BRAGD: Constrained Multi-Label POS Tagging for Faroese

Annika Simonsen[1] **Barbara Scalvini**[2] **Uni Johannesen**[2]
Iben Nyholm Debess[2] **Hafsteinn Einarsson**[1] **Vésteinn Snæbjarnarson**[1,3]
[1]University of Iceland [2]University of the Faroe Islands [3]University of Copenhagen
{annika, hafsteinne, vesteinns}@hi.is, {barbaras, unij, ibennd}@setur.fo

Abstract

We present the first multi-label part-of-speech (POS) tagger for Faroese using linguistically-informed constraints, addressing the data sparsity problem inherent in *compound tag* approaches. We propose the BRAGD tagset, which decomposes compound morphological tags into independent features (word class, gender, number, case, etc.). The BRAGD tagset is the third iteration of a tagset previously released for Faroese, with substantial modifications that are better aligned with Faroese grammar. We annotate the previously released *Sosialurin* corpus with the tagset, as well as a new annotated out-of-domain test corpus of 500 sentences from more varied and contemporary texts. To train the tagger, we use a constrained loss function that dynamically masks morphologically invalid features based on the word class (noun, verb, adjective, etc.). We fine-tune a Scandinavian transformer language model using the constrained multi-label loss, achieving an overall accuracy of 97.5%. We find that models trained with multi-label loss perform better, converge faster, and show significantly lower error rates on out-of-domain data than single-label approaches or previously reported methods for Faroese POS tagging. This confirms that the multi-label approach learns robust morphological patterns rather than memorizing domain-specific tag distributions. We release models, code, and the systematically revised Sosialurin-BRAGD corpus, featuring the new BRAGD tagset and a new out-of-domain evaluation corpus from diverse and contemporary text types.

Keywords: part-of-speech, multi-label, linguistic constraints, low-resource, Faroese

https://github.com/Maltoknidepilin/BRAGD

1. Introduction

Part-of-speech (POS) tagging is the task of automatically assigning grammatical labels to words in text, identifying whether each word is a noun, verb, adjective, etc., along with its morphological properties such as case, gender, and number. For morphologically rich languages like Faroese, POS tagging provides foundational infrastructure for downstream applications such as syntactic parsing, information extraction, grammar and spell checking, and named entity recognition. While large language models offer general-purpose text processing capabilities (Blevins et al., 2023), dedicated POS taggers remain essential for Faroese for several reasons: they provide a cost-effective and computationally efficient solution that runs locally without dependency on commercial API services; they enable community control over model development and maintenance; and they serve important roles in linguistic research and language education, where explicit morphological analysis is fundamental to studying and teaching Faroese grammar.

Faroese is a North Germanic language spoken by approximately 50,000 people in the Faroe Islands (Statistics Faroe Islands, 2025). Like its close relative, Icelandic, it exhibits complex inflectional patterns including four-case declension, three-gender noun systems, and intricate verbal and adjectival conjugation (Thráinsson et al., 2012), where a single adjective can take 29 unique forms depending on case, number, gender, degree and declension (Rúnarsson et al., 2023). This morphological richness poses significant challenges for accurate tagging, particularly given the language's limited computational resources.

A recent approach to Faroese POS was a neural method adapted from Icelandic (Hafsteinsson, 2020). Hafsteinsson (2020) trained a Faroese version of ABLTagger (Steingrímsson et al., 2019), a two-stage LSTM architecture.[1] The single-label approach in ABLTagger does not make the most use of the sparse training data: many compound tags appear rarely, which motivated the use of a multi-label approach in this work.

Inspired by recent work for Icelandic (Snæbjarnarson et al., 2022), we propose a constrained multi-label approach. We expand the 371 compound tags from Hafsteinsson (2020) to 651 compound tags and then decompose them into 73 binary dimensions representing independent morphological features (word class, gender, number, case, etc.). While a compound tag like "noun-masculine-singular-genitive-indefinite" might appear 100 times in the training corpus, the individual features "masculine", "singular", and "genitive" each appear thousands of times across

[1]The Faroese ABLTagger in (Hafsteinsson, 2020) outputs a distribution over 371 compound morphological tags and achieved 91.4% accuracy on the Sosialurin corpus.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11650–11668
11–16 May 2026.

Token / tags	Detailed change
Tann PRIOR: PDMSN NEW: RMSND	**Word class:** Article ↪ Masc, Sg, Nom, Def *[R: new main category with attributes]*
væl PRIOR: DN NEW: DNd	**Word class:** Adverb ↪ does not govern case, no degree *[d: explicit no-degree marker]*
dámdi PRIOR: VAMSN NEW: LWMSN	**Word class:** Past Participle ↪ Weak, Masc, Sg, Nom *[LW: category restructuring]*
sangarin PRIOR: SMSNA NEW: SMSNAr	**Word class:** Noun ↪ Masc, Sg, Nom, Def *[r: explicit no-proper marker]*
hevur PRIOR: VIPS3 NEW: VIAPS3	**Word class:** Verb ↪ Indic, Active, Pres, Sg, 3rd person *[A: new voice subcategory]*
sungið PRIOR: VANSN NEW: VUAtnp	**Word class:** Verb ↪ Sup, Active, no-tense, no-number, no-person *[UAtnp: new supine, new voice subcategory, explicit none-markers]*

Table 1: Multi-label decomposition comparing Hafsteinsson (2020) with BRAGD for "*Tann væl dámdi sangarin hevur sungið*" ("The well-liked singer has sung"). The key changes shown are the addition of a new article main category, the restructuring of Past Participles, dividing the Mood subcategory into separate Mood and Voice subcategories in Verbs, adding the supine verb form, and explicit non-feature markers (r, p).

different word classes, providing a richer training signal. Table 1 illustrates this decomposition using a Faroese example sentence (See Appendix A for the full tagset).

The constrained multi-label approach masks morphologically invalid features based on word class during both training and inference, since not all morphological features are valid for all word classes. For example, verbs do not have gender, and adjectives do not have person marking. Our loss function uses the word class during training to dynamically mask out morphologically irrelevant feature dimensions, enforcing linguistic constraints during training. At inference time, the approach ensures that predictions always produce valid feature combinations for the predicted word class.

We systematically revise the Sosialurin corpus (Hansen et al., 2004; Hafsteinsson, 2020) with corrections to a total of 27,426 out of 117,322 tags, correcting annotation errors inherited from Icelandic conventions to create the improved Sosialurin-BRAGD corpus that better reflects Faroese grammar (see Table 1 and Section 3 for examples, Appendix A for the full tagset, and Appendix B for a detailed comparison with prior tagsets).

We also release a new 500-sentence manually-annotated corpus from diverse contemporary texts (2019-2022) to evaluate out-of-domain performance. By fine-tuning ScandiBERT (Snæbjarnarson et al., 2023), a transformer model trained on Scandinavian languages including Faroese, we achieve 97.5% overall accuracy, a 7.6% improvement over ABLTagger when re-evaluated on the same corpus (89.9%, see Table 4).

2. Related work

2.1. Faroese POS tagging

Foundational work in Faroese POS tagging was conducted by Hansen et al. (2004), who created the Sosialurin corpus from newspaper text. They manually tagged a portion of the corpus and used this to train a TnT tagger (Brants, 2000), which they used to machine-tag the remaining corpus before manual correction to create a gold standard. The final corpus contains over 110,000 words from 221 articles using a tagset based on Icelandic methods (Pind et al., 1991), with 390 unique compound tags.

Hafsteinsson (2020); Hafsteinsson and Ingason (2020) advanced Faroese POS tagging by adapting ABLTagger (Steingrímsson et al., 2019), a bi-LSTM recurrent neural-network tagger developed for Icelandic. Hafsteinsson (2020) reduced the original 390-tag Sosialurin tagset through systematic revisions based on Icelandic conventions, achieving 91.4% overall accuracy on their tagset by making use of a morphological lexicon. However, their single-label approach treats each morphological analysis as an atomic compound tag, which does not directly allow joint optimization.

While universal tagset frameworks such as the Universal POS tagset (Petrov et al., 2012) and Universal Dependencies (de Marneffe et al., 2021) have enabled cross-lingual comparison and transfer—including for Faroese, where Tyers et al. (2018) created a synthetic UD treebank for cross-lingual dependency parsing—these frameworks prioritize cross-linguistic consistency over language-specific granularity. Both the original Sosialurin tagset and Hafsteinsson (2020)'s revision inherit Icelandic grammatical conventions that do not fully reflect Faroese grammar. Our BRAGD tagset instead prioritizes grammatical fidelity to Faroese, capturing distinctions such as mediopassive voice and absolute superlative that are absent from both universal and prior Faroese-specific schemes but essential for accurate morphological analysis.

2.2. Multi-label Morphological Tagging

Structured prediction methods have long incorporated constraints into morphological modeling. Conditional random fields (CRFs; Lafferty et al., 2001) pioneered the integration of output label dependencies into sequence labeling, and CRF-based morphological taggers such as MarMoT (Müller et al., 2013) demonstrated the value of encoding feature dependencies for morphologically rich languages. Cotterell and Heigold (2017) further combined neural character-level representations with CRF-factored outputs for cross-lingual morphological transfer. Our constrained loss function follows a similar principle in a neural setting: rather than modeling pairwise label transitions, we enforce hard constraints on which morphological features are valid for each word class, integrating linguistic knowledge directly into the training objective. More specifically, multi-label approaches to morphological tagging model multiple linguistic features jointly rather than as independent decisions (Chrupała et al., 2008). Neural approaches have proven particularly effective: Müller and Schuetze (2015) showed that word representations enable robust joint prediction of multiple morphological features, substantially improving handling of rare word forms. Silfverberg and Drobac (2018) demonstrated that explicitly modeling sub-label dependencies in neural architectures improves prediction accuracy, especially for morphologically complex languages. This approach has also been used for Icelandic POS-tagging (Snæbjarnarson et al., 2022). Our work builds on these insights by integrating linguistic constraints about valid feature combinations directly into the loss function through dynamic masking.

3. The BRAGD Tagset and Corpus

We revise Hafsteinsson (2020)'s tagset and tagged corpus[2] (itself a revision of the original Sosialurin corpus by Hansen et al. (2004)) to create Sosialurin-BRAGD[3], incorporating systematic corrections and a restructured tagset (renamed BRAGD tagset) that better reflects Faroese grammar. The conversion was carried out by native Faroese linguists, who analyzed the tagset to identify necessary changes for proper Faroese grammar representation and then defined the 15 word classes. After establishing this mapping, regular expressions were used to replace the old tags with the new tagset.

[2]We started with the corpus file available at https://github.com/hinrikur/FAR-GOLD/blob/master/correction/finished/sosialurin-revised.txt.

[3]BRAGD refers to the traditional Faroese practice of marking sheep: a *bragd* is a distinctive notch or incision cut into a sheep's ear for identification.

Our revision process involved two complementary efforts: first, structural modifications to the tagset itself to better represent Faroese grammar (Section 3.1), and second, systematic correction of annotation errors where the previous tagset had been misapplied in the corpus (Section 3.2). The Sosialurin-BRAGD corpus contains 6,099 sentences with 117,322 word-level tokens (including punctuation). A full overview of the BRAGD tagset is given in Appendix A with an overview of differences to the prior tagsets in Appendix B.

3.1. BRAGD Tagset Development

We restructure the tagset in Hafsteinsson (2020) to better align with Faroese grammatical categories. The changes are based on canonical reference works in Faroese linguistics, including overview works, grammar books, and dictionaries, e.g. Thráinsson et al. (2012), Andreasen and Dahl (1997), and Poulsen et al. (1998). The original tagset from Hansen et al. (2004) was based on an Icelandic tagset and Icelandic grammatical conventions and lacks a clear linguistic justification within Faroese grammar. Hafsteinsson (2020) addresses some of the discrepancies between the original tagset and Faroese grammatical structure. However, several of these issues remained unresolved. The remaining gaps in the tagset stem from historical developments rather than principled linguistic decisions. Many of the modifications introduced in the BRAGD tagset aim to address these shortcomings. We list these below.

Pronouns and articles. To better align the tagset with the syntactic realities of the Faroese language, we make the following structural changes to pronouns and articles:

Establishment of the Article Category (R): We introduced a new main category for articles, a class previously omitted, also discussed in Hafsteinsson (2020). While articles in Faroese often share orthographic forms with pronouns and numerals (and were previously tagged as such, see Section 3.2), they serve distinct syntactic functions in Faroese that justify this new classification.

Refining Pronoun Subcategories: The "indefinite demonstrative" pronoun subgroup (PB) is eliminated. These instances are reclassified in the corpus as indefinite pronouns (PI) or demonstrative pronouns (PD) based on syntactic function.

Pronoun Restructuring: We separate Gender from the Person subcategory into a new Gender class for pronouns and add explicit 3rd-person marking to the Person subcategory. This separation ensures that pronouns follow the same multi-dimensional structure as other categories, enabling more consistent cross-category learning.

Verbs. The following changes were made to the Verb Category:

Added Voice: We remove the Middle voice marker (E, labeled 'Medium' in the previous tagset) from the Verb Mood subcategory and instead add a Voice subcategory for Verbs with Active (A) and Mediopassive (M) Voice. Mediopassive is a distinct grammatical category in Faroese with specific syntactic properties (sometimes referred to as "-st form"), and treating it as mood obscures important grammatical distinctions[4].

Participle Reorganization: Past participles are moved from the V-verb category to a new L-past participle category. Morphologically, Faroese participles function like adjectives, inflecting for gender, number, and case, represented by subcategories in all tagsets. Furthermore, past participles have strong/weak declension, which is reflected in a new subcategory. Given that participles are non-finite verb forms and lack mood, that specific subcategory was eliminated.

Added Supine Form: We added the supine to the Verb category. Supine forms are orthographically identical to past participle forms, but serve different syntactic and semantic functions.

Adverb Reordering. We swap positions 2 and 3 in the adverb category (Category/Case governor $\leftrightarrow$ Degree) for consistency with other word classes. This ensures that degree features occupy parallel positions across word classes (adjectives, adverbs) that share this property.

The BRAGD tagset is not yet exhaustively aligned with Faroese grammar; remaining gaps are discussed in the Limitations (See Section 8).

3.2. Sosialurin-BRAGD Development

After defining the BRAGD tagset, we annotated the corpus by combining systematic conversion with targeted correction of annotation errors in the earlier version. The revision therefore involved two kinds of changes: (i) harmonisation changes required to map existing annotations into the new tagset, and (ii) grammar-driven reanalysis where the earlier annotation did not adequately reflect Faroese usage. In total, we applied **31,185** tag changes, of which **22,535** were applied automatically and **8,650** were made through manual review (including partially automatic procedures where context rules allowed some occurrences to be auto-skipped); **21,214** changes reflect tagset harmonisation and **9,971** reflect grammar-driven reanalysis (Table 2).

[4]We note that Hafsteinsson (2020) independently made a similar revision in unpublished subsequent work, though we compare only to his published tagset.

Reason	Automatic	Manual	Total
Tagset	20,375	839	21,214
Grammar	2,160	7,811	9,971
Total	22,535	8,650	31,185

Table 2: Tag changes in Sosialurin-BRAGD. Rows categorize why the change was made (tagset harmonisation vs. grammatical correction), while columns indicate the processing method (automatic vs. manual). Because some tokens were revised more than once during the correction process, totals reflect logged edit operations rather than unique final token differences.

In addition to rule-based conversion and targeted manual correction of ambiguous forms, we used iterative retagging to identify further inconsistencies in the corpus. Using a preliminary tagger trained on the updated corpus, we repeatedly retagged the corpus and manually reviewed all tokens where the existing tag disagreed with the retagged output; in later passes, previously reviewed tokens, both kept and changed, were automatically skipped to improve efficiency and maintain decision consistency. Across four such retagging passes, **3,411** tags were corrected (included in the manual, grammar-driven portion of Table 2) and **2,441** additional tokens were manually verified without change. Because this review was triggered by tagger–corpus disagreement, the procedure selectively targeted tagger-detectable inconsistencies: systematic errors may persist where the tagger agrees with an incorrect annotation, and corrections may therefore be concentrated in categories to which the tagger is especially sensitive.

To complement the process-based overview in Table 2, we also computed an independent corpus-level diff between Hafsteinsson (2020)'s revised corpus and Sosialurin-BRAGD and aggregated the resulting differences by main category (Table 3). The two tables therefore capture different aspects of the revision and are not expected to yield identical totals: Table 2 is derived from revision logs and reports edit operations, allowing classification by reason (tagset vs. grammar) and method (automatic vs. manual), whereas Table 3 is based only on the final differences between the two corpus versions and therefore shows where changes are concentrated by main category. Because some tokens were revised more than once during the correction process, sometimes across multiple passes or in relation to different grammatical features, Table 2 may count multiple edit operations for a single token. The totals in Table 3 are therefore lower, since that table reflects only the final net differences between corpus versions. In net terms, the final corpus differs from the earlier revised version by 27,426 to-

Hafsteinsson (2020)	BRAGD	Changes
Pron. (P)	Pron. (P)	8,998
Conj. (C)	Conj. (C)	7,933
Nouns (S)	Nouns (S)	2,144
Past part. (VA)	Past part. (L)	1,744
Past part. (VA)	Verbs (V)	1,358
Adj. (A)	Adj. (A)	727
Verbs (V)	Verbs (V)	618
Foreign (F)	Nouns (S)	610
Pron. (P)	Articles (R)	568
Adverbs (D)	Conj. (C)	544
Adverbs (D)	Adverbs (D)	416
Verbs (V)	Adj. (A)	307
Abbrev. (T)	Nouns (S)	291
Other		1,168
Total		27,426

Table 3: Number of tag changes grouped by main category, computed from a token-level alignment between Hafsteinsson (2020) and Sosialurin-BRAGD. Unlike Table 2, these counts reflect net differences between corpus versions rather than logged revision operations.

kens, corresponding to 23.4% of all tokens.

The largest concentrations of change in Table 3 mainly reflect the structural revisions discussed in Section 3.1, especially the reorganisation of pronominal features, the refinement of conjunction subcategories, and the redistribution of former past participle tags across the new participle and verb analyses. By contrast, other frequent shifts primarily reflect corpus-level reanalysis of ambiguous forms in context, including article–pronoun disambiguation, reassignment of form-identical items between adverb and conjunction functions, and corrections within adverbial subtypes. Thus, while some high-frequency changes are direct consequences of adopting the BRAGD tagset, others represent genuine annotation corrections made during corpus revision. A finer-grained breakdown of feature-level changes within main categories (e.g. gender, case, pronoun type, and conjunction category) is provided in Appendix C.

3.3. Out-of-Domain Evaluation Data

To know how well a POS-tagger trained on the *Sosialurin-BRAGD* data performs on out-of-domain text, we annotate 500 sentences (approximately 10,000 tokens) from the BLARK corpus for Faroese (Simonsen et al., 2022). This dataset represents a significant domain shift from the training data. While *Sosialurin-BRAGD* contains newspaper text from the late 1990s to the early 2000s, the *OOD-BRAGD* evaluation set includes diverse text types from 2019–2022, including newspaper articles, blogs, essays, content from the Faroese Statistical Office website (hagstova.fo), and publications from MEGD (Faroese Disability Organizations). This evaluation allows us to assess the model's robustness across different genres and contemporary language use.

Three native Faroese-speaking linguists manually evaluated model predictions on OOD-BRAGD. The annotators evaluated different sentences to maximize coverage, but maintained regular communication throughout, jointly discussing ambiguous or challenging cases to ensure tagging consistency. For the annotation of the out-of-domain data, predictions were first produced automatically using the multi-label POS model (see Section 4) and then manually reviewed. After manual review, we introduced additional refinements, such as adding supine mood and possessive pronouns, to both the tagset and the corpus. These revisions were subsequently propagated to OOD-BRAGD to ensure that it consistently reflects the most recent version of the training data and the tagset.

4. Training POS-taggers

We fine-tune the ScandiBERT (Snæbjarnarson et al., 2023) model for single and multi-label POS classification (see App. F). ScandiBERT has 12 encoder layers (hidden size 768, 12 attention heads per layer). We take the final hidden states (the per-token outputs of the last transformer layer) and feed them through a token-classification head (0.1 dropout and a linear projection) to obtain logits over 73 labels. Each output dimension corresponds to a specific morphological feature, allowing prediction of word class and detailed morphological properties.

4.1. Feature Encoding

To ensure consistent dimensionality across all tokens, we encode the BRAGD tags as 73-dimensional binary vectors with explicit "no-feature" markers (no-gender, no-case, no-number, etc.) for word classes where certain features are inapplicable. For example, personal pronouns in first and second person are marked with "no-gender" while third-person pronouns have explicit gender marking. This representation ensures that every token has a defined value for all morphological dimensions, preventing the model from erroneously assigning gender, number, or case features to words that do not carry them, even when such features would otherwise be expected given the word class.

4.2. Constrained Multi-Label Loss

We use a constrained loss function that integrates corpus-derived linguistic constraints directly into

the learning objective. The key insight is that not all morphological features are valid for all word classes. For instance, verbs do not have gender, and adjectives do not have person. This ensures that the model only needs to learn to predict applicable labels and prevents a large class imbalance from predicting properties as 'not-present'. At inference, we follow a similar pattern: we first predict the word class and then decode only the relevant feature labels.

The tagset defines which morphological features are relevant for each word class (see Appendix A). This mapping drives the constrained loss: for each token, we compute cross-entropy only for the features (including the word class itself) that are valid for that token's **gold word class**.

Formally, let $\mathcal{D}$ be a batch of sequences. For a sequence $\mathbf{x} \in \mathcal{D}$ and token position i, let τ_i denote the gold word class and $\Phi(\tau_i)$ the set of morphological features active for that class (e.g., gender, number, case for nouns; tense, mood for verbs, including the word class label itself). The model p exposes one output head per feature; we write $p^\phi(\mathbf{x}_i \mid \mathbf{x}_{<i})$ for the logits of head $\phi \in \Phi(\tau_i)$, where τ_i is the word class at position i, and y_i^ϕ for the corresponding gold label for the given feature ϕ. We define the per-token loss as the mean cross-entropy over active features:

$$\ell(\mathbf{x}, i) = \sum_{\phi \in \Phi(\tau_i)} \mathrm{CE}\Big(p^\phi(\mathbf{x}_i \mid \mathbf{x}_{<i}),\ y_i^\phi\Big), \quad (1)$$

and the batch loss as the mean over all tokens:

$$\mathcal{L}(\mathcal{D}) = \frac{1}{\sum_{\mathbf{x} \in \mathcal{D}} |\mathbf{x}|} \sum_{\mathbf{x} \in \mathcal{D}} \sum_{i=1}^{|\mathbf{x}|} \ell(\mathbf{x}, i). \quad (2)$$

The per-token loss sums the cross-entropy terms over all active features, so tokens with richer morphology (e.g., nouns, verbs) contribute proportionally more gradient than morphologically simpler classes (e.g., punctuation). At inference time, since τ_i is unavailable, we first predict the word class $\hat{y}_i^\tau = \arg\max\, p^\tau(\mathbf{x}_i \mid \mathbf{x}_{<i})$ and then decode only the heads in $\Phi(\hat{y}_i^\tau)$. This approach ensures that morphological features are only predicted for linguistically valid combinations.

4.3. Training Configuration

We train and evaluate all models using 10-fold cross-validation. Each fold was trained using the AdamW optimizer with a learning rate of 2×10^{-5} and batch size of 8, continuing for up to 100 epochs or until validation performance did not improve for 3 consecutive epochs. The best-performing checkpoint was selected for each fold.

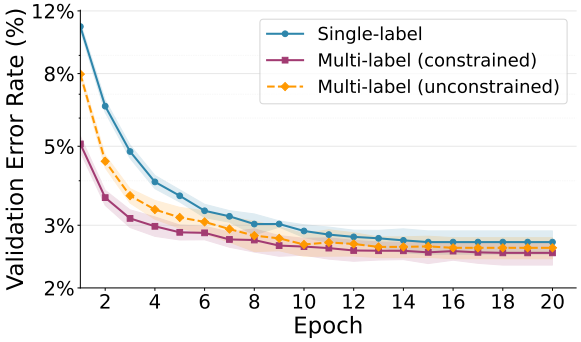

Figure 1: Validation error rate (log scale) on Sosialurin-BRAGD as a function of training epoch for single-label, multi-label (constrained), and unconstrained multi-label models. Error is measured on the full compound tag (all morphological features must be correct). Shaded regions show the standard deviation over 10-fold cross-validation.

The batch size was selected by comparing 4, 8, 16, 32, and 64; we chose 8 as it provided the best balance between training stability and computational efficiency. The learning rate of 2×10^{-5} follows standard practice for fine-tuning BERT-based models and was not extensively tuned. We did not use gradient clipping, learning rate warmup, or weight decay, as initial experiments showed stable training without these techniques.

Using the BRAGD tagset, training dynamics revealed that all configurations converged within a similar range of approximately 11–12 epochs on average. The multi-label constrained formulation converged in an average of 11.5 epochs (range: 5–18), compared to 12.3 epochs (range: 10–15) for single-label. The unconstrained variant converged in 10.8 epochs (range: 8–16). Rather than conferring a substantial convergence advantage, the multi-label formulation's primary benefit lies in its constrained loss function, which computes gradients only for valid morphological feature combinations. A visualization of this is given in Figure 1.

All models were trained and evaluated on a single NVIDIA RTX 3090 GPU. Training time per epoch was approximately 1 minute for the single-label models and 2 minutes for the multi-label models.

5. Results

We train and evaluate the BRAGD POS-tagger using 10-fold cross-validation. We also evaluate each checkpoing on the out-of-domain BRAGD corpus. The metrics we consider are accuracy at the word level (all labels correct) and word class level, as well as F1 scores per label.

We evaluate against two baseline systems, both used for earlier Faroese POS-taggers: TnT (Brants, 2000) and the Faroese ABLTagger (Hafsteinsson,

	Method	Overall	Noun	Verb	Adj.	Adverb	Pron.	Num.	Con.	R Article	Partic.	Abbr.	For.	Punc.	Sym.
Sos. (ID)	TnT	$84.9_{0.4}$	$\mathbf{99.7}_{0.1}$	$98.8_{0.2}$	$95.0_{0.6}$	$98.2_{0.2}$	$97.0_{0.5}$	$97.4_{0.8}$	$98.5_{0.2}$	$89.5_{1.8}$	$87.5_{2.5}$	$99.5_{1.1}$	$\mathbf{87.6}_{9.4}$	$\mathbf{100.0}_{0.0}$	$\mathbf{100.0}_{0.0}$
	ABL-Tag.	$89.9_{0.3}$	$93.8_{1.9}$	$97.8_{0.3}$	$89.4_{0.9}$	$97.6_{0.2}$	$96.7_{0.5}$	$91.3_{1.4}$	$98.7_{0.2}$	$92.0_{1.2}$	$87.9_{3.4}$	$\mathbf{100.0}_{0.0}$	$80.1_{4.3}$	$\mathbf{100.0}_{0.0}$	$98.3_{0.2}$
	Sc.B-sing	$97.3_{0.2}$	$99.6_{0.1}$	$\mathbf{99.8}_{0.1}$	$96.1_{0.5}$	$\mathbf{99.1}_{0.1}$	$98.6_{0.3}$	$98.5_{0.8}$	$\mathbf{99.5}_{0.1}$	$96.2_{0.9}$	$95.3_{1.0}$	$95.5_{4.0}$	$79.1_{11.9}$	$\mathbf{100.0}_{0.0}$	$95.1_{8.0}$
	Sc.B-multi	$\mathbf{97.5}_{0.2}$	$\mathbf{99.7}_{0.1}$	$\mathbf{99.8}_{0.0}$	$\mathbf{96.4}_{0.6}$	$\mathbf{99.1}_{0.1}$	$\mathbf{98.7}_{0.4}$	$\mathbf{98.7}_{0.6}$	$\mathbf{99.5}_{0.1}$	$96.6_{1.1}$	$\mathbf{96.0}_{0.4}$	$96.5_{2.5}$	$80.2_{11.1}$	$\mathbf{100.0}_{0.0}$	$96.6_{7.4}$
	Sc.B-unc.	$97.4_{0.2}$	$99.6_{0.1}$	$\mathbf{99.8}_{0.1}$	$\mathbf{96.4}_{0.8}$	$\mathbf{99.1}_{0.2}$	$\mathbf{98.7}_{0.3}$	$\mathbf{98.7}_{0.5}$	$\mathbf{99.5}_{0.1}$	$\mathbf{96.7}_{0.8}$	$95.3_{0.6}$	$96.2_{2.9}$	$77.3_{11.6}$	$\mathbf{100.0}_{0.0}$	$96.6_{7.4}$
OOD	TnT	$74.9_{0.1}$	$99.0_{0.1}$	$98.5_{0.1}$	$93.8_{0.2}$	$97.8_{0.1}$	$97.1_{0.2}$	$98.9_{0.2}$	$98.0_{0.1}$	$89.6_{0.8}$	$87.1_{1.1}$	$\mathbf{100.0}_{0.0}$	—	$\mathbf{100.0}_{0.0}$	$\mathbf{100.0}_{0.0}$
	ABL-Tag.	$75.9_{0.4}$	$93.3_{1.6}$	$91.4_{0.4}$	$76.8_{1.1}$	$95.4_{0.2}$	$95.3_{0.4}$	$52.5_{2.9}$	$96.8_{0.2}$	$90.3_{1.4}$	$19.5_{6.3}$	$95.8_{0.7}$	$18.4_{5.3}$	$\mathbf{100.0}_{0.0}$	$90.8_{0.3}$
	Sc.B-sing	$95.8_{0.2}$	$99.1_{0.2}$	$\mathbf{99.7}_{0.1}$	$95.5_{0.4}$	$98.6_{0.1}$	$99.0_{0.1}$	$\mathbf{99.9}_{0.1}$	$99.2_{0.1}$	$\mathbf{98.1}_{0.4}$	$93.8_{0.8}$	$72.2_{3.7}$	$\mathbf{71.1}_{4.2}$	$\mathbf{100.0}_{0.0}$	$95.8_{0.6}$
	Sc.B-multi	$\mathbf{96.2}_{0.2}$	$99.1_{0.1}$	$\mathbf{99.7}_{0.1}$	$\mathbf{95.6}_{0.2}$	$\mathbf{98.8}_{0.2}$	$99.1_{0.1}$	$99.6_{0.3}$	$\mathbf{99.3}_{0.1}$	$97.9_{0.5}$	$\mathbf{94.8}_{0.5}$	$70.7_{10.6}$	$69.6_{3.4}$	$\mathbf{100.0}_{0.0}$	$94.8_{1.0}$
	Sc.B-unc.	$96.0_{0.1}$	$\mathbf{99.2}_{0.1}$	$\mathbf{99.7}_{0.0}$	$95.4_{0.4}$	$98.5_{0.1}$	$\mathbf{99.2}_{0.2}$	$99.6_{0.3}$	$99.2_{0.1}$	$98.0_{0.6}$	$94.6_{1.1}$	$76.1_{7.8}$	$70.6_{3.7}$	$\mathbf{100.0}_{0.0}$	$94.7_{0.9}$

Table 4: Results from training and evaluating on the *Sosialurin-BRAGD* dataset (10-fold cross-validation) and *OOD-BRAGD* corpora. Overall indicates the composite accuracy. Word class columns show F1 scores. Subscripts show standard deviations. Best results in each column are **bolded**.

2020). For TnT, we used the NLTK (Bird et al., 2009) implementation (nltk.tag.tnt.TnT) with its default settings (no backoff), trained on our training splits and evaluated on the held-out test sets. For the Faroese ABLTagger, we refactored the original implementation to use PyTorch.[5]

5.1. Overall Performance

Table 4 shows our models' performance compared to the baseline methods. On the Sosialurin-BRAGD corpus using 10-fold cross-validation, our multi-label ScandiBERT model achieves 97.5% overall accuracy, outperforming the single-label ScandiBERT (97.3%), ABLTagger (89.9%), and TnT (84.9%). On the out-of-domain evaluation set, the multi-label model achieves 96.2% accuracy, demonstrating strong generalization to contemporary text types. The difference between 96.2% and 95.8% for multi-label versus single-label corresponds to a 9.5% (1 - 0.038/0.042) lower error rate. The performance improvement over ABLTagger is even more pronounced on out-of-domain data (96.2% vs 75.9%), suggesting that the multi-label approach learns more robust morphological representations.

We tested significance using paired Wilcoxon signed-rank tests and Nadeau-Bengio corrected paired t-tests with Holm-Bonferroni correction (See Appendix D). The constrained multi-label model significantly outperforms single-label on both in-domain and OOD data ($p = 0.010$, Cohen's $d > 0.9$), and significantly outperforms the unconstrained variant on OOD data ($p = 0.010$). All ScandiBERT fine-tuned models significantly outperform ABLTagger and the TnT-tagger (NB $p < 0.001$ and Wilcoxon $p = 0.010$).

5.2. Per-Class Performance

Table 6 shows per-word-class performance. The model achieves high performance across most word classes. On Sosialurin, the highest F1 scores are for Verbs (99.8%), Noun (99.7%), Conjunction (99.5%), and Pronoun (98.7%). The most challenging word class is Foreign words (80.2%), though this rare class shows high variance across folds due to limited training examples. On the out-of-domain corpus, performance remains strong with Verbs (99.7%), Number (99.6%), Conjunction (99.3%), and Noun (99.1%) achieving the highest scores. Foreign words drop to 69.6% F1 on OOD data, and Abbreviations show a notable domain gap (96.5% ID vs 70.7% OOD). Comparing single-label and multi-label approaches, the multi-label model shows improvements across most word classes, with the largest accuracy gains on Participle (95.7% vs 94.5%) and Number (98.4% vs 97.6% on Sosialurin-BRAGD).

5.3. Impact of Rare Classes on Macro F1

Three word classes (Unanalyzed word, Web address, and Symbol) have extremely limited representation in the corpus. Unanalyzed word and Web address together account for only 8 tokens (0.007% of the corpus), while Symbol has 35 tokens (0.03%). These classes show high variance in performance across folds due to insufficient training data, though the multi-label model handles them better than traditional approaches on average.

5.4. Morphological Feature Performance

Table 5 presents detailed per-label accuracy and F1 scores. Analysis of morphological feature performance reveals that most errors occur in morphologically ambiguous contexts or with rare word classes. Rare word classes such as Abbreviations and Foreign words remain challenging for all models, though the multi-label approach shows more stable cross-domain performance. These error patterns are consistent with known challenges in Faroese morphology and suggest that the model has learned linguistically meaningful representations rather than memorizing surface patterns.

[5]The original implementation at https://github.com/hinrikur/far-ABLTagger/ used an older neural-network training library, DyNet (Neubig et al., 2017), which we did not manage to run.

Label	In-domain						OOD					
	Single		Multi		Unc.		Single		Multi		Unc.	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Word Class	**85.4**$_{4.2}$	96.3$_{1.1}$	84.1$_{3.2}$	**96.8**$_{1.1}$	83.3$_{1.5}$	96.4$_{1.1}$	98.8$_{0.1}$	94.1$_{0.5}$	**98.8**$_{0.1}$	93.8$_{0.7}$	98.8$_{0.0}$	**94.4**$_{0.6}$
Subcategories	98.7$_{0.1}$	90.8$_{1.0}$	**98.7**$_{0.2}$	**97.4**$_{1.1}$	98.7$_{0.2}$	97.0$_{1.0}$	98.7$_{0.1}$	90.2$_{1.2}$	**98.8**$_{0.1}$	**96.0**$_{2.2}$	98.7$_{0.2}$	95.2$_{1.9}$
Gender	97.9$_{0.3}$	78.6$_{0.2}$	**98.1**$_{0.4}$	**98.2**$_{0.3}$	98.1$_{0.3}$	98.1$_{0.3}$	96.0$_{0.3}$	77.1$_{0.2}$	**96.3**$_{0.2}$	**96.4**$_{0.2}$	96.1$_{0.3}$	96.2$_{0.3}$
Number	98.9$_{0.2}$	74.2$_{0.1}$	**99.1**$_{0.1}$	**98.9**$_{0.2}$	99.1$_{0.1}$	98.9$_{0.2}$	98.3$_{0.2}$	73.7$_{0.1}$	**98.7**$_{0.1}$	**98.3**$_{0.2}$	98.6$_{0.2}$	98.3$_{0.3}$
Case	98.0$_{0.3}$	81.6$_{0.4}$	**98.2**$_{0.2}$	**98.0**$_{0.3}$	98.2$_{0.3}$	98.0$_{0.5}$	97.2$_{0.2}$	80.4$_{0.3}$	**97.6**$_{0.3}$	**97.2**$_{0.4}$	97.6$_{0.2}$	97.2$_{0.4}$
Article	99.4$_{0.2}$	66.4$_{0.1}$	**99.5**$_{0.1}$	**99.5**$_{0.1}$	99.5$_{0.1}$	99.4$_{0.1}$	98.7$_{0.2}$	66.1$_{0.1}$	98.9$_{0.2}$	98.7$_{0.3}$	**99.0**$_{0.2}$	**98.9**$_{0.2}$
Proper Noun	99.3$_{0.1}$	66.3$_{0.1}$	**99.5**$_{0.2}$	**99.3**$_{0.2}$	99.4$_{0.2}$	99.2$_{0.2}$	97.1$_{0.2}$	64.5$_{0.2}$	**97.7**$_{0.1}$	97.0$_{0.1}$	**97.7**$_{0.2}$	**97.0**$_{0.2}$
Degree	98.4$_{0.3}$	71.9$_{7.6}$	**99.0**$_{0.2}$	**92.0**$_{9.6}$	98.8$_{0.3}$	86.2$_{10.2}$	98.0$_{0.3}$	64.8$_{0.3}$	**98.9**$_{0.2}$	**78.7**$_{0.2}$	98.5$_{0.1}$	78.4$_{0.2}$
Declension	96.9$_{0.8}$	73.5$_{0.7}$	99.1$_{0.4}$	98.5$_{0.8}$	**99.2**$_{0.4}$	**98.6**$_{0.8}$	96.5$_{0.4}$	73.8$_{0.2}$	**99.2**$_{0.2}$	**99.0**$_{0.2}$	99.2$_{0.2}$	98.5$_{0.4}$
Mood	99.3$_{0.2}$	75.1$_{6.6}$	**99.3**$_{0.2}$	**93.8**$_{7.6}$	99.2$_{0.2}$	93.5$_{6.9}$	99.4$_{0.1}$	76.6$_{0.8}$	**99.5**$_{0.1}$	**96.2**$_{1.5}$	99.5$_{0.1}$	95.8$_{1.1}$
Voice	99.8$_{0.1}$	64.7$_{5.3}$	99.9$_{0.1}$	96.3$_{10.6}$	**99.9**$_{0.1}$	**96.4**$_{10.6}$	99.7$_{0.1}$	66.6$_{0.1}$	**100.0**$_{0.0}$	**99.9**$_{0.2}$	100.0$_{0.0}$	99.7$_{0.3}$
Tense	99.3$_{0.2}$	74.6$_{0.1}$	**99.4**$_{0.2}$	**99.4**$_{0.2}$	99.3$_{0.2}$	99.3$_{0.2}$	99.4$_{0.1}$	74.6$_{0.1}$	**99.5**$_{0.1}$	**99.5**$_{0.1}$	99.4$_{0.1}$	99.4$_{0.1}$
Person	**99.0**$_{0.1}$	78.6$_{0.9}$	98.9$_{0.1}$	97.5$_{0.9}$	98.9$_{0.2}$	**97.6**$_{0.9}$	99.0$_{0.2}$	78.5$_{0.5}$	**99.2**$_{0.1}$	**98.5**$_{0.2}$	99.0$_{0.1}$	98.1$_{0.3}$
Definiteness	97.2$_{1.0}$	65.5$_{0.5}$	**99.0**$_{0.5}$	**98.9**$_{0.6}$	98.9$_{0.5}$	98.7$_{0.6}$	98.2$_{0.8}$	66.0$_{0.3}$	99.3$_{0.5}$	99.3$_{0.5}$	**99.4**$_{0.5}$	**99.4**$_{0.5}$

Table 5: Per-label accuracy and F1 scores comparing single-label, multi-label (constrained), and unconstrained ScandiBERT models on in-domain (Sosialurin-BRAGD, 10-fold CV) and out-of-domain (OOD-BRAGD) datasets. Values show mean$_{\text{std}}$ across folds. We ignore the *unknown word* and *web* tags in the word-type macro F1 results, as these have only a handful of labels.

Word Class	In-domain						OOD					
	Single		Multi		Unc.		Single		Multi		Unc.	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Noun	99.7$_{0.1}$	99.6$_{0.1}$	**99.7**$_{0.1}$	**99.7**$_{0.1}$	99.7$_{0.1}$	99.6$_{0.1}$	99.0$_{0.2}$	99.1$_{0.2}$	98.9$_{0.2}$	99.1$_{0.1}$	**99.2**$_{0.2}$	**99.2**$_{0.1}$
Adjective	96.3$_{1.1}$	96.1$_{0.5}$	**96.4**$_{0.8}$	**96.4**$_{0.6}$	96.0$_{1.1}$	96.4$_{0.8}$	**95.6**$_{0.5}$	95.5$_{0.4}$	94.8$_{0.6}$	**95.6**$_{0.2}$	94.4$_{0.7}$	95.4$_{0.4}$
Pronoun	**98.7**$_{0.3}$	98.6$_{0.3}$	98.6$_{0.5}$	**98.7**$_{0.4}$	98.7$_{0.3}$	98.7$_{0.3}$	99.1$_{0.3}$	99.0$_{0.1}$	**99.2**$_{0.1}$	99.1$_{0.1}$	99.1$_{0.1}$	**99.2**$_{0.2}$
Number	97.6$_{1.2}$	98.5$_{0.8}$	98.4$_{0.6}$	98.7$_{0.6}$	**98.4**$_{0.8}$	**98.7**$_{0.5}$	**100.0**$_{0.1}$	**99.9**$_{0.1}$	99.8$_{0.2}$	99.6$_{0.3}$	99.6$_{0.3}$	99.6$_{0.3}$
Verbs	**99.8**$_{0.1}$	99.8$_{0.1}$	99.8$_{0.1}$	**99.8**$_{0.0}$	99.8$_{0.1}$	99.8$_{0.1}$	99.7$_{0.1}$	99.7$_{0.1}$	**99.8**$_{0.1}$	**99.7**$_{0.1}$	99.7$_{0.1}$	99.7$_{0.0}$
Participle	94.5$_{1.6}$	95.3$_{1.0}$	95.7$_{1.1}$	**96.0**$_{0.4}$	**96.0**$_{1.6}$	95.3$_{0.6}$	91.3$_{1.4}$	93.8$_{0.8}$	95.0$_{1.0}$	**94.8**$_{0.5}$	**95.4**$_{2.1}$	94.6$_{1.1}$
Adverb	98.9$_{0.3}$	99.1$_{0.1}$	**99.1**$_{0.2}$	**99.1**$_{0.1}$	99.0$_{0.4}$	99.1$_{0.2}$	98.9$_{0.3}$	98.6$_{0.1}$	**99.1**$_{0.2}$	**98.8**$_{0.2}$	98.8$_{0.3}$	98.5$_{0.1}$
Conjunctions	99.6$_{0.1}$	99.5$_{0.1}$	**99.6**$_{0.1}$	**99.5**$_{0.1}$	99.6$_{0.2}$	99.5$_{0.1}$	99.2$_{0.1}$	99.2$_{0.1}$	99.2$_{0.1}$	**99.3**$_{0.1}$	**99.3**$_{0.2}$	99.2$_{0.1}$
Foreign words	**79.4**$_{16.9}$	79.1$_{11.9}$	76.2$_{17.4}$	**80.2**$_{11.1}$	75.1$_{17.7}$	77.3$_{11.6}$	**63.5**$_{3.4}$	**71.1**$_{4.2}$	**63.5**$_{3.4}$	69.6$_{3.4}$	60.5$_{2.8}$	70.6$_{3.7}$
Abbreviation	95.3$_{5.8}$	95.5$_{4.0}$	**96.7**$_{3.9}$	**96.5**$_{2.5}$	95.4$_{4.6}$	96.2$_{2.9}$	**100.0**$_{0.0}$	72.2$_{3.7}$	97.7$_{7.3}$	70.7$_{10.6}$	99.2$_{2.4}$	**76.1**$_{7.8}$
Punctuation	100.0$_{0.0}$	100.0$_{0.0}$	**100.0**$_{0.0}$	**100.0**$_{0.0}$	100.0$_{0.0}$	100.0$_{0.0}$	**100.0**$_{0.0}$	**100.0**$_{0.0}$	**100.0**$_{0.0}$	**100.0**$_{0.0}$	**100.0**$_{0.0}$	**100.0**$_{0.0}$
Symbol	**94.2**$_{12.5}$	95.1$_{8.0}$	**94.2**$_{12.5}$	**96.6**$_{7.4}$	**94.2**$_{12.5}$	**96.6**$_{7.4}$	**92.3**$_{0.0}$	**95.8**$_{0.6}$	91.9$_{1.2}$	94.8$_{1.0}$	**92.3**$_{0.0}$	94.7$_{0.9}$
R Article	97.2$_{1.0}$	96.2$_{0.9}$	97.7$_{0.9}$	96.6$_{1.1}$	**97.9**$_{1.2}$	**96.7**$_{0.8}$	98.2$_{0.8}$	**98.1**$_{0.4}$	98.4$_{0.8}$	97.9$_{0.5}$	**98.6**$_{1.0}$	98.0$_{0.6}$

Table 6: Per-word-class accuracy and F1 scores comparing single-label, multi-label (constrained), and unconstrained ScandiBERT models on in-domain (Sosialurin-BRAGD, 10-fold CV) and out-of-domain (OOD-BRAGD) datasets. Values show mean$_{\text{std}}$ across folds.

5.5. Comparison with Baselines

Our multi-label ScandiBERT model achieves 97.5% accuracy on Sosialurin-BRAGD, a 7.6 percentage point improvement over ABLTagger (89.9%), representing a 75.2% error reduction (1 - 0.025/0.101). On out-of-domain data, the improvement is even more substantial: 96.2% vs 75.9%, a 20.3 percentage point gain and 84.2% error reduction (1 - 0.038/0.241). This substantial gain can be attributed to several complementary factors: ScandiBERT's multilingual pre-training provides robust representations for Faroese, while the multi-label framework enables joint optimization across morphological features rather than treating them as independent atomic units. The constrained loss function enforces linguistic validity, and our systematic corpus revisions create more consistent training data. The multi-label model shows only a 1.3 pp drop from ID (97.5%) to OOD (96.2%), while ABLTagger drops 14.0 pp (89.9% to 75.9%), demonstrating superior generalization. While we build on the valuable corpus work of Hafsteinsson (2020) and Hansen et al. (2004), our results show that methods designed specifically for Faroese morphology, leveraging multilingual pre-training and constrained multi-label classification, substantially outperform approaches adapted from Icelandic.

6. Discussion

Our results demonstrate that neural multi-label approaches can achieve high accuracy for morphologically complex low-resource languages when combined with appropriate architectural choices and training procedures. The training corpus is relatively small (117,322 words, compared to approximately 590,000 tokens for the Icelandic IFD corpus (Pind et al., 1991; Loftsson et al., 2009) and 500,000 tokens for Swedish (Megyesi, 2001)), yet the model achieves strong performance. This is likely due to the effectiveness

of the ScandiBERT pre-training, which leverages data from related Scandinavian languages. The cross-lingual transfer from Icelandic, Danish, Norwegian, and Swedish provides robust representations that require less Faroese-specific fine-tuning data. The substantial performance gap for ABLTagger between in-domain (89.9%) and out-of-domain (75.9%) data contrasts with our multi-label model's modest 1.3 pp gap (97.5% ID, 96.2% OOD). ABLTagger drops 14.0 pp while our model drops only 1.3 pp. This suggests that single-label approaches overfit to domain-specific tag distributions, while multi-label learning captures more generalizable morphological patterns.

Error analysis. Manual evaluation reveals four systematic error patterns. **Verb-preposition compound ambiguity:** the model confuses lexicalized compounds (e.g., *tilkallaður* from *at tilkalla*) with separable verb phrases (e.g., *framkomin* from *koma fram*), as these forms are morphologically identical. **Case assignment:** the accusative-dative distinction is challenging with prepositions like *í*, which governs accusative for motion (*í húsið*, "into the house") but dative for location (*í húsinum*, "in the house"). **Gender in compounds:** the model must identify the rightmost noun to assign gender (e.g., *leigubilur*, masc., from *bilur*), but struggles when head nouns are rare or obscured by subword tokenization. **Rare word classes:** Abbreviations (96.5% $\rightarrow$ 70.7%) and Foreign words (80.2% $\rightarrow$ 69.6%) drop substantially across domains, while major classes remain stable (Verb F1: 99.8% $\rightarrow$ 99.7%; Noun F1: 99.7% $\rightarrow$ 99.1%).

Transfer, Inheritance and Representation The systematic corrections detailed in Section 3 reveal that the baseline ABLTagger with a more Icelandic-based tagset does more than just underperform; it applies a grammatical framework that often overlooks native Faroese distinctions, such as the mediopassive voice and specific syntactic functions. This inherited framing reflects a long-standing academic relationship, rooted in 19th-century orthographic standards (Petersen et al., 2025). The 84.2% reduction in out-of-domain error demonstrates that these shared foundations can lead to a systematic misrepresentation of Faroese morphology in modern NLP. For small language communities, this development has profound consequences for linguistic autonomy. A tool should not merely "perform" well on standardized tests; it must accurately reflect the categories actually present in the language to remain useful for linguists, teachers, and native speakers. Without community-driven, linguistically informed oversight, cross-lingual transfer risks "normalizing" the categories of the source language in future training data. Our results suggest that while transfer is a valuable and practical starting point, the ultimate goal should be to build tools that serve as accurate mirrors of a language's unique identity rather than transferring external grammatical norms, however convenient.

6.1. Future Directions

This work opens several directions for Faroese language technology and morphological analysis. Further refinement of the tagset would ensure full alignment with Faroese grammar. The multi-label framework naturally extends to **related morphological tasks** such as lemmatization and dependency parsing by adding output dimensions for lemma prediction or syntactic relations. The shared morphological representations should transfer effectively to these tasks. Given the high cost of expert annotation for low-resource languages, **active learning** could reduce annotation requirements. Our constrained loss function provides a natural uncertainty metric: tokens where the predicted word class yields conflicting feature predictions are likely annotation candidates. Future work could investigate **robustness to dialectal variation**, including orthographic differences and regional varieties within Faroese, particularly for social media text, supporting broader documentation of Faroese language use.

7. Conclusion

We present the first neural multi-label POS tagger for Faroese, achieving 97.5% accuracy on in-domain data and 96.2% on out-of-domain texts—an 84.2% reduction in error rate compared to ABLTagger. By decomposing 651 compound tags into 73 binary dimensions and enforcing linguistic validity through a constrained loss function, we address the data sparsity inherent in morphologically rich low-resource languages while ensuring that predictions respect Faroese grammar. The significant advantage of the constrained model over its unconstrained counterpart on out-of-domain data ($p = 0.010$) confirms that these linguistically motivated constraints provide genuine generalization benefits, not just theoretical appeal.

Beyond the tagger, we contribute the BRAGD tagset, the revised Sosialurin-BRAGD corpus and OOD corpus—an effort driven by native speakers to move Faroese morphological analysis away from inherited Icelandic conventions and toward categories that reflect the language on its own terms. For low-resource languages, our work demonstrates that combining multilingual pre-training with language-specific structural constraints can yield robust taggers even with limited annotated data, provided the linguistic framework is developed in close collaboration with the language community.

8. Limitations

Several important limitations should be considered when interpreting our results and applying this work:

Lack of Ablation Studies. We do not provide systematic ablation studies isolating the contribution of each component (ScandiBERT pre-training, multi-label formulation, constrained loss, corpus improvements). While the combined system achieves strong performance, quantifying individual contributions would guide future work on other low-resource languages. The similar convergence speed of multi-label and single-label (11–12 epochs) suggests that the accuracy gains come from the multi-label formulation itself rather than faster convergence. Our unconstrained multi-label variant partially isolates the effect of the constrained loss, showing that it contributes significantly to out-of-domain performance ($p = 0.010$, Wilcoxon corrected), but further ablations separating the effects of corpus improvements from architectural choices would provide more complete evidence.

Tagset Incomparability. Our systematic corpus revisions and 73-dimensional tagset improve linguistic validity but complicate direct comparison with Hafsteinsson (2020)'s 371-tag system. While we re-evaluate ABLTagger on our revised corpus, subtle annotation differences may affect performance comparisons. We provide detailed change logs and conversion scripts to enable replication, but perfect comparability with prior work remains challenging.

Temporal and Domain Gaps. Training data from 1990s-2000s newspaper text may not capture contemporary Faroese usage, particularly neologisms, borrowed terms, and evolving grammatical patterns. While out-of-domain evaluation on 2019-2022 BLARK texts provides some evidence of temporal robustness, comprehensive evaluation across decades would better assess generalization. Social media, technical documentation, and other genres remain unrepresented.

Rare Category Performance. Word classes with fewer than 50 training examples (unanalyzed words, web addresses, symbols) show unreliable performance due to insufficient data. The constrained loss function helps but cannot overcome fundamental data sparsity. Applications requiring high accuracy on these categories would need targeted data collection or alternative approaches.

Tagset Completeness. The BRAGD tagset does not yet fully align with Faroese grammar. For example, prepositions are still grouped in the Adverb category, a classification inherited from Hafsteinsson (2020) that is not congruent with Faroese grammatical conventions (see Appendix B). This and other remaining gaps will be addressed in future versions of the tagset.

9. Ethical Statement

Language Preservation and Revitalization. This work directly supports the preservation and promotion of Faroese, a low-resource language spoken by approximately 50,000 people. Accurate morphological analysis is foundational infrastructure for language technology, enabling applications in education, translation, and corpus linguistics that support language vitality. We emphasize that technology alone cannot preserve a language, but appropriate language technology can empower speakersand facilitate language transmission.

Resource Accessibility. We commit to releasing our improved Sosialurin-BRAGD corpus, trained models, and evaluation code under open licenses to maximize accessibility for researchers and the Faroese language community.

Annotation Labor. Three native Faroese speakers contributed substantial expert knowledge to manual evaluation and corpus revision. All annotators were compensated fairly for their time and expertise.

Evaluation Limitations and Error Impact. While 97.5% accuracy represents substantial progress, 2.5% error rate means approximately 1 in 40 words receives incorrect morphological analysis. For high-stakes applications (e.g., educational materials, official documentation), human review remains essential. We caution against deploying this technology in contexts where errors could have serious consequences for speakers or learners.

10. Acknowledgements

The authors would like to thank the creators of the original *Sosialurin* corpus, as well as Zakaris Svabo Hansen for providing historical context on his foundational work. Special thanks also go to Hinrik Hafsteinsson for his essential work in updating the corpus and its accompanying tagset. We sincerely thank Haukur Barri Símonarson who helped us conceptualize this work.

AS is supported by the European Commission under grant agreement no. 101135671. VS is supported by the Pioneer Centre for AI, DNRF grant number P1. This work is additionally supported by NordPlus funding for the Faroese Megaword

Corpus project (NPLA-2023/10060). We thank the reviewers for their constructive comments, which significantly improved the quality and clarity of this manuscript.

11. References

Yvonne Adesam and Aleksandrs Berdicevskis. 2021. Part-of-speech tagging of Swedish texts in the neural era. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 200–209, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.

Eibe Bouckaert, Remco R.and Frank. 2004. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. In *Advances in Knowledge Discovery and Data Mining*, pages 3–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Thorsten Brants. 2000. TnT – A Statistical Part-of-Speech Tagger. In *Sixth Applied Natural Language Processing Conference*, pages 224–231, Seattle, Washington, USA. Association for Computational Linguistics.

Grzegorz Chrupała, Georgiana Dinu, and Josef van Genabith. 2008. Learning Morphology with Morfette. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual Character-Level Neural Morphological Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7(1):1–30.

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.

Hinrik Hafsteinsson. 2020. A Faroese part-of-speech tagger built with Icelandic methods. Data preparation, training and evaluation. Master's thesis, The University of Iceland, September.

Hinrik Hafsteinsson and Anton Karl Ingason. 2020. Developing a Faroese PoS-tagging solution using Icelandic methods. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 481–490, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Zakaris Svabo Hansen, Heini Justinussen, and Mortan Ólason. 2004. Marking av teldutøkum tekstsavni [Tagging of a digital text corpus].

Per E. Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for Norwegian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 30–40, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 282–289. Morgan Kaufmann.

Hrafn Loftsson, Ida Kramarczyk, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2009. Improving the PoS tagging accuracy of Icelandic text. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages

103–110, Odense, Denmark. Northern European Association for Language Technology (NEALT).

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with Words at the National Library of Sweden – Making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

Beáta Megyesi. 2001. Comparing Data-Driven Learning Algorithms for PoS Tagging of Swedish. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Thomas Müller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Thomas Müller and Hinrich Schuetze. 2015. Robust Morphological Tagging with Word Representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 526–536, Denver, Colorado. Association for Computational Linguistics.

Claude Nadeau and Yoshua Bengio. 1999. Inference for the Generalization Error. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. DyNet: The Dynamic Neural Network Toolkit.

Robert Östling and Jörg Tiedemann. 2017. Continuous multilinguality with language vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.

Hjalmar P. Petersen, Christer Lindqvist, Jógvan í Lon Jacobsen, Zakaris Svabo Hansen, and Sissal M. Rasmussen, editors. 2025. *Føroysk Málsøga 1*. Nám, Tórshavn.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörgen Pind, Friðrik Magnússon, and Stefán Briem. 1991. *Íslensk orðtíðnibók [The Icelandic Frequency Dictionary]*. The Institute of Lexicography, University of Iceland, Reykjavík, Iceland.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Miikka Silfverberg and Senka Drobac. 2018. Sublabel dependencies for Neural Morphological Tagging – The Joint Submission of University of Colorado and University of Helsinki for VarDial 2018. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 37–45, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Annika Simonsen, Sandra Saxov Lamhauge, Iben Nyholm Debess, and Peter Juel Henrichsen. 2022. Creating a Basic Language Resource Kit for Faroese. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4637–4643, Marseille, France. European Language Resources Association.

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.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a Low-Resource Language via Close Relatives: The Case Study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

Statistics Faroe Islands. 2025. Population. Statistics Faroe Islands. Accessed December 2024.

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger

with a morphological lexicon and a lexical category identification step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1161–1168, Varna, Bulgaria. INCOMA Ltd.

Francis Tyers, Mariya Sheyanova, Aleksandra Martynova, Pavel Stepachev, and Konstantin Vinogorodskiy. 2018. Multi-source synthetic treebank creation for improved cross-lingual dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 144–150, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

12. Language Resource References

Paulivar Andreasen and Árni Dahl. 1997. *Mállæra*. Føroya Skúlabókagrunnur, Tórshavn.

Certainly AI. 2020. *Nordic BERT Models*. Danish, Norwegian, and Swedish BERT models trained from scratch.

Jóhan Hendrik Winther Poulsen, Marjun Simonsen, Jógvan í Lon Jacobsen, Anfinnur Johansen, Zakaris Svabo Hansen, and Ragnar Sigrunarson. 1998. *Føroysk orðabók*. Orðabókagrunnurin.

Rúnarsson, Kristján and Jákupsson, Heðin and Hansen, Zakarias Svabo. 2023. *góður*. Føroyski bendingargrunnurin [Faroese Morphological Database]. Accessed: 2025-10-20.

Höskuldur Thráinsson, Hjalmar Páll Petersen, Jógvan í Lon Jacobsen, and Zakaris Svabo Hansen. 2012. *Faroese. An overview and reference grammar*, 3 edition. Faroe University Press/Linguistic Institute of Iceland, Tórshavn/Reykjavík.

A. Appendix: BRAGD tagset

Table 7: The BRAGD tagset used for Sosialurin-BRAGD and OOD-BRAGD

No	Category	Analytical symbol
1	Word class	S-substantive (noun)
2	Gender	M-masculine, F-feminine, N-neuter, g-no gender
3	Number	S-singular, P-plural, n-no number
4	Case	N-nominative, A-accusative, D-dative, G-genitive, c-no case
5	Definiteness	A-definite (with suffixed article), a-indefinite (without suffixed article)
6	Proper noun	P-Proper Noun, r-not proper noun
1	Word class	R-article
2	Gender	M-masculine, F-feminine, N-neuter
3	Number	S-singular, P-plural
4	Case	N-nominative, A-accusative, D-dative, G-genitive
5	Definiteness	I-indefinite, D-definite
1	Word class	A-adjective
2	Degree	P-positive, C-comparative, S-superlative, A-absolute superlative, d-no degree
3	Declension	S-strong, W-weak, e-no declension
4	Gender	M-masculine, F-feminine, N-neuter, g-no gender
5	Number	S-singular, P-plural, n-no number
6	Case	N-nominative, A-accusative, D-dative, G-genitive, c-no case
1	Word class	P-pronoun
2	Subcategory	D-demonstrative, E-possessive, I-indefinite, P-personal, Q-interrogative, X-reflexive
3	Gender	M-masculine, F-feminine, N-neuter, g-no gender
4	Person	1-1st pers, 2-2nd pers, 3-3rd pers, p-no person
5	Number	S-singular, P-plural, n-no number
6	Case	N-nominative, A-accusative, D-dative, G-genitive, c-no case
1	Word class	N-numeral
2	Category	C-cardinal number, D-Date/indeclinable number, R-number preceding numeral
3	Gender	M-masculine, F-feminine, N-neuter, g-no gender
4	Number	S-singular, P-plural, n-no number
5	Case	N-nominative, A-accusative, D-dative, G-genitive, c-no case
1	Word class	V-verb (except for past participle)
2	Mood	I-infinitive, M-imperative, N-indicative, S-subjunctive, U-supine
3	Voice	A-active, M-mediopassive, v-no voice
4	Tense	P-present, A-past, t-no tense
5	Number	S-singular, P-plural, n-no number
6	Person	1-1st person, 2-2nd person, 3-3rd person, p-no person
1	Word class	L-past participle
2	Declension	S-strong, W-weak
3	Gender	M-masculine, F-feminine, N-neuter
4	Number	S-singular, P-plural
5	Case	N-nominative, A-accusative, D-dative, G-genitive
1	Word class	D-adverb
2	Category	N-does not govern case(adverb), G-governs case(preposition), I-interjection/exclamation
3	Degree	C-comparative, S-superlative, A-absolute superlative, d-no degree
1	Word class	C-conjunction
2	Category	C-coordinative, S-subordinative, I-infinitive (only "at" infinitive), R-relative
1	Word class	F-foreign word
1	Word class	X-unanalysed word
1	Word class	T-abbreviation
2	Category	S-acronym/abbreviation, T-short form (one word)
1	Word class	W-e-mail, web address
1	Word class	K-Punctuation
2	Category	E-end of sentence, C-comma, Q-quotes, O-other
1	Word class	M-symbol

B. Appendix: Tagset Evolution

Table 8 traces the evolution of the Faroese POS tagset across three iterations: the original tagset by Hansen et al. (2004), the revised tagset by Hafsteinsson (2020), and our BRAGD tagset. The most significant structural changes in BRAGD are: (1) the introduction of explicit no-feature markers (lowercase letters such as g, n, c) that enable fixed-dimensional multi-label classification; (2) the separation of the past participle into its own word class (L), reflecting its distinct adjectival morphology; (3) the promotion of the article to a full word class (R) with inflectional features; (4) adding strong/weak declension to past participles; (5) the extraction of Voice from verb Mood into an independent feature dimension; and (6) adding the supine to verbs (see Section 3.1 for justification).

Table 8: Evolution of the Faroese POS tagset across three iterations. Cells highlighted in green denote additions, blue denotes modifications, and ~~red~~ denotes removals. The rightmost column shows the current BRAGD tagset.

Word class	Feature	Hansen et al. (2004)	Hafsteinsson (2020)	BRAGD
S Noun	Gender	M, F, N, X	M, F, N, X	M, F, N, g
	Number	S, P	S, P	S, P, n
	Case	N, A, D, G	N, A, D, G	N, A, D, G, c
	Article	A	A	A, a
	Proper	P, L	P	P, r
BRAGD adds explicit no-feature markers (g, n, c, a, r) to enable multi-label classification where every feature dimension has a value. X (unspecified gender) is replaced by g (no gender) for consistency with the marker convention. L (place name) was merged into P (proper noun) in Hafsteinsson (2020).				
R Article	Gender	—	—	M, F, N
	Number	—	—	S, P
	Case	—	—	N, A, D, G
	Definite	—	—	I, D
The free-standing article is introduced as its own word class in BRAGD, with full inflectional features. In earlier tagsets, articles were not tagged separately.				
A Adj.	Degree	P, C, S	P, C, S	P, C, S, A, d
	Declen.	S, W, I	S, W, I	S, W, e
	Gender	M, F, N	M, F, N	M, F, N, g
	Number	S, P	S, P	S, P, n
	Case	N, A, D, G	N, A, D, G	N, A, D, G, c
A (absolute superlative) and d (no degree) are added to Degree. I (indeclinable) is renamed to e (no declension) for consistency with the lowercase no-feature convention.				
P Pron.	Subcat.	D, I	D, B, E, I, P, Q, R	D, E, I, P, Q, X
	Gender	M, F, N, 1, 2	M, F, N, 1, 2	M, F, N, g
	Person	(combined)	(combined)	1, 2, 3, p
	Number	S, P	S, P	S, P, n
	Case	N, A, D, G	N, A, D, G	N, A, D, G, c
Hafsteinsson (2020) greatly expanded the subcategories. BRAGD further refines these: B (indefinite demonstrative) is merged into I (indefinite), R (relative) is dropped as a pronoun subtype, and X (reflexive) is added. Gender and Person, previously combined into one positional column, are split into separate feature dimensions to enable independent prediction. Third person (3) is added explicitly.				
N Num.	Subcat.	C, O	C, D, P, R	C, D, R
	Gender	M, F, N	M, F, N	M, F, N, g
	Number	S, P	S, P	S, P, n
	Case	N, A, D, G	N, A, D, G	N, A, D, G, c
O (ordinal) was replaced by D (date/indeclinable) and new subtypes P (percentage) and R (preceding numeral) in Hafsteinsson (2020). BRAGD drops P (percentage), which was rare and better handled as other numerals.				
V Verb	Mood	I, M, N, S, P, E	I, M, N, S, P, E	I, M, N, S, U

continued on next page

Table 8 continued

Word class	Feature	Hansen et al. (2004)	Hafsteinsson (2020)	BRAGD
	Voice	—	—	A, M, v
	Tense	P, A	P, A	P, A, t
	Number	S, P	S, P	S, P, n
	Person	1, 2, 3	1, 2, 3	1, 2, 3, p
Major restructuring: P (present participle) is removed from Mood as it is reclassified. E (medium/mediopassive) is extracted into a new Voice feature (M), adding both A (active), M (mediopassive), and v (no voice) in the new voice feature dimension. U (supine) is added as a mood value.				
	Mood	A	A	—
	Declen.	—	—	S, W
L/**V** Partic.	Gender	M, F, N	M, F, N	M, F, N
	Number	S, P	S, P	S, P
	Case	N, A, D, G	N, A, D, G	N, A, D, G
The past participle is promoted to its own word class L. In earlier tagsets, participles were tagged as verbs with mood = A and then took adjectival features (gender/number/case). Separating them simplifies the verb schema and reflects their distinct morphological behavior (adjectival agreement).				
D Adv.	Subcat.	~~N, A, D, G~~	N, G, I	N, G, I
	Degree	C, S	C, S	C, S, A, d
In Hansen et al. (2004), the second feature was "case governor" with four case values. Hafsteinsson (2020) simplified this to three categories: N (no case), G (governs case), I (interjection). BRAGD retains this and adds A (absolute superlative) and d (no degree) to the Degree feature. Feature order is swapped: subcategory first, then degree.				
C Conj.	Subcat.	I, R	I, R	C, S, I, R
C (coordinative) and S (subordinative) are added, providing a basic syntactic distinction that was previously unmarked.				
T Abbr.	Subcat.	—	S, T	S, T
Hafsteinsson (2020) added S (acronym) and T (short form) to distinguish abbreviation types.				
K Punc.	Subcat.	—	E, C, Q, O	E, C, Q, O
Punctuation was not a separate word class in Hansen et al. (2004). Added in Hafsteinsson (2020).				
~~**E** Prep.~~	Subcat.	N, A, D, G	—	—
Prepositions were absorbed into the adverb class (D) in Hafsteinsson (2020), with the "governs case" (G) value replacing the separate word class.				
~~**I** Interj.~~	—	(no features)	—	—
Interjections were absorbed into adverbs (D) with subcategory value I in Hafsteinsson (2020).				
Unchanged across all three versions:				
F (Foreign word), **X** (Unanalyzed word), **M** (Symbol), **W** (Web/email address) — these word classes have no morphological features in any version. M and W were added in Hafsteinsson (2020) and remain unchanged in BRAGD.				

C. Appendix: Feature-level Tag Changes

To show which morphological features were most frequently revised, we compute feature-level change counts from a token-by-token comparison between Hafsteinsson (2020)'s revised corpus and Sosialurin-BRAGD, using the same alignment as in Table 3. We count only tokens whose tags differ in a meaningful way between the two versions, excluding cases that are equivalent under predictable tagset-format conversion. We further restrict the analysis to tokens whose main category remains the same, so that the differences can be interpreted as within-category feature changes rather than changes in word class. Table 9 reports the resulting frequencies by feature dimension.

D. Appendix: Statistical Testing Methodology

To rigorously compare model performance, we apply two complementary statistical tests to the 10-fold cross-validation results, following recommendations from Demšar (2006) and Bouckaert (2004).

Feature	Count
Conjunction category	7,933
Person	5,043
Pronoun category	2,661
Declension	2,198
Number	1,590
Gender	1,358
Proper	1,293
Case	1,190
Adverb case governance	376
Tense	341
Mood	312
Degree	254
Definiteness	181
Numeral category	45
Voice	17
Abbreviation category	9
Token pairs counted	22,727

Table 9: Counts of feature-level changes in aligned tokens whose main category is unchanged, excluding predictable tagset-format equivalences.

Test Selection. We use **paired Wilcoxon signed-rank tests** (two-sided) as our primary test. This non-parametric test is appropriate because: (1) with only 10 paired observations (one per fold), normality of differences cannot be reliably assessed; (2) it makes no distributional assumptions; and (3) it is robust to outliers. As a secondary test, we use the **Nadeau-Bengio corrected paired t-test** (Nadeau and Bengio, 1999), which accounts for the non-independence of cross-validation folds (folds share training data). The correction adjusts the variance estimate by the factor $(1/k + n_{\text{test}}/n_{\text{train}})$, where $k = 10$ folds and $n_{\text{test}}/n_{\text{train}} = 1/9$ for 10-fold CV, and uses a t-distribution with $k - 1 = 9$ degrees of freedom.

Multiple Comparison Correction. We perform 5 pairwise comparisons per evaluation setting (in-domain and OOD), for a total of 10 tests. We apply the Holm-Bonferroni step-down procedure separately within each evaluation setting to control the family-wise error rate at $\alpha = 0.05$.

Effect Size. We report Cohen's d (mean difference divided by pooled standard deviation) as a standardized effect size measure. Following standard conventions: $|d| < 0.2$ is negligible, $0.2 \leq |d| < 0.5$ is small, $0.5 \leq |d| < 0.8$ is medium, and $|d| \geq 0.8$ is large.

Results. Table 10 reports the pairwise comparisons. The constrained multi-label model significantly outperforms the single-label model on both in-domain ($p = 0.010$, Wilcoxon corrected; $p = 0.021$, NB corrected) and OOD ($p = 0.010$; $p = 0.003$) evaluations, with large effect sizes ($d > 0.9$). The constrained model also significantly outperforms the unconstrained variant on OOD data ($p = 0.010$ Wilcoxon corrected; $d = 1.46$), while the in-domain difference is significant by Wilcoxon ($p = 0.020$) but not by the more conservative NB test ($p = 0.089$). All ScandiBERT models are significantly better than both ABLTagger and TnT ($p < 0.001$ NB corrected; $p = 0.010$ Wilcoxon corrected).

E. Appendix: Reproducibility Details

Hardware and Software. All experiments were conducted on a single NVIDIA RTX 3090 GPU (24 GB VRAM). Software versions: Python 3.10, PyTorch 2.1, Hugging Face Transformers 4.35, ScandiBERT model identifier `vesteinn/ScandiBERT`.

Hyperparameters. Table 11 lists the complete set of hyperparameters used for all ScandiBERT-based models.

Comparison	Δ	Wilcoxon p	W p (corr.)	NB p	NB p (corr.)	Cohen's d
In-domain (Sosialurin-BRAGD, 10-fold CV)						
Multi (constr.) vs Single	$+0.18$	0.002	0.010^*	0.011	0.021^*	0.92
Multi (constr.) vs Unc.	$+0.08$	0.020	0.020^*	0.089	0.089	0.45
Multi (constr.) vs ABLTagger	$+7.55$	0.002	0.010^*	<0.001	$<0.001^*$	28.1
Multi (constr.) vs TnT	$+12.6$	0.002	0.010^*	<0.001	$<0.001^*$	38.7
Single vs ABLTagger	$+7.37$	0.002	0.010^*	<0.001	$<0.001^*$	27.0
Out-of-domain (OOD-BRAGD)						
Multi (constr.) vs Single	$+0.39$	0.002	0.010^*	0.002	0.003^*	2.33
Multi (constr.) vs Unc.	$+0.19$	0.002	0.010^*	0.016	0.016^*	1.46
Multi (constr.) vs ABLTagger	$+20.3$	0.002	0.010^*	<0.001	$<0.001^*$	70.4
Multi (constr.) vs TnT	$+21.3$	0.002	0.010^*	<0.001	$<0.001^*$	181
Single vs ABLTagger	$+19.9$	0.002	0.010^*	<0.001	$<0.001^*$	67.4

Table 10: Pairwise model comparisons using paired Wilcoxon signed-rank tests and Nadeau-Bengio corrected paired t-tests. Δ is the mean accuracy difference in percentage points. Corrected p-values use Holm-Bonferroni correction (5 comparisons per setting). * indicates significance at $\alpha = 0.05$ after correction.

Parameter	Value
Pre-trained model	`vesteinn/ScandiBERT`
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	8 sentences
Max epochs	100
Early stopping patience	3 epochs
Dropout (classification head)	0.1
Gradient clipping	None
Learning rate warmup	None
Weight decay	0
Label dimensions	73

Table 11: Hyperparameters for ScandiBERT-based models (single-label, multi-label constrained, and unconstrained variants). All variants use identical hyperparameters; only the loss function differs.

Data Splits. 10-fold cross-validation splits were created at the sentence level using scikit-learn's `KFold` (n_splits $= 10$, shuffle $=$ True, random_state $= 42$), ensuring reproducible partitioning with 90% of sentences for training and 10% for validation in each fold. The same splits were used for all models.

Training Time. Training time per epoch was approximately 1 minute for single-label models and 2 minutes for multi-label models. Total training time per fold ranged from 10–20 minutes depending on convergence. The full 10-fold evaluation for one model configuration required approximately 2–3 hours.

Evaluation Protocol. *Composite accuracy* requires all label dimensions to match for a token to be counted as correct. For the multi-label model, this means all 73 binary predictions must match the gold labels. Per-word-class macro F1 excludes the *Unanalyzed word* and *Web email or address* classes (8 tokens combined, 0.007% of corpus) from the macro average, as these classes are too rare for meaningful F1 computation.

Model Variants. Three ScandiBERT configurations were evaluated:

- **Single-label**: Standard classification over 651 compound tags.
- **Multi-label (constrained)**: 73-dimensional binary classification with dynamic feature masking based on gold (training) or predicted (inference) word class.
- **Multi-label (unconstrained)**: Same 73-dimensional output but without feature masking; the loss is a simple sum of cross-entropy losses across all 73 dimensions.

F. Scandinavian Language Models

Neural approaches to POS tagging have transformed NLP for Nordic languages (Malmsten et al., 2020; Östling and Tiedemann, 2017; Kummervold et al., 2021; Adesam and Berdicevskis, 2021). Early neural architectures employed bidirectional LSTMs with character-level representations (Plank et al., 2016), while transformer-based models like BERT brought further advances (Vaswani et al., 2017; Devlin et al., 2019). However, massively multilingual models like mBERT underperform on Nordic languages due to insufficient representation (Certainly AI, 2020), motivating language-specific models like NorBERT (Kutuzov et al., 2021) and Swedish BERT (Malmsten et al., 2020).

Most relevant to our work, Snæbjarnarson et al. (2023) developed ScandiBERT specifically for cross-lingual transfer within Scandinavian languages. By training on mixed corpora of Danish, Norwegian, Swedish, Icelandic, and Faroese, ScandiBERT achieved substantial improvements over XLM-RoBERTa on multiple tasks, including POS tagging. Notably, Snæbjarnarson et al. (2023) found that including Faroese in the tokenizer vocabulary slightly reduced downstream performance compared to a variant without Faroese, hypothesizing that the reduced subword overlap with other Scandinavian languages limits transfer. Despite this trade-off, ScandiBERT's focused Scandinavian pretraining makes it a strong foundation for Faroese morphological analysis.