Hassle-free POS-Tagging for the Alsatian Dialects

Delphine Bernhard, Anne-Laure Ligozat

Hassle-free POS-Tagging for the Alsatian Dialects

Delphine Bernhard[1], **Anne-Laure Ligozat**[2]
LiLPa, Université de Strasbourg, France[1]
LIMSI-CNRS, Orsay & ENSIIE, Evry, France[2]

Abstract

In this article, we present a method to perform POS-tagging for the Alsatian dialects, relying on tools developed for the German language. The results show that a simple transposition of closed-class words already leads to great improvements over results obtained for the original texts.

1 Introduction

The Alsatian dialects are spoken in the Alsace region, located in the North-East of France, next to the German border. They belong to the Franconian and Alemannic language families [4]. According to a recent study, 43% of the Alsatian population still speaks the regional dialect [5]. However, the proportion of Alsatian speakers is decreasing regularly since the 1960s, to the benefit of the French language.

Alsatian poses several important challenges for computational tools:

- There is no standard and widely acknowledged writing convention;
- Alsatian is actually a continuum of dialects, with lexical and pronunciation variants;

- Digital text corpora and resources are scarce.

Given these constraints, it is difficult to develop text processing tools for the Alsatian dialects using standard methods, which require either large amounts of annotated text or trained professionals. In this article, we focus on one of the first building blocks in any text processing system and present a simple yet effective technique for performing a basic morphosyntactic analysis of Alsatian texts. In this method, we rely on the closeness between Alsatian and standard German: closed-class words (determiners, pronouns, prepositions, conjunctions) and auxiliary verbs are transposed into their German translations and German Part-Of-Speech (POS) taggers are subsequently applied.

2 Related Work

Non-standard writing is an issue for NLP tools, which are generally built for standard data, untainted by spelling errors or unusual phenomena. This problem has been addressed from different perspectives in the context of POS tagging.

Giesbrecht and Evert [2] evaluate German POS taggers on Web documents and show that the performance of taggers is lower on such kind of data. They also note that the performance is highly dependent on the text genre and that some Web-specific genres, such as forums, are harder to process.

When annotated data is available, it is possible to train a new tagger. This approach has for instance been used by Dipper [1] who trains the TreeTagger based on different versions of a corpus of Middle High German texts.

However, when annotated data is scarce or non-existent, it is necessary to use methods relying on tools and resources developed for a similar language. Hana et al. [3] describe a tagger for Old Czech based on two strategies. The first consists in transforming a corpus of modern Czech so that it looks like Old Czech and then train a POS tagger. The tagger can then be applied to a modernised old Czech corpus, which, when converted

back to the original form, constitutes a tagged old Czech corpus. The second strategy consists in using a morphosyntactic resource for old Czech, in order to approximate the emission probabilities for the tagger.

We reuse some of the ideas proposed by Hana et al.: (i) we partially transform the Alsatian texts into German and (ii) we apply this transformation to the most frequent words, i.e. closed class words. The transformations could not be based on sound and graphemic changes as done by Hana et al., since these changes are not systematic in Alsatian (see Figure 1, which displays some graphical variants found in several online lexicons for two Alsatian lexemes).

French	**German**	**English**	**Alsatian variants**
cuisine	Küche	kitchen	Kuch, Kucha, Kische, Khésche, Kùch, Kücha, Kuche, Kiche, Kuchi
lundi	Montag	monday	Mondàà, Mantig, Mandig, Mondàà, Mondoe, Mondàj, Maandi, Mandi

Figure 1: Alsatian variants for two lexemes, as found in several online lexicons.

3 Approach

In order to assess how well Alsatian texts can be processed by German POS taggers, we have collected a corpus of 4 texts:

- *Alsace*: an article from a local newspaper, *L'Alsace*, entitled "Zìmlig beschta Frìnd" and dated February 18th 2012.[1]
- *Duttlenheim*: a summary of a theater play posted on the website of a theater company, in the year 2004.[2]
- *Hoflieferant*: a page from the theater play entitled "D'r Hoflieferant" by Gustave Stoskopf (1906).

[1] http://www.lalsace.fr/actualite/2012/02/18/zimlig-beschta-frind
[2] http://theatreduttlenheim.free.fr/html/annee2004.htm

- *Wikipedia*: an article from the alemannic Wikipedia, on the topic of the Alsatian museum of Strasbourg, retrieved on October 30th, 2012. [3]

Document	# tokens	# transpositions	# unknown lemmas before		# unknown lemmas after	
Alsace	320	103	194	60.6%	102	31.9%
Duttlenheim	166	31	61	36.7%	35	21.1%
Hoflieferant	230	39	74	32.2%	44	19.1%
Wikipedia	399	143	248	62.2%	139	34.8%

Table 1: Corpus statistics.

These texts were semi-automatically annotated using a simplified set of POS tags: ADJ (Adjective), ADP (Preposition / Postposition), ADV (Adverb), CARD (Cardinal number), CONJ (Conjunction), DET (Determiner), FM (Foreign Material), ITJ (Interjection), N (Noun), PRN (Pronoun), PRT (Particle), V (Verb), Punctuation. To do so, we manually translated the texts into approximate German and then corrected the tags provided by the Stanford POS Tagger (see Figure 3). A similar semi-automatic method for obtaining a gold standard was employed by Giesbrecht and Evert [2].

POS-tagging of the original texts was performed by applying the German TreeTagger [6] and Stanford POS Tagger [7] to two different kinds of inputs :

- Manually tokenised raw texts, without further preprocessing ;
- Texts where closed class words and auxiliary verbs were transposed into the equivalent word forms in German, e.g. *àwer* → *aber*, *fer* → *für*, *isch* → *ist*, etc. Figure 2 displays an example Alsatian sentence, before and after the transposition. The German gloss shares 5 words with the transposed Alsatian sentence.

[3] `http://als.wikipedia.org/wiki/Els\%C3\%A4ssisches_Museum_\%28Stra\%C3\%9Fburg\%29`

ALS	Brüchsch	kenn	Angscht	ze han	for	mich	, papa .
TRANS	Brüchsch	keine	Angscht	zu haben	für	mich	, papa .
GER-GLOSS	Brauchst	keine	Angst	zu haben	für	mich	, Vater.
ENG-GLOSS	Need	no	fear	to have	for	me	, dad
FRE-GLOSS	As besoin	pas	peur	à avoir	pour	moi	, papa .

Figure 2: Example transposition (TRANS) from Alsatian (ALS) and GERMAN (GER-GLOSS), English (ENG-GLOSS) and French (FRE-GLOSS) glosses.

Overall, the lexicon used for the transposition contains 107 entries. The lexicon only includes unambiguous entries as one Alsatian form may correspond to several German word forms. Table 1 displays the number of tokens in the corpus, as well as the number and percentage of unknown lemmas (according to the TreeTagger), before and after the transposition.

4 Results

In order to evaluate the results, we compare the automatically annotated texts against our manual annotations. The more detailed tags provided by the TreeTagger and the Stanford Tagger are automatically mapped to our simplified tagset. Results are given in Tables 2 for the TreeTagger and 3 for the Stanford Parser.

The accuracy of the TreeTagger on the original texts is quite low: under 0.70 for all texts. Yet, after transposition, its accuracy is much higher, since the minimum is 0.78 for the Hoflieferant text. We calculated its accuracy for non-transposed words only, in order to evaluate whether tagging performance improved only on transposed words, or if other words were also impacted. Non-transposed words are also better analysed: for example, the tagging accuracy improves from 0.71 to 0.77 for non-transposed words in the Wikipedia text.

The Stanford Tagger has slightly better results than the TreeTagger, probably because of its tagging model, but the fact that the manual an-

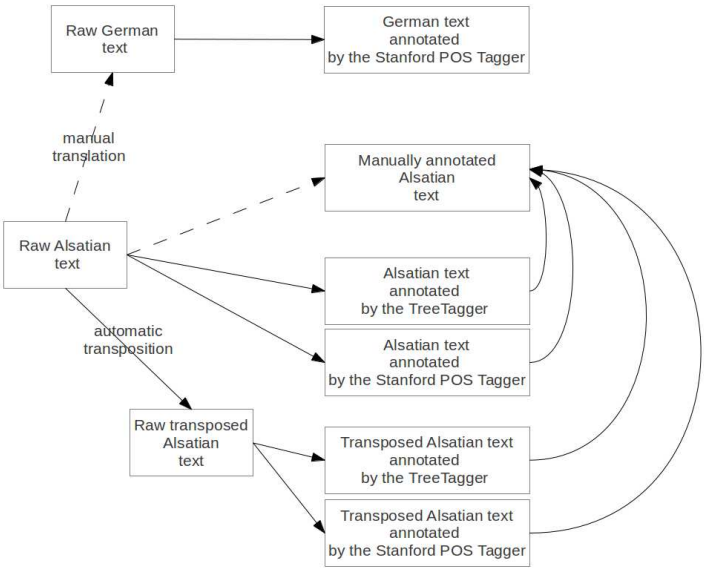

Figure 3: POS-tagging of raw and transposed Alsatian texts

notation was based on the Stanford Tagger annotations could also lead to a small bias. Results are also substantially improved on the transposed files: for example, the tagging accuracy raises from 0.53 to 0.85 on the Wikipedia file. The accuracy for non-transposed words is also improved, as with the TreeTagger.

5 Conclusion and Perspectives

The results show that a simple approach which consists in transposing closed-class words already leads to considerable gains in the tagging accuracy. However, there is still room for improvement, in order to reach performance levels above 95%. One possibility that we wish to explore is

Document	Original file	Original file, non-transposed words only	Transposed file	Transposed file, non-transposed words only
Alsace	0.48	0.68	0.79	0.74
Duttlenheim	0.67	0.78	0.86	0.83
Hoflieferant	0.64	0.72	0.78	0.75
Wikipedia	0.50	0.71	0.83	0.77

Table 2: TreeTagger accuracy

Document	Original file	Original file, non-transposed words only	Transposed file	Transposed file, non-transposed words only
Alsace	0.56	0.74	0.86	0.83
Duttlenheim	0.77	0.83	0.88	0.86
Hoflieferant	0.67	0.75	0.82	0.80
Wikipedia	0.53	0.70	0.85	0.80

Table 3: Stanford PoS Tagger accuracy

the identification of Alsatian and German cognates, which could then be automatically transposed. Another possibility would consist in integrating word class information found in available Alsatian lexicons in the tagging process.

References

[1] Stefanie Dipper. Morphological and Part-of-Speech Tagging of Historical Language Data: A Comparison. *Journal for Language Technology and Computational Linguistics*, 26(2):25–37, 2011.

[2] Eugenie Giesbrecht and Stefan Evert. Is Part-of-Speech Tagging a

Solved Task? An Evaluation of POS Taggers for the German Web as Corpus. In *Proceedings of the 5th Web as Corpus Workshop (WAC5)*, 2009.

[3] Jirka Hana, Anna Feldman, and Katsiaryna Aharodnik. A Low-budget Tagger for Old Czech. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH '11)*, pages 10–18, 2011.

[4] Dominique Huck, Arlette Bothorel-Witz, and Anemone Geiger-Jaillet. L'Alsace et ses langues. Éléments de description d'une situation sociolinguistique en zone frontalière. *Aspects of Multilingualism in European Border Regions: Insights and Views from Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol*, page 13–100, 2007.

[5] OLCA / EDinstitut. Étude sur le dialecte alsacien, 2012.

[6] Helmut Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49, 1994.

[7] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, HLT-NAACL'03, page 173–180, 2003.