**Topic:** Frisian POS-tagger

**IDs in Google Sheets:**

**Author:** Hans Van de Velde

**What exists:**

* Written corpora of modern, middle and old Frisian and several (transcribed) speech corpora. FA is modernising its existing linguistic research infrastructure (Taaldatabank Fries), incorporating these datasets in a new research infrastructure, Poarte ta it Frysk.
* At this moment, there is a limited amount of tagged data available and all language data have to be tagged manually

**What must be created anew:**

* Part-of-speech tagger for Frisian. In the first phase we will focus on present-day Frisian, future developments will be POS-tagging for middle and old Frisian. For the development, existing tools such as FROG, TreeTagger and spaCy will be used, which will keep the development costs relatively low. The POS-tagger will also be used to build a language detector for Frisian.
* A reusable route to build a POS-tagger for low resource languages

**Targeted/Actual users:**

* linguists with theoretical interests;
* researchers of Frisian and other West-Germanic languages;
* developers of tools for Frisian language users.

**Actual use (quantify!):**

The modernization and enrichment of the corpora will increase their use by researchers and developers (a new target group). Over the past six months, the Taaldatabank Fries had around 100 unique visitors, twenty of them running at least twenty requests. With more data, the use will increase.

**Social Impact** **(concrete examples):**

Better analysis of written and spoken corpora is crucial for the further development of language tools for Frisian and will improve the quality of search algorithms and translation services. The provincial government has a tradition of funding the development of such tools (see taalweb.frl for products developed by Fryske Akademy) and will continue to do so (see Bestuursafspraak Friese Taal en Cultuur). The language detector will help in obtaining data on the use of Frisian in social media and on the Internet, which will help in developing a more adequate language policy.

**Estimate in PMs (try to justify):**

Total 7 PM, of which 5 PM will be funded by CLARIAH+. Extra costs will be covered by Fryske Akademy (see full project description below for details).

|  |  |  |
| --- | --- | --- |
|  | **corpus to be used** | **time period** |
| Plain model transfer / Dutch | any Dutch corpus | 1 week |
| Plain model transfer / Middle Frisian | Midfrysk corpus | 1 week |
| Retrieving modern lemmas for historical words | Midfrysk corpus | 1 month |
| Checking accuracy & correcting tagged corpus |  | 1 month |
| Annotation projection from Dutch | Oersetter corpus | 1 month |
| Checking accuracy & correcting tagged corpus |  | 1 month |
| Using the ‘Fryske standertwurdlist’ | any Frisian corpus | 2 weeks |
| Checking accurary & correcting tagged corpus |  | 1 month |
| Writing paper(s) |  | 1 month |
|  |  | 7 months |

**Lead + PMs:**

* Hans Van de Velde (FA): project leader (covered by FA)
* Gosse Bouma (RUG) (1 PM)
* 6 PM developers at FA: Wilbert Heeringa / Derk Drukker / Eduard Drenth.

**Participants + PMs:**

FA: will receive 5 PM

RUG: FA will make an arrangement with RUG for the work of Gosse Bouma

**A POS tagger for West Lauwers Frisian**

**1. Introduction**

While a wealth of tools has been developed for large (inter)national languages like English, Spanish and French, this is not obviously the case for smaller regional minority languages, such as West Lauwers Frisian, a language having about 450,000 speakers in the province of Friesland in the Northwest of the Netherlands. For example, for the TreeTagger parameter files are provided for only a few minority languages (Catalan, Galician, Mongolian), and the smallest of them (Galician) is still spoken by 2.4 million speakers.

While small languages are struggling with restricted possibilities for developing language tools, there is a need to develop an efficient strategy that makes the development of tools feasible, for example by taking a route via a larger and related language for which the desired tools have already been developed. In fact, for each (minority) language at least a POS tagger should be available, a tool by which each word in a corpus is marked up as corresponding to a particular part of speech, based on both its definition and its context (definition Wikipedia). The main purpose of using POS tags is “disambiguation”. When searching for the verb ‘ship’ in large corpus, one would like to get only those sentences that include ‘ship’ being used as a verb. Without POS tags one needs to go check all sentences that include ‘ship’ and find the ones with the verb ‘ship’ manually .

Phrase chunking, a natural language process that separates and segments of a sentence into its subconstituents, such as noun, verb and prepositional phrases, builds on top of POS tags. These tags in turn can be used as features for higher level tasks such as building parse trees, which can in turn be used for named entity resolution, coreference resolution, sentiment analysis and question answering.

POS tagging is fundamental for textometry, a discipline in which knowledge is derived from corpora without predefined information models. MacMurray & Leenhardt (2011) describe textometry as an approach in which “a text possesses its own internal structure that would be difficult to analyze by manual means alone. By applying statistical and probabilistic calculations directly to the textual units of comparable texts in a corpus it becomes possible to analyze patterns and trends that would otherwise be obscured by the quantity of the textual units.” (p. 606).

For example, by looking at frequencies of POS categories using POS-tagged corpora Hirst & Feiguina (2007) presented a method for authorship discrimination that is based on the frequency of bigrams of syntactic labels that arise from partial parsing of the text. With this method the authors obtained a high accuracy on discrimination of the work of Anne and Charlotte Brontë (Brontë 1847, Brontë 1848, Brontë 1853), both alone and combined with other classification features.

While Hirst & Feiguina (2007) focused on determining the authorship of texts, Nerbonne & Wiersma (2006), Lauttamus et al. (2007), Wiersma et al. (2010), and Nerbonne et al. (2010) measured the impact of L1 on L2 syntax in second language acquisition on the basis of corpora of English of Finnish Australians. They presented an application of a technique from language technology to tag a corpus automatically and to detect syntactic differences between two varieties of Finnish Australian English, one spoken by the first generation and the other by the second generation. The technique compares frequencies of trigrams of part-of-speech categories as indicators of syntactic distance between the varieties and then examine potential effects of language contact. The frequency vectors were compared and analyzed by using a permutation test, which resulted in both a general measure of difference and a list with the *n-*grams that are most responsible for the difference. The findings showed syntactic ‘contamination’ from Finnish in the English of the adult first-generation speakers of Finnish ethnic origin. The results show that we can attribute some interlanguage features in the first generation to Finnish substratum transfer.

Sanders (2007) extended the method and its application. He extended the method by using leaf-path ancestors of Sampson (2000) instead of trigrams, which captures internal syntactic structure – every leaf in a parse tree records the path back to the root. The corpus used for testing is the International Corpus of English, Great Britain (Nelson et al., 2002), which contains syntactically annotated speech of Great Britain. The speakers were grouped into geographical regions based on place of birth. Sanders showed that dialectal variation in eleven British regions from the International Corpus of English, Great Britain (ICE-GB) is detectable by the algorithm, using both leaf-ancestor paths and trigrams.

A practical application that we currently envisage is the use of the POS tagger in a language detector. This detector is used for detecting whether a Twitter message is written in Frisian or Dutch. Basically this detector determines for each message the proportion of words found in the Frisian word list and the proportion of words found in a Dutch word list (e.g. CELEX). Since there is overlap the sum of the two number will be higher than 100%. However, we would like to find out whether the ratio of the two proportions differs per part of speech. If a Twitter message contains both Frisian and Dutch words, is, for example, the proportion of Frisian nouns relatively higher than the proportions of Frisian verbs?

Due to the size of the corpora, the only viable tagging option is an automatic annotation. There are several papers that focus on tagging (historical) texts via a closely-related language. This approach is promising when developing a POS-tagger for low-resource languages like Frisian.

In the simplest case Frisian texts are tagged by taggers which are trained on the basis of tagged corpora of closely-related languages or language varieties such as Dutch, Middle Frisian or maybe even German. Scherrer, who tagged Alsatian data using a German POS-tagger obtained an accuracy of 48% to 77%. Tjong Kim Sang (2016) tagged 17th-century Dutch texts with a modern Dutch POS-tagger and got accuracies of 62.8% and 63.7%. The low accuracies show that this approach is not enough, but it can serve as a baseline to which other approaches can be compared.

Results can be improved by adapting the spelling of corpora prior to tagging them with a tagger. For example, Hupkes & Bod (2016) investigated the POS-tagging of 17th-century Dutch texts. They found that “that modernizing the spelling of corpora prior to tagging them with a tagger trained on contemporary Dutch results in a large increase in accuracy, but that spelling normalization alone is not sufficient to obtain state-of-the-art results.”

Yarowsky et al. (2001), Bentivogli et al. (2004), Moon and Baldridge (2007), Van Huyssteen & Pilon (2009), Hupkes & Bod (2016) and Agić et al (2016) transferred annotation using using parallel sentences of parallel corpora of closely-related languages. Using this approach the goal is to augment an unannotated target sentence with syntactic annotations projected from one or more source sentences through word alignments. A POS-tagger then is trained on a corpus which was automatically annotated by projecting (automatically assigned) POS-tags via word alignments from a contemporary corpus.

Tjong Kim Sang (2016) also POS-tags 17th-century Dutch texts. He translates the texts to modern Dutch word by word and subsequently uses Frog, a POS tagger for modern Dutch. He evaluated four word-by-word translation methods. In the first method, the machine translation system Moses was used with two versions of the Dutch Statenvertaling bible, one from the year 1637 and one from 1888. The second approach used the Integrated Language Bank (GTB), an online collection of historical dictionaries, with links of historical words to their modern counterparts. The lexicon service makes it possible to retrieve modern lemmas for historical words. The third used a lexicon that was learned from two versions of a Dutch bible, one from the seventeenth century and one from the nineteenth century. The fourth method employed orthographic rules learned from the learned lexicon. The rules converted historical character sequences to their modern equivalent. The lexicon-based methods outperformed the method that used orthographic rules.

**2.** **Towards a POS-tagger for West Lauwers Frisian**

In order to develop a Frisian POS-tagger we first need a tagged corpus and second a POS tagger that will be trained using the corpus. We consider at least the following POS-taggers:

* Frog (https://languagemachines.github.io/frog/)
* TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/)
* spaCy (https://spacy.io/usage/linguistic-features)

Regarding the corpora, we will investigate multiple approaches.

**2.1 Plain model transfer**

In the simplest case we can use an approach that is known as 'plain model transfer'. The idea is to pretend that languages A and B are in fact the same. A model is trained on A and applied to B without change. We see three possibilities:

a. To train a tagger using a Dutch corpus and use it for tagging Frisian texts.

b. To train a tagger using a Middle Frisian corpus in order to tag modern Frisian texts. The Middle Frisian corpus available at the Fryske Akademy was originally annotated by Willem Visser. The corpus can be searched via a web interface at <https://web2.fa.knaw.nl/corpus-frontend/frysk/search> where the user can select ‘Midfrysk’ as a language variety. A REST interface is available as well (BlackLab / Apache Lucene, see also: http://inl.github.io/BlackLab/blacklab-server-overview.html).

c. Scherrer & Rabus (2017) proposed to combine existing resources from the etymologically close Slavic languages Russian, Ukrainian, Slovak, and Polish and adapt them to Rusyn. Similarly, we can develop a multi-source tagger that uses both Dutch and Middle Frisian training data.

**2.2** **Retrieving modern lemmas for historical words**

A further improvement may be obtained by respelling historical texts. Most approaches proposed by Tjong Kim Sang’s (2016) assume the existence of parallel corpora. There are only a handful modern Frisian translations of the Middle Frisian texts. However, together with the POS tags modern Frisian lemmas are given for the Middle Frisian words.

**2.3 Annotation projection from Dutch**

In this approach we use a parallel corpus used for the online Frisian-Dutch translation program ‘Oersetter’ [translator] which is available at: <https://taalweb.frl/oersetter> or <http://oersethelp.nl/>. The corpus consists of 120 Frisian and 120 Dutch texts. The Frisian texts have an average word count of 25,019.39 words (3,002,327 words in total), and the Dutch texts have an average word count of 42,143.73 words ( 5,057,248 words in total).

A pilot experiment has been carried out by Gosse Bouma (University of Groningen). First he aligned 1000 Dutch sentences from the Oersetter corpus with the corresponding Frisian sentences using fast\_align, a simple, fast and unsupervised word aligner (see: https://github.com/clab/fast\_align, Dyer et al. (2013)). Then, the Dutch words got POS tags assigned with spaCy. Finally, the POS-tags of the Dutch words were copied to the corresponding Frisian words.

Looking at the results, the aligner works well for word for word translations. The aligner was not able to process swaps correctly. The POS-tagger did not always assign the right tags to the Dutch words. From the 1000 sentences a little more than 300 sentences were completely aligned. Therefore, POS tags were assigned to all words in this subset of sentences.

We see the following possibilities for improvement:

* Training of fast\_align on all training data (160K sentences).
* Using other word aligners, for example efmaral (https://github.com/robertostling/efmaral, Östling & Jörg Tiedemann (2016)) or berkeleyaligner (https://code.google.com/archive/p/berkeleyaligner/).
* Respelling the Frisian text, especially of ' t, ' n and ' e, words ending on *st*, words ending on *sto*.
* Using other POS-taggers.

**2.4 Using the ‘Fryske standertwurdlist’ [Frisian standard word list]**

This list contains all Frisian lemmas with their morphological forms. All words are POS-tagged and the coding adheres to Universal Dependencies (UD), a framework for cross-linguistically consistent grammatical annotation (see: <http://universaldependencies.org/> and Petrov et al. (2011)).

Using the word list we will assign a POS tag to each word in the Frisian texts. In order to correctly assign a POS tag to a word that has multiple and different POS tags in the word list, the context in which they occur needs to be considered. Since this cannot be done automatically, those words will be marked so that they can easily be tagged manually afterwards. We will also try to use statistical information from Dutch trigrams to tag Frisian.

**3. Planning**

|  |  |  |
| --- | --- | --- |
|  | **corpus to be used** | **time period** |
| Plain model transfer / Dutch | any Dutch corpus | 1 week |
| Plain model transfer / Middle Frisian | Midfrysk corpus | 1 week |
| Retrieving modern lemmas for historical words | Midfrysk corpus | 1 month |
| Checking accuracy & correcting tagged corpus |  | 1 month |
| Annotation projection from Dutch | Oersetter corpus | 1 month |
| Checking accuracy & correcting tagged corpus |  | 1 month |
| Using the ‘Fryske standertwurdlist’ | any Frisian corpus | 2 weeks |
| Checking accurary & correcting tagged corpus |  | 1 month |
| Writing paper(s) |  | 1 month |
|  |  | 7 months |

The table does not necessarily show the order by which the tasks are carried out. CLARIAH+ will fund 5 PM, FA the rest.

**4. Output**

A tagger for modern Frisian freely available under GPL (see: http://www.gnu.org/copyleft/gpl.html), a paper in an A journal.

**5. Literature**

Agić, Ž., Johannsen, A., Plank, B., Alonso, H. M., Schluter, N., & Søgaard, A. (2016). Multilingual projection for parsing truly low-resource languages. Transactions of the Association for Computational Linguistics, 4, 301-312.

Bentivogli, L., Forner, P., and Pianta, E. (2004). Evaluating cross-language annotation transfer in the multisemcor corpus. In Proceedings of the 20th international conference on Computational Linguistics, page 364. Association for Computational Linguistics.

Brontë, A. (1847). Agnes Grey. London: T.C. Newby. Published under pseudonym Acton Bell.

Brontë, A. (1848). The Tenant of Wildfell Hall. London: T.C. Newby. Published under pseudonym Acton Bell.

Brontë, C. (1853). Villette. London: Smith, Elder & Company. Published under pseudonym Currer Bell.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In Proc. of NAACL.

Hirst, G. and Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts, Literary & Linguistic Computing 22(4): 405–419.

Hupkes, D., & Bod, R. (2016). POS-tagging of Historical Dutch. In LREC.

MacMurray, E. and Leenhardt, M. (2011). Textometry and Information Discovery: A New Approach to Mining Textual Data on the Web. Proceedings of the ICAI Conference 2011, Las Vegas, pp. 605-611.

Lauttamus, T., Nerbonne, J. and Wiersma, W. (2007). Detecting syntactic contamination in emigrants. The English of Finnish Australians, SKY Journal of Linguistics 21: 273–307.

Moon, T. and Baldridge, J. (2007). Part-of-speech tagging for middle english through alignment and projection of parallel diachronic texts. In EMNLP-CoNLL, pages 390–399.

Nerbonne, J., Lauttamus, T., Wiersma, W. and Opas-Hänninen, L. L. (2010). Applying Language Technology to Detect Shift Effects. In Norde, M., de Jonge, B. and Hasselblatt, C. (eds.), Language Contact. New Perspectives. Amsterdam: Benjamins, 2010. Series IMPACT: Studies in Language and Society, pp. 27–44.

Nerbonne, J. and Wiersma, W. (2006). A Measure of Aggregate Syntactic Distance. In Nerbonne, J. and Hinrichs, E. (eds.) Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006, pp. 82–90.

Petrov, S., Das, D., & McDonald, R. (2011). A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Östling, R., & Tiedemann, J. (2016). Efficient word alignment with markov chain monte carlo. The Prague Bulletin of Mathematical Linguistics, 106(1), 125-146.

Sanders, N. C. (2007). Measuring Syntactic Difference in British English. In: Proceedings of the ACL 2007 Student Research Workshop. Madison: Omnipress, pp. 1–6.

Scherrer, Y., & Rabus, A. (2017). Multi-source morphosyntactic tagging for Spoken Rusyn. In Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) (pp. 84-92).

Tjong Kim Sang, E. (2016). Improving part-of-speech tagging of historical text by first translating to modern text. In International Workshop on Computational History and Data-Driven Humanities (pp. 54-64). Springer, Cham.

Van Huyssteen, G. B. and Pilon, S. (2009). Rule-based conversion of closely-related languages: a dutch-to-afrikaans convertor.

Wiersma, W., Nerbonne, J. and Lauttamus, T. (2010). Automatically Extracting Typical Syntactic Differences from Corpora, Literary and Linguistic Computing 26(1): 107–124.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In: Proceedings of the first international conference on Human language technology research, pages 1–8. Association for Computational Linguistics.