Cross-Lingual Dependency Parsing for Faroese with UUparser: Project Report

Xingming Li

March 16, 2024

1 Introduction

Cross-lingual dependency parsing is a crucial task with applications in natural language processing (NLP) and multilingual models. Parsing Faroese presents unique challenges due to its low resource. The project aims to explore cross-lingual dependency parsing using UUparser, a versatile parser known for its cross-lingual capabilities. It investigates the effectiveness of single-source and double-source models with different transfer languages for parsing Faroese. It also checks into the impact of publication year of the text for single-source models and varying proportions of Icelandic and Norwegian Nynorsk for double-source models. The primary research question focuses on determining the optimal configurations for cross-lingual dependency parsing for Faroese using UUparser. The report provides detailed insights into the performance, strengths, and challenges of these configurations.

2 Related Work

Previous research, such as Tyers et al. (2018), has investigated methods for improving cross-lingual dependency parsing by creating synthetic treebanks from multiple annotated data sources. These approaches leverage diverse linguistic resources to enhance parsing accuracy for low-resource languages. While various techniques, including transfer learning and novel architectures, have been explored, challenges persist in optimizing parsing models for languages with unique characteristics and limited resources. Tyers et al. (2018) underscores the potential of synthetic treebanks in improving cross-lingual parsing performance, providing an inspiration for the present project's exploration of cross-lingual dependency parsing for Faroese using UUparser.

3 Data and Method

3.1 Faroese

Faroese is a North Germanic language spoken by approximately 70,000 people, primarily in the Faroe Islands, an autonomous territory of Denmark. Ginsburgh and Weber (2011) has highlighted it belongs to the Insular Scandinavian branch of the Germanic language family, closely related to Icelandic and Western Norwegian dialects. Faroese exhibits rich inflectional morphology, with nouns, adjectives, and verbs inflecting for case, number, and gender. Additionally, it features complex syntactic structures, including free word order and a rich system of subordinate clauses. Faroese is considered a low-resource language in the context of NLP. This means that there is limited annotated data, linguistic resources, and computational tools available for developing language technologies such as parsers. Cross-lingual approaches offer a solution by leveraging annotated data from related languages to improve parsing performance.

3.2 UUparser

UUparser serves as the foundation for the experiments of the project. At its core, UUparser relies on pre-trained multilingual word embeddings and a neural network architecture designed to process input sentences and capture universal syntactic dependencies. The model predicts universal dependency labels adhering to a standardized annotation scheme, facilitating seamless knowledge transfer between languages. Multilingual training exposes the model to diverse linguistic patterns, encouraging the learning of shared syntactic structures. Additionally, transfer learning techniques are employed to adapt knowledge from high-resource languages to low-resource languages, enhancing performance on languages with limited annotated data.

3.3 Experimental Setup

The experimental setup focuses on few-shot cross-lingual dependency parsing for Faroese using UUparser. The datasets include Universal Dependencies (UD) treebanks of Icelandic-IcePaHC, Danish-DDT, Norwegian-Bokmaal, Norwegian-Nynorsk, and Swedish-LinES as training datasets, with the Faroese-FarPaHC treebank serving as the target for parsing, focusing on the Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS) for its development set at the optimal iteration for each model.

Experiment	**TRF 1**	**Training Sentence**	**TRF 2**	**Training Sentence**
Single-Source	Icelandic-IcePaHC	500		
	Danish-DDT	500		
	Norwegian-Bokmaal	500		
	Norwegian-Nynorsk	500		
	Swedish-LinES	500		
Double-Source	Icelandic-IcePaHC	250	Danish-DDT	250
	Icelandic-IcePaHC	250	Norwegian-Bokmaal	250
	Icelandic-IcePaHC	250	Norwegian-Nynorsk	250
	Icelandic-IcePaHC	250	Swedish-LinES	250
Varying Ratios	Icelandic-IcePaHC	400	Norwegian-Nynorsk	100
	Icelandic-IcePaHC	300	Norwegian-Nynorsk	200
	Icelandic-IcePaHC	200	Norwegian-Nynorsk	300
	Icelandic-IcePaHC	100	Norwegian-Nynorsk	400

Table 1: Numbers of Sentences Selected from Treebanks for Training

For the Faroese dataset, the training set consists of 100 sentences from Gospel of St. John (Edition: 1937, published in 1936), and the development set comprises 300 sentences from Acts of the Apostles (Edition: 1937, published in 1928).[1] For each set of experiments, the specific treebank data volumes for training are shown in Table 1. In the single-source experiments involving transfer languages, each transfer language is represented by a training set containing 500 sentences and a development set with 300 sentences. The double-source experiments entail combining the Icelandic-IcePaHC treebank with four other training treebanks of different transfer languages. Each combined set has a training set containing 500 sentences (250 sentences each) and a development set with 300 sentences for each treebank. The experiments examining different proportions of Icelandic and Norwegian Nynorsk involve combining the Icelandic-IcePaHC treebank with the Norwegian-Nynorsk treebank. The training set consists of 500 sentences with varying proportions (Icelandic accounts for 80%, 60%, 40% and 20% respectively), and the development set comprises 300 sentences for each treebank. All sentences mentioned above have been selected from the beginning of the treebanks.

Additionally, since the first 500 training sentences extracted from the Icelandic-IcePaHC training set were published in the year 1150, to further explore the influence of temporal variations, two additional subsets consisting of 500 training sentences each were chosen from the Icelandic-IcePaHC training set, originating from the years 1908 and 2008, respectively.[2]

This setup ensures that all experiments have been trained on a total of 600 sentences, including 100 same Faroese sentences, facilitating a robust and convincing comparison.

Experiments	**Models**	**UAS**	**LAS**
Single-Source	Icelandic	71.68	63.94
	Danish	66.60	56.79
	Norwegian Bokmaal	56.77	56.84
	Norwegian Nynorsk	67.16	57.89
	Swedish	67.15	57.42
Double-Source	Icelandic-Danish	70.91	62.98
	Icelandic-Norwegian Bokmaal	71.05	63.06
	Icelandic-Norwegian Nynorsk	71.58	63.26
	Icelandic-Swedish	69.68	61.82
Varying Ratios	Icelandic 80%-Norwegian Nynorsk 20%	70.80	63.13
	Icelandic 60%-Norwegian Nynorsk 40%	70.63	62.91
	Icelandic 40%-Norwegian Nynorsk 60%	70.04	61.95
	Icelandic 20%-Norwegian Nynorsk 80%	69.19	60.75
Varying Ages	Icelandic 1908	69.36	61.79
	Icelandic 2008	69.84	62.37

Table 2: UAS and LAS for the Faroese Development Set at the Optimal Iteration for Each Model (Treebank names have been omitted and ”Icelandic” represents the model trained by first 500 sentences from UD_Icelandic-IcePaHC treebank which were published in 1150)

[1] https://universaldependencies.org/treebanks/fo_farpahc/index.html

[2] https://universaldependencies.org/treebanks/is_icepahc/index.html

4 Results

The results obtained from the experiments are shown in Table 2. In the single-source models, utilizing Icelandic-IcePaHC as the transfer language treebank yielded promising results, achieving a UAS of 71.68 and LAS of 63.94 at its peak performance. Meanwhile, the Norwegian-Nynorsk transfer language treebank exhibited more notable performance than the other three, reaching its zenith with a UAS of 67.16 and LAS of 57.89.

Moving on to the double-source models, combinations such as ”Icelandic-Danish”, ”Icelandic-Norwegian Bokmaal”, ”Icelandic-Norwegian Nynorsk”, and ”Icelandic-Swedish” are analyzed. Each combination’s impact on UAS and LAS scores is considered, shedding light on the advantages of the ”Icelandic-Norwegian Nynorsk” model which achieved a UAS of 71.58 and LAS of 63.26.

Additionally, combined models with varying proportions, such as ”Icelandic 80%-Norwegian Nynorsk 20%”, ”Icelandic 60%-Norwegian Nynorsk 40%”, ”Icelandic 40%-Norwegian Nynorsk 60%”, and ”Icelandic 20%-Norwegian Nynorsk 80%”, are evaluated. These experiments provide valuable insights into the influence of different language proportions in multi-source models on the overall parsing performance, especially with the increase in Icelandic data we can see the UAS and LAS scores are also increasing.

In addition to the aforementioned experiments, the influence of temporal variations on the performance of cross-lingual dependency parsing models is also investigated. The results show that parsing Faroese using texts published in 1150 still achieved the highest UAS and LAS scores.

To complement the detailed analysis of individual models, learning curves are presented in Figure 1, illustrating how LAS scores evolve over epochs for each model. These visual representations offer a dynamic perspective on the model’s convergence and performance development throughout the training process and show that throughout the training process, single-source model with Icelandic published in 1150 and double-source model with Icelandic and Norwegian Nynorsk consistently demonstrate superior performance compared to other models, achieving high scores across multiple epochs.

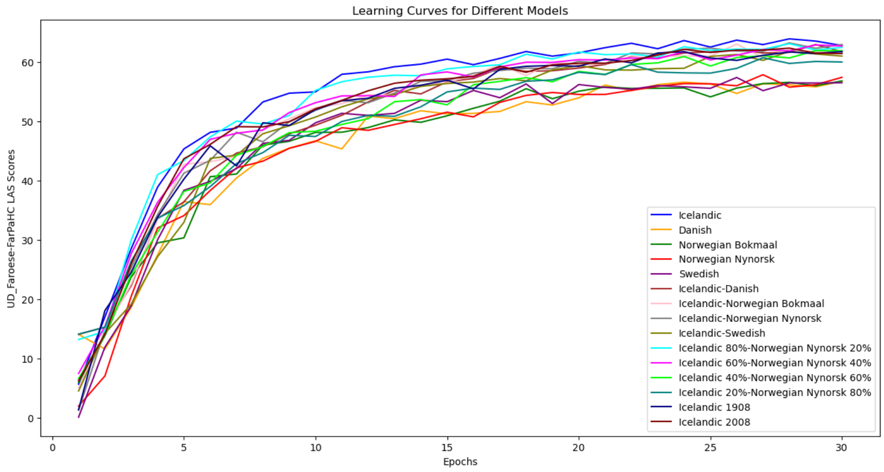
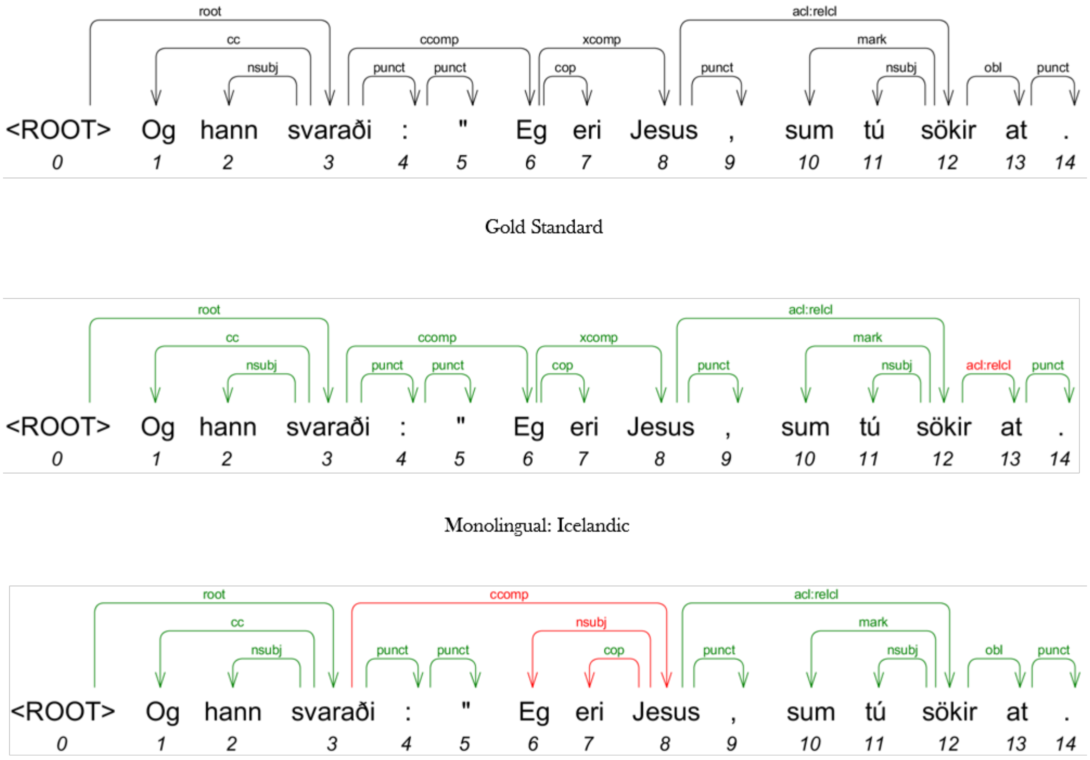


Figure 1: Learning Curves Illustrating the Development of LAS Scores (Treebank names have been omitted and ”Icelandic” represents the model trained by first 500 sentences from UD_Icelandic-IcePaHC treebank which were published in 1150)

5 Qualitative Evaluation

In comparing the parses across different models, especially between single-source model with Icelandic and double-source model with Icelandic and Norwegian Nynorsk, notable differences were observed in the dependency structures for certain sentences. For instance, the sentence in Figure 2 involves complex syntactic constructions (i.e. nested clauses). The single-source model with Icelandic as the transfer language tended to produce more accurate parses by correctly annotating all nested clauses.



Gold Standard

Monolingual: Icelandic

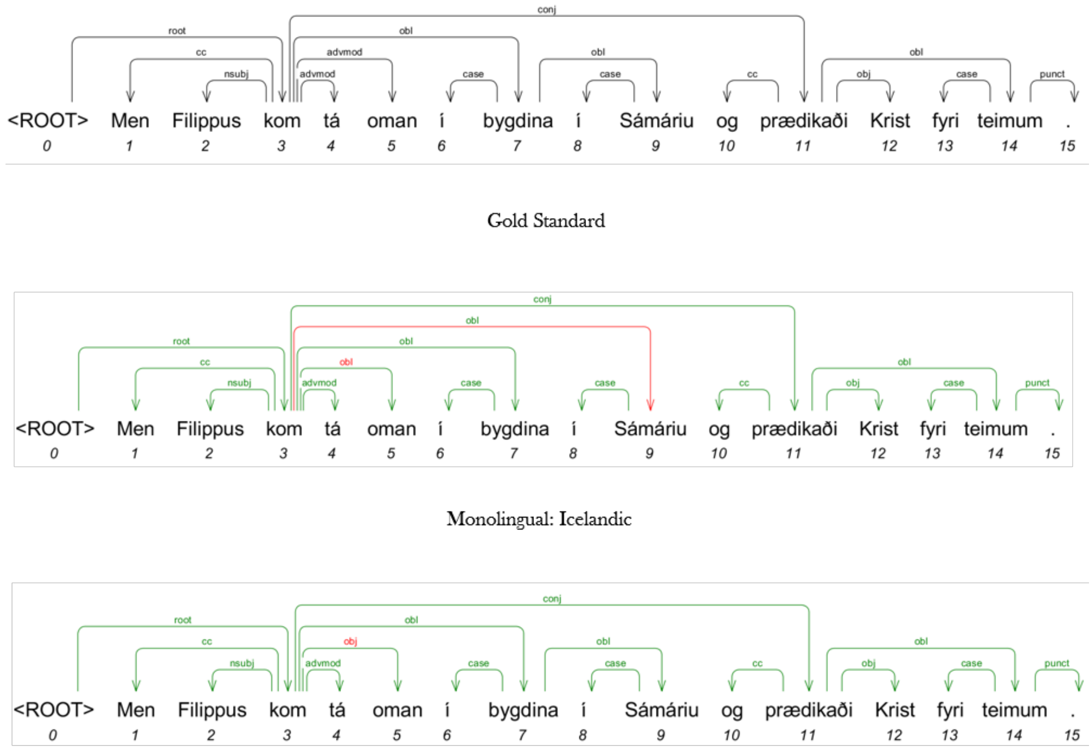
Multilingual: Icelandic + Norwegian Nynorsk

Figure 2: Sentence 272 in Development Set of UD_Faroese-FarPaHC (This sentence means "And he answered: 'I am Jesus, whom you are persecuting.")

However, as shown in Figure 3, in sentences with simpler structures, the double-source model sometimes outperformed the single-source model by providing more nuanced and contextually appropriate dependency relations. These discrepancies highlight the importance of considering both syntactic complexity and linguistic diversity when evaluating cross-lingual dependency parsing models.

6 Discussion

The results of the experiments shed light on several important aspects of cross-lingual dependency parsing for Faroese using UUparser. Firstly, the performance of single-source models varied significantly depending on the choice of transfer language. While Icelandic emerged as the most effective transfer language as expected, Norwegian Nynorsk also



Gold Standard

Monolingual: Icelandic

Multilingual: Icelandic + Norwegian Nynorsk

Figure 3: Sentence 231 in Development Set of UD_Faroese-FarPaHC (This sentence means ”But then Philip came down to the city of Samaria and preached Christ to them.”)

demonstrated promising results, indicating the importance of linguistic similarity and shared syntactic structures between the transfer and target languages.

Moreover, the double-source models showcased the potential benefits of leveraging multiple transfer languages. Combinations such as ”Icelandic-Norwegian Nynorsk” yielded notable performance in parsing accuracy, suggesting that incorporating diverse linguistic knowledge can enhance the model’s ability to capture the nuances of syntactic structures.

The experiments examining varying proportions of Icelandic and Norwegian Nynorsk provided valuable insights into the impact of data distribution on parsing performance. Results indicate that increasing the proportion of Icelandic data generally leads to improved parsing accuracy, highlighting the significance of high-quality Icelandic training data in cross-lingual parsing tasks for Faroese.

Additionally, the analysis of temporal variations revealed intriguing findings regarding the influence of publication year on parsing performance. Being trained on texts dating back to the year 1150, the single-source Icelandic model consistently outperformed those trained on more recent data. This unexpected result might attribute to the text genres and underscores the robustness of historical linguistic knowledge and its relevance in contemporary NLP tasks.

The experiments highlight the importance of careful selection of transfer languages, data distribution, and temporal considerations in optimizing cross-lingual dependency parsing models for Faroese. By leveraging diverse linguistic resources and exploring historical texts, we can enhance the accuracy and robustness of parsing models, facilitating more effective natural language understanding across different languages and time periods.

A primary limitation of the project lies in the generalizability of findings to other languages, as the effectiveness of cross-lingual dependency parsing models for Faroese may not directly translate to languages with different linguistic structures, resource availability, or typological characteristics. Additionally, the impact of data quality and annotation guidelines on parsing performance could introduce biases or errors, potentially limiting the reliability of the project's conclusions.

7 Conclusion

In conclusion, the project provides insights into cross-lingual dependency parsing for Faroese using UUparser. Through a series of experiments, it explored the effectiveness of single-source and double-source models with different transfer languages, examined the impact of language and treebank proportions, and investigated temporal variations in training data. The findings demonstrate that leveraging transfer languages such as Icelandic and Norwegian Nynorsk can significantly improve parsing accuracy. Moreover, the analysis highlights the importance of data distribution, with increased proportions of Icelandic data leading to enhanced parsing performance. Furthermore, the unexpected influence of historical texts on parsing accuracy underscores the value of resorting diverse linguistic resources, including texts dating back centuries. By integrating historical knowledge into parsing models, it can achieve more robust and accurate natural language understanding.

Overall, the project contributes to the advancement of cross-lingual dependency parsing techniques and stresses the importance of linguistic diversity and historical knowledge in natural language processing. Future work can aim to further explore the potential of incorporating additional linguistic resources and historical texts to enhance parsing performance for Faroese and other low-resource languages.

References

Ginsburgh, V. and Weber, S. (2011), *How many languages do we need? The economics of linguistic diversity*, Princeton University Press.

Tyers, F., Sheyanova, M., Martynova, A., Stepachev, P. and Vinogorodskiy, K. (2018), Multi-source synthetic treebank creation for improved cross-lingual dependency parsing, *in* 'Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)', pp. 144–150.