OldBERTur: Named Entity Recognition for Medieval Icelandic

Pontus Henningsson, Eva Pettersson, Erik Lenas
Swedish National Archives, Uppsala University, Swedish National Archives
pontus.henningsson@riksarkivet.se, eva.pettersson@lingfil.uu.se, erik.lenas@riksarkivet.se

Abstract

We present OldBERTur, a Named Entity Recognition (NER) model for Old Icelandic available in two variations, one for normalised texts, and one for diplomatic texts. Using a BERT-based model architecture, we fine-tune an existing BERT language model, and due to training data scarcity, we employ multiple training configurations, including pre-training domain adaptation, sentence-level data resampling, and modern Icelandic data augmentation; achieving a 93 F1 score for normalised texts, and 79 for diplomatic texts. We find that additional training configurations, such as resampling entity-annotated Old Icelandic texts, significantly improve performance in low-resource settings, while the effectiveness of added training configurations diminishes as the available training data increases. Our models can be used to automatically identify and classify person and location names in texts sourced from the rich Icelandic medieval literary tradition. Our models, along with their data and code, are made publicly available to allow for reuse and future research into medieval Scandinavian NLP and beyond.

Keywords: Named Entity Recognition, NER, Old Icelandic, Natural Language Processing, NLP, Digital Humanities

1. Introduction

Old Icelandic, the written language of Iceland during the Middle Ages,[1] contains some of the most important literary works written in a vernacular language in Scandinavia (Clunies Ross, 2000). The medieval Icelandic literary texts, including the Poetic and Prose Eddas and numerous sagas, written mainly between the 12th and 15th century, have been studied extensively for many decades; playing a key role in the understanding of the medieval Icelandic and wider Scandinavian world along with its languages and poetry, as well as providing a tantalising glimpse into pre-Christian Scandinavia and its mythology (Acker and Larrington, 2002; Clover and Lindow, 2019; Tómasson, 2006).

Considerable efforts have been made to digitise many of these Old Icelandic texts, with, for instance, the Medieval Nordic Text Archive (Menota, 2025) offering free access to medieval Nordic manuscripts. Despite increased digitisation, Old Icelandic still remains relatively unexplored by modern Natural Language Processing (NLP) techniques. While Old Icelandic and modern Icelandic are relatively similar due to the conservative nature of Icelandic and its focus on language preservation (Hilmarsson-Dunn and Kristinsson, 2010; Árnason and Leonard, 2011), there are still differences in older texts that may be hard for NLP tools trained on modern language to analyse. This is particularly true for Named Entity Recognition (NER), the task of automatically extracting named entities, such as persons and places, since naming conventions tend to shift over time. The lack of dedicated NER tools for Old Icelandic thus limits the possibilities for researchers to extract and analyse people, places, and their relationships, at scale from the literary Old Icelandic tradition. Furthermore, seeing as Icelandic, and indeed Old Icelandic, is a language with rich and complex morphology (Loftsson and Rögnvaldsson, 2007), off-the-shelf methods from related languages, such as NLP tools created for other Scandinavian languages, cannot necessarily be directly applied to the Old Icelandic material.

In this paper, we present and discuss OldBERTur, an openly available BERT-based (Devlin et al., 2019) NER model for Old Icelandic available in two variations, one for normalised transcriptions of Old Icelandic and one for diplomatic transcriptions. We provide the code, data, and models as freely available resources.[2]

2. Related Work

Cultural heritage institutions, such as archives, libraries, and museums, have increasingly made their historical material available online through digitisation. These institutions now provide instant digital access to a multitude of historical content, for

[1] Often also referred to as Old Norse, we use the term Old Icelandic to signify the main language used for Icelandic literary production in medieval Iceland. For a deeper contextual discussion of Old Norse-Icelandic, the reader is directed to Clover and Lindow (2019); Clunies Ross (2000).

[2] Normalised NER Model: huggingface.co/Riksarkivet/oldbertur-normalised-old-icelandic-ner, diplomatic NER model: huggingface.co/Riksarkivet/oldbertur-diplomatic-old-icelandic-ner, code and data: github.com/phenningsson/Medieval-Icelandic-NER.

Proceedings of LT4HALA 2026 @ LREC 2026, pages 469–481.

instance scanned images of historical documents and digital editions featuring transcriptions (Birnbaum et al., 2017; Kasperowski et al., 2024). While this digitisation serves a way to make the material more accessible and findable, as well as a way to preserve and reduce the need for its physical use; digitisation has also created an abundance of machine-readable historical material (Ehrmann et al., 2023). This "Big Data of the Past" as coined by Kaplan and di Lenardo (2017), combined with modern computational methods hold great potential for enabling increased digitisation, accessibility, and analysis of the past at scale (Ehrmann et al., 2023; Nockels et al., 2024).

One area where historical texts and computational methods increasingly converge is through the development of NER models for historical texts. The HIPE shared tasks (Ehrmann et al., 2020, 2022) established benchmarks for NER models for historical newspapers in English, French, German, Swedish, and Finnish; with the results demonstrating that neural NER models, in particular those relying on domain-adapted BERT-models, show solid performance for historical NER. Notably, the HIPE tasks exhibit how fine-tuning and adapting a NER-model for both language and domain can improve its performance, especially when supplied with sufficient training data (Ehrmann et al., 2020, 2022).

2.1. NLP for Icelandic

Icelandic is a North Germanic language belonging to the West Scandinavian branch together with Norwegian and Faroese, while Danish and Swedish make up the East Scandinavian languages (Hovdhaugen et al., 2000). While relatively less-resourced compared to major European languages, Icelandic has some key NLP resources. The IceNLP toolkit (Loftsson, 2019) serves as a comprehensive rule-based toolbox for analysis of contemporary Icelandic texts, including a tokeniser, a part-of-speech tagger, NER, and more. IceBERT (Snæbjarnarson et al., 2022), a RoBERTa-based (Liu et al., 2019) model trained on 16GB of modern Icelandic text, achieves impressive performance on many downstream NLP tasks, including NER, with an F1 score of 91.43 on the MIM-GOLD-NER dataset (Ingólfsdóttir et al., 2020), and subsequently improved to 92.73 (Guðjónsson et al., 2021). However, computational tools for older versions of the Icelandic language remain scarce.

2.2. Medieval NER

Multiple studies have showcased that neural NER models can perform well on medieval texts. Torres Aguilar (2022) created an annotated corpus and NER models for multilingual medieval charters written in Spanish, French, and Latin from the 10th to 15th century, consisting of ∼2,3 million tokens and ∼177k annotated entities. Leveraging pre-trained models and neural networks through both Bi-LSTM-CRF and BERT-based approaches, the study achieves F1 scores over 94 for all developed models. Torres Aguilar and Stutzmann (2021) created an annotated corpus of around ∼500k tokens stemming from ∼1.2k charters written in medieval French from the 13th–14th century, using similar architectures as Torres Aguilar (2022) for the NER models. The results show competitive performance, with F1 scores averaging well above 90. Similarly, Díez Platas et al. (2021) developed a NER model for 12th–15th century Spanish on a corpus consisting of medieval literature and administrative texts containing more than 3k entities, achieving F1 scores in the 74–87 range, using a rule-based approach relying on linguistic features and gazetteers.

However, there are some challenges inherent to applying computational methods to medieval texts. Many medieval texts, especially texts written in vernacular languages, showcase frequent variations in orthography, regular use of context-dependent nicknames or titles, dialectal nuances, and use of abbreviations which often require domain knowledge to resolve (Aguilar, 2025; Clérice et al., 2024; Schoen and Saretto, 2022). Despite difficulties and diachronic changes, previous studies show that it is possible to create competent NER models for medieval texts when given enough data and computational support, although previous work has mainly focused on medieval Romance languages.

For Old Scandinavian languages in particular, Besnier and Mattingly (2021) created a small entity-annotated dataset consisting of the Old Icelandic poem Völuspá; something which can be significantly expanded upon, leaving a research gap that this study sets out to address.

3. Data

Despite the mass digitisation of historical material, not all languages are served the same. Similar to contemporary low-resource languages, many historical languages are faced with a lack of texts and data available for computational methods and analysis (Tekgürler, 2025). Historical corpora annotated with named entities are generally concentrated to the 19th–20th century; large corpora from other time periods are scarce (Ehrmann et al., 2023). Indeed, while there are digital editions of Old Icelandic texts available online, annotated corpora for NER is limited. We therefore use data augmentation to complement our Old Icelandic training data, further discussed in the sections below. We use

the traditional Beginning-Inside-Outside (BIO) tagging, and follow the CoNLL 2002 format (Tjong Kim Sang, 2002) for the annotated entities.

3.1. Medieval Nordic Text Archive

The Medieval Nordic Text Archive (Menota, 2025) provides free access to digital editions of medieval Nordic texts, often with multiple levels of transcription: *facsimile*, which copies the manuscript as faithfully as possible; *diplomatic*, which resolves abbreviations and reduces allographic variation while retaining many features of the original text; *normalised*, which resolves abbreviations and standardises orthography, spelling, and punctuation to a modern editorial standard (Haugen, 2004). We use only diplomatic and normalised transcriptions as training data; facsimile transcriptions are too difficult for our limited training data.

Texts available through Menota are encoded in a customised TEI (TEI Consortium, 2025) XML format, called Menotic XML (Haugen, 2004). There are 57 available Old Icelandic manuscripts (∼860k words total). Out of these 57 texts, nine are entity-annotated on a diplomatic level, and out of these nine texts annotated on a diplomatic level, seven are also annotated on a normalised level. Thus, we have nine texts for our diplomatic Menota corpus, and seven texts for our normalised Menota corpus, with the annotation carried out by experts and available through Menota (Haugen, 2019). All the entity-annotated texts contain information on person and place names as entities, and we only take person and place name entities into account. The specific texts used are listed in the Appendix.

Entities were extracted using Python, and invalid Unicode characters in the diplomatic transcriptions were substituted to their closest representation in the Medieval Unicode Font Initiative (MUFI) (Haugen, 2015), or, when not possible, we used their normalised Menota equivalent. Words split up due to line breaks were merged to ensure complete words and entities. All in all, we ended up with two Menota datasets, one consisting of diplomatic entities, and one consisting of normalised entities.

To illustrate the transcription levels used in this study, the below example shows a parallel text using different transcription levels from *Alexander saga* from Menota, where person names are marked in *italic* and place names in <u>underline</u>.

Diplomatic: eınn aꝼ hans ſcaꞇꞇkonungom er neꝼnꝺꝛ *pʜllıppuſ*. hann reð ꝼyr <u>grıklanꝺı</u>. ꝺroꞇꞇnıng kona hanſ heꞇ *olımpıaſ*. Son aꞇꞇo þau þann er *alexanꝺer* heꞇ. Sa maðꝛ ꝩar með hırð konungſenſ er *nepꞇanabuſ* heꝼer heıꞇıð.

Normalised: Einn af hans skattkonungum er nefndr *Philippus*. Hann réð fyr <u>Grikklandi</u>. Dróttning kona hans hét *Olympias*. Son áttu þau þann, er *Alexander* hét. Sá maðr var með hirð konungsins, er *Neptanabus* hefir heitit.

The diplomatic transcription retains letter forms from the original manuscript, such as ꞇ(t), ꝼ(f), ꝺ(d), ꝛ(r), and long-s(ſ), while the normalised transcription substitutes these to their modern equivalents. Furthermore, the normalised transcription capitalises entities, which the diplomatic crucially does not, meaning that the diplomatic texts not only exhibit a greater variation of characters and orthography, but also showcase a lack of an often important textual clue for identifying a named entity for both humans and machines alike. For instance, compare the diplomatic <u>grıklanꝺı</u> with the normalised <u>Grikklandi</u>, or *nepꞇanabuſ* with *Neptanabus*.

3.2. Icelandic Parsed Historical Corpus

The Icelandic Parsed Historical Corpus (IcePaHC) (Wallenberg et al., 2024), presented and discussed in detail by Rögnvaldsson et al. (2012), is an open and freely available corpus of parsed historical texts written in Icelandic, ranging from the 12th century to modern day, in normalised form. While not explicitly entity-annotated, entities can be extracted from parsing annotations, where proper nouns are tagged with *NPR* (Noun Proper) together with its case suffix, and each token is represented with both surface form and lemma (e.g., "NPR-A Þorlák-þorlákur"). Named entities were extracted through a script that takes all *NPR* tags and its context, for instance extracting adjacent proper noun sequences such as the person "Hallur Gissurarson" from the sequence "(NPR-N Hallur-hallur) (NPR-N Gissurarson-gissurarson)".

This approach, while effective for a high-recall extraction of entities, does require manual verification, both to correct missed or mistaken entity annotations, but also to distinguish between names of persons and locations in ambiguous cases. To this end, six texts from IcePaHC, written in the 13th century and not present in the Menota corpora, were automatically extracted using the script, and then manually verified by the lead author of this paper. This process resulted in an extension of the normalised Old Icelandic NER data, comparable to normalised versions of the Menota texts.

3.3. MIM-GOLD-NER

To further augment our Old Icelandic data, we use the MIM-GOLD-NER dataset (Ingólfsdóttir et al., 2020), a contemporary Icelandic corpus of over 1 million tokens with over 48k annotated entities. This dataset has been used in previous studies to

benchmark Icelandic NER performance (Guðjónsson et al., 2021; Snæbjarnarson et al., 2022), and significantly extends our training data. To meet our needs, we filter the corpus to only contain our target entities, i.e., persons and locations, whereas other entity types are converted to non-entities.

3.4. Corpus Statistics

As seen in Table 1, our resulting Old Icelandic corpora (Menota normalised, IcePaHC, and Menota diplomatic) are relatively small with limited entity counts. The entity density (the percentage of tokens that represent entities) ranges between 3.3% and 5.5%, meaning that around 95% of the tokens are non-entities. There is also a significant class imbalance, where around 85% of the entities in the Old Icelandic corpora are person entities. It is also worth noting, that the diplomatic Menota corpus is almost twice the size of the normalised Menota corpus, and seeing as there is a general lack of entity-annotated transcriptions of Old Icelandic, this diplomatic corpus forms the basis of our diplomatic training data. For normalised data on the other hand, IcePaHC provides additional data comparable to Menota's normalised texts, and these two corpora together form the basis of our normalised training data. MIM-GOLD-NER adds a substantial amount of entities along with a more balanced entity distribution (63% persons, 37% locations), but is only used for training since our target domain is Old Icelandic. The Old Icelandic texts used in this study span from ∼1210 to 1700, with the vast majority of texts dating to the 13th and 14th century. All Old Icelandic texts used are written between ∼1210–1399 except for one, Íslendingabók AM 113 b fol., which is a later 17th century copy of a medieval manuscript which has since been lost. The full name and details about each text and the corpus it belongs to is available in the Appendix.

Corpus	Tokens	Per.	Loc.	Dens.
Normalised				
Menota	73,258	2,197	368	3.6%
IcePaHC	91,383	4,119	702	5.5%
Diplomatic				
Menota	138,404	5,803	913	5.1%
Modern (augmentation)				
MIM-GOLD	>1M	15,587	9,002	3.3%

Table 1: Corpus statistics. Per. = Person, Loc. = Location, Dens. = entity density, the percentage of tokens representing entities.

4. Methods

We use the aforementioned *IceBERT* by Snæbjarnarson et al. (2022), a RoBERTa-based model (Liu et al., 2019) trained on modern Icelandic, as our base language model for the NER task. This base model is already pre-trained on contemporary Icelandic texts and has achieved competitive performance for Icelandic NER, serving as a suitable base for our Old Icelandic NER models. However, we hypothesise that IceBERT may struggle with Old Icelandic diplomatic texts due to their orthographic variation, and that domain adaptive pre-training may improve performance since domain adaptation has shown improved performance in previous studies (Gururangan et al., 2020; Schweter and Baiter, 2019).

4.1. Domain Adaptation

We fine-tune IceBERT using Masked Language Modelling (MLM) on an unlabelled corpus of diplomatic Old Icelandic texts derived from 15 works in Menota consisting of ∼34k sentences and ∼385k words. This corpus consists of our nine entity-annotated diplomatic texts without their annotations, i.e. just the texts themselves, as well as six additional unannotated diplomatic texts from the same period from Menota. The 15 diplomatic texts used for MLM domain-adaptation are listed in the Appendix. Following Gururangan et al. (2020), we use task-adaptive pre-training (TAPT) to facilitate model familiarity with Old Icelandic diplomatic texts prior to being fine-tuned for NER. Around 12% of characters in the diplomatic texts are characters that appear in neither our modern nor normalised Old Icelandic NER corpus, featuring characters such as: ꝛ, ꝑ, ꝺ, ꝓ.

We train the model for 8 epochs (Liu et al., 2019), batch size of 32, 256 maximum sequence length, 3e-5 learning rate (Devlin et al., 2019), and 6% warm up (Gururangan et al., 2020). The MLM masking probability is set to the standard 15% (Devlin et al., 2019; Gururangan et al., 2020; Liu et al., 2019). Training was performed on a Quadro RTX 5000 GPU using FP16 mixed precision. This resulted in a fine-tuned IceBERT model created for diplomatic Old Icelandic texts. For diplomatic NER, we thus use two variations: IceBERT off-the-shelf; and our domain adapted version of IceBERT. For normalised texts, we use the original IceBERT as our base model seeing as the normalised transcriptions follow more modern editorial standards.

4.2. Sentence-Level Resampling

Following Wang and Wang (2022), we apply sentence-level resampling to address the data

scarcity and class imbalance in our Old Icelandic corpora. We leverage *sentence-level Count weighted by Rareness* (sCR), which calculates an expansion factor for each sentence based on the number of entity tokens and the rareness weight for each entity type, where under-represented entity types are given a higher weight. We perform sCR on the normalised Old Icelandic data (Menota + IcePaHC), but only on the diplomatic texts derived from Menota. We strictly use sCR on our training data, not on our development or test set. Table 2 shows the sCR on our training sets, along with our other training data configurations.

4.3. Training Data Configurations

We experiment with several training data configurations to investigate the effect of each component. For normalised NER, we combine Menota (M), IcePaHC (I), and MIM-GOLD-NER (MIM), with additional sCR resampling of Menota and IcePaHC (marked by superscript R, ^{R}). For diplomatic NER, we use the same combinations as above, but only Menota is resampled since it is the only diplomatic source available. We also train models using only MIM-GOLD-NER to evaluate potential transfer learning from modern to Old Icelandic. Our training data configurations are presented in Table 2.

Config.	Per.	Loc.	Total
Normalised			
M	1,486	180	1,666
M^{R}	7,421	1,041	8,462
M + I	4,283	542	4,825
$(M + I)^{R}$	22,330	2,929	25,259
M + I + MIM	19,870	9,544	29,414
$(M + I)^{R}$ + MIM	37,917	11,931	49,848
Diplomatic			
M	3,969	472	4,441
M^{R}	20,485	2,622	23,107
M + I	8,088	1,174	9,262
M^{R} + I	24,604	3,324	27,928
M + MIM	19,556	9,474	29,030
M^{R} + MIM	36,072	11,624	47,696
M + I + MIM	23,675	10,176	33,851
M^{R} + I + MIM	40,191	12,326	52,517
Modern			
MIM	15,587	9,002	24,589

Table 2: Training data configurations. M = Menota, I = IcePaHC, MIM = MIM-GOLD-NER. ^{R} = sCR resampling. Per. = Person, Loc. = Location.

4.4. NER Training

We view NER as a token classification task, adding a token classification head on top of IceBERT. We use the BIO tagging scheme with two entity types; person and location. We therefore have five labels: B-person, I-person, B-location, I-location, and O. Using the recommendations of RoBERTa (Liu et al., 2019) and BERT (Devlin et al., 2019) for fine-tuning along with what performed the best during hyperparameter experimentation based on F1 scores achieved on the development set, we train all NER models for 5 epochs with 2e-5 learning rate, batch size of 16, and maximum sequence length of 256 tokens. We leverage a 10% warm-up ratio and 0.01 weight decay. To address class imbalance, we use class weights for weighted cross-entropy loss (He and Garcia, 2009); 0.1 for non-entities (O) and 30.0 for entity classes. We employ early stopping with a patience of 3 and select the checkpoint with the highest development F1 score for evaluation.

4.5. Evaluation

We use a stratified split in our evaluation due to our entity-class imbalance; a naive random split could result in evaluation sets with too few location entities, which could hinder a generalisable and robust evaluation of the NER models' performance. Sentences containing location entities are thus split 50/50 between training and evaluation, and the evaluation set is split equally between development and test set. Sentences without location entities use a standard 70/15/15 split. This stratified approach allows for both the development and test set to contain a sufficient amount of location entities to enable reliable evaluation of model performance while still retaining location entities in the training data.

To reflect our target domain, development and test sets are derived exclusively from Old Icelandic data (normalised: Menota + IcePaHC; diplomatic: Menota). The test sets are strictly held out on the sentence level, and thus never seen during training. They are, however, derived from the same manuscripts as the training data. While this is due to the lack of available entity-annotated Old Icelandic texts, this also implies that the reported scores may benefit from scribal conventions within these manuscripts, which means that the models' generalisability to completely unseen manuscripts remains to be fully evaluated (discussed further in Limitations, section 8). We evaluate our models on the test sets using the standard metrics of precision, recall, and micro-F1 score using the seqeval library (Nakayama, 2018), which evaluates on entity-level and adheres to the evaluation of Tjong Kim Sang (2002), where both entity boundary span and entity type needs to be correct in order for the prediction to be viewed as correct. As an example, the entire multi-token entity "Björn Leifsson" needs to be entirely and correctly matched to give a correct result. There are two diplomatic NER model vari-

ations: one using the original IceBERT language model, the other using the domain-adapted IceBERT model fine-tuned through MLM for diplomatic Old Icelandic texts; normalised NER uses the original IceBERT as language model. Our evaluation sets are presented in Table 3.

Split	Person	Location	Total	Density
Normalised				
Dev	1,036	265	1,301	5.1%
Test	997	263	1,260	5.0%
Diplomatic				
Dev	912	210	1,122	5.5%
Test	922	231	1,153	5.4%

Table 3: Evaluation sets. Dev = Development set, Test = Test set.

5. Results

In the sections below, we present our results for normalised and diplomatic NER, followed by ablation studies examining the contributions of domain adaptation, resampling, and modern Icelandic data augmentation.

5.1. Normalised NER Results

Table 4 presents the results for normalised Old Icelandic NER. Our best model achieves an F1 score of 0.93 (93) by combining resampled Old Icelandic data from Menota and IcePaHC, and augmented with modern Icelandic NER data from MIM-GOLD-NER. All training data configurations maintain a high recall, while precision is the most varied. Each component added to the configuration contributes positively to model performance. With a baseline of an F1 score at 0.87 using only Menota, adding IcePaHC yields a minor improvement to 0.90, whereas sCR resampling results in a greater gain, achieving an F1 of 0.92. The best performing model uses the maximum amount of available data of Old Icelandic resampling and modern Icelandic data augmentation with an F1 score of 0.93. Unsurprisingly, using only modern Icelandic data has the lowest performance in terms of F1 score.

5.2. Diplomatic NER Results

Table 5 displays the results for diplomatic Old Icelandic NER, comparing the results of the original IceBERT and our domain-adapted IceBERT. The difficulty of the diplomatic NER task is reflected in the models' performance, with the best model, the domain-adapted model using the maximum available data, achieving an F1 score of 0.79 (14 points lower than the normalised model). This showcases the genuine difficulty of diplomatic texts and its frequent orthographic variations which are significantly more diverse compared to the normalised Old Icelandic texts. While the recall is relatively consistent across data configurations, between 0.77 and 0.85 (excluding the MIM-only configuration), the precision changes drastically. The lowest-resource configuration, using the Menota-only training configuration with the original IceBERT model, achieves an F1 score of 0.17, indicating that the model over-predicts entities mistakenly when dealing with unfamiliar medieval orthographic forms. Precision improves progressively with increased data, and the domain-adapted IceBERT model outperforms the original across all configurations.

Training Data	P	R	F1	Loc.	Per.
M	0.82	0.94	0.87	0.84	0.88
M^R	0.85	**0.95**	0.90	0.87	0.90
M + I	0.86	**0.95**	0.90	0.87	0.90
$(M + I)^R$	0.89	**0.95**	0.92	**0.90**	**0.93**
M + I + MIM	0.89	0.94	0.91	0.89	0.92
$(M + I)^R$ + MIM	**0.90**	**0.95**	**0.93**	0.89	**0.93**
MIM	0.84	0.84	0.84	0.86	0.83

Table 4: NER results on normalised Old Icelandic. P = Precision, R = Recall, M = Menota, I = IcePaHC, MIM = MIM-GOLD-NER. Superscript R indicates sCR resampling. Loc. and Per. show F1 scores for Location and Person entities respectively.

Training Data	P	R	F1	Loc.	Per.
Original IceBERT					
M	0.17	0.80	0.28	0.24	0.29
M^R	0.45	0.84	0.58	0.54	0.59
M + I	0.33	0.80	0.47	0.36	0.50
M^R + I	0.55	0.83	0.66	0.63	0.67
M + MIM	0.57	0.77	0.65	0.54	0.68
M^R + MIM	0.73	0.81	0.77	0.70	0.79
M^R + I + MIM	0.72	0.82	0.77	0.69	0.78
MIM	0.38	0.16	0.22	0.17	0.23
Domain-adapted IceBERT					
M	0.28	0.83	0.42	0.39	0.42
M^R	0.54	**0.85**	0.66	0.65	0.66
M + I	0.42	0.84	0.56	0.51	0.57
M^R + I	0.61	0.84	0.71	0.68	0.72
M + MIM	0.59	0.82	0.69	0.59	0.71
M^R + MIM	0.75	0.81	0.78	**0.73**	0.79
M^R + I + MIM	**0.77**	0.81	**0.79**	**0.73**	**0.80**
MIM	0.37	0.29	0.32	0.21	0.34

Table 5: NER results on diplomatic Old Icelandic. P = Precision, R = Recall, M = Menota, I = IcePaHC, MIM = MIM-GOLD-NER. Superscript R indicates sCR resampling. Loc. and Per. show F1 scores for Location and Person entities respectively.

5.3. Ablation Study

In this section, we present how our training configurations impacted the performance of NER on normalised and diplomatic Old Icelandic texts.

5.3.1. Effects of Resampling

Figure 1 illustrates how the sCR resampling consistently improves performance across all the training data configurations for both normalised and diplomatic texts. While resampling provides marginal gains to the normalised configurations (+2–3 F1), it does provide significant improvements in the low-resource diplomatic configurations where data is scarce, notably with the Menota-only diplomatic configurations gaining a +24–30 F1 score through resampling. sCR helps address class imbalance and performance, and this is especially prominent when training data is limited.

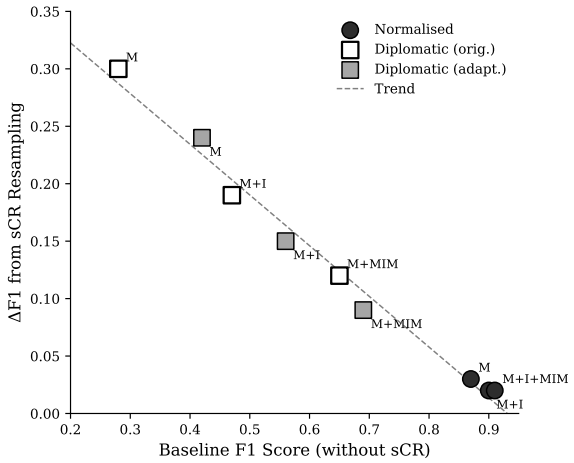

Figure 1: Relationship between baseline F1 and F1 improvement (Δ) from sCR resampling. Diplomatic texts benefit the most from resampling.

5.3.2. Effects of Domain Adaptation

Figure 2 illustrates how the domain-adapted IceBERT model outperforms the original IceBERT model on every training data configuration in terms of F1 score, where the difference in performance generally decreases as the amount of training data provided increases. The domain adaptation is most valuable in low-resource scenarios, where the domain-adapted language model's familiarity with the conventions of Old Icelandic diplomatic texts compensates for its limited NER training data, while the original IceBERT model progressively catches up as the amount of training data increases; the NER training itself successfully acts as implicit domain adaptation when given sufficient data.

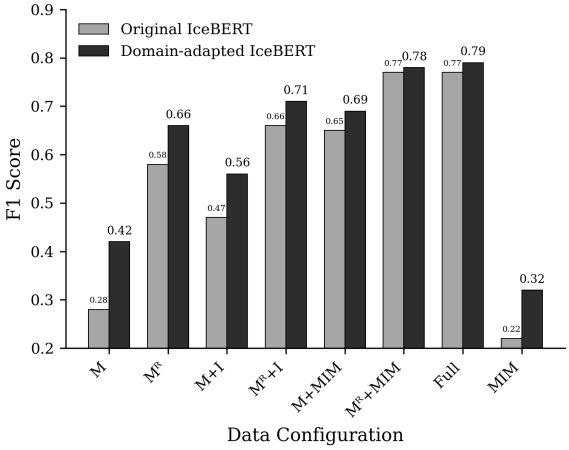

Figure 2: Effect of domain adaptation on diplomatic NER. "Full" relates to the data configuration of M^R + I + MIM which is the largest configuration.

5.3.3. Effects of Modern Icelandic Data Augmentation

Figure 3 illustrates the effect of adding modern Icelandic training data through the MIM-GOLD-NER dataset, where the lowest-resource diplomatic setting gains the biggest jump in performance, increasing F1 between +37 and +27 respectively, while the normalised settings sees an increase of a mere +0.01 F1. Similar to the results displayed in both Figure 1 and Figure 2 above, the impact of added data configurations, in this case the addition of modern Icelandic, diminishes as other, more domain-specific Old Icelandic training data increase. Interestingly, the big jump in performance for the lowest-resource diplomatic configurations suggests that there is some transfer learning from modern Icelandic to diplomatic Old Icelandic texts.

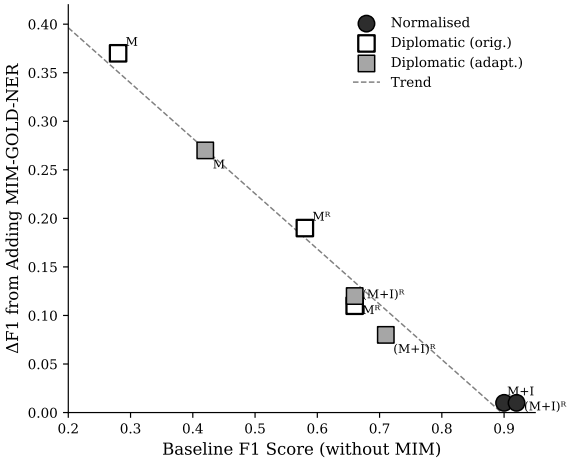

Figure 3: Relationship between baseline F1 and Δ improvement from adding modern Icelandic training data; diplomatic, low-resource, configurations benefits the most.

6. Discussion

Our experiments show competent NER performance for Old Icelandic despite its low-resource nature of NER-annotated data, with F1 scores of 93 for normalised and 79 for diplomatic texts. Our results are in line with previous medieval NER work, with F1 scores in the range of 74–94+ (Torres Aguilar, 2022; Torres Aguilar and Stutzmann, 2021; Díez Platas et al., 2021), and modern Icelandic F1 scores in the range 91–93 (Snæbjarnarson et al., 2022; Guðjónsson et al., 2021).

Looking at the difference between the F1 scores for person and location entities, person consistently has a higher F1 score than locations. Due to the class imbalance, this is expected, where around 85% of entities are persons in the Old Icelandic training data; the model simply sees far more person examples than locations during training. The resampling strategy of sCR is more effective for location entities compared to person entities, highlighting that the sCR strategy is working as intended since it privileges the resampling of under-represented entities (e.g., the domain-adapted diplomatic M vs. M^R gains +26 F1 points for location compared to +24 for person). Similarly, modern data augmentation aids the F1 scores of location in particular, which is likely due to an increase in location entities in the training data, as well as the MIM-GOLD-NER dataset having a more balanced entity distribution with around 37% entities being locations, balancing the person-focused Old Icelandic training data (e.g., domain-adapted M^R location F1 increases from 0.65 to 0.73 when adding MIM-GOLD-NER). However, the location F1 for normalised texts actually peaks at 0.90 with $(M + I)^R$ and slightly decreased to 0.89 when MIM-GOLD-NER is added. This suggests that the modern location names, often differing a lot to medieval place names, may dilute the model's performance ever so slightly when training data is abundant.

Across the ablation studies, the impact of training configurations are inverse to the baseline performance; low-resource configurations benefit greatly from each added configuration, while high-resource configurations show diminishing returns as the NER fine-tuning itself provides implicit domain adaptation. Conversely, when training data is scarce, each technique listed above compensates for the lack of data in complementary ways; resampling data helps address class imbalance, domain adaptation familiarises the model with Old Icelandic diplomatic orthography, and data augmentation of modern Icelandic provides additional entity examples to learn from. This is particularly evident in the precision-recall scores, where recall remains stable but precision varies considerably, where low-resource diplomatic training data configurations show an over-prediction on unfamiliar medieval orthographic forms.

6.1. Challenge of Diplomatic Texts

The 14-point performance gap between the best performing normalised and diplomatic model represents not only the fundamental challenge of working with medieval texts using different transcription levels, but also the linguistic distance between diplomatic Old Icelandic and modern Icelandic. Approximately 12% of all characters in the diplomatic corpus do not occur in the normalised Old Icelandic or modern Icelandic corpora, and the zero-shot results of models trained on only the modern Icelandic data are clearly indicative of this difficulty; MIM-only training achieves an F1 of 0.84 for normalised texts, but only 0.22-0.32 on diplomatic texts, where the domain-adapted model outperforms the original. This stark difference demonstrates how normalisation bridges the gap between medieval and modern texts through standardisation, while diplomatic texts preserve greater complexity and nuance, with the two different transcription levels likely appealing to different users. Our domain adaptation approach for diplomatic texts showcases consistent improvements, and the best F1 performance, across all configurations. However, while domain adaptation is valuable, it is not a complete substitute for domain-specific annotated NER training data, especially as the fine-tuning NER process itself acts as a way of adapting the model to its target domain in our experiments.

6.2. Error Analysis

To better understand our models' behaviour, we conducted a qualitative analysis on the predictions of our best-performing normalised and diplomatic NER models on their respective test set. The diplomatic model, unsurprisingly, showcases more errors than the normalised, which relates to their F1 score gap of 14 points. Both models primarily struggle with detection errors of false positives and negatives rather than entity span or classification errors; false positives and negatives constitute 67% of the normalised model's errors and 90% of the diplomatic errors. For both models, false positives are the most frequent, accounting for over half of all errors. The normalised model often predicts demonyms, such as "Gautar" (Geats), as entities (which is not necessarily wrong, but it is not what it is trained for), while the diplomatic model frequently mistakenly predicts common words as entities, like "ðalí". False negatives are relatively rare for the normalised model with 11%, while the diplomatic model's second largest error category is false negatives with 38%. This likely relates to the increased orthographic variation in the diplomatic texts, as

well as the lack of capitalised entities which might make it more difficult for the model to latch onto textual clues of an entity. In terms of classification errors, the normalised model errors account for 12% of correct entity span but wrong type, with only 4% for the diplomatic model. Span boundary errors constitute 21% of normalised errors and mainly include missed patronyms, with 6% for diplomatic. Full per entity precision and recall scores are presented in the Appendix.

6.3. Implications

Our results enable researchers working with normalised Old Icelandic texts to apply our NER model to extract person and place names, for instance for social network analysis of Icelandic saga literature, or mapping the spatial settings of medieval Icelandic texts. The diplomatic NER model, while requiring more manual verification, can allow for NER on texts that are only available in diplomatic digital editions, reducing the potential preprocessing burden of scholars needing NER for Old Icelandic while simultaneously preserving the texts' diplomatic textual form. The successful cross-temporal transfer of modern Icelandic NER data to normalised Old Icelandic texts demonstrates that modern language resources can be used to aid historical ones, suggesting that cross-temporal transfer for historical NLP development can be useful to augment low-resource historical training data. Finally, our work contributes to a growing ecosystem of NER tools crafted for medieval texts (Torres Aguilar and Stutzmann, 2021; Torres Aguilar, 2022; Besnier and Mattingly, 2021; Díez Platas et al., 2021), where we build on this to include capabilities for Old Icelandic.

6.4. Future Work

Future work could involve expanding the corpus of NER-annotated Old Icelandic texts, where a systematic annotation of Menota manuscripts, potentially involving our NER models and manual verification to speed up the annotation process, would provide a significant amount of additional training and evaluation data. Leveraging the many medieval Scandinavian manuscripts stored in Menota could also be used to explore the multilingual use of medieval NER models for Scandinavian languages. Old Norwegian, featuring prominently in Menota, shares significant linguistic similarities with Old Icelandic, and combining the two could make our NER models more robust. Finally, integrating the NER models and their output with entity linking could transform them from entities into linked data, where person names could be linked to name registries and place names to geographical gazetteers. Authority lists could be created in order to link an entity and its many spelling variations to one singular form through the use of clustering. The entities from this work could, for instance, be integrated as part of the Norse World project, a database of the Scandinavian world and spatiality sourced from medieval literature (Petrulevich et al., 2020). Similarly, investigating transfer learning from other Germanic languages, or indeed medieval Latin, could potentially provide a beneficial way of improving our models and their robustness, especially considering that medieval Latin was often used alongside vernacular languages in the Middle Ages. Finally, future work could include facsimile transcriptions since they are the most orthographically diverse and thereby more challenging, and can help shed further light on how orthographic variation affects NER performance.

7. Conclusion

We have presented OldBERTur, a NER system for Old Icelandic, achieving F1 scores of 93 for normalised and 79 for diplomatic transcriptions. Our experiments showcase competitive performance despite limited training data through domain adaptation, sentence-level resampling, and modern Icelandic data augmentation. Across ablation studies, we find that the effectiveness of additional data configurations is the inverse of the baseline performance; low-resource configurations show substantial benefit to added training configurations, higher-resource configurations show diminishing returns as the NER fine-tuning itself acts as an implicit domain adaptation. The 14-point F1 gap between our best performing normalised and diplomatic model reflects the difficulty of working with medieval orthographic conventions, especially for diplomatic texts where many characters do not appear in normalised Old Icelandic or modern Icelandic. We release our models, code, and data as open resources to contribute and support further research into medieval Scandinavian NLP and beyond.

8. Limitations

Our range of entities is restricted to only person and location names, and we perform flat rather than nested NER. For instance, "Björn Leifsson úr Ási" is captured as two entities rather than one; "Björn Leifsson" (person) and "Ási" (place). Furthermore, we evaluate our models on held-out test sets derived from the same manuscripts that we used for training. While this is due to the low-resource nature of Old Icelandic NER data, and we use stratified splitting to ensure a robust evaluation, we do not evaluate on completely unseen manuscripts or textual genres. While this reflects our target domain, it does imply that the generalisability of our models to other Old Icelandic texts, especially those with different scribal standards or from different time periods, remains to be evaluated upon. The Old Icelandic diplomatic training data is sourced exclusively from Menota, while the normalised data combines both Menota and IcePaHC. The asymmetry in training data means that the diplomatic model might be less diverse and generalisable.

9. Ethical Considerations

We work with publicly available historical texts with no modern personal data involved. As with all annotation, it is interpretative, and there may be potential biases in the NER annotations stemming from the annotators individual interpretation of how or what constitutes an entity and/or how it should be annotated according to an annotation guideline. However, this is reflective of the nature of annotation, and we consider this to not be an ethical risk.

10. Acknowledgements

This work was supported by Riksbankens Jubileumsfond (grant IN21-0003). We are also deeply grateful to Alexandra Petrulevich for providing the key conceptual starting point.

11. Bibliographical References

Paul Acker and Carolyne Larrington. 2002. *The Poetic Edda: Essays on Old Norse Mythology*. Routledge.

Sergio Torres Aguilar. 2025. TRIDIS: A Comprehensive Medieval and Early Modern Corpus for HTR and NER. *arXiv preprint arXiv:2503.22714*.

Kristján Árnason and Stephen Pax Leonard. 2011. Language ideology and standardisation in iceland. In *Standard Languages and Language Standards in a Changing Europe*, pages 91–96, Noregur. Novus Forlag.

Clément Besnier and William Mattingly. 2021. Named-Entity Dataset for Medieval Latin, Middle High German and Old Norse. *Journal of Open Humanities Data*, 7(23):1–5.

David J Birnbaum, Sheila Bonde, and Mike Kestemont. 2017. The Digital Middle Ages: An Introduction. *Speculum*, 92(S1):S1–S38.

Thibault Clérice, Ariane Pinche, Malamatenia Vlachou-Efstathiou, Alix Chagué, Jean-Baptiste Camps, Matthias Gille Levenson, Olivier Brisville-Fertin, Federico Boschetti, Franz Fischer, Michael Gervers, et al. 2024. CATMuS Medieval: A Multilingual Large-scale Cross-century Dataset in Latin Script for Handwritten Text Recognition and Beyond. In *International Conference on Document Analysis and Recognition*, pages 174–194. Springer.

Carol J. Clover and John Lindow. 2019. *Old Norse-Icelandic Literature: A Critical Guide*. Cornell University Press, Ithaca, NY.

Margaret Clunies Ross. 2000. *Old Icelandic Literature and Society*. Cambridge: Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mª Luisa Díez Platas, Salvador Ros Muñoz, Elena González-Blanco, Pablo Ruiz Fabo, and Elena Álvarez Mellado. 2021. Medieval Spanish (12th–15th centuries) Named Entity Recognition and Attribute Annotation System Based on Contextual Information. *Journal of the Association for Information Science and Technology*, 72(2):224–238.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named Entity Recognition and Classification in Historical Documents: A Survey. *ACM Computing Surveys*, 56(2):1–47.

Maud Ehrmann, Matteo Romanello, Alex Flückiger, and Simon Clematide. 2020. Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers. In *CEUR Workshop Proceedings*, (No. 2696). CEUR-WS.

Maud Ehrmann, Matteo Romanello, Sven Najem-Meyer, Antoine Doucet, and Simon Clematide. 2022. Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 423–446, Cham. Springer International Publishing.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Odd Einar Haugen. 2004. Parallel Views: Multilevel Encoding of Medieval Nordic Primary Sources. *Literary and Linguistic Computing*, 19(1):73–91.

Haibo He and Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Amanda Hilmarsson-Dunn and Ari Páll Kristinsson. 2010. The language situation in Iceland. *Current Issues in Language Planning*, 11(3):207–276.

Even Hovdhaugen, Carol Henriksen, Fred Karlsson, and Bengt Sigurd. 2000. *The History of Linguistics in the Nordic Countries*. Societas Scientiarum Fennica.

Frédéric Kaplan and Isabella di Lenardo. 2017. Big Data of the Past. *Frontiers in Digital Humanities*, 4:12.

Dick Kasperowski, Karl-Magnus Johansson, and Olof Karsvall. 2024. Temporalities and Values in an Epistemic Culture: Citizen Humanities, Local Knowledge, and AI-Supported Transcription of Archives. *Archives and Manuscripts*, 51(2):3–22.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Hrafn Loftsson and Eiríkur Rögnvaldsson. 2007. IceNLP: A Natural Language Processing Toolkit for Icelandic. In *Interspeech 2007*, pages 1533–1536.

Joseph Nockels, Paul Gooding, and Melissa Terras. 2024. The Implications of Handwritten Text Recognition for Accessing the Past at Scale. *Journal of Documentation*, 80(7):148–167.

Alexandra Petrulevich, Agnieszka Backman, and Jonathan Adams. 2020. Medieval macrospace through gis: the norse world project approach. *The Cartographic Journal*, 57(1):18–27.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984. European Language Resources Association (ELRA).

Jenna Schoen and Gianmarco E Saretto. 2022. Optical Character Recognition (OCR) and medieval Manuscripts: Reconsidering Transcriptions in the Digital Age. *Digital Philology: A Journal of Medieval Cultures*, 11(1):174–206.

Stefan Schweter and Johannes Baiter. 2019. Towards Robust Named Entity Recognition for Historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 96–103, Florence, Italy. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Merve Tekgürler. 2025. LLMs for Translation: Historical, Low-Resourced Languages and Contemporary AI Models. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 227–237, Albuquerque, New Mexico. Association for Computational Linguistics.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, pages 1–4. Association for Computational Linguistics.

Sergio Torres Aguilar. 2022. Multilingual Named Entity Recognition for Medieval Charters Using Stacked Embeddings and BERT-based Models. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 119–128, Marseille, France. European Language Resources Association.

Sergio Torres Aguilar and Dominique Stutzmann. 2021. Named Entity Recognition for French Medieval Charters. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 37–46, NIT Silchar, India. NLP Association of India (NLPAI).

Sverrir Tómasson. 2006. Old Icelandic Prose. In Daisy Neijmann, editor, *A History of Icelandic Literature*, volume 5 of *Histories of Scandinavian Literature*, pages 64–173. University of Nebraska Press and The American-Scandinavian Foundation, Lincoln, Nebraska and London.

Xiaochen Wang and Yue Wang. 2022. Sentence-Level Resampling for Named Entity Recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2151–2165, Seattle, United States. Association for Computational Linguistics.

12. Language Resource References

Ásmundur Alma Guðjónsson, Hrafn Loftsson, and Jón Friðrik Daðason. 2021. Icelandic NER API - ensemble model (21.09). CLARIN-IS.

Odd Einar Haugen. 2015. MUFI character recommendation, version 4.0. Medieval Unicode Font Initiative.

Odd Einar Haugen. 2019. The Menota Handbook: Guidelines for the Electronic Encoding of Medieval Nordic Primary Sources. Gen. ed. Odd Einar Haugen. Version 3.0. Bergen: Medieval Nordic Text Archive.

Svanhvít Lilja Ingólfsdóttir, Ásmundur Alma Guðjónsson, and Hrafn Loftsson. 2020. MIM-GOLD-NER – named entity recognition corpus (21.09). CLARIN-IS.

Hrafn Loftsson. 2019. IceNLP Natural Language Processing Toolkit. CLARIN-IS.

Menota. 2025. Medieval Nordic Text Archive. Menota Consortium, hosted at CLARINO Bergen Centre, University of Bergen.

Hiroki Nakayama. 2018. seqeval: A Python Framework for Sequence Labeling Evaluation. GitHub.

TEI Consortium. 2025. TEI P5: Guidelines for electronic text encoding and interchange. Version 4.10.2. Text Encoding Initiative Consortium.

Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2024. Icelandic parsed historical corpus (IcePaHC) 2024.03. CLARIN-IS.

Appendix

Text	Date	Transcription
Alexanders saga (AM 519a 4to)	∼1280	Dipl., Norm.
Njáls saga fragm. (AM 162B θ fol)	1300–1350	Dipl., Norm.
Njáls saga fragm. (AM 162B κ fol)	1325–1375	Dipl., Norm.
Drauma-Jóns saga (AM 657a-b 4to)	1350–1399	Dipl., Norm.
Íslendingabók (AM 113b fol)	1650–1700	Dipl., Norm.
Læknísbók (AM 655 XXX 4to)	1250–1300	Dipl., Norm.
Óláfs saga helga (Lbs fragm. 82)	1258–1264	Dipl., Norm.
Vǫluspá (AM 544 4to)	1290–1360	Dipl. only
Codex Wormianus (AM 242 fol)	∼1350	Dipl. only

Table 6: Menota texts used for NER training and evaluation. Dipl. = Diplomatic, Norm. = Normalised. Seven texts have both diplomatic and normalised transcriptions; two are diplomatic only. The normalised texts were used for the normalised NER model, and diplomatic for the diplomatic NER model.

Text	Date
Jarteinabók	1200-1220
Þorláks saga helga	1200-1220
Sturlunga saga	1271-1284
Þetubrot Egils Sögu	∼1250
Jómsvíkinga saga	1250-1275
Morkinskinna	1200-1275

Table 7: IcePaHC texts used for normalised NER training and evaluation. All texts are in normalised form and date to the 13th century.

Text	Date
Alexanders saga (AM 519a 4to)	∼1280
Njáls saga fragm. (AM 162B θ fol)	1300–1350
Njáls saga fragm. (AM 162B κ fol)	1325–1375
Brennu-Njáls saga (AM 132 fol)	1320–1350
Drauma-Jóns saga (AM 657a-b 4to)	1350–1399
Egils saga (AM 132 fol)	1320–1350
Finnboga saga (AM 132 fol)	1330–1370
Íslendingabók (AM 113b fol)	1650–1700
Kormáks saga (AM 132 fol)	1330–1370
Læknísbók (AM 655 XXX 4to)	1250–1300
Laxdæla saga (AM 132 fol)	1320–1350
Óláfs saga helga (Lbs fragm. 82)	1258–1264
Víga-Glúms saga (AM 132 fol)	1330–1370
Vǫluspá (AM 544 4to)	1290–1360
Codex Wormianus (AM 242 fol)	∼1350

Table 8: Menota texts used for MLM domain adaptation. This corpus consists of the nine entity-annotated diplomatic texts (without their annotations) and six additional unannotated diplomatic texts.

Model	Entity	P	R	F1	Support
Normalised: (M + I)^R + MIM					
	Location	0.83	0.96	0.89	263
	Person	0.92	0.95	0.93	997
	micro avg	0.90	0.95	0.93	1,260
Diplomatic: M^R + I + MIM (domain-adapted)					
	Location	0.79	0.68	0.73	232
	Person	0.76	0.85	0.80	922
	micro avg	0.77	0.81	0.79	1,154

Table 9: Per-entity precision (P), recall (R), and F1 scores for the best-performing normalised and diplomatic NER models on their respective test sets. Support shows the number of entities in the test set.