Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2025”

April 23–25, 2025

Faroese Corpus Development: Strategies for Low-Resource Languages Corpora

Konstantin Satdarov
National Research University Higher School of Economics / Moscow, Russia
kesatdarov@edu.hse.ru

Darya Kharlamova
National Research University Higher School of Economics / Moscow, Russia
dasha.kh18@gmail.com

Andrey Yakuboy
National Research University Higher School of Economics / Moscow, Russia
aiyakuboy@edu.hse.ru

Alisa Lezina
National Research University Higher School of Economics / Moscow, Russia
aalezina@edu.hse.ru

Abstract

This paper presents the development of a comprehensive morphologically annotated corpus for Faroese, a low-resource Germanic language. We describe the creation of a large corpus of contemporary news texts automatically annotated using a custom-trained SpaCy model. The study demonstrates the effectiveness of creating linguistic resources for low-resource languages using minimal initial data. We trained a Transformer-based morphological parsing model on the small but high-quality OFT treebank using 5-fold cross-validation, achieving significant accuracy in morphological tagging and lemmatization. Manual evaluation confirms satisfactory performance of the automatic annotation, though certain challenges remain in distinguishing homonymous word forms across different parts of speech. This research provides a methodological framework for developing comprehensive linguistic resources for other low-resource languages with minimal initial data requirements.

Keywords: Faroese language, morphological annotation, corpus linguistics, low-resource languages, SpaCy, Tsakorpus, corpus development

DOI: 10.28995/2075-7182-2025-23-312-322

Создание фарерского корпуса: подходы для малоресурсных языков

Константин Сатдаров
Национальный исследовательский университет «Высшая школа экономики» / Москва, Россия
kesatdarov@edu.hse.ru

Дарья Харламова
Национальный исследовательский университет «Высшая школа экономики» / Москва, Россия
dasha.kh18@gmail.com

Андрей Якубой
Национальный исследовательский университет «Высшая школа экономики» / Москва, Россия
aiyakuboy@edu.hse.ru

Алиса Лезина
Национальный исследовательский университет «Высшая школа экономики» / Москва, Россия
aalezina@edu.hse.ru

Аннотация

В статье описывается разработка морфологически размеченного корпуса для фарерского языка – малоресурсного германского языка. Описывается создание большого корпуса современных новостных текстов с автоматической разметкой, выполненной с помощью специально обученной модели SpaCy. Мы демонстри-

руем эффективность создания лингвистических ресурсов для малоресурсных языков при минимальном количестве начальных данных. На основе небольшого, но качественно размеченного набора данных OFT была обучена морфологическая модель на базе архитектуры Transformer с использованием кросс-валидации, что позволило достичь высокой точности в морфологической разметке и лемматизации. Ручная оценка подтверждает удовлетворительное качество автоматической разметки, хотя остаются определенные трудности в различении омонимичных словоформ, относящихся к разным частям речи. Данное исследование предлагает методологическую основу для разработки лингвистических ресурсов для других малоресурсных языков при минимальных требованиях к количеству начальных данных.

Ключевые слова: фарерский язык, морфологическая разметка, корпусная лингвистика, малоресурсные языки, SpaCy, Tsakorpus, разработка корпусов

1 Introduction

It is our deep belief that everyone in the world of linguistic studies would benefit if it were easy to develop and deploy automated corpora for minor and endangered languages. This would give theory-grounded researchers (e.g. typologists) more easily accessible data and statistics-based methodology to back up their claims, and the practise-oriented developers would face less difficulties developing linguistic resources (like parsers, automated translation systems, etc.) for the wide support of minor languages. This would, in turn, lead to more chances of preserving endangered languages. However, even given the modern automatisation tools, it is still challenging to create a voluminous and reliable corpus without pouring countless money and efforts into manual annotation. This slows down the industry significantly.

Despite being the primary language of the Faroe Islands, Faroese remains among the low-resource languages in terms of available linguistic data and tools, save for some. This is the main reason why we chose this language as a test subject to develop a new corpus encompassing both historical and modern texts and an automated UD-compliant markup. The goal of this research is to create such a corpus working from the ground up, using as little resources as possible, as if Faroese was a minor language mainly studied through field research and sporadic Internet presence. If this proves to be successful, we will know the exact number of resources one needs to build a reliable base corpus with a universally acknowledged markup system for almost any language.

2 Related work

Currently, there are three major resources for the Faroese language available:

1. OFT [16] is based on texts from the Faroese Wikipedia. The corpus contains approximately 10,000 tokens across 1,208 sentences, with rich morphological annotation. Trosterud's finite-state morphological analyzer and constraint grammar [15] was used to markup the majority of the data, followed by manual verification of the morphological features and addition of dependency annotations according to Universal Dependencies 2.0 guidelines. The morphological tags were then automatically converted to UD-compatible features using a lookup table and set overlap procedure. While the treebank provides high-quality morphological information, its limited size and reliance on Wikipedia as the sole source results in certain limitations, such as an abundance of copular constructions and minimal representation of first and second person forms.
2. FarPaHC [2], a corpus of 40.000 tokens in size, was created through the automatic conversion of the Faroese Parsed Historical Corpus [8] from the Penn Treebank format. While the treebank provides broader coverage than OFT, the automatic conversion process has resulted in less detailed morphological annotations. Notably, the absence of lemma information in FarPaHC limits its utility for training lemmatization models, a crucial component for creating searchable corpora.
3. The news text corpus integrated into the GiellaLT linguistic infrastructure [17]. This corpus consists of approximately 25 million tokens with automatic morphological annotation performed using the system's finite-state grammar. The corpus is built on top of the Korp corpus engine [4] and includes a collection of news texts that have been automatically processed and annotated. However, several limitations affect its utility: the interface may present challenges

for researchers, the grammar-based disambiguation system is still in early development stages, and copyright restrictions prevent full corpus downloads, allowing only partial access to the dataset.

4. Regarding available morphological parsing tools, both Stanza [12] and UDPipe2 [14] offer neural processing pipelines for Faroese, trained on the FarPaHC treebank. However, these tools are limited by their training data: the absence of lemmatization capabilities and rather poor morphological tagset.
5. The GiellaLT system also provides a rule-based finite-state morphological and syntactic parser [15], which consists of a morphological analyzer based on a comprehensive lexicon and morphophonological rules, a constraint grammar disambiguator, and a dependency parser. The morphological component excels in handling regular inflectional patterns and compounds, while the constraint grammar manages morphological disambiguation, especially within noun phrases and prepositional constructions. However, this approach struggles with novel vocabulary, such as loanwords, and relies on a proprietary tagset that needs conversion for integration with multilingual pipelines. The system also shows limitations in handling irregular verbs and certain adjectival forms, including comparatives and superlatives. Moreover, its disambiguation component is still in early development, affecting analysis reliability.

It can be seen that there were no systematic attempts at training a Transformer-based morphological tagger, so our research will cover this aspect, employing the novel neural network along with the usage of SpaCy [7], a powerful tool in NLP that is capable of processing minor languages (to show that, there are official guides on how to train a custom pipeline applicable to low-resource languages). In the current research, we mainly used the OFT treebank and FarPaHC treebank — the former as the source of training data, and the latter as the reference for historical subcorpus markup.

3 Data preprocessing

We decided to use Faroese news websites as the source of our data since they were the most readily available source of contemporary Faroese texts online. The distribution of texts from different sources was as follows in the Table 1:

Website	n, texts
dagur.fo	24796
nordlysid.fo	1246
foroya-landsstyri.fo	4195
in.fo	315

Table 1: Sources and the number of texts gathered from them.

After preprocessing, which included the removal of non-Faroese texts (e.g. news articles with English or Danish translations), the combined total of texts from these sources amounted to 30732. For every entry we also have gathered the following metadata: date of publication, heading, link, author and tags.

Concerning the tags, we faced a problem: although a few sources were used for data retrieval, each of them contained specific tags that were close in meaning, but not interchangeable. Since we wanted to preserve search by tag, we manually unified the closest tags into broader categories.

For instance, Annar ítróttur (‘Other sports’), Fótbóltur (‘Footbal’l), Hondbóltur (‘Handball’), Ítróttur (‘Sports’) were all merged into Ítróttur. On a governmental site foroyalandsstyri.fo tags were representing different ministries (e.g. Umhvørvismálaráðið - Ministry of the Environment, Heilsumálaráðið - Ministry of Health, etc.), so we merged them into Landsstýristíðindi (Government News). However, even though we have merged tags into broader categories, we have preserved the source tags as

searchable secondary tags. As a result, the end corpus supports 8 primary tags and 23 secondary tags. The markup process is illustrated in Figure 1: the secondary tags highlighted with the same color were unified under one primary tag.

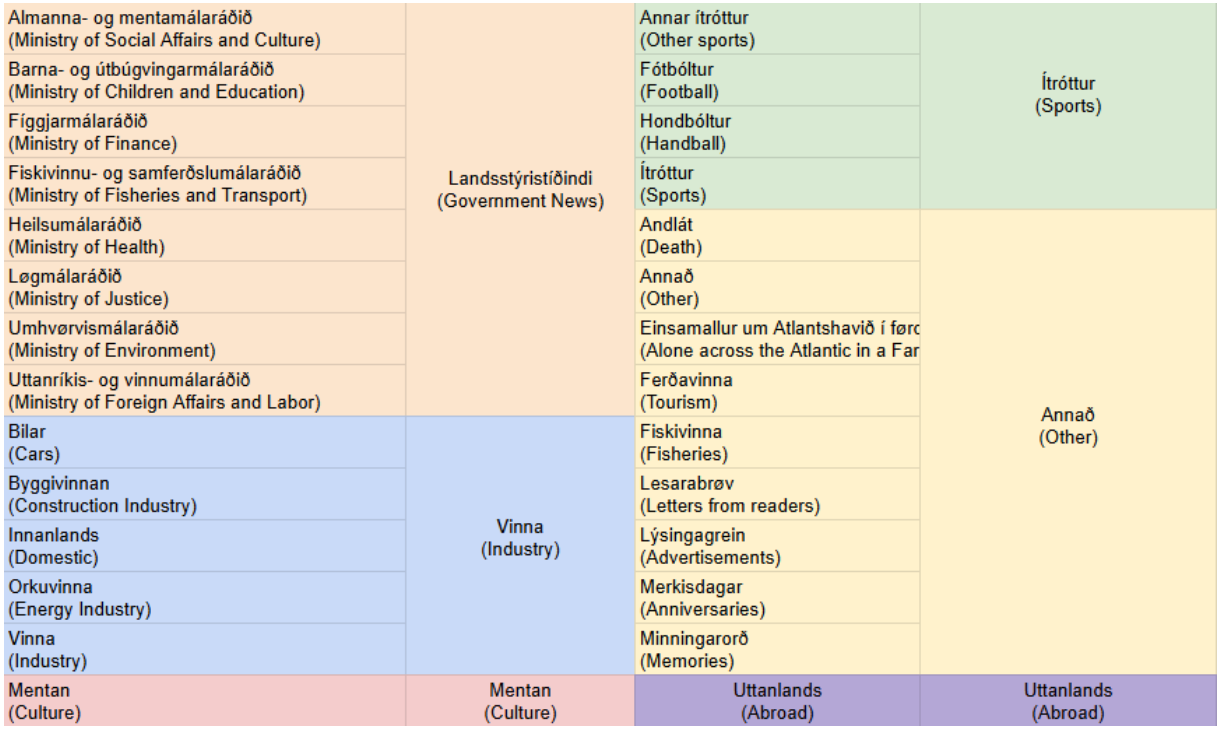

Figure 1: The manual process of tags unification.

4 Models

To expedite the speed of the manual annotation, we trained a Transformer-based morphological parsing model that is capable of performing UD markup of the Faroese texts using a guide provided by SpaCy and recommended for use when working with low-resource languages.

As we did not have enough resources to collect a sufficiently large and consistent corpus that could be used to train a model, we resorted to using OFT Faroese UD treebank. While the dataset was developed as a data collection aimed at testing rather than training, we used it due to the fact that it was a small, but curated and manually checked data collection. This made it the perfect choice for proving the possibility of training an automated SpaCy model on a small amount of data. In the dataset that we chose, there were only 1208 sentences, making it a very small collection.

Due to the size of the data, we trained the model using 5-fold cross-validation [10] to prove the validity of the metrics. Before we started training, we prepared the data as follows: we united every five sentences together and used them as separate data points so that the model could learn how to handle cross-sentence barriers. After that, we separated out the test dataset (73 texts, each containing 5 sentences from the original dataset).

Overall, we performed three experiments. For the first one, we used a default SpaCy-configuration file that was obtained through the official SpaCy config generation utility. We set up the training pipeline to contain the following in-built SpaCy components: "tok2vec", "tagger", "morphologizer", "trainable_lemmatizer", "parser". Here we did not use Transformer, because we wanted to understand the impact this computationally costly component has on the morphological tagging - a task that was known to be solved effectively without the use of neural networks by previous works (see Related Works above). We arrived at the following results for the dev dataset and the test dataset (see Table 2).

Fold number	lemma_acc	tag_acc	pos_acc	morph_acc	dep_uas	dep_las	sents_f
1	0.77982	0.9263	0.90896	0.82514	0.83162	0.74274	0.99099
2	0.81331	0.9442	0.91449	0.83158	0.81443	0.73625	0.98235
3	0.81328	0.92973	0.90015	0.84214	0.83289	0.75165	0.98246
4	0.81188	0.94621	0.91675	0.82591	0.82301	0.74226	0.96512
5	0.82651	0.94319	0.91961	0.83601	0.84533	0.77172	0.94737
Mean	0.80896	0.937926	0,911992	0.832156	0.829456	0.748924	0.973658
Test	0.8190	0.9396	0.9105	0.8296	0.8207	0.7294	0.9699

Table 2: The results for the heuristics-based SpaCy pipeline

Having obtained the baseline, we performed the second experiment once again using pre-generated SpaCy config and the following pipeline: "transformer", "trainable_lemmatizer", "tagger", "morphologizer", "parser". Bert-base-multilingual-uncased [6] was used as the baseline Transformer architecture. We once again performed the same 5-fold cross-validation and arrived at the following results (see Table 3).

Fold number	lemma_acc	tag_acc	pos_acc	morph_acc	dep_uas	dep_las	sents_f
1	0.83065	0.96468	0.94737	0.85914	0.8596	0.796	0.96491
2	0.82105	0.95102	0.93878	0.84558	0.85297	0.78229	0.94461
3	0.84253	0.96939	0.95044	0.8812	0.88889	0.82317	0.98817
4	0.82692	0.97584	0.95492	0.8864	0.89864	0.8356	1
5	0.84552	0.9639	0.95523	0.88592	0.88654	0.80779	0.97281
Mean	0.833334	0.964966	0.949348	0.871648	0.877328	0.80897	0.9741
Test	0.8398	0.9657	0.9508	0.8808	0.8918	0.8135	0.9591

Table 3: The results for the Transformer-based SpaCy pipeline

For the final experiment, we performed a bayes-based hyperparameter optimization sweep using the WandB library [3]. To maximize the morph_micro_f metric that reflects the quality of intra-tag F-score, we tried 6 different combinations of hyperparameters. See the different hyperparameter combinations that were explored, results and the reported target metric in Tables 4 and 5 along with Figures 2 and 3 depicting the target metric alignment with different parameter combinations and the hyperparameter importances and correlations with the target metric.

Figure 2 showcases the importance of every metric and its' correlation with the result. The metric's importance was received through WandB logging system. It shows the degree of usefulness of each hyperparameter in predicting the end result (namely, micro_f_score). Correlation with the result indicates the presence of a link between the value of a given parameter and the prediction. The correlation is linear and ranges from -1 to 1, with -1 reflecting negative correlation, 1 signifying positive correlation, and 0 indicating the absence of correlation.

Number of combination/ hyperparameters	components.-morphologizer.label_smoothing	components.tagger.label_smoothing	components.trainable_lemmatizer.-top_k	components.trainable_lemmatizer.-top_k.min_tree_freq	training.-dropout	training.optimizer.learn_rate
1	0.2	0.1	2	3	0.028	0.010
2	0	0.2	5	5	0.195	0.005
3	0	0.1	5	3	0.181	0.002
4	0.2	0.1	3	5	0.263	0.008
5	0.2	0	3	2	0.120	0.004
6	0	0.2	3	5	0.454	0.004

Table 4: The hyperparameter combinations that were explored

Combination	lemma_acc	tag_acc	pos_acc	morph_acc	dep_uas	sents_f
1	0.81004	0.90005	0.8775	0.81394	0.76503	0.96266
2	0.79365	0.91235	0.87955	0.81394	0.74348	0.95
3	0.80594	0.91645	0.8816	0.81292	0.78351	0.92683
4	0.78751	0.91235	0.85392	0.80574	0.75715	0.9465
5	0.82232	0.91645	0.8775	0.82727	0.78957	0.93061
6	0.79263	0.90313	0.88365	0.83137	0.80146	0.975

Table 5: The metrics for different hyperparameter combination

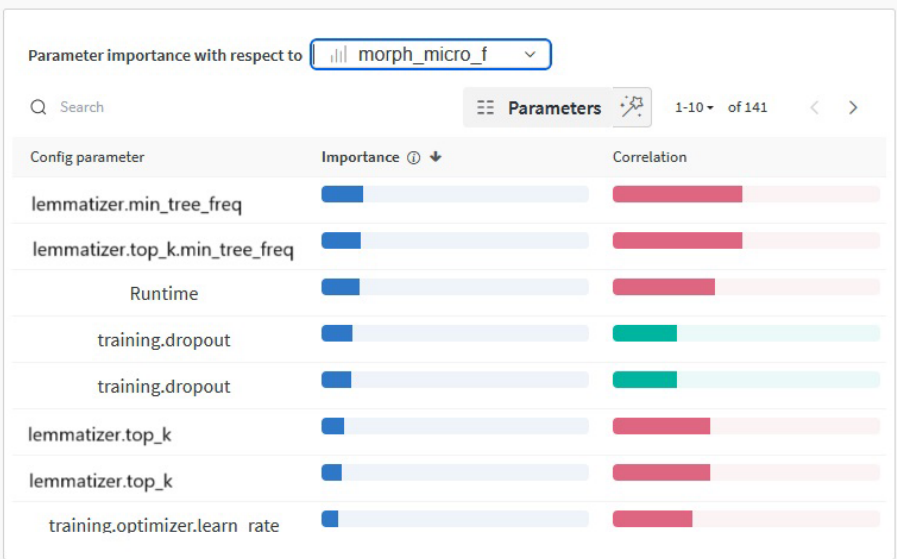

Figure 2: Hyperparameter importance for morph_micro_f.

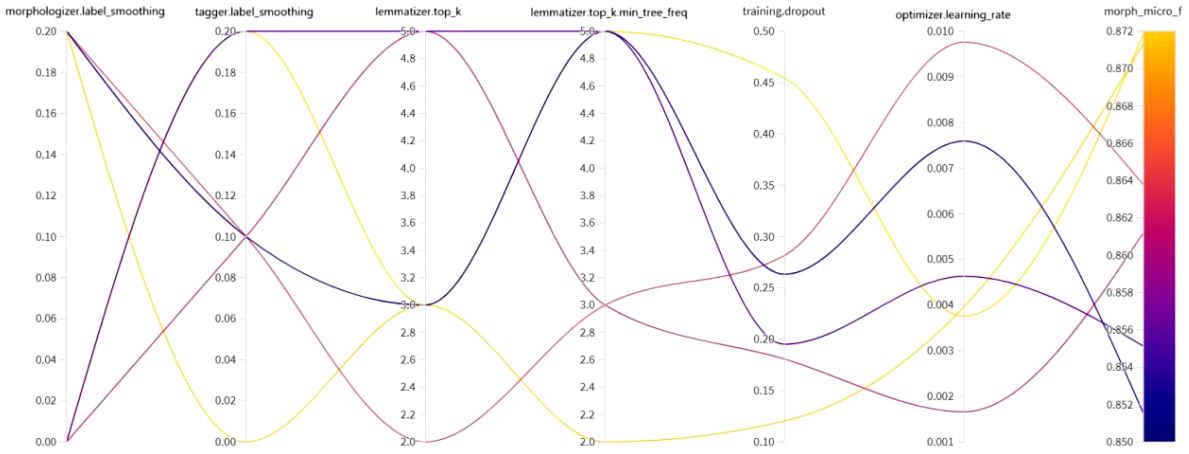

Figure 3: Hyperparameters alignment with different target metric scores.

5 Corpora

Several open-source corpus engines are available for creating searchable corpora from morphologically annotated texts, including NoSketchEngine [13; 9], ANNIS [11], Korp [4], and Tsakorpus [1]. However, none of these platforms natively support corpus creation from CONLL-U files. For our Faroese corpora, we selected Tsakorpus due to the platform's modular design and ElasticSearch configurations reducing resource needs while sustaining comparable performance to alternatives. Its data model, which stores annotated texts in JSON files, accommodates any morphological tagset and metadata schema, making it suitable for conversion from various annotation formats.

Our resource consists of two independent corpora: a large corpus of news texts (approximately 9 million tokens) automatically annotated using our SpaCy model, and a smaller corpus of biblical texts (about 3,000 tokens) with manual morphological annotation. Both corpora were initially annotated in CONLL-U format and converted to Tsakorpus-compatible JSON format using a custom-built universal converter. This converter can process both single CONLL-U files, automatically detecting and separating document boundaries, and directories containing multiple CONLL-U files, with one file per document. Document-level metadata can be extracted either directly from the dataset or from external files. The converter processes morphological annotation by extracting word forms, lemmas, Universal POS tags, and grammatical features, while syntactic markup is omitted. The converter's language-independent design makes it suitable for potential integration into the Tsakorpus platform to facilitate the creation of searchable corpora from other Universal Dependencies treebanks.

We optimized the corpus interface for researchers' usability. One of the most convenient features is the grammar selection table. It allows users unfamiliar with UD tags to select them from categorized lists organized by linguistic categories such as parts of speech, case, number, tense and more. The search functionality includes queries by word form, lemma, grammatical features, multi-word and negative queries, and full-text search. Users can create subcorpora based on metadata: the news corpus supports filtering by source, publication year range, and primary or secondary tags, while the biblical corpus enables chapter-based selection. The platform also provides statistical analysis tools for examining subcorpora.

Both corpora are freely available at www.farcorpus.fikl.ru. The interface is currently available only in English, but the platform's localization mechanisms allow for addition of languages natively supported by Tsakorpus, such as Russian, French, German, and Hebrew. Adding other interface languages would require more effort but remains feasible through the platform's built-in localization framework.

6 Manual quality evaluation

As stated above, the corpus currently includes two types of annotation: manual and automatic. Apparently, the former is seen as an example on which we can rely in order to evaluate the quality of the automatic one. The reason for this is that detailed manual analysis allows to create almost flawless annotation meeting the demands.

Currently, the most concerning weakness is that it is still prone to confuse parts of speech such as nouns, verbs, adjectives and pronouns. However, given high metrics pertaining to the morphological tagging (pos_acc, morph_acc in Table 3) this is not strongly observed on the training data. Thus, we will be able to understand the scale of this issue no sooner than we obtain a bigger testing dataset, the works for developing which are already underway.

This can be explained by the fact that many word-forms of different parts of speech in Faroese have the same endings, which renders the analysis more elaborate. For instance, the -ur ending is both a nominal ending (Nom. Sg., for instance) and a verbal one (3Sg). Thus, though we say that lemmatization and morphological description are fairly good, there are still data that need clarifying and checking. Indeclinable lexemes and frequently attested word-forms have proven to be annotated accurately.

Despite the fact that automatic annotation still needs improvement, we can still claim that the overall performance of the model seems satisfactory. For some sentences, the model can provide accurate morphological descriptions and lemmas of word-forms. The following figure exemplifies an accurately lemmatized sentence, and we agree with the morphological description it provides.

```
1   Og      og      CCONJ   CC      _       3       cc      _       _
2   í       í       ADP     Pr      _       3       case    _       _
3   stundini        stund   NOUN    N       Case=Dat|Definite=Def|Gender=Fem|Number=Sing    4       obl     _       _
4   fór     fara    VERB    V       Mood=Ind|Number=Sing|Person=3|Tense=Past        0       ROOT    _       _
5   spitalskusjúkan spitalskusjúka  NOUN    N       Case=Nom|Definite=Def|Gender=Fem|Number=Sing    4       nsubj   _       _
6   av      av      ADP     Pr      _       7       case    _       _
7   honum   honum   PRON    Pron    Case=Dat|Gender=Masc|Number=Sing|Person=3|PronType=Prs  4       obl     _       SpaceAfter=No
8   ,       ,       PUNCT   CLB     _       11      punct   _       _
9   og      og      CCONJ   CC      _       11      cc      _       _
10  hann    hann    PRON    Pron    Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs  11      nsubj   _       _
11  varð    verða   VERB    V       Mood=Ind|Number=Sing|Person=3|Tense=Past        4       conj    _       _
12  reinur  reinur  ADJ     A       Case=Nom|Definite=Ind|Gender=Masc|Number=Sing   11      xcomp   _       SpaceAfter=No
13  .       .       PUNCT   CLB     _       11      punct   _       SpaceAfter=No
```

Figure 4: An accurately lemmatized sentence.

The next figure illustrates a sentence which should be further corrected in terms of lemmatization and morphological description.

```
1   Og  og  CCONJ   CC  _   3   cc  _   _
2   tá  tá  ADV Adv _   3   advmod  _   _
3   ið  ið  SCONJ   CS  _   0   ROOT    _   _
4   Jesus   Jesus   PROPN   N   Case=Nom|Gender=Masc|Number=Sing    6   nsubj   _   _
5   sá  sá  ADV V   _   6   advmod  _   _
6   trúgv   trúgvur ADJ A   Case=Nom|Definite=Ind|Gender=Fem|Number=Sing    3   dep _   _
7   teirra  teirra  PRON    Pron    Case=Acc|Number=Sing|Person=3|PronType=Prs  3   obj _   SpaceAfter=No
8   ,   ,   PUNCT   CLB _   22  punct   _
9   sigur   siga    VERB    V   Mood=Ind|Number=Sing|Person=3|Tense=Pres    22  nsubj   _   _
10  hann    hann    PRON    Pron    Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs  9   nsubj   _   _
11  við við ADP Pr  _   13  case
12  hin hin DET Det Case=Acc|Gender=Fem|Number=Sing 13  det _   _
13  giktsjúka   giktsjúki   NOUN    N   Case=Acc|Definite=Ind|Gender=Fem|Number=Sing    9   obl _   SpaceAfter=No
14  :   :   PUNCT   CLB _   19  punct   _   _
15  «   «   PUNCT   PUNCT   _   16  punct   _   SpaceAfter=No
16  Sonur   Sonur   NOUN    N   Case=Nom|Definite=Ind|Gender=Masc|Number=Sing   19  dep _   _
17  mín mín PRON    Pron    Case=Gen|Number=Sing|Person=1|PronType=Prs  16  punct   _   SpaceAfter=No
18  ,   ,   PUNCT   CLB _   19  punct   _   _
19  syndir  synda   VERB    V   Mood=Ind|Number=Sing|Person=3|Tense=Pres    20  amod    _   _
20  tínar   tínar   PRON    N   Case=Nom|Number=Plur|Person=3|PronType=Prs  9   nsubj   _   _
21  eru vera    AUX V   Mood=Ind|Number=Plur|Tense=Pres 22  cop _   _
22  fyrigivnar  fyrigivnur  ADJ A   Case=Nom|Definite=Ind|Gender=Masc|Number=Plur   3   dep _   SpaceAfter=No
23  !   !   PUNCT   CLB _   22  punct   _   SpaceAfter=No
24  »   »   PUNCT   PUNCT   _   0   ROOT    _   SpaceAfter=No
```

(VERB)
(NOUN)
GEN
MASC
(ADJ)
(DET)
(NOUN)
(DET — tín)
(VERB; VerbForm=Part)

Figure 5: A sentence with suggested corrections.

7 Discussion of the results

Over the course of this research, we managed to achieve remarkable results in every sphere of corpus creation. Let us discuss our findings and the outcomes of the research.

First of all, we have shown that even though Faroese tends to be viewed as a low-resource language, obtaining a significant number of texts from the open access webpages is more than feasible. This method of corpus creation can be viewed as viable, providing an insight into contemporary speech used in different social situations. This method could be used to create more robust corpora for minor languages, because the texts there are not solicited by the researchers.

Second of all, we have created a Transformer-based model capable of reliably processing faroese texts and providing full UD-compliant annotations while using a small corpus of high-quality data. We have performed three experiments. The first experiment showed that the effective high-quality morphological markup was achievable, even given the limited, but high-quality dataset. It also helped point out the areas where the model was the weakest: namely, dep_las metrics. The second experiment demonstrated the importance of using Transformer-based architecture in training language processing pipelines. As shown by the data in Tables 1-4, the introduction of Transformer allowed to significantly improve the scores for the majority of the metrics. As we employed cross-validation in both experiments, we have demonstrated the reproducibility of our results. As for the last experiment, we have shown that the explored hyperparameters do not impact the results significantly. However, the tuning still led to some improvements in the test metrics.

Third of all, we have managed to create a usable corpus that was adequately processed, using only a small and easily achievable number of resources. We have also chosen a corpus engine easy to deploy and maintain that has a rich search and analysis capabilities and rich functionality. Overall, this research shows that it is indeed possible to create a corpus and a model with a very small number of resources. To us, this opens a lot of possibilities for minor language researchers, because, as shown by the authors of [16] and confirmed by the present research, an adequate language dataset can be collected by using an automated model and correcting the errors there. As for the base number of sentences required to train such a model, we have demonstrated that the default SpaCy pipeline can be trained on approximately 1200 sentences - a moderate dataset that can be collected over the course of multiple field researches. Thus, it stands to reason that using such a pipeline a lot of minor languages can be processed and the data on them made publicly available for the advance of the linguistic studies.

However, there are some limitations that need to be addressed in the subsequent research. First, due to the differences in markup systems, we could not directly compare the quality of our model against the existing solutions, like [15]. This issue can be addressed by converting the tags used in each system and using the conversions for meaningful comparative analysis. Second, as we used the treebank initially aimed at testing, we would like to change this as soon as possible. At the moment, we are working to create a new small Faroese training corpus of a similar size that will be used to retrain the model from scratch. This will allow to use the OFT treeback for its intended purpose of benchmarking.

Lastly, while SpaCy provided an effective framework for our research due to its ease of setup, moderate computational requirements, and streamlined deployment capabilities, several alternative approaches merit consideration for future work. The selection of SpaCy allowed us to focus on data collection and annotation quality rather than complex infrastructure development, which was crucial given our limited initial resources. However, more sophisticated frameworks such as Stanza [12] and UDPipe2 [14] have demonstrated superior performance in morphological and syntactic analysis across numerous languages and could potentially yield improved results for Faroese with appropriate training data. Additionally, recent advances in transformer architectures such as XLM-RoBERTa [5] offer promising alternatives that could enhance tagging accuracy through their cross-lingual transfer capabilities.

We did not pursue these approaches at this stage due to the substantial computational resources required for fine-tuning such models and because current multilingual models lack explicit Faroese language data.

An additional limitation that we would like to address is the possibility of using data from the closely related languages. The Faroese language belongs to the North Germanic language branch. However, unlike the Mainland Scandinavian languages, i.e. Danish, Swedish, and Norwegian, Faroese and Icelandic are characterised by arguably complex morphology; therefore, the range of the donor languages is limited to Icelandic. Nonetheless, despite being partially intelligible in writing, they also exhibit morphological differences. For instance, the Faroese verbs do not conjugate for person in plural, whereas the Icelandic verbs do. Another limitation concerns morphonological alternations, which appear to be different in both languages.

We should also briefly mention that we have endeavoured to work with the language data as if we did not have other opportunities and sources. Our work showcases what can be done supposing any limitations arise during the creation of a corpus for a low-resource language.

8 Conclusion

Over the course of this work, we managed to create a Transformer-based model for automated UD-compliant Faroese texts markup and developed a corpus of modern and historical Faroese texts using as little pre-existing resources as possible. This research stands to confirm that given modern linguistic technologies and language learning models and pipelines capabilities, it is possible to create a full-fledged corpus and a model without manually annotating copious amounts of language data.

It is possible to bring the research even further and explore the possibilities of data augmentation strategies through the integration of rule-based morphological parsers or machine-readable dictionaries. Such approaches could potentially provide a cost-effective method for expanding the training data volume and, consequently, enhancing model performance. Additionally, we could explore the options of implementing syntactic annotation into the existing corpus. While the development of such corpora presents additional complexities, such resources could substantially advance Faroese language re-search.

References

[1] Arkhangelskiy Timofey A. Electronic corpora of the albanian, kalmyk, lezgian, and ossetic languages // Automatic Documentation and Mathematical Linguistics. — 2012. — Vol. 46. — P. 118–123.

[2] Arnardóttir Þórunn, Hafsteinsson Hinrik, Sigurðsson Einar Freyr, Bjarnadóttir Kristín, Ingason Anton Karl, Jónsdóttir Hildur, Steingrímsson Steinþór. A Universal Dependencies Conversion Pipeline for a Penn-format Constituency Treebank // Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020). — Barcelona, Spain, 2020. — P. 16–25.

[3] Biewald Lukas. Experiment Tracking with Weights and Biases. — 2020. — Access mode: https://www.wandb.com/

[4] Borin Lars, Forsberg Markus, Roxendal Johan. Korp – the corpus infrastructure of Språkbanken // Proceedings of LREC 2012. — Istanbul: ELRA, 2012. — P. 474–478.

[5] Conneau A. et al. Unsupervised cross-lingual representation learning at scale //arXiv preprint arXiv:1911.02116. – 2019.

[6] Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // CoRR. — 2018. — Vol. abs/1810.04805. — Access mode: http://arxiv.org/abs/1810.04805

[7] Honnibal Matthew, Montani Ines, Van Landeghem Sofie, Boyd Adriane. spaCy: Industrial-strength natural language processing in python. — Zenodo, Honolulu, HI, USA, 2020.

[8] Ingason Anton Karl, Rögnvaldsson Eiríkur, Sigurðsson Einar Freyr, Wallenberg Joel C. Faroese Parsed Historical Corpus (FarPaHC) 0.1. — CLARIN-IS, 2012.

[9] Kilgarriff Adam, Baisa Vít, Bušta Jan, Jakubíček Miloš, Kovář Vojtěch, Michelfeit Jan, Rychlý Pavel, Suchomel Vít. The Sketch Engine: ten years on // Lexicography. — 2014. — Vol. 1, No. 1. — P. 7–36.

[10] Kohavi Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection // Proceedings of the 14th International Joint Conference on Artificial Intelligence. — Montreal, Canada, 1995. — Vol. 14, No. 2. — P. 1137–1145.

[11] Krause Thomas, Zeldes Amir. ANNIS3: A new architecture for generic corpus query and visualization // Digital Scholarship in the Humanities. — 2016. — Vol. 31, No. 1. — P. 118–139.

[12] Qi Peng, Zhang Yuhao, Zhang Yuhui, Bolton Jason, Manning Christopher D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. — Online, 2020. — P. 101–108.

[13] Rychlý Pavel. Manatee/Bonito-A Modular Corpus Manager // RASLAN. — 2007. — Vol. 2007. — P. 65–70.

[14] Straka Milan. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task // Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. — Brussels, Belgium, 2018. — P. 197–207.

[15] Trosterud Trond. A constraint grammar for Faroese // Constraint Grammar and robust parsing: Proceedings of the NODALIDA 2009 workshop. — 2009. — Vol. 8. — P. 1–7.

[16] Tyers Francis M., Sheyanova Mariya, Martynova Alexandra, Stepachev Pavel, Vinogradovsky Konstantin. Multi-source synthetic treebank creation for improved cross-lingual dependency parsing // Proceedings of the Second Workshop on Universal Dependencies (UDW 2018). — 2018. — P. 144–150.

[17] Wiechetek Linda, Hiovain-Asikainen Katri, Mikkelsen Inga Lill Sigga, Moshagen Sjur, Pirinen Flammie, Trosterud Trond, Gaup Børre. Unmasking the Myth of Effortless Big Data - Making an Open Source Multilingual Infrastructure and Building Language Resources from Scratch // Proceedings of the Thirteenth Language Resources and Evaluation Conference. — Marseille, France, 2022. — P. 1167–1177.