Oersetter: Frisian-Dutch Statistical Machine Translation

Maarten van Gompel and Antal van den Bosch, Centre for Language Studies, Radboud University Nijmegen and Anne Dykstra, Fryske Akademy

> **Abstract**

In this paper we present a statistical machine translation (SMT) system for Frisian to Dutch and Dutch to Frisian. A parallel training corpus has been established, which has subsequently been used to automatically learn a phrase-based SMT model. The translation system is built around the open-source SMT software Moses. The resulting system, named *Oersetter*, is released as a website for human end users, as well as a web service for software to interact with. We here discuss the workings, setup and performance of our system, which to our knowledge is the very first Frisian-Dutch SMT system.

> Introduction

In the past decade Machine Translation (MT) has gained considerable ground, not in the least through the rising popularity of web-based services such as Google Translate, which has seen a steady increase in its array of supported languages. Most current systems such as Google Translate are based on statistics derived from parallel translated texts. Whilst far from perfect, statistical machine translation has become a useful tool as it has given any user the means to at least understand the gist of previously indecipherable texts.

Frisian, however, has not yet made an appearance on this stage. Nevertheless, an automated translation system may be a valuable aid in certain circumstances and help reduce the effort of a human translator. A tool such as the one proposed here might be worthwhile for non-Frisian speakers who seek to understand a Frisian text. Lesser used languages such as Frisian pose extra challenges for machine translation, as it is often difficult to collect sufficient data to train the statistical model.

Machine translation can roughly be divided into two approaches: rule-based machine translation and statistical machine translation (or broader: example-based machine translation). In the rule-based approach linguistic experts compile a database of translation rules, often syntactic in nature. Translation in this approach then generally begins with a syntactic (and to some extent semantic) analysis, yielding a more abstract

representation that is converted using the rules in the database to an appropriate abstract structure in the target language. This then acts, together with a bilingual lexicon, as a template for generating the final translation.

In the literature the rule-based approach has largely been superseded by statistical machine translation, which has proven to deliver superior translation quality, especially with the advent of phrase-based SMT (Köhn, Och, Marcu, 2003). This is a data-driven approach, in which a system automatically learns how to translate from one language into another by means of a parallel training corpus, i.e. a collection of texts that are translations of one another. On the basis of the corpus an SMT system models statistical conditional probabilities of certain sequences of words in the two languages being translations of each other, and generates a translation model. The approach uses no handcrafted linguistically-motivated rules, and is in essence language independent.

In this study, we aim to apply phrase-based SMT to the language pair Frisian-Dutch, in order to generate systems that translate in both directions. First we discuss the workings of SMT in more detail, then we present the data we collected for the system, and finally we discuss the results we obtained in an experimental setup in which we test on translations the system was not trained on.

> Statistical Machine Translation

A good translation should faithfully convey the meaning of the original, and it should be rendered in fluent natural language. In statistical machine translation, two distinct statistical models represent these two aspects. The *translation model* is used to compute the likelihood of a sentence being faithful to the original meaning, and the *target language model* imposes a maximally fluent natural word order on the resulting translation in the target language by scoring typical, predictable word order as more probable than uncommon or malformed combinations. The central processing component of an SMT system is the *decoder*, which computes probabilities according to at least these two models for a huge number of possible translation hypotheses. This constitutes a vast search problem in which countless hypotheses are tested and compete against one another for the best probability score according to the joint statistical models. The translation chosen is the hypothesis found to attain the best score. Due to the size and complexity of the search problem, and the need to keep time and memory requirements manageable, considerable pruning of the search takes place. It is quite possible that the selected translation is found in what is called a *local maximum*, and not necessarily

a *global maximum*, the theoretically highest achievable score if the search space were explored exhaustively.
The translation model, language model and possibly other included models each contribute according to a certain weight in establishing the probability score for each translation hypothesis. These weights are parameters to the system and can be optimised empirically through *minimum error rate training* (MERT) (Och, 2003). This procedure tests various weight values and evaluates the system on a "development set" of sentence pairs not included in the training material. MERT is an iterative process that will eventually choose weights for the various models that minimise the error ratio on the development set.
The statistical probabilities for translation model and language model are automatically learned from example data. The input for the translation model is a parallel corpus of Dutch-Frisian texts, assembled specifically for this project. The final translation model takes the form of a *phrase translation table*, which maps phrases, i.e. consecutive word *n*-grams, in the source language to a scored distribution of phrases in the target language. A short and simplified excerpt of the Frisian to Dutch phrase-translation table is shown below:

brûke kin	→	kan aanwenden (0.5), kan gebruiken (0.5)
bus nei de stêd	→	bus naar de stad (1.0)
bus nei hûs	→	bus naar huis (1.0)
de sneinske klean	→	de zondagse kleren (1.0)
de sluting	→	de sluiting (0.5), het sluiten (0.5)

The phrases need not be sound linguistic units, but simply emerge from the data as being likely translations that often co-occur in translated sentences. To be able to count such co-occurrences, the Frisian-Dutch parallel texts have first been sentence-aligned, meaning that sentences that are translations are neatly grouped. Subsequently, a word alignment was built using GIZA++ (Och and Ney, 2003), linking each Frisian word to a Dutch word. This then is the basis upon which phrases can be extracted and a phrase-translation table is finally formed. When presented with input to be translated, it is the job of the decoder to again form coherent sentences in the target language on the basis of all translated phrase fragments for that input sentence.
The target language model is a *trigram model* trained with maximum-likelihood estimation. For a given string of words, such as a hypothesised translation, it yields a pseudo-probabilistic score corresponding to the likelihood of that sequence.

> Data

Two datasets have been compiled for the construction of the Frisian-Dutch system. The first is the Dutch-Frisian parallel corpus that the translation model is trained on. The second model is the Frisian language model used in Dutch to Frisian translation. A third dataset, the Dutch language model, for the reverse direction, did not need any special effort as it was compiled from other readily available sources for Dutch.

A statistical machine translation system learns to translate from the data it is trained on. This implies that the choice of domain is an important one, and that it is best to mix different domains if a generic translation system is the goal. If a system is trained on for instance only judiciary texts, then such a system may perform well on legal texts, but if suddenly confronted with another genre then translation quality will likely turn out worse.

Frisian is largely a spoken language. Though Frisian has a literary tradition, the number of Frisian novels cannot by far compete against the number of Dutch novels. Compared to Dutch, the number of non-literary texts is also relatively small. The Province of Fryslân is a strong advocate of spoken and written Frisian. Many provincial, and also municipal, official texts are in Dutch and in Frisian. It is inherent to a small language like Frisian that written Frisian does not cover as many domains as Dutch. The situation concerning bilingual Dutch and Frisian texts is even worse. Consequently, we have to make do with what is available. The Frisian-Dutch parallel corpus at the moment contains a number of novels, technical texts and official texts from the provincial and municipal authorities. The corpus at this moment is relatively small and far from well balanced. We are still digitising texts to add to the corpus.

Another constraint is that the latest Frisian spelling reform was introduced in 1980. Since we aim at translations in the current spelling, we can only include texts that date from after 1980. Frisian does not have a fully-fledged standard yet, which means that some Frisian texts may contain words from one of the three major dialects. As a result, some translations may be partly in dialect. The Fryske Akademy is working on a Standard for Frisian. As soon as the Standard has been established, the corpus will be standardised.

In addition, the parallel corpus has been expanded with 3,141 excerpts from the Nederlandse Volksverhalenbank[1], a digital collection of about 26 thousand Dutch and 16 thousand Frisian folktales maintained at the Meertens Institute, Amsterdam, the Netherlands. As most folktales in the database have a Dutch summary, we were able to select 3,146 parallel text fragments

1 http://www.verhalenbank.nl

where the number of words in the Dutch summary is larger than 80% of the number of words in the Frisian text, producing pairs such as

> *Ien dy't graech in poppe ha woe, hong in kikkert oan 'e doar. Dan kaem de eibert.*
> Iemand die graag een kind wil hebben, hangt een kikker aan de deur. Dan komt de ooievaar.

The reason to include material from the Volksverhalenbank is that we wanted to expand the corpus in a fairly easy and quick way. A serious drawback of the Volksverhalenbank is that the material is in the spelling used before 1980. In the example above 'graech' and 'kaem' would for instance be spelled 'graach' and 'kaam' in the current spelling. In our analyses below we have not considered the influence of the two different spelling systems.

After sentence-alignment using the *uplug* software (Tiedemann, 1999), we end up with a total of 44,503 sentence pairs, containing 701,782 words of Frisian and 673,277 words of Dutch, including punctuation tokens. This is not a large corpus in SMT standards; for European language pairs, EU corpora are available with tens of millions of words.[2]

The corpus source for the Frisian language model is monolingual in nature and consists of 594,975 sentences and 10,043,516 words, making it considerably larger than the parallel corpus. The Frisian portion of the parallel corpus is also included in this language model. The corpus contains texts from 1980 onwards. The FA tried to cover as many domains as possible. Yet, a major part of the corpus inevitably consists of literary texts.

> Results

Evaluation of Machine Translation quality is not trivial. When asking multiple people for a translation of a sentence, multiple results may be produced that may all be good translations. Although human evaluation is always to be preferred, it is often impractical in evaluating machine translation system performance. In this study we evaluate our MT system using automated metrics such as BLEU (Papineni et al, 2003). These metrics compute a measure of overlap between the translation produced by the system and one or more provided by a human.

To obtain these human-translated reference sentences, we keep a number of the sentence pairs in our parallel corpus apart for testing, which means

2 E.g. see the Open Corpus, http://opus.lingfil.uu.se/

that we do not include them to train the system. Similarly, another part of our sentence pairs is held out as a development set for the purpose of optimising the system's parameters in MERT training. Such separate sets are necessary for a fair and unbiased evaluation. This results in 43,252 sentence pairs for training, 1,000 for the test set, and 250 for the development set.
We conducted various experiments to assess the quality of the translations across a variety of evaluation metrics common in MT literature. The results are shown in Table 1. For BLEU, METEOR and NIST, the higher scores are the better ones. TER, WER and PER are error rates and therefore here lower scores are better. Comparison against an unoptimised baseline was made to measure the impact of the language model and the MERT parameter optimisation.

Dutch>Frisian	**BLEU**	**METEOR**	**NIST**	**TER**	**WER**	**PER**
1. Unoptimised	0.4609	0.6456	8.7035	0.38	0.4059	0.3201
2. +Frisian LM	0.4921	0.6588	8.9552	0.38	0.4043	0.3106
3. +MERT	0.4892	0.654	8.7543	0.37	0.3818	0.323
4. +LM+MERT	**0.523**	**0.673**	**9.124**	**0.36**	**0.3782**	**0.2991**
Frisian>Dutch	**BLEU**	**METEOR**	**NIST**	**TER**	**WER**	**PER**
5. Unoptimised	0.4971	0.6628	8.9216	0.36	0.3823	0.3005
6. +Dutch LM	0.4914	0.6627	8.9179	0.37	0.4031	0.3016
7. +MERT	**0.5104**	**0.6673**	**8.9804**	**0.34**	**0.3567**	0.3008
8. +LM+MERT	0.5008	0.6658	8.9549	0.36	0.3938	**0.2992**

Table 1: MT Evaluation results on Dutch to Frisian (top) and Frisian to Dutch (bottom).

The second experiment ("2.") in Table 1 demonstrates the positive effect of the extended Frisian Language Model, as contrasted to the first experiment ("1."), which only uses a language model generated on the Frisian part of the much smaller parallel corpus data itself. Experiment 3 on the third line shows the positive effect of MERT optimisation, and experiment 4 combines both techniques, MERT optimisation as well as an extended Frisian Language Model, and this combination gives the overall best result for Dutch to Frisian translation. For that reason this configuration has been selected to use in the final *Oersetter* system.
Surprisingly, the Dutch Language model used in experiment 6 shows a deterioration over the baseline. The language model used here is a massive language model generated from the largest (monolingual) Dutch corpora available. Although vast (over 500 million words), this corpus is completely disjoint from the comparatively tiny Dutch portion of the

parallel corpus, whereas the language model in experiment 5 is based solely on this Dutch portion of the parallel corpus. The fact that these do not intersect may account for the drop below baseline in experiment 6. Additional learning curve experiments show a clear picture of the impact of the amount of training data. In Figure 1 we see that translation quality overall improves as more training data are added. We expect this trend to continue in a log-linear fashion as more parallel training data are added.

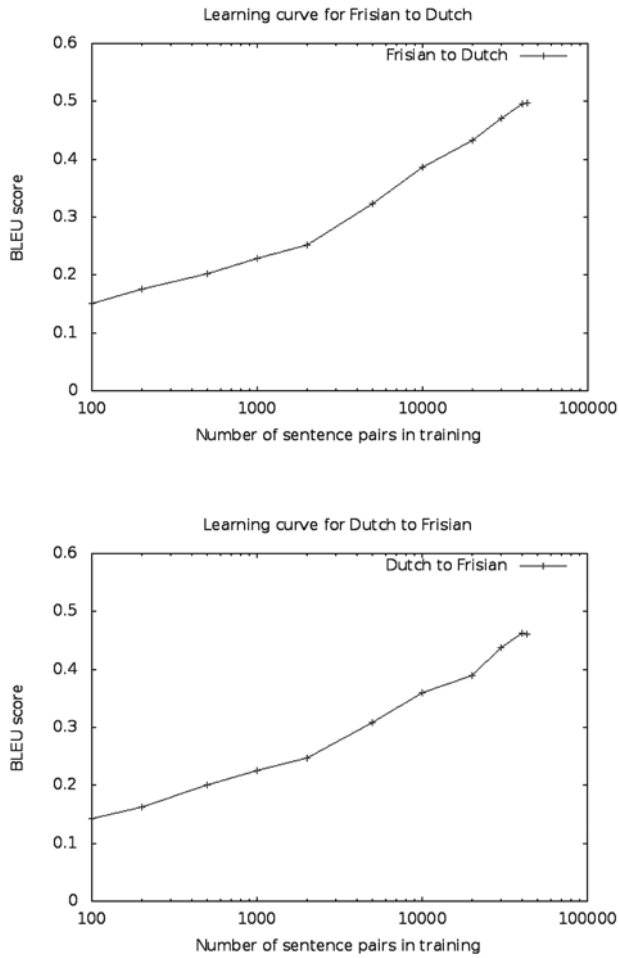

Figure 1: Learning curves for Frisian to Dutch (top), and Dutch to Frisian (bottom). BLEU score as a function of the number of sentence pairs in the training set.

The above analysis gives a fair impression of the system, but it is not as transparent as some example translations produced by the Dutch to Frisian system:

Correct translations:

A. *Recht boven mij is een spin bezig met een web.*
Rjocht boppe my is in spin dwaande mei in web.

B. *Men komt in aanraking met andere mensen, men krijgt ander werk.*
Men komt yn oanrekking mei oare minsken, men kriget oar wurk.

C. *Ik zou mij in een andere naam niet zo goed thuis hebben gevoeld.*
Ik soe my yn in oare namme net sa goed thús field hawwe.

D. *We bulderden van het lachen.*
Wat laken wy.

Incorrect translations:

E. *Twee mannen haalden eens twee visjes in de stad.*
Twa kammeraden wienen op Sinteklaesfreed togearre nei stêd.

We note that despite the modest size of the parallel training corpus, the translation scores are high in comparison to typical scores obtained in Dutch-English experiments. However, these scores can not be readily compared with the ones from experiments on other corpora language pairs. To provide some reference: In the SMT literature, a popular parallel corpus for research is the Europarl corpus. This contains the proceedings of the European Parliament in several major European languages. In version 3 of this corpus (Tiedemann, 2009) we count 1,313,076 sentence pairs for Dutch-English, on which we achieve a BLEU score of 0.233 for Dutch to English. In this present study, in contrast, we trained on a modest 44,503 sentence pairs. The reason that translation quality is this high can be attributed to a large extent to the fact that Dutch and Frisian are very closely related languages, often following similar syntactic patterns. Sentences A and B in the above excerpt illustrate this; they follow a word-by-word pattern. Sentence C exhibits a minor but mandatory change in word-order (*hebben gevoeld field hawwe*). Sentence D shows an entirely different structure. Here a larger phrasal fragment, the entire sentence even, has been found in the phrase-translation table and has been mapped to its translation. It may also occur that the system learns the

wrong phrase translation, as illustrated in sentence E where the phrase *"mannen haalden eens twee visjes in de"* has been erroneously mapped to *"kammeraden wienen op Sinteklaesfreed togearre nei"*, due to an incorrect sentence alignment stemming from the folktale material included in the training set.

We provide some technical notes on how the system proposed in this study is made available, and what software has been used. A test version of the *Oersetter* system is for the time being online accessible through *http://fa.demo.textinfo.nl*. At the core of the machine translation system lies the decoder Moses (Köhn et al., 2007). A Moses server is running for each of the translation directions. Interaction with Moses is mediated through an MT experiment framework written as part of the Colibri project, a currently on-going PhD research project at the Centre for Language Studies of Radboud University, focussing on machine translation. We used CLAM (van Gompel, 2012) to quickly build a RESTful web service for the system[3], allowing other software to interact with it. Finally, a custom web-interface interacting with the underlying web service was built to provide a user-friendly front-end. All software used and developed is open source.[4]

> Conclusion

This study has shown the viability of Dutch to Frisian and Frisian to Dutch statistical machine translation. Work has been done to assemble a parallel corpus. A monolingual Frisian corpus of about 10 million words has been used for the generation of a language model, which proves beneficial compared to using only the Frisian material in the parallel corpus. The collected corpus and language model prove successful in SMT; trained on about 44.5 thousand sentences, and tested in both directions between the language pairs, considerably higher scores are obtained in automatic evaluation metrics than an English-Dutch system trained on over a million sentences. The scores are about as high as the highest scores obtained on the main European language pairs reported in the literature. We conclude that an important factor contributing to the overall high results is the close similarity between Dutch and Frisian, which makes the job of translation easier as less reorderings are needed and words and phrases are more easily mapped to their similar counterparts.

3 A full RESTful specification of the *Oersetter* web service can be found at http://webservices.ticc.uvt.nl/oersetter/info/

4 Software is licensed under the GPL v.3; see http://www.gnu.org/copyleft/gpl.html

> **Acknowledgements**

We are grateful to Dolf Trieschnigg for extracting the Dutch-Frisian translations from the Nederlandse Volksverhalendatabank and to Wytske Rypma, Janneke Spoelstra and Doete Venema for establishing the parallel corpus.

> **Literature**

A. Dykstra and J. Reitsma, 'De struktuer en de ynhâld fan 'e Taaldatabank fan it Frysk'. *It Beaken* 55 (1993), pp. 55-82.

M. van Gompel, 'CLAM: Computational Linguistics Application Mediator. Documentation', *ILK Technical Report 12-02* (Tilburg 2012).

P. Köhn, F.J. Och, D. Marcu, 'Statistical phrase-based translation', *NAACL '03 Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology* - Volume 1 (2003) pp. 48-54.

P. Köhn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst, 'Moses: Open source toolkit for statistical machine translation', *ACL '07 Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions* (2007) pp. 177-180.

F.J. Och, 'Minimum Error Rate Training in Statistical Machine Translation', *ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* - Volume 1 (2003) pp. 160-167.

F.J. Och and H. Ney, 'A Systematic Comparison of Various Statistical Alignment Models', *Computational Linguistics,* volume 29, number 1 (2003) pp. 19-51.

K. Papineni, S. Roukos, T. Ward and Wei-jing Zhu, 'BLEU: a Method for Automatic Evaluation of Machine Translation', *ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (2002) pp. 311-318.

J. Tiedemann, 'Uplug - a modular corpus tool for parallel corpora', *Technical Report 17, Department of Linguistics, University of Uppsala* (1999).

J. Tiedemann, 'News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces', in: N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov (eds.), *Recent Advances in Natural Language Processing*, volume 5, pp. 237-248. (Amsterdam/Philadelphia 2009).