Theoretical Linguistics Rivals Embeddings in Language Clustering for Multilingual Named Entity Recognition

Sakura Imai[1], **Daisuke Kawahara**[1], **Naho Orita**[1], **Hiromune Oda**[2]
[1]Waseda University [2]The University of Tokyo
sakura_imai@toki.waseda.jp, {dkw, orita}@waseda.jp
hiromuneoda@g.ecc.u-tokyo.ac.jp

Abstract

While embedding-based methods have been dominant in language clustering for multilingual tasks, clustering based on linguistic features has not yet been explored much, as it remains baselines (Tan et al., 2019; Shaffer, 2021). This study investigates whether and how theoretical linguistics improves language clustering for multilingual named entity recognition (NER). We propose two types of language groupings: one based on morpho-syntactic features in a nominal domain and one based on a head parameter. Our NER experiments show that the proposed methods largely outperform a state-of-the-art embedding-based model, suggesting that theoretical linguistics plays a significant role in multilingual learning tasks.

1 Introduction

Language clustering has been used to facilitate an effective cross-lingual transfer for low-resource languages in various tasks, such as machine translation (Tan et al., 2019). While the majority of recent clustering approaches depend on embeddings from language models, linguistic knowledge has not yet been exploited enough. Previous studies have merely used descriptive typological features (Oncevay et al., 2020) and a coarse language family classification as baselines (Shaffer, 2021). We argue that there is large room for improvement in language clustering using linguistics knowledge.

This study examines two language classifications based on theoretical linguistics and tests their effectiveness in multilingual NER. Multilingual NER is selected because comparison models are available from Shaffer (2021), namely an embedding-based classification and a language family classification. Although there are datasets available for NER in various languages (Tedeschi et al., 2021; Adelani et al., 2021; Rahimi et al., 2019), our study focuses on Indo-European languages because there is a rich body of research in theoretical linguistics.

Our classification approaches draw on morpho-syntactic parameters proposed primarily in theoretical syntax. The first classification is based on a language tree created by Ceolin et al. (2021), which reflects various morpho-syntactic parameters in a nominal domain. The second classification uses the head parameter (Chomsky, 1981), which indicates the "head" of a phrase in relation to its complements. We select these parameters because NER is a task that identifies mentions and types of named entities that are mostly nouns.

We show that clustering languages based on such parameters results in more effective language groupings beyond the state-of-the-art embedding-based method. Moreover, our clustering approaches demonstrate comparable or better performance than a model trained with all Indo-European languages (hence regardless of a substantial difference in the data size). These results suggest that theoretical linguistics has a promising potential in multilingual NLP tasks.

2 Related Work

In the current age of globalization, collecting information using various languages is getting more important than ever. Multilingual models have gained increasing attention for this purpose. Recently, pre-trained large-scale multilingual models using neural networks, such as Multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2020), have provided competitive results. However, the amount of labeled data available for fine-tuning these multilingual models is highly skewed toward "major" languages. In fact, there are more than 2,000 low-resource languages with little or no labeled data (Joshi et al., 2020).

To alleviate the problem with low-resource languages, cross-lingual transfer learning has been proposed (Artetxe and Schwenk, 2019). The aim of this method is to adapt a language model trained

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics - Student Research Workshop, pages 139–151

with high-resource languages to low-resource languages. Various transfer learning methods have been proposed. For example, Patil et al. (2022) proposed a technique using subword units (byte pair encoding (Sennrich et al., 2016)). Ri and Tsuruoka (2022) investigated which conditions make cross-lingual transfer learning possible by conducting artificial language experiments.

Language clustering is another kind of transfer learning method mainly used in machine translation. Tan et al. (2019) compared clustering by language family and by embeddings and reported that the embedding-based clustering better improved translation accuracy. Oncevay et al. (2020) proposed a language clustering method that integrates syntactic features of WALS (Dryer and Haspelmath, 2013) and embeddings from machine translation models. As for NER, Shaffer (2021) compared clustering by language family and by embeddings and reported that the embedding-based clustering outperformed language family clustering. In sum, clustering by linguistic prior was used as baselines, and these baselines did not attain better results than the ones with embeddings.

Other than language clustering, linguistic knowledge has been widely used in various NLP tasks (O'Horan et al., 2016; Gerz et al., 2018; Ponti et al., 2019). For example, some approaches use typological or phylogenetic features in multilingual fine-tuning for cross-lingual transfer (Lin et al., 2019; Pires et al., 2019; Dhamecha et al., 2021; de Vries et al., 2022). Likewise, language family information or typological features, such as word order, have been used in various kinds of multilingual tasks, such as machine translation (Saleh et al., 2021; Chronopoulou et al., 2022), dependency parsing (Ammar et al., 2016), and pre-training (Fujinuma et al., 2022).

Crucially, however, the linguistic information used in all these studies is limited to the extent of language family and typological features which are directly observable. No studies using more profound linguistic knowledge have been conducted. Therefore, it remains to be seen whether and to what extent linguistic knowledge other than linguistic family and typological features could help improve clustering for multilingual tasks.

3 Language Clustering using Parameters of Theoretical Linguistics

3.1 Linguistic Parameters

As shown in Section 2, multiple studies have attempted to use linguistic priors for multilingual NLP tasks. However, the knowledge used in these studies remains descriptive and unable to represent the internal nature of language.

Thus, we use "linguistic parameters" proposed by Chomsky (1981) in theoretical linguistics for our clustering to capture the characteristics of language that cannot be seen superficially and cannot be captured by phylogenetic comparison of languages. As seen in Sections 3.3 and 3.4, linguistic parameters are morpho-syntactically more detailed and abstract than typological features in WALS that have been used in the previous studies. We apply these parameters to our clustering methods and conduct experiments on multilingual NER.

3.2 Selection of Tasks and Languages

This study selects NER as the target task for comparison with Shaffer's (2021) study, which tried to improve the performance of multilingual NER by clustering languages based on embeddings and language family.

We use 25 languages that belong to the Indo-European language family because there is a sufficient amount of annotated data available for NER, and there is a rich body of literature in theoretical linguistics.

Table 1 lists the languages used in this study. Each language is represented by its ISO 639-1 language code[1], which is summarized in Appendix (Table 10). In the previous study (Shaffer, 2021), sub-families such as Celtic were not used, despite that their NER data are available. To conduct more comprehensive experiments, we select languages from a broader range of sub-families.

3.3 Clustering based on Nominal Parameters

NER is a task that identifies and classifies entities in texts. Since the named entities are mostly represented as noun phrases, clustering languages by features related to a noun phrase would be effective for training. Thus, we focus on morpho-syntactic parameters that capture cross-linguistic similarities and differences in a nominal domain.

[1] http://www.infoterm.info/standardization/iso_639_1_2002.php

Sub-family	Languages	Shaffer (2021)
Romance	ro, fr, es, pt, it, scn	fr, es, it
Germanic	af, nl, de, is, en, da, no, fo	de, en, da
Greek	el	-
Slavic	bg, pl, ru, sl, hr	ru
Indo-Iranian	ps, mr, hi	hi
Celtic	cy, ga	-

Table 1: The languages used in this study and Shaffer (2021).

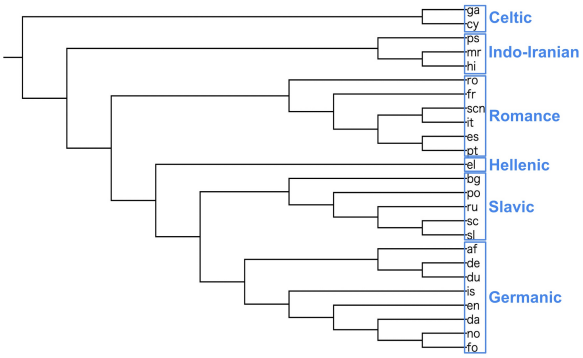

Figure 1: Language tree by Ceolin et al. (2021).

To cluster languages by nominal parameters, we use a language tree proposed by Ceolin et al. (2021). They classified Indo-European languages based on 94 morpho-syntactic parameters in a nominal domain. An example of nominal parameters, "grammaticalized gender" is shown in (1).

(1) a. il libro
the.MASC book.MASC

b. la macchina
the.FEM car.FEM

In languages such as Italian, the gender of definite articles varies depending on the gender of nouns as seen in (1a, 1b).

This parameter is just one example and many other types of parameters are considered in (Ceolin et al., 2021): e.g., the presence/absence of the definite article added to the relative clause and the presence/absence of genitive markings using an adposition. These parameters have often been discussed in theoretical syntax, but many of them are not included in descriptive studies, such as WALS. The relevant language tree is shown in Figure 1, which was created by Ceolin et al. (2021) based on the inter-lingual distances.[2]

To make clusters, we incrementally combine subfamilies close to each other in the language tree. For example, to create 3 clusters, we first combine

[2] https://github.com/AndreaCeolin/Boundaries

#	Sub-family
1	Germanic, Slavic, Hellenic, Romance
2	Indo-Iranian
3	Celtic

Table 2: Clustering by Figure 1 (number of clusters: 3).

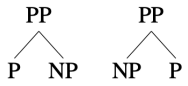

Figure 2: Head-initial (left) and head-final (right) of pre/postpositional phrase (PP).

Germanic and Slavic because they are close to each other in the tree (Figure 1). Hellenic and then Romance are merged into the German-Slavic group. Celtic and Indo-Iranian remain as independent clusters. Table 2 summarizes these 3 clusters. For our experiments, the number of clusters is determined by the elbow method described in Section 4.2.

3.4 Clustering based on the Head Parameter

To identify named entities in text, a language model may use contextual information surrounding the noun phrases. Since a noun phrase is often a part of a verb phrase as an object or a part of an adpositional phrase (i.e., a pre/postpositional phrase) that represents location, clustering languages by this kind of structural information may lead to a more effective clustering.

Based on this hypothesis, the same 25 Indo-European languages are clustered by the head parameter. The head parameter determines where the head (the "core" element) of a phrase is placed in the phrase structure. For example, in the case of a pre/postpositional phrase (PP), if it is head-initial, the head, i.e., the preposition (P), precedes the noun phrase (NP), and vice versa (see Figure 2).

The crucial difference from previous descriptive work such as WALS is that the word order of modifiers (e.g., adverbs for verbs and adjectives for nouns) is irrelevant, but the order of the head (e.g., V in VP) and its complement (e.g., NP for V in VP) is crucial under the head parameter. This is different from the word order classifications in WALS, where the order of the head is no more or less significant than that of modifiers and the notion of head is much less clear. Thus, the head parameter offers a simpler and more abstract framing of word order in a phrase, which crucially focuses on the position of the head and its complement in a phrase. Table 3 shows the classification based on the head

Head Parameter	Sub-Family
Mainly Head-Initial	Romance, Slavic, Germanic, Greek, Celtic
Mainly Head-Final	Indo-Iranian

Table 3: Clustering based on the head parameter (number of clusters: 2).

parameter.

4 NER Experiments

We conduct experiments on NER using the two clustering methods described in Section 3.

4.1 Experimental Setup

There are several datasets available for NER experiments, such as WikiNEuRal (Tedeschi et al., 2021) and MasakhaNER (Adelani et al., 2021). Among them, we select the WikiAnn dataset[3] (Rahimi et al., 2019) because it has an extensive coverage of Indo-European languages, where these languages have been well-documented in theoretical linguistics. The WikiAnn dataset consists of Wikipedia articles for 176 languages that are automatically annotated with three types of named entities: LOC (location), PER (person), and ORG (organization).

An overview of our experiments is shown in Figure 3. First, the training sets of all languages in a cluster are concatenated and fed into a pre-trained language model for fine-tuning. We use XLM-RoBERTa-base[4] (Conneau et al., 2020) as the pre-trained language model. This model has 270M parameters and was trained on 2.5TB of CommonCrawl data in 100 languages. Then, the evaluation set of each language in the cluster is used to evaluate and calculate an F1 score. We perform this evaluation for each cluster using the seqeval framework (Nakayama, 2018) three times and calculate the mean F1 score and standard deviation. For all experiments, we set the batch size to 32, the maximum length of the input to 512, and the learning rate to 5e-5 and conduct three epochs of fine-tuning. We use NVIDIA V100 SXM2 on ABCI[5] as our computing resource, and the average time cost for fine-tuning is approximately one hour.

In our experiments, we select three classifications as baselines. The first is monolingual in which each language is taken as a single cluster. The second is a clustering based on embeddings, and the last is Indo-European all languages (IE-all). Since all the target languages shown in Table 1 are phylogenetically classified into the Indo-European family, using "language family" for clustering corresponds to using a single cluster consisting of all languages in this study.

4.2 Clustering based on Embeddings

We use the embedding-based clustering method proposed by Shaffer (2021) for comparison. An overview of embedding-based clustering is shown in Figure 4.

First, a pre-trained language model is fine-tuned with a language identification task using the WikiAnn training sets. We trained XLM-RoBERTa-base for 3 epochs, setting the batch size to 32, the random seed to 42, and the learning rate to 5e-5. Following Shaffer (2021), we tried a single seed for this preliminary experiment. Language identification is the task of predicting which language the input text is written. We use all 25 languages in Table 1.

Next, each sentence in the WikiAnn validation sets is given to the fine-tuned XLM-RoBERTa model to obtain embeddings from the [CLS] tokens. Based on the obtained embeddings, clustering is performed recursively by agglomerative clustering. We then label the cluster for each input sentence and choose the most frequent cluster for each language among its sentences.

Table 4 shows the resulting clusters using 1,000 and 10,000 samples from the validation set for each language in the WikiAnn dataset. 1,000 and 10,000 are the maximum number of inputs from the validation sets, respectively. For languages that have the validation samples for less than the limits, all samples are used to obtain embeddings.

The optimal number of clusters is determined to be 3 by the elbow method (Thorndike, 1953) when comparing with the clustering method using the nominal parameters described in Section 3 (see Section 5.1 for the experimental results with other numbers of clusters {2, 4, 5}). The elbow method is used to align our embedding-based method with Shaffer's (2021) study, to make a comparison with the clusterings by the nominal parameters. The number of clusters is aligned to 2 to generate clusters when compared with the clustering method using the head parameter.

[3] https://huggingface.co/datasets/wikiann
[4] https://huggingface.co/xlm-roberta-base
[5] https://abci.ai/

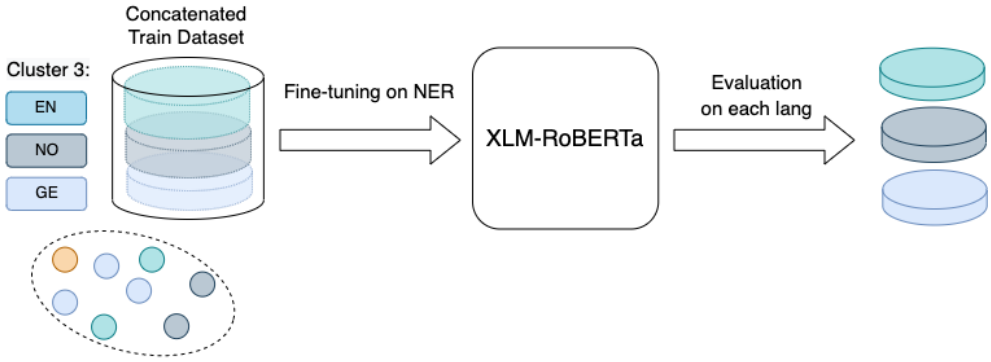

Figure 3: Outline of our experiments on named entity recognition.

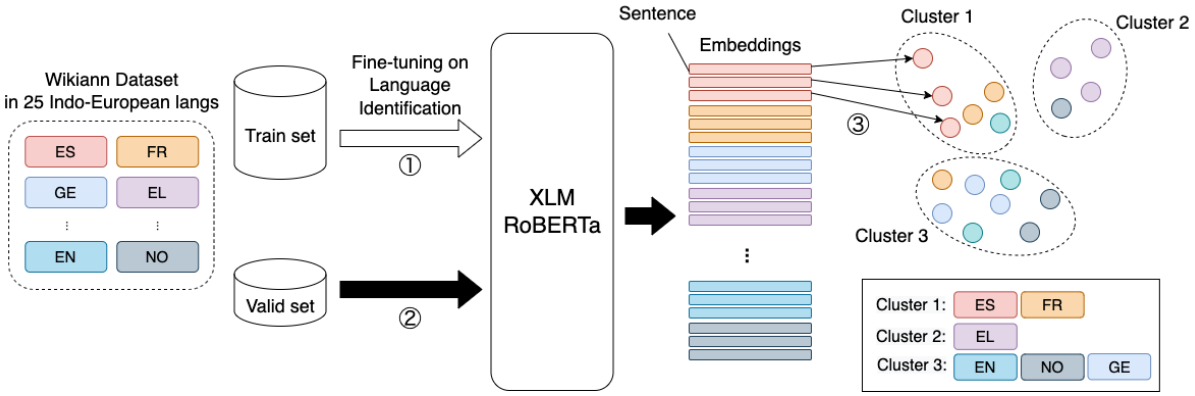

Figure 4: Overview of embedding-based clustering a la Shaffer (2021): The details of this method are described in Section 4.2.

#	Languages	
	1,000 samples	10,000 samples
1	cy, ga, ps, mr, hi, ro, fr, bg, pl, ru, sl, hr, af, nl, de, is, en, da, no, fo	ga, ro, fr, es, pt, it, scn, pl, sl, hr, de, en
2	es, pt, it, scn	mr, hi, ru, af, nl, is, da, no, fo
3	el	cy, ps, el, bg

Table 4: Embedding-based clustering results when using 1,000 and 10,000 samples from validation sets from the WikiAnn dataset (number of clusters: 3).

4.3 Results

Table 5 shows the comparisons in the NER evaluations of monolinguals and the clusterings using the nominal parameters, embeddings (1,000 and 10,000 samples), and all languages in Indo-European family (IE-all). Table 6 shows the results with the head parameter.

We first compare the NER evaluations of the clusterings based on the morpho-syntactic parameters and embeddings. The NER evaluations using the nominal parameters (Table 5) show that the clustering by the nominal parameters is superior to that of by embeddings. More than 70% of all the target languages attained the better scores. The clustering based on the head parameter (Table 6) outperformed the embedding-based clusterings as well, achieving the best scores in 80% of the target languages.

We then compare our methods using morpho-syntactic parameters with a model using all the Indo-European languages (IE-all). As for the number of languages that achieved the best score, 11 languages attained better scores with the clustering by the nominal parameters. This is slightly lower than the scores with the IE-all, which was 14 languages (Table 5). The clustering based on the head parameter scored the best in that approximately 70% of all the target languages outperformed the model with the IE-all (Table 6).

5 Analysis

5.1 Quantitative Analysis

Our parameter-based methods significantly outperformed the embedding-based method as in Section

lang	#train	mono	3 clusters noun	#1000	#10000	IE-all
cy	10,000	91.09	91.57	91.73	92.42	**92.95**
ga	1,000	76.51	**85.72**	84.11	84.43	84.90
ps	100	0.00	**55.92**	54.68	53.32	52.22
mr	5,000	85.50	86.96	87.93	**88.58**	88.34
hi	5,000	86.06	86.89	89.18	**89.90**	89.47
ro	20,000	92.64	**94.32**	94.04	93.98	94.18
fr	20,000	88.99	91.04	90.74	90.53	**91.05**
es	20,000	89.19	91.51	91.34	90.52	**91.63**
pt	20,000	90.24	92.11	91.79	91.43	**92.15**
it	20,000	90.79	**92.22**	91.93	91.52	92.06
scn	100	1.18	80.08	75.58	77.12	**81.04**
el	20,000	90.07	**91.21**	90.40	90.07	91.04
bg	20,000	92.48	93.25	92.64	93.34	**93.42**
pl	20,000	89.86	91.34	91.12	91.22	**91.43**
ru	20,000	88.52	89.96	89.32	**90.02**	89.88
sl	15,000	93.02	**93.89**	93.65	93.88	93.86
hr	20,000	90.90	92.05	91.88	**92.06**	92.02
af	5,000	89.06	91.19	91.51	90.75	**91.80**
nl	20,000	90.64	**92.59**	91.74	92.17	92.49
de	20,000	87.47	88.59	88.13	88.31	**88.70**
is	1,000	73.98	87.54	86.75	87.44	**88.29**
en	20,000	82.27	84.12	**84.22**	83.97	84.01
da	20,000	91.73	**93.15**	92.59	93.03	93.04
no	20,000	91.98	93.32	93.14	93.24	**93.49**
fo	100	0.00	86.61	86.35	87.44	**87.69**

Table 5: Nominal parameters clustering evaluations (F1). Each score is the mean over 3 training runs. The highest score for each language is indicated in **bold**.

lang	#train	mono	2 clusters head	#1000	#10000	IE-all
cy	10,000	91.09	**93.15**	92.22	91.88	92.95
ga	1,000	76.51	**85.37**	84.11	84.38	84.90
ps	100	0.00	**55.92**	55.31	55.02	52.22
mr	5,000	85.50	86.96	**88.71**	88.29	88.34
hi	5,000	86.06	86.89	**89.48**	89.42	89.47
ro	20,000	92.64	**94.43**	94.04	94.17	94.18
fr	20,000	88.99	**91.10**	90.74	90.56	91.05
es	20,000	89.19	**91.66**	91.34	90.52	91.63
pt	20,000	90.24	92.00	91.79	91.43	**92.15**
it	20,000	90.79	92.03	91.93	91.52	**92.06**
scn	100	1.18	77.04	75.58	77.12	**81.04**
el	20,000	90.07	**91.49**	90.91	91.28	91.04
bg	20,000	92.48	**93.60**	93.22	93.34	93.42
pl	20,000	89.86	91.38	91.12	91.33	**91.43**
ru	20,000	88.52	89.86	89.77	**89.98**	89.88
sl	15,000	93.02	**93.97**	93.65	93.95	93.86
hr	20,000	90.90	**92.27**	91.88	92.07	92.02
af	5,000	89.06	91.70	91.30	91.73	**91.80**
nl	20,000	90.64	**92.56**	91.90	92.23	92.49
de	20,000	87.47	**89.06**	88.13	88.61	88.70
is	1,000	73.98	88.04	87.28	87.63	**88.29**
en	20,000	82.27	**84.37**	84.22	84.02	84.01
da	20,000	91.73	**93.39**	92.76	92.91	93.04
no	20,000	91.98	93.46	93.05	93.34	**93.49**
fo	100	0.00	88.21	87.58	**88.70**	87.69

Table 6: Head parameter clustering evaluations (F1). Each score is the mean over 3 training runs. The highest score for each language is indicated in **bold**.

4.3. This suggests that the parameters in theoretical linguistics have a yet-to-be-explored potential in multilingual NLP. This section provides some more detailed analysis that supports this claim.

Clustering results First, we observe some unstable results in the embedding-based clustering. Table 4 shows that the resulting clusters greatly differ depending on the number of samples used to obtain embeddings. Thus, the embedding-based clustering could lead to inconsistent results and may not always be the most effective method.

The elbow method Moreover, we found that the optimal number of clusters determined by the elbow method did not result in the best performance in the embedding-based approach. For example, while the elbow method identified 3 clusters as optimal, the best scores were obtained when the number of clusters was 5 with 10,000 samples. This indicates that the optimal number of clusters obtained by the elbow method may not always be the most effective one, at least in NER.[6] Thus, we examine the results with different numbers of clusters. In particular, we compare clustering by embeddings and by the nominal parameters.[7] Tables 7 and 8 show the resulting clusters obtained by the embedding-based clustering when $k = 2, 3, 4, 5$ and Table 9 shows the NER results using these clusters and the results using the nominal parameters.

Sample size In the results of embedding-based clustering, the clustering with 10,000 samples always outperforms the clustering with 1,000 samples, regardless of the number of clusters. Thus, the following compares clustering by the nominal parameters and by the embeddings with 10,000 samples. Overall, clustering by the nominal parameters achieved better scores than by embeddings, except in the case of 5 clusters. When the number of the clusters is 5, 11 languages achieved better scores in the nominal parameters while 13 languages did so in the embedding-based clustering. We think this difference is due to the biased distribution in Cluster #1 of the embedding-based clustering (Table 8), i.e., 18 languages out of 25 languages are clustered together, while the clusters obtained by the nom-

[6] Shaffer (2021) also used the elbow method to determine the number of clusters (which was 4) but their experiments did not test other numbers of clusters.

[7] While there are only 2 clusters available in the head-parameter classification (i.e., either head-initial or head-final), we could test different numbers of clusters using the nominal parameters.

#	The number of clusters			
	2	3	4	5
1	cy, ps, mr, hi, el, bg, ru, af, nl, is, da, no, fo	cy, ps, el, bg	cy, ps, el, bg	cy, ps, bg
2	ga, ro, fr, es, pt, it, scn, pl, sl, hr, de, en	ga, ro, fr, es, pt, it, scn, pl, sl, hr, de, en	ga, ro, fr, es, pt, it, scn, pl, sl, hr, de, en	ga, ro, fr, es, pt, it, scn, pl, sl, hr, de, en
3	-	mr, hi, ru, af, nl, is, da, no, fo	mr, hi, af, nl	mr, hi, af, nl
4	-	-	ru, is, da, no, fo	ru, is, da, no, fo
5	-	-	-	el

Table 7: Embedding-based clustering with different cluster numbers (using 1,000 samples).

#	The number of clusters			
	2	3	4	5
1	cy, ga, ps, mr, hi, ro, fr, el, bg, pl, ru, sl, hr, af, nl, de, is, en, da, no, fo	cy, ga, ps, mr, hi, ro, fr, bg, pl, ru, sl, hr, af, nl, de, is, en, da, no, fo	cy, ga, ps, mr, hi, ro, fr, pl, ru, sl, hr, af, nl, de, is, en, da, no, fo	cy, ga, ps, mr, hi, ro, fr, pl, sl, hr, af, nl, de, is, en, da, no, fo
2	es, pt, it, scn	es, pt, it, scn	es, pt, it, scn	es, pt, it, scn
3	-	el	el	el
4	-	-	bg	bg
5	-	-	-	ru

Table 8: Embedding-based clustering with different cluster numbers (using 10,000 samples).

inal parameters distribute relatively evenly (Cluster #1{Germanic, Slavic}, #2{Hellenic}, #3{Romance}, #4{Indo-Iranian}, #5{Celtic}). Despite of this difference in the training data, clustering by nominal parameters achieved comparable results.

NER results with IE-all We have also run the NER experiments using all the Indo-European languages (see IE-all in Tables 5 and 6). Since this contains the largest training samples in our experiments, the performance would have been better than the other methods using clusters that normally contain the smaller training data. However, the nominal parameters showed comparable results, and the head parameter outperformed better than the IE-all. Together with the comparison results from the embedding-based method above, we argue that the parameters from theoretical linguistics have a potential to mitigate the data sparsity problem that has been present in the multilingual NLP tasks.

Methodological compatibility Another point to note is that some languages seem to be more compatible with a particular method than others. For example, one of low-resource languages, Pashto (ps) and some high-resource languages, such as Romanian (ro) and Danish (da), showed the best scores when using the clusters obtained by our parameter-based approach. On the other hand, Siciliano (scn) with the IE-all and relatively low-resource languages such as Marathi (mr) and Hindi (hi) with the embedding-based clustering demonstrated the best scores. These results indicate that different methods might have captured different aspects of languages regardless of the amount of data and that linguistic properties effective in clustering may differ depending on language.

5.2 Qualitative Analysis

This section attempts to provide some qualitative analysis based on the predictions obtained in the NER evaluations. We use the prediction data in English from our results of the head parameter clustering (Table 3) and the embedding-based clustering with 10,000 samples (Table 4). In the following examples, **h** indicates a prediction result from the head parameter clustering, which is correct. The notation **e** indicates a prediction from the embedding-based clustering, which is incorrect.

In (2h), the named entity representing an organization (ORG) “Allen Fieldhouse” appears after the preposition “at”. It is clearly predictable to English speakers that words representing location (LOC) or ORG appear after “at”, while it is less likely with words describing person (PER). However, the type of entity was not correctly predicted with the embedding-based clustering (2e). The correct prediction in (2h) seems reasonable if identification of the head along with its complement could facilitate inferring the contexts where a named entity occurs.

lang	#train	2 clusters			3 clusters			4 clusters			5 clusters		
		noun	#1000	#10000	noun	#1000	#10000	noun	#1000	#10000	noun	#1000	#10000
cy	10,000	91.57	**92.22**	91.88	91.57	91.73	**92.42**	91.57	91.73	**91.98**	91.57	91.27	**92.64**
ga	1,000	**85.72**	84.11	84.38	**85.72**	84.11	84.43	**85.72**	84.11	84.53	**85.72**	84.11	85.13
ps	100	53.97	**55.31**	55.02	**55.92**	54.68	53.32	**55.92**	54.68	55.37	**55.92**	52.97	53.54
mr	5,000	88.34	**88.71**	88.29	86.96	87.93	**88.58**	86.96	87.38	**88.09**	86.96	87.38	**88.13**
hi	5,000	**90.09**	89.48	89.42	86.89	89.18	**89.90**	86.89	88.66	**89.70**	86.89	88.66	**88.98**
ro	20,000	**94.32**	94.04	94.17	**94.32**	94.04	93.98	93.69	**94.04**	94.02	93.69	94.04	**94.06**
fr	20,000	**91.01**	90.74	90.56	**91.04**	90.74	90.53	90.39	**90.74**	90.52	90.39	**90.74**	90.32
es	20,000	**91.38**	91.34	90.52	**91.51**	91.34	90.52	90.96	**91.34**	90.52	90.96	**91.34**	90.52
pt	20,000	**92.14**	91.79	91.43	**92.11**	91.79	91.43	91.57	**91.79**	91.43	91.57	**91.79**	91.43
it	20,000	**92.16**	91.93	91.52	**92.22**	91.93	91.52	91.54	**91.93**	91.52	91.54	**91.93**	91.52
scn	100	76.54	75.58	**77.12**	**80.08**	75.58	77.12	76.77	75.58	**77.12**	76.77	75.58	**77.12**
el	20,000	91.18	90.91	**91.28**	**91.21**	90.40	90.07	**91.18**	90.40	90.07	**90.07**	90.07	**90.07**
bg	20,000	**93.44**	93.22	93.34	93.25	92.64	**93.34**	**93.18**	92.64	92.48	**93.19**	92.58	92.48
pl	20,000	**91.45**	91.12	91.33	**91.34**	91.12	91.22	91.19	91.12	**91.23**	91.18	91.12	**91.24**
ru	20,000	**90.01**	89.77	89.98	89.96	89.32	**90.02**	**89.97**	89.18	89.66	**89.81**	89.18	88.52
sl	15,000	93.79	93.65	**93.95**	93.89	93.65	**93.88**	**93.93**	93.65	93.61	93.78	93.65	**93.81**
hr	20,000	**92.12**	91.88	92.07	92.05	91.88	**92.06**	91.91	91.88	**92.14**	**91.97**	91.88	91.91
af	5,000	91.16	91.30	**91.73**	91.19	**91.51**	90.75	**91.46**	90.73	91.18	**91.37**	90.73	91.14
nl	20,000	**92.62**	91.90	92.23	**92.59**	91.74	92.17	**92.26**	90.86	92.14	92.14	90.86	**92.20**
de	20,000	88.51	88.13	**88.61**	**88.59**	88.13	88.31	88.25	88.13	**88.33**	88.25	88.13	**88.38**
is	1,000	**87.65**	87.28	87.63	**87.54**	86.75	87.44	**87.92**	86.51	87.77	87.51	86.51	**87.71**
en	20,000	84.11	**84.22**	84.02	84.12	**84.22**	83.97	83.75	**84.22**	83.89	83.83	**84.22**	83.89
da	20,000	**93.10**	92.76	92.91	**93.15**	92.59	93.03	**93.00**	92.43	92.78	92.92	92.43	**92.99**
no	20,000	**93.48**	93.05	93.34	**93.32**	93.14	93.24	**93.31**	92.79	93.27	**93.24**	92.79	93.17
fo	100	87.01	87.58	**88.70**	86.61	86.35	**87.44**	**88.70**	87.72	87.76	86.78	87.72	**88.33**

Table 9: Nominal parameter clustering evaluations for the number of clusters {2, 3, 4, 5} (F1). Each score is the mean over 3 training runs. In each number of clusters, the highest score for each language is indicated in **bold**.

(2) h. His 46 points tied the record for most points scored by an opponent at **Allen Fieldhouse**.
ORG

e. ... an opponent at **Allen Fieldhouse**.
PER

In (3e), a named entity consisted of three words "Arlington National Cemetery" was wrongly predicted to be split into ORG and LOC. This indicates that the named entity is not correctly identified as the complement of "in." Given this, we conjecture that clustering by the head parameter can be helpful in correctly predicting the position of the head in the phrase. Specifically, learning from the sequences of a P-head followed by its NP complement may have facilitated identifying the span of the named entity.

(3) h. He died in 1887 and was buried in **Arlington National Cemetery**.
ORG

e. ... in **Arlington** **National Cemetery**.
ORG **LOC**

5.3 Annotation Errors in the WikiAnn Dataset

When examining the incorrect predictions in English data, we found that the WikiAnn dataset contains some non-negligible annotation errors. From our sampling-based examination, we estimate that approximately 1% of annotation errors could be included in the WikiAnn dataset. Examples of the annotation errors found in the WikiAnn dataset are shown in (4) and (5). In (4), *Cleveland, Ohio* is not an organization name. In (5), although *Sanremo* is a named entity indicating location, the unnecessary brackets "[[" could have caused an error in its annotation.

(4) He was born in **Cleveland , Ohio**.
ORG

(5) Washhouse in [[Sanremo, **Italy**, ...
LOC

Since the annotations of the WikiAnn dataset were machine-generated, some errors could have occurred in its process. However, these annotation errors need to be revised to improve the reliability of NER evaluations.

6 Conclusion

We have proposed two language clustering methods based on the morpho-syntactic parameters proposed in theoretical linguistics. We showed that these clustering methods outperformed the embedding-based clustering in multilingual NER with Indo-European languages. We have also compared the model using all the Indo-European languages as the training data. Despite the large difference in the data size, our approach outperformed this model as well. These results suggest that parameters in theoretical linguistics have a potential utility in multilingual NLP tasks and that this direction is worth exploring.

Future work will extend this approach to other language families as well as different multilingual tasks, such as machine translation. Another direction would be to probe the clusters derived from the embedding-based method to explore features that might not have been captured by our approach or any approaches that make use of explicit linguistic features.

Limitations

The morpho-syntactic parameters used in this study are just a fraction of various other linguistic parameters that have been proposed in theoretical syntax (e.g., Roberts 2019). A set of optimal language parameters for language clustering may vary depending on the target task. It remains to be seen whether and how various parameters in theoretical linguistics could improve different NLP tasks. For example, cross-lingual transfer learning may be performed more effectively by carefully tailoring the linguistic parameters to a particular task, like what we have done for NER.

Related to the above point, one limitation of our approach would be the fact that some languages have not yet been investigated well in theoretical linguistics, particularly some underdocumented or endangered languages. Even as for well-documented languages in theoretical linguistics, some parameters still remain controversial, such as the so-called NP/DP parameter (e.g., Bošković 2012). Thus, our approach proceeds in tandem with the advancement of theoretical linguistics.

Ethics Statement

We used a freely available dataset and a pre-trained model from the Hugging Face Hub for our experiments. We selected a pre-trained model with an appropriate size (XLM-RoBERTa-base) given our purpose of use. We needed to perform many rounds of clustering and fine-tuning for the pre-trained model. Therefore, we set preliminary experiments beforehand with a smaller sample size for each step to ensure that the experiments could be performed effectively.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP21H04901.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Željko Bošković. 2012. *On NPs and Clauses*, page 179–246. De Gruyter Mouton, Berlin, Boston.

Andrea Ceolin, Cristina Guardiano, Giuseppe Longobardi, Monica Alexandrina Irimia, Luca Bortolussi, and Andrea Sgarro. 2021. At the boundaries of syntactic prehistory. *Philosophical Transactions of the Royal Society B*, 376.

Noam Chomsky. 1981. *Lectures on Government and Binding*. De Gruyter, Berlin, Germany.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2022. Language-family adapters for multilingual neural machine translation. *ArXiv*, abs/2209.15236.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

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.

Tejas Dhamecha, Rudra Murthy, Samarth Bharadwaj, Karthik Sankaranarayanan, and Pushpak Bhattacharyya. 2021. Role of Language Relatedness in Multilingual Fine-tuning of Language Models: A Case Study in Indo-Aryan Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8584–8595, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Katharina Kann. 2022. Match the script, adapt if multilingual: Analyzing the effect of multilingual pretraining on cross-lingual transferability. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1500–1512, Dublin, Ireland. Association for Computational Linguistics.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406, Online. Association for Computational Linguistics.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Ryokan Ri and Yoshimasa Tsuruoka. 2022. Pretraining with artificial language: Studying transferable knowledge in language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7302–7315, Dublin, Ireland. Association for Computational Linguistics.

Ian Roberts. 2019. *Parameter Hierarchies and Universal Grammar*. Oxford University Press.

Fahimeh Saleh, Wray Buntine, Gholamreza Haffari, and Lan Du. 2021. Multilingual neural machine translation: Can linguistic hierarchies help? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1313–1330, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kyle Shaffer. 2021. Language clustering for multilingual named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 40–45, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973, Hong Kong, China. Association for Computational Linguistics.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robert L. Thorndike. 1953. Who belongs in the family? *Psychometrika*, 18:267–276.

A Appendix

The summary of the languages used in our experiments is shown in Table 10.

Table 11 shows the NER evaluations of head parameter-based clustering with standard deviation scores in parentheses.

Tables 12 and 13 represent the NER evaluations when we set the number of clusters to {2, 3} and {4, 5}, respectively, with standard deviations in parentheses (see Section 5.1 for the details).

ISO 639-1 Code	Language	Sub-family
cy	Welsh	Celtic
ga	Irish	
ps	Pashto	Indo-Iranian
mr	Marathi	
hi	Hindi	
ro	Romanian	Romance
fr	French	
es	Spanish	
pt	Portuguese	
it	Italian	
scn	Siciliano	
el	Greek	Hellenic
bg	Bulgarian	Slavic
pl	Polish	
ru	Russian	
sl	Slovenian	
hr	Serbo-Croatian	
af	Afrikaans	Germanic
nl	Dutch	
de	German	
is	Icelandic	
en	English	
da	Danish	
no	Norwegian	
fo	Faroese	

Table 10: The summary of language codes mentioned in this paper, along with the sub-families they belong to.

			2 clusters			
lang	#train	mono	head	#1000	#10000	family
cy	10,000	91.09 (0.30)	**93.15 (0.03)**	92.22 (0.37)	91.88 (0.37)	92.95 (0.45)
ga	1,000	76.51 (1.18)	**85.37 (0.54)**	84.11 (0.61)	84.38 (0.21)	84.90 (0.45)
ps	100	0.00 (0.00)	**55.92 (2.84)**	55.31 (1.40)	55.02 (0.76)	52.22 (1.31)
mr	5,000	85.5 (0.03)	86.96 (0.39)	**88.71 (0.66)**	88.29 (0.40)	88.34 (0.37)
hi	5,000	86.06 (0.54)	86.89 (0.30)	**89.48 (0.42)**	89.42 (0.80)	89.47 (0.45)
ro	20,000	92.64 (0.11)	**94.43 (0.27)**	94.04 (0.12)	94.17 (0.11)	94.18 (0.03)
fr	20,000	88.99 (0.14)	**91.10 (0.09)**	90.74 (0.09)	90.56 (0.13)	91.05 (0.15)
es	20,000	89.19 (0.12)	**91.66 (0.31)**	91.34 (0.10)	90.52 (0.19)	91.63 (0.02)
pt	20,000	90.24 (0.06)	92.00 (0.22)	91.79 (0.07)	91.43 (0.06)	**92.15 (0.06)**
it	20,000	90.79 (0.21)	92.03 (0.12)	91.93 (0.11)	91.52 (0.07)	**92.06 (0.10)**
scn	100	1.18 (1.67)	77.04 (1.46)	75.58 (1.20)	77.12 (1.63)	**81.04 (2.88)**
el	20,000	90.07 (0.15)	**91.49 (0.05)**	90.91 (0.08)	91.28 (0.17)	91.04 (0.09)
bg	20,000	92.48 (0.07)	**93.60 (0.17)**	93.22 (0.11)	93.34 (0.03)	93.42 (0.11)
pl	20,000	89.86 (0.08)	91.38 (0.11)	91.12 (0.04)	91.33 (0.10)	**91.43 (0.17)**
ru	20,000	88.52 (0.14)	89.86 (0.16)	89.77 (0.07)	**89.98 (0.12)**	89.88 (0.02)
sl	15,000	93.02 (0.04)	**93.97 (0.27)**	93.65 (0.19)	93.95 (0.10)	93.86 (0.16)
hr	20,000	90.90 (0.22)	**92.27 (0.03)**	91.88 (0.12)	92.07 (0.13)	92.02 (0.04)
af	5,000	89.06 (0.09)	91.70 (0.31)	91.30 (0.57)	91.73 (0.31)	**91.80 (0.20)**
nl	20,000	90.64 (0.15)	**92.56 (0.23)**	91.90 (0.10)	92.23 (0.07)	92.49 (0.07)
de	20,000	87.47 (0.10)	**89.06 (0.32)**	88.13 (0.05)	88.61 (0.06)	88.70 (0.01)
is	1,000	73.98 (2.36)	88.04 (0.40)	87.28 (0.39)	87.63 (0.56)	**88.29 (0.77)**
en	20,000	82.27 (0.14)	**84.37 (0.15)**	84.22 (0.23)	84.02 (0.06)	84.01 (0.09)
da	20,000	91.73 (0.11)	**93.39 (0.27)**	92.76 (0.09)	92.91 (0.15)	93.04 (0.03)
no	20,000	91.98 (0.13)	93.46 (0.16)	93.05 (0.07)	93.34 (0.20)	**93.49 (0.09)**
fo	100	0.00 (0.00)	88.21 (1.52)	87.58 (1.09)	**88.70 (1.22)**	87.69 (0.75)

Table 11: Head parameter clustering evaluations (F1): Each score is the mean over 3 training runs, with a standard deviation in parentheses. The highest score for each language is indicated in **bold**.

lang	#train	2 clusters noun	2 clusters #1000	2 clusters #10000	3 clusters noun	3 clusters #1000	3 clusters #10000
cy	10,000	91.57 (0.12)	**92.22 (0.37)**	91.88 (0.37)	91.57 (0.12)	91.73 (0.03)	**92.42 (0.59)**
ga	1,000	**85.72 (0.04)**	84.11 (0.61)	84.38 (0.21)	**85.72 (0.04)**	84.11 (0.61)	84.43 (0.60)
ps	100	53.97 (3.36)	**55.31 (1.40)**	55.02 (0.76)	**55.92 (2.84)**	54.68 (1.07)	53.32 (2.06)
mr	5,000	88.34 (0.35)	**88.71 (0.66)**	88.29 (0.40)	86.96 (0.39)	87.93 (0.31)	**88.58 (0.44)**
hi	5,000	**90.09 (0.29)**	89.48 (0.42)	89.42 (0.80)	86.89 (0.30)	89.18 (0.54)	**89.90 (0.29)**
ro	20,000	**94.32 (0.10)**	94.04 (0.12)	94.17 (0.11)	**94.32 (0.05)**	94.04 (0.12)	93.98 (0.12)
fr	20,000	**91.01 (0.04)**	90.74 (0.09)	90.56 (0.13)	**91.04 (0.03)**	90.74 (0.09)	90.53 (0.02)
es	20,000	**91.38 (0.18)**	91.34 (0.10)	90.52 (0.19)	**91.51 (0.08)**	91.34 (0.10)	90.52 (0.19)
pt	20,000	**92.14 (0.12)**	91.79 (0.07)	91.43 (0.06)	**92.11 (0.10)**	91.79 (0.07)	91.43 (0.06)
it	20,000	**92.16 (0.15)**	91.93 (0.11)	91.52 (0.07)	**92.22 (0.12)**	91.93 (0.11)	91.52 (0.07)
scn	100	76.54 (0.92)	75.58 (1.20)	**77.12 (1.63)**	**80.08 (2.69)**	75.58 (1.20)	77.12 (1.63)
el	20,000	91.18 (0.21)	90.91 (0.08)	**91.28 (0.17)**	**91.21 (0.01)**	90.40 (0.11)	90.07 (0.15)
bg	20,000	**93.44 (0.07)**	93.22 (0.11)	93.34 (0.03)	93.25 (0.02)	92.64 (0.07)	**93.34 (0.15)**
pl	20,000	**91.45 (0.09)**	91.12 (0.04)	91.33 (0.10)	**91.34 (0.02)**	91.12 (0.04)	91.22 (0.05)
ru	20,000	**90.01 (0.08)**	89.77 (0.07)	89.98 (0.12)	89.96 (0.18)	89.32 (0.06)	**90.02 (0.04)**
sl	15,000	93.79 (0.10)	93.65 (0.19)	**93.95 (0.10)**	93.89 (0.22)	93.65 (0.19)	**93.88 (0.09)**
hr	20,000	**92.12 (0.11)**	91.88 (0.12)	92.07 (0.13)	92.05 (0.07)	91.88 (0.12)	**92.06 (0.11)**
af	5,000	91.16 (0.16)	91.30 (0.57)	**91.73 (0.31)**	91.19 (0.37)	**91.51 (0.40)**	90.75 (0.17)
nl	20,000	**92.62 (0.02)**	91.90 (0.10)	92.23 (0.07)	**92.59 (0.17)**	91.74 (0.12)	92.17 (0.16)
de	20,000	88.51 (0.04)	88.13 (0.05)	**88.61 (0.06)**	**88.59 (0.13)**	88.13 (0.05)	88.31 (0.13)
is	1,000	**87.65 (0.23)**	87.28 (0.39)	87.63 (0.56)	**87.54 (0.24)**	86.75 (0.39)	87.44 (0.16)
en	20,000	84.11 (0.29)	**84.22 (0.23)**	84.02 (0.06)	84.12 (0.09)	**84.22 (0.23)**	83.97 (0.05)
da	20,000	**93.10 (0.11)**	92.76 (0.09)	92.91 (0.15)	**93.15 (0.18)**	92.59 (0.15)	93.03 (0.10)
no	20,000	**93.48 (0.06)**	93.05 (0.07)	93.34 (0.20)	**93.32 (0.13)**	93.14 (0.02)	93.24 (0.06)
fo	100	87.01 (0.90)	87.58 (1.09)	**88.70 (1.22)**	86.61 (0.59)	86.35 (1.26)	**87.44 (0.66)**

Table 12: Nominal parameter clustering evaluations with the number of clusters {2, 3} (F1): Each score is the mean over 3 training runs, with a standard deviation in parentheses. The highest score for each language is indicated in **bold**.

lang	#train	4 clusters noun	4 clusters #1000	4 clusters #10000	5 clusters noun	5 clusters #1000	5 clusters #10000
cy	10,000	91.57 (0.12)	91.73 (0.03)	**91.98 (0.42)**	91.57 (0.12)	91.27 (0.34)	**92.64 (0.13)**
ga	1,000	**85.72 (0.04)**	84.11 (0.61)	84.53 (0.27)	**85.72 (0.04)**	84.11 (0.61)	85.13 (0.81)
ps	100	**55.92 (2.84)**	54.68 (1.07)	55.37 (0.69)	**55.92 (2.84)**	52.97 (2.53)	53.54 (2.79)
mr	5,000	86.96 (0.39)	87.38 (0.86)	**88.09 (0.19)**	86.96 (0.39)	87.38 (0.86)	**88.13 (0.52)**
hi	5,000	86.89 (0.30)	88.66 (0.37)	**89.70 (0.09)**	86.89 (0.30)	88.66 (0.37)	**88.98 (0.38)**
ro	20,000	93.69 (0.04)	**94.04 (0.12)**	94.02 (0.08)	93.69 (0.04)	94.04 (0.12)	**94.06 (0.13)**
fr	20,000	90.39 (0.03)	**90.74 (0.09)**	90.52 (0.21)	90.39 (0.03)	**90.74 (0.09)**	90.32 (0.14)
es	20,000	90.96 (0.13)	**91.34 (0.10)**	90.52 (0.19)	90.96 (0.13)	**91.34 (0.10)**	90.52 (0.19)
pt	20,000	91.57 (0.06)	**91.79 (0.07)**	91.43 (0.06)	91.57 (0.06)	**91.79 (0.07)**	91.43 (0.06)
it	20,000	91.54 (0.06)	**91.93 (0.11)**	91.52 (0.07)	91.54 (0.06)	**91.93 (0.11)**	91.52 (0.07)
scn	100	76.77 (1.32)	75.58 (1.20)	**77.12 (1.63)**	76.77 (1.32)	75.58 (1.20)	**77.12 (1.63)**
el	20,000	**91.18 (0.13)**	90.4 (0.11)	90.07 (0.15)	**90.07 (0.15)**	90.07 (0.15)	**90.07 (0.15)**
bg	20,000	**93.18 (0.10)**	92.64 (0.07)	92.48 (0.07)	**93.19 (0.10)**	92.58 (0.03)	92.48 (0.07)
pl	20,000	91.19 (0.02)	91.12 (0.04)	**91.23 (0.09)**	91.18 (0.10)	91.12 (0.04)	**91.24 (0.05)**
ru	20,000	**89.97 (0.15)**	89.18 (0.18)	89.66 (0.02)	**89.81 (0.20)**	89.18 (0.18)	88.52 (0.14)
sl	15,000	**93.93 (0.18)**	93.65 (0.19)	93.61 (0.02)	93.78 (0.06)	93.65 (0.19)	**93.81 (0.06)**
hr	20,000	91.91 (0.06)	91.88 (0.12)	**92.14 (0.10)**	**91.97 (0.09)**	91.88 (0.12)	91.91 (0.17)
af	5,000	**91.46 (0.70)**	90.73 (0.05)	91.18 (0.12)	**91.37 (0.31)**	90.73 (0.05)	91.14 (0.34)
nl	20,000	**92.26 (0.11)**	90.86 (0.17)	92.14 (0.04)	92.14 (0.14)	90.86 (0.17)	**92.20 (0.15)**
de	20,000	88.25 (0.09)	88.13 (0.05)	**88.33 (0.07)**	88.25 (0.21)	88.13 (0.05)	**88.38 (0.06)**
is	1,000	**87.92 (0.83)**	86.51 (0.09)	87.77 (0.40)	87.51 (0.37)	86.51 (0.09)	**87.71 (0.25)**
en	20,000	83.75 (0.19)	**84.22 (0.23)**	83.89 (0.14)	83.83 (0.03)	**84.22 (0.23)**	83.89 (0.03)
da	20,000	**93.00 (0.05)**	92.43 (0.08)	92.78 (0.09)	92.92 (0.10)	92.43 (0.08)	**92.99 (0.04)**
no	20,000	**93.31 (0.11)**	92.79 (0.00)	93.27 (0.06)	**93.24 (0.07)**	92.79 (0.00)	93.17 (0.13)
fo	100	**88.70 (1.58)**	87.72 (0.82)	87.76 (1.06)	86.78 (2.33)	87.72 (0.82)	**88.33 (0.28)**

Table 13: Nominal parameter clustering evaluations for the number of clusters {4, 5} (F1): Each score is the mean over 3 training runs, with a standard deviation in parentheses. The highest score for each language is indicated in **bold**.