corpus (Koehn, 2005) as a source for parallel sentences. It notably includes Romanian sentences, with their translation in 20 European languages such as English or Spanish, that we will use in our experiments.

(Perez, 2012) released the treebank Romanian Syntactic Annotated Corpus (RSAC) of 67,686 tokens (punctuations excluded) and 3,587 sentences from various sources (JRC-Acquis, Wikipedia, 1984, textbook exercises and translations from FrameNet). Two other corpora exist: MULTEXT-East, a multilingual corpus extracted from the novel *1984*, sentence-aligned and with morphosyntactic annotations; a corpus of 36,150 tokens has also been annotated with dependencies in the context of the BALRIC project.[1] However, these corpora can not be used for our purpose: the former does not contain information about dependencies and, as reported by (Calacean and Nivre, 2009), the latter consists in simple sentences of only 8.9 tokens in average, and no diacritics,[2] therefore not representing actual Romanian language. Recently, a Romanian part was added to the 1.2 version of the Universal Dependencies corpus (McDonald et al., 2013), but it consists of only 633 sentences.

All these corpora are much smaller than the corpora usually used to train supervised models. For instance, the Penn Treebank contains more than 1,000,000 English tokens, the French part of the Universal Dependency Treebank contains 400,000 tokens and Europarl has about 2,000,000 English-French sentence pairs (380,000 for English-Romanian). Romanian can, therefore, be considered, comparatively, as a resource-poor language.

Only a few studies have addressed the issue of Romanian supervised tagging and parsing. Morfette (Chrupala et al., 2008) predicts jointly the inflectional morphology and lemmatization of Romanian text; the tagset it uses includes PoS information. A Romanian dependency parser has been developed by (Calacean and Nivre, 2009), it is trained however on the BALRIC corpus that, as explained above, is not representative of Romanian. Finally, (Colhon and Simionescu, 2012) experimented with a Maximum Entropy approach to build a supervised parser on an earlier version of RSAC.

3. Cross-lingual PoS tagger

In this section, we describe our experiments in transferring annotations to develop a PoS tagger for Romanian. We are particularly interested in evaluating the impact of the source language.

Transfer Method Our approach is based on the method introduced by Täckström et al. (2013) that combines two sources of information to automatically label a Romanian corpus: *token constraints* extracted from word alignments with *type constraints* extracted from crowd-sourced dictionaries. Following (Wisniewski et al., 2014), we use an history-based model trained in the ambiguous learning framework, and the following feature templates: lowercase words, prefixes of size up to 5 and suffixes of size up to 7 in a context window of size 2, the spelling pattern of the current word, the last two predicted tags and their combination, and the combination of the last tag with the current word. This is a generic feature set that is used in many works on PoS tagging and does not take into account the specificities of Romanian.

The transfer method of Täckström et al. (2013) relies on *type constraints* to filter out the tags transferred through alignment links. We extracted these constraints from both the English[3] and Romanian[4] WIKTIONARY: the latter covers a large vocabulary but only contains a few inflection tables and consequently does not contain most wordforms, while the former covers a smaller vocabulary but with complete inflection tables. At the end, we extracted PoS information for 405,125 wordforms and mapped these tags to the universal PoS tags (Petrov et al., 2012) used in all other corpora.[5] the resulting lexicon covers 69% of the types in RSAC, while coverage of 80% to 90% have been achieved

Experimental Evaluation To train a cross-lingual tagger for Romanian, we consider 300,000 Romanian sentences aligned with their translation in English, French, Italian and Spanish, extracted from the Europarl parallel corpus (Koehn, 2005). The taggers for

[1] http://www.phobos.ro/roric

[2] According to (Calacean and Nivre, 2009), the annotated texts are automatically saved using the ASCII character encoding, which explains the absence of the five Romanian diacritics (ă, â, î, ș and ț) from the treebank.

[3] http://en.wiktionary.org

[4] http://ro.wiktionary.org

[5] The mapping and the dictionary extracted from WIKTIONARY can be downloaded from https://perso.limsi.fr/wisniews/weakly/.

Source	en	fr	it	es	fr+it+es	avg(fr,it,es)
Weakly	79.6	79.1	79.1	79.1	79.3	79.1
Weakly+rules	82.0	82.7	81.8	82.7	82.5	82.4
Supervised				88.8		

Table 1: Accuracies of fully and weakly supervised taggers from various sources.

	NOUN	VERB	ADJ	ADV	DET	CONJ-ADP	PRT-PRON-NUM
Supervised	94	81	76	68	82	97	84
Transfer	84	81	75	54	73	92	67

Table 2: Accuracy for some PoS tags of the supervised and transferred (from fr+it+es) taggers.

the various source languages considered are trained on the Universal Dependency Treebank 2.0 (McDonald et al., 2013) (UDT). As advocated by (Wisniewski et al., 2014), we add handcrafted rules to describe the possible PoS tags of 15 frequent clitics and function words, the tokenization and annotation of which differ between datasets. For monolingual Romanian, we use RSAC, mapped to universal PoS tags and divided into RSAC-train and RSAC-test parts. A baseline supervised tagger is trained on the former and all models are evaluated on the latter.

Tables 1 and 2 report the performance, evaluated by the usual accuracy, of weakly and fully supervised taggers and the performance for each label. These results show that using English as a source language yields roughly similar results to Romance languages, even if it does not belong to the same linguistic family. Concerning the multi-source experiment, the final model has a lower accuracy than French or Spanish, but still higher than the average one. Considering that using multiple sources dispenses from a language-specific analysis to estimate the best source language, we consider that this loss is negligible compared to the cost of choosing Italian as source. Overall, the transferred taggers are outperformed by the supervised model, by 6 points, which is comparable to typical losses of such systems (Täckström et al., 2013; Wisniewski et al., 2014).

More precisely, Figure 1 shows the learning curve of the supervised tagger, which allows us to estimate the number of target sentences that must be annotated to reach the same error rates as cross-lingual models: the best cross-lingual model is equivalent to a supervised model trained on only 363 sentences. This observation strongly challenges the interest of cross-lingual transfer: for Romanian, only a very limited amount of annotated data is required to outperform a tagger trained on transferred annotations.

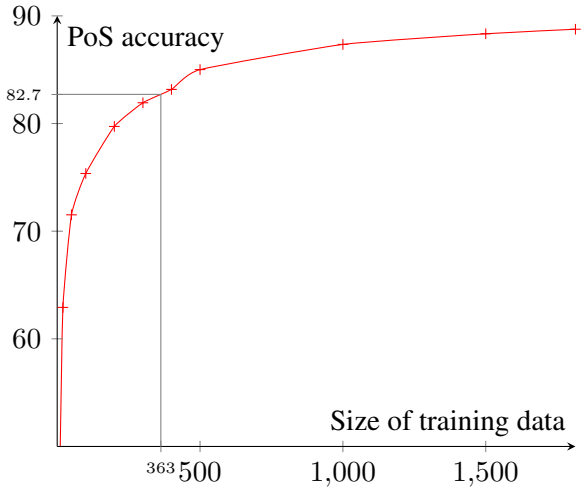

Figure 1: Accuracy of supervised taggers trained with different data sizes.

Error Analysis Analyzing the errors by categories reveals that adverbs are particularly poorly predicted by the supervised tagger, and even more by the Romance cross-lingual one. This can be correlated with the fact that, according to (Haspelmath and Tadmor, 2009), ADV is the category with the most Slavic loanwords (20%): such adverbs rarely benefit from the knowledge gathered on the remaining 80% and their prediction must be trained separately, therefore on even smaller data, which strongly degrades the model. When the main source of PoS knowledge is Romance, this effect unsurprisingly increases.

A detailed analysis also reveals that the word '*a*' has a high error rate. This is consistent with its actual ambiguity (among DET, VERB or PRT) and may also be partly explained, in the case of cross-lingual models, by the absence of some Romanian language constructs from other Romance languages: for instance, the infinitive '*a avea*' in Romanian aligns with '*avoir*' in French, which incorrectly biases PoS prediction of '*a*'

towards VERB.

This experiment suggests that the performance of such cross-lingual taggers mostly suffers from annotation and tokenization discrepancies, which was already pointed out by (Wisniewski et al., 2014). However, in most cases, those issues can easily be solved by a few handwritten corrections (e.g. our 15 extra rules, that improve the scores by 3 points) and since they represent an important part of the prediction errors, alleviating them may turn weakly supervised taggers into truly competitive solutions for low-resourced languages.

4. Cross-lingual dependency parser

We conduct a similar study on dependency parsing, considering the framework proposed by McDonald et al. (2011) to train parsers in Romanian from various combinations of source languages.

Transfer Method We briefly present here McDonald et al. (2011)'s algorithm. The transfer process starts with a delexicalized model transfer (Zeman and Resnik, 2008; McDonald et al., 2013): assuming all languages are annotated using a common PoS tagset, a model considering only PoS features is trained on a source language and used to parse directly Romanian. Using a common representation enables combination of multiple sources with raw treebank concatenation.

This crude approach has proven effective for many languages even if it is hindered by the lack of lexical information. To overcome this limit, McDonald et al. (2011) propose to relexicalize the model, in a second step: a small set of unannotated target data is first annotated by the delexicalized model, then used as training data for a new, lexicalized, model. As a final step, a parallel unannotated corpus is used to ensure parsing agreement of source and target models and rerank target hypotheses otherwise.

Formally, for each sentence pair, a *pseudo* reference is chosen, out of the k-best parsing hypotheses of the target sentence, by selecting the hypothesis which best aligns with the source parse according to an ALIGN metrics; the target sentence is then processed in the standard learning strategy with this new reference. For source and target parses y_s and y_t, ALIGN is computed as the sum of SCORE over the source-target pairs of edges $(s_{(i)}, s_{(j)})$ and $(t_{(i)}, t_{(j)})$, where i denotes the head of modifier j:

$$\text{SCORE}(y_s, y_t, (s_{(i)}, s_{(j)}), (t_{(i)}, t_{(j)})) = \begin{cases} +1 & \text{if } (s_{(i)}, s_{(j)}) \in y_s \text{ and } (t_{(i)}, t_{(j)}) \in y_t \\ -1 & \text{if } (s_{(i)}, s_{(j)}) \in y_s \text{ and } (t_{(i)}, t_{(j)}) \notin y_t \\ -1 & \text{if } (s_{(i)}, s_{(j)}) \notin y_s \text{ and } (t_{(i)}, t_{(j)}) \in y_t \\ 0 & \text{otherwise} \end{cases}$$

Intuitively, this metrics rewards parses with edges that strictly correspond to an edge in the source language through two alignment links, and penalizes every other pattern. Romanian being morphologically richer than other Romance languages, we use a slightly modified version of this metrics that accounts for frequent many-to-one alignments by adding a small reward ($\frac{1}{4}$) for edges between tokens aligned to the same token, instead of penalizing them. This models the fact that such tokens typically depend from each other, while keeping their impact on the parse quality low. The following metrics ensues:

$$\text{SCORE}(y_s, y_t, (s_{(i)}, s_{(j)}), (t_{(i)}, t_{(j)})) = \begin{cases} +1 & \text{if } (s_{(i)}, s_{(j)}) \in y_s \text{ and } (t_{(i)}, t_{(j)}) \in y_t \\ +\frac{1}{4} & \text{if } (s_{(i)}, s_{(j)}) \in y_s \text{ and } t_{(i)} == t_{(j)} \\ +\frac{1}{4} & \text{if } s_{(i)} == s_{(j)} \text{ and } (t_{(i)}, t_{(j)}) \in y_t \\ -1 & \text{if } (s_{(i)}, s_{(j)}) \in y_s \text{ and } (t_{(i)}, t_{(j)}) \notin y_t \\ -1 & \text{if } (s_{(i)}, s_{(j)}) \notin y_s \text{ and } (t_{(i)}, t_{(j)}) \in y_t \\ 0 & \text{otherwise} \end{cases}$$

Experimental Evaluation In all experiments, we train transition-based dependency parsers with the arc-eager transition system, an averaged perceptron, beam search of size 8 and early update, using our own implementation based on the recommendations of (Goldberg et al., 2013). We use universal PoS and the feature templates of (Zhang and Nivre, 2011), without labels and with decision history of size 8. These features, designed for English, have not been tailored to the specificities of Romanian.

We remove dependency annotations from RSAC-train to use it both as tagger trainset and parser relexicalization data. Source models are trained on UDT and Europarl is PoS annotated with supervised taggers and truncated to 80,000 sentences to limit the bias towards projection. The supervised parser is trained on annotated RSAC-train, thus enabling the comparison with the model relexicalized on the same data. All methods are evaluated on RSAC-test with gold PoS, in order to alleviate the loss due only to low quality taggers.

Table 3 presents the performance of the supervised model and successive cross-lingual ones and Table 4

Source	en	fr	it	es	fr+it+es	avg(fr,it,es)
Delexicalized	55.6	60.8	61.5	61.2	61.7	61.2
Relexicalized	57.4	61.8	62.1	62.1	61.6	62.0
Full transfer	65.7	67.0	66.9	67.1	67.1	67.0
Supervised				82.7		

Table 3: Performance (in UAS) of supervised and cross-lingual parsers from various sources.

	NOUN	VERB	ADJ	ADV	DET	CONJ	ADP	PRON	PRT	NUM
Supervised	89	79	86	70	93	73	88	74	91	78
Transfer	78	48	73	52	79	50	81	62	65	79

Table 4: UAS by child PoS tag of the supervised and transferred (from fr+it+es) parsers.

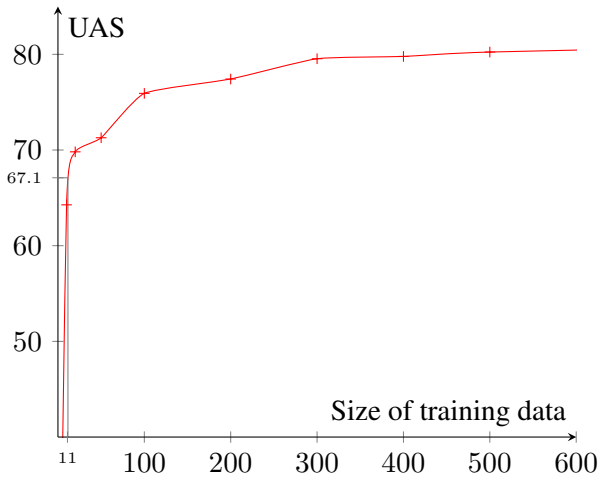

Figure 2: UAS of supervised parsers trained with different data sizes.

shows the attachment score for each PoS tag. It results that for dependency parsing, Romance languages are better sources than English, and both the achieved scores and the gain of relexicalization and projection are comparable to those obtained in (McDonald et al., 2011) and (McDonald et al., 2013). Contrary to the previous section, here using multiple sources brings indeed extra knowledge, and errors can not be attributed to one category in particular: both open (VERB, ADV) and closed (CONJ, PRT) classes suffer greatly from weak supervision.

However, Figure 2 provides a visual comparison of the best cross-lingual model with supervised ones, and it shows that the former one has performance equivalent to supervision by less than 15 sentences. This surprising result reveals that, even if it improves the performance of a delexicalized model, for the considered dataset, the method of McDonald et al. (2011) only captures very few information and the annotation effort it saves is actually quite small.

5. Discussion

Experimental results of our cross-lingual PoS tagger and dependency parser correlate with intuitions offered by language similarities: Romance sources are individually better than English (at least slightly) and multi-source transfer consistently improves the scores over an average single source. However, in both cases cross-lingual performance are disappointing when quantitatively compared to supervised models. This is particularly true for dependency parsing.

An important source of errors is the annotation scheme discrepancy, which is a known issue (McDonald et al., 2013). A promising step has been taken to solve it by the Universal Dependencies project and the recent apparition of Romanian in version 1.2, but this is still a work in progress and the Romanian treebank has still to grow bigger to become usable in realistic conditions.

This discrepancy only explains partly the low performance, though. To enlighten this, we notice that in both systems the state-of-the-art methods we have used only consider parts of the existing cross-lingual information: word alignments for tagging, and for parsing, word alignments and syntactic similarity between related languages. This ignores the linguistic knowledge that resides in wordforms, and with regards to the fact that related languages are generally at least partly mutually intelligible, we believe that the systems could be greatly improved by accounting for their strong lexical similarities. The fact that, despite a closer relatedness, taggers transferred from English still yield comparable performance to taggers transferred from Romance languages, supports this affirmation.

In addition, some syntactic structures cannot be fully encoded in PoS tags and word alignments only. For instance the varying prevalence of constructs such as

'*like*' versus '*please to*', or the preferred used of subordinates instead of completives in Romanian show that syntactic phenomena can not be fully represented through the noised channel that word alignments are. This also explains the varying transfer efficiency on both tasks: while PoS projection only supposes that linked words (i.e. co-occurring ones) share a same PoS tag, which is generally true even for unrelated languages, dependency transfer systems make the strong, and seemingly wrong, assumption that related languages have similar syntax.

Finally, such methods oversee the fact that a large part of the predictions are done on closed classes (33% of RSAC tokens) or easy dependencies (typically 50% of the tokens are linked with a neighbor), while such information could already be efficiently learned on even small annotated target datasets; otherwise is seems like we are wasting large amounts of cross-lingual data to indirectly get these easy parameters in a weakly supervised way. Some evidences, such as Slavic loanwords, are nevertheless difficult to learn by transfer. We think that this issue could be tackled by developing true multi-source transfer in which languages from which the annotations are transferred would not be pre-selected but the model will automatically select the relevant information from each source.

Further experiments should also deepen our error analysis, and study for instance if source, target and cross-lingual have the same error typology, or if high performance in source languages are due to good prediction of phenomena that simply do not exist in the target language, e.g. French definite articles.

In a nutshell, apart from annotation scheme issues, which are to be handled, we feel that state-of-the-art methods do not fully capitalize on cross-lingual knowledge. They underestimate the amount of available information and should gather finer knowledge sources, e.g. lexical ones, and be more selective in picking the information where it exists: in small annotated data and other sources.

6. Conclusion

The purpose of this study was to quantify the benefits of annotating target data, compared to restricting to transfer methods, word alignments and automatic data acquisition.

Our results show that cross-lingual PoS taggers, while below the performance of available supervised ones, still yield interesting results, equivalent to annotating about 350 sentences. On the contrary, the underlying structures involved in dependency parsing are too complex to be encoded in word alignments and rely on strong language-specific components. 15 annotated sentences in Romanian convey indeed more linguistic knowledge than large amounts of parallel and monolingual data in other languages.

We therefore recommend investigating hybrid methods combining cross-lingual transfer with a small amount of target annotated data, in order to learn the missing language-specific structures. Such methods would also benefit from leveraging lexical similarities among languages in the same family, and from appropriate handling of multiple sources.

In future work, we intend to evaluate the effects of introducing fine-grained lexical similarity weighting in multi-source cross-lingual systems.

7. Acknowledgements

This work has been partly funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 645452 (QT21) and the French *Direction générale de l'armement*.

8. Bibliographical References

Calacean, M. and Nivre, J. (2009). A data-driven dependency parser for romanian. In *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT 7)*, pages 65–76.

Chrupala, G., Dinu, G., and van Genabith, J. (2008). Learning morphology with morfette. In *LREC*.

Colhon, M. and Simionescu, R. (2012). Deriving a statistical syntactic parsing from a treebank. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, page 34.

Goldberg, Y., Zhao, K., and Huang, L. (2013). Efficient implementation of beam-search incremental parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 628–633.

Haspelmath, M. and Tadmor, U., (2009). *Loanwords in the World's Languages: A Comparative Handbook*, chapter 8. Walter de Gruyter.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

Lacroix, O., Aufrant, L., Wisniewski, G., and Yvon, F. (2016). Frustratingly easy cross-lingual transfer for transition-based dependency parsing. In *NAACL*.

McDonald, R., Petrov, S., and Hall, K. (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 62–72.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K.,

Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *Knowledge and Data Engineering, IEEE Transactions on*, 22(10):1345–1359.

Perez, C.-A. (2012). Linguistic resources for the processing of the romanian language. In *University Al. I. Cuza of Iasi*.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation*, LREC'12, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Roegiest, E. (2006). *Vers les sources des langues romanes: un itinéraire linguistique à travers la Romania*. Acco.

Straka, M., Hajič, J., Straková, J., and Hajič Jr, J. (2015). Parsing universal dependency treebanks using neural networks and search-based oracle. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, page 208.

Täckström, O., Das, D., Petrov, S., McDonald, R., and Nivre, J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Wisniewski, G., Pécheux, N., Gahbiche-Braham, S., and Yvon, F. (2014). Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1779–1785.

Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In *IJCNLP*, pages 35–42.

Zhang, Y. and Nivre, J. (2011). Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 188–193.