The Hidden Folk: Linguistic Properties encoded in Multilingual Contextual Character Representations

Anonymous ACL submission

Abstract

To gain a better understanding of the linguistic information encoded in character-based language models, we probe the multilingual contextual CANINE model. We design a range of phonetic probing tasks in six Nordic languages, including Faroese as an additional zero-shot instance. The results show that phonetic information such as consonant voicing and vowel roundness are indeed encoded in the character representations and that this information is transferred to a similar zero-shot language.

1 Introduction

Subword and character sequence information is crucial in state-of-the-art neural language models (Sennrich et al., 2016; Kudo, 2018) because it improves their generalization and robustness capabilities (Xue et al., 2022; Tay et al., 2021). Additionally, character-level features are beneficial for morphologically rich and low-resource languages (Papay et al., 2018; Riabi et al., 2021). However, there is a lack of interpretability methods for character-based models. Only a few approaches have tried to understand the linguistic information encoded in character embeddings and cross-lingual approaches must be evaluated more rigorously by considering typology and linguistic distance (Artetxe et al., 2020). Therefore, in this work, we analyze how much phonetic information is encoded in contextualized multilingual character embeddings from the CANINE language model (Clark et al., 2022).

Based on six Nordic languages, we extract phonetic features for characters in context through unsupervised grapheme-to-phoneme alignment and design a set of probing tasks. We explore the character representations in two different evaluation scenarios, a traditional train/test split scenario and a leave-one-letter-out scenario. We find that phonetic information on a global level (e.g., vowel and consonant detection) is encoded accurately in the character representations and more mixed results are achieved on lower-level probing tasks such as consonant voicing and manner, or vowel height and roundness. We also see that this information is transferred to the related zero-shot language Faroese. Our code is available online[1].

2 Related Work

We discuss previous work in this area by focusing on existing multilingual character language models and the interpretability of these models.

Building Character-Level Models Subword and character-level information is exhibiting great benefits for computational language models (Bostrom and Durrett, 2020; Zhang et al., 2021). With the rise of multilingual models, pre-trained simultaneously on 100+ languages, these are also being adapted for and augmented with character-level information (Xue et al., 2022; Tay et al., 2021). We choose to work with the CANINE model by Clark et al. (2022), because it is a strongly performing neural encoder which operates directly on character sequences, i.e., without explicit tokenization or vocabulary, and incorporates a pre-training strategy operating directly on characters.

Understanding Character-Level Models While a wide range of probing tasks have been studied for contextualized word representations (e.g. Liu et al. 2019), there is limited work on directly probing character representations. Recent work probes word-level representation with respect to their knowledge about characters. For instance, Kaushal and Mahowald (2022) predict the presence of a particular character in a token showing that large models robustly encode this information across various scripts. Additionally, Itzhak and Levy (2021) test the "spelling abilities" of language models showing that the embedding layers of RoBERTa and GPT-2 learn the internal char-

[1] URL omitted for anonymity.

acter composition of whole words to a surprising extent, without seeing the characters coupled with the tokens during training.

Specifically on character-level models, Boldsen et al. (2022) compare perceptual representations to character embeddings. Their cross-lingual analysis shows that character representations correlate with phonological representations for languages using an alphabetic script and implies a relationship between the information encoded in the embeddings and the orthographic transparency of the languages. Furthermore, Hahn and Baroni (2019) probe character models in a cognitively realistic task on data with removed word boundaries showing that recurrent LMs learn morphological, syntactic and semantic aspects even on unsegmented text. These findings encourage the exploration of character and phoneme-level learning.

3 Contextualized Character Embeddings

We extract character embeddings (in the context of full words) from the CANINE model. In this section, we present the multilingual data and the embedding extraction.

Data Since character-level features are important for morphologically rich and low-resource languuages (Lauscher et al., 2020; Garrette and Baldridge, 2013), we choose a set of six Nordic languages for our experiments: Danish (da), Swedish (sv), Norwegian (nb), Finnish (fi), Icelandic (is) and Faroese (fo). Five of the languages are included in the training data of the character language model (da sv, nb, fi and is). Additionally, we use Faroese to test performance of multilingual zero-shot embeddings. The starting point for extracting character embeddings is a frequency list for each language (see Table 1). We select the 10000 most frequent words of every language and then randomly sample 3000 of these words and retrieve embeddings for all characters in these 3000 words. This implies that more frequent characters in a given language will be better represented in the embeddings. Note that word length will affect the number of character embeddings extracted.

Model We extract contextualized embeddings from the CANINE model (Clark et al., 2022). CANINE is a neural encoder which operates directly on character-level without requiring an explicit tokenization strategy or a pre-defined vocabulary. We choose this model since it showed superior performance on multilingual downstream tasks. CANINE has been trained on data from 104 languages.[2] While it performs well on NLP tasks, it has not yet been explored which type of linguistic information is encoded in these pre-trained character representations. We use the HuggingFace checkpoint of the CANINE model with autoregressive character loss.[3] We input words to the model, and use the last hidden state of a character in the context of the word it occurs in as a contextualized character embedding ($d = 768$).

4 Phonetic Feature Extraction

In order to extract phonetic features from characters, we need to know how a specific letter should be pronounced. We do that by aligning the characters with the string of phones as given by Wikipron. As the number of letters and phones may not match, we align them using `m2m-aligner` (Jiampojamarn et al., 2007), an unsupervised model that is based on Expectation-Maximization. Then, we use the `ipapy` toolkit to obtain phonetic features for each phone. We describe this process below.

Pipeline We use the aligner to obtain phonetic features for characters in all six languages. First, we obtain a dictionary for each language from Wikipron (Lee et al., 2020),[4] and then, we align graphemes to phones using `m2m-aligner`.[5] Wikipron includes both phonemic and phonetic representations of words, which they refer to broad and narrow, respectively. As the model for extracting features works at a phonetic level, we use the phonetic representations (narrow). In the next step, we use the `ipapy`[6] toolkit to extract phonetic features for each phone.

Finally, the IPA features are merged with the CANINE character representations. This process results in one dataset per language, consisting of 6067 characters in 899 words for Danish, 700 characters in 135 words for Faroese, 4698 characterss in 745 words for Finnish, 268 characters in 43 words for Icelandic, 302 characters in 57 words

[2]The CANINE model is pretrained on on the multilingual Wikipedia data of mBERT.

[3]https://huggingface.co/google/canine-c

[4]Please find the size of the dictionaries in Appendix A.2.

[5]We evaluated the alignments of `m2maligner` by using a manually aligned dictionary. This is available for the Danish language (Juul, 2010), where $\sim 42,000$ words and their phonetic transcriptions are aligned. Results show that the word error rate is below $< 2.5\%$.

[6]https://github.com/pettarin/ipapy

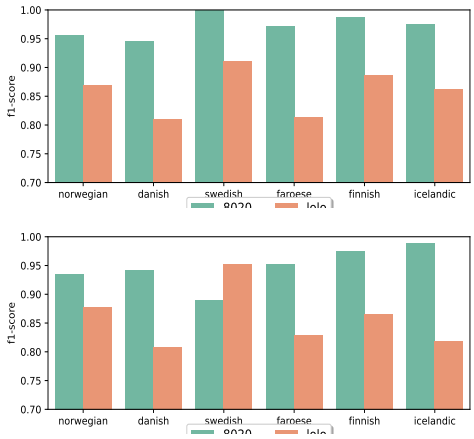

Figure 1: Weighted F1-scores for two global features: vowel prediction (top) and consonant predictions (bottom). Exact numbers in Appendix A.4.

for Norwegian, and 312 character in 58 words for Swedish. The reason for the drastic decrease in samples for some of the language is the small size of the pronunciation dictionaries.

5 Probing

In this section, we describe the probing tasks and the evaluation scenarios that we devise to test the extracted character embeddings.

We design 23 tasks to investigate the phonological knowledge encoded in the multilingual character representations. The tasks are split into three categories: global features (e.g., is this character a vowel or not?), consonant features (e.g., is the manner of articulation of this consonant plosive or not?[7], and vowel features (e.g., is this vowel pronounced as a rounded vowel?[8]

The probing tasks have the structure $F : X \rightarrow Y$, where given a set of character representations X, we want to find the best mapping F that relates X to a set of target features Y using a supervised Logistic Regression classification model.

As discussed by Hewitt and Liang (2019), it is important to take into account the expressivity of a probing task, since overly expressive probes, i.e., too many possible mappings for $F : X \rightarrow Y$, does not reveal much about the internal feature representations. Therefore, we test the all probing classifiers in two evaluation scenarios: (i) 80/20, a random 80% training and 20% test split of the data, and (ii) LoLo, a leave-one-letter-out training and test split. The 80/20 setup implies that the same characters (in different contexts) can appear train *and* test split, resulting in a simpler probing task, whereas the LoLo setup ensures zero-shot learning for the specific character of which all representations in all contexts are held out during training.

[7] In Danish, "b" in *peber* (['pʰewɐ]) vs. *åben* (['ɔːb̥m̩])

[8] In Danish, "o" as in ballon ([b̥a'lʌŋ]) (unrounded) vs. blod (['b̥loðˀ])(rounded)

6 Results

To understand the implicit phonetic information encoded in contextual representations, we present weighted F1-scores for a selection of features.[9] Most of the individual probing tasks show F1-scores above 0.5, which means that the models generally perform better than a random baseline.

In Figure 1, we observe the results for two global features, where we predict whether a character is a vowel (top) and whether a letter is a consonant (bottom). In each plot, we report the results of both evaluation scenarios. Faroese performs very similar to Danish, even though the CANINE training data does not include any Faroese data. Furthermore, Norwegian and Swedish are the languages with the smallest difference between different validation methodologies (80/20 vs. LoLo).

Figure 2 shows the LoLo performance of each model for all characters and languages. The models were trained to predict whether the contextual character embedding had the label `global_type_vowel`, meaning that the character in this specific context is pronounced as a vowel. Taking a closer look at Danish and Norwegian, for instance, the F1-scores are relatively similar. This was expected given the similarities between the two languages in the written form. Although the results are similar, if we zoom in to specific character, we observe that the performance for vowels is systematically worse for Danish, which is reasonable given the complex nature of the Danish vowel system (Trecca et al., 2018).

Figure 3 shows the weighted F1-scores of predicting whether a character in a given context is a plosive or not. This figure shows that our LoLo evaluation methodology is suitable, but we must consider certain aspects. We should strive for balanced training sets—meaning that the number of both positive and negative instances should be relatively similar— in order to get interpretable, and not misleading, results. The figure shows some characters for which the F1-scores are low for most languages. These are cases in which certain letters

[9] See Appendix A.4 for the full results.

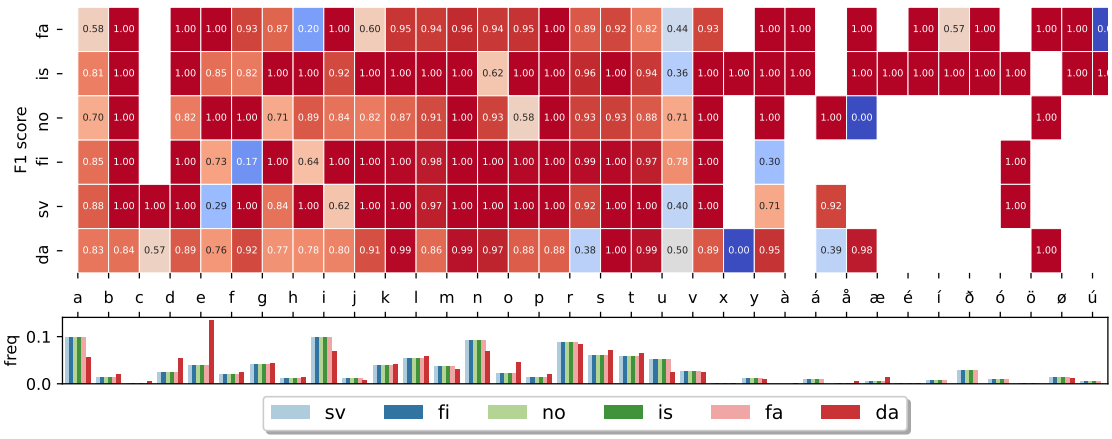

Figure 2: Heatmap for **global type vowel**. The barplot shows the frequency of the given character.

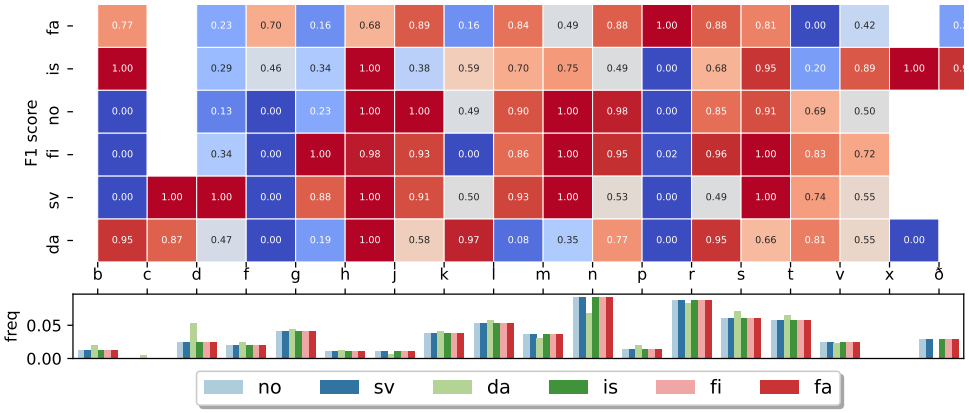

Figure 3: Heatmap for **consonant manner plosive**. The barplot shows the frequency of the given character.

do not have certain phonological features.

7 Conclusions & Future Work

In this work, we design a probing mechanism to better understand information encoded in contextual character representations, for which we use two validation mechanisms. The first one, where the training data is divided into training and testing set, and the second one where we train one model for each character. The reasoning behind this is that a letter representation, even though it occurs in a different context, would have a similar representation and therefore, it would involve a kind of data leakage to the testset. We can imagine the example of the letter "a" in the words "tram" and "gas" in English. The contexts are rather different, but we would expect the representation of the letter "a" to be relatively similar, as it is pronounced similarly. Therefore, the LoLo evaluation allows us to test the representations of a character in *any* context in a zero-shot scenario.

We use Wikipron data as linguistic knowledge and we align graphs and phones using an automatic aligning mechanism. We validate the aligner for Danish on manually aligned data, but other languages may have their own specific challenges. We use a wide range of phonetic probing tasks to accommodate language-specific particularities.

It is crucial that the number of positive and negative instances is checked when probing each phonetic feature. As expected, many pronunciation-related features do not occur in specific languages, and thus, this results in phonetic features with only one class, which will not shed any light on such phonetic properties. Hence, future research should address the design of these probing tasks as much as the results. Finally, since some of the datasets were relatively small, using a grapheme-to-phoneme model could increase the number of instances. Besides, these analyses could be extended to more languages to cover the full spectrum of orthographic depth.

References

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

Sidsel Boldsen, Manex Agirrezabal, and Nora Hollenstein. 2022. Interpreting character embeddings with perceptual representations: The case of shape, sound, and color. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6819–6836, Dublin, Ireland. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia. Association for Computational Linguistics.

Michael Hahn and Marco Baroni. 2019. Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text. *Transactions of the Association for Computational Linguistics*, 7:467–484.

Zakaris Svabo Hansen, Heini Justinussen, and Mortan Ólason. 2004. Marking av teldutøkum tekstsavni [tagging of a digital text corpus]. *URL http://ark. axeltra. com/index. php*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Itay Itzhak and Omer Levy. 2021. Models in a spelling bee: Language models implicitly learn the character composition of tokens. *arXiv preprint arXiv:2108.11193*.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379, Rochester, New York. Association for Computational Linguistics.

Holger Juul. 2010. K-a-tt-e-p-i-n-er: om komplekse bogstav-lyd-forbindelser i danske ord. *NyS*, 39:10–32.

Ayush Kaushal and Kyle Mahowald. 2022. What do tokens know about their characters and how do they know it? *arXiv preprint arXiv:2206.02608*.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1):121–163.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. 2020. Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Sigríður Ólafsdóttir, Auður Pálsdóttir, Starkaður Barkarson, and Ásdís Björg Björgvinsdóttir. 2022. Word frequency list from the icelandic corpus for academic words (orðtíðnilisti málheildar fyrir íslenskan námsorðaforða - MÍNO). CLARIN-IS.

Sean Papay, Sebastian Padó, and Ngoc Thang Vu. 2018. Addressing low-resource scenarios with character-aware embeddings. In *Proceedings of the Second Workshop on Subword/Character LEvel Models*,

pages 32–37, New Orleans. Association for Computational Linguistics.

Arij Riabi, Benoît Sagot, and Djamé Seddah. 2021. Can character-based language models improve downstream task performances in low-resource and noisy language scenarios? In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 423–436, Online. 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.

Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2021. Charformer: Fast character transformers via gradient-based subword tokenization. In *International Conference on Learning Representations*.

Fabio Trecca, Dorthe Bleses, Thomas O Madsen, and Morten H Christiansen. 2018. Does sound structure affect word learning? an eye-tracking study of danish learning toddlers. *Journal of Experimental Child Psychology*, 167:180–203.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Xinsong Zhang, Pengshuai Li, and Hang Li. 2021. AMBERT: A pre-trained language model with multi-grained tokenization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 421–435, Online. Association for Computational Linguistics.

A Appendix

A.1 Frequency Lists

Table 1 contains the details about the frequency lists used for each language.

Language	Reference & Source
Danish	DSL frequency list by Jørg Asmussen
Faroese	Sosialurin corpus by Hansen et al. (2004)
Finnish	Parole corpus frequency list by the Institure for the Languages of Finland
Icelandic	Icelandic Corpus for Academic Words by Ólafsdóttir et al. (2022)
Norwegian	Kelly List by Kilgarriff et al. (2014)
Swedish	Kelly List by Kilgarriff et al. (2014)

Table 1: Data sources of the frequency lists for all six languages.

A.2 Pronunciation Dictionaries

Table 2 shows the size (word count) of the pronunciation dictionaries used in this work to train the `m2n-aligner` in all six languages.

Language	Words
Danish	8,219
Faroese	1,118
Finnish	80,377
Icelandic	464
Norwegian (bokmål)	604
Swedish	372

Table 2: Pronunciation dictionary size for all six languages, obtained from Lee et al. (2020).

A.3 Average word length and standard deviation

Table 3 shows the average word length and standard deviation for each language in the CANINE embeddings.

Language	mean	std.
Danish	7.5398	3.0756
Faroese	7.8557	3.6023
Finnish	7.7544	2.8365
Icelandic	8.4372	3.4093
Norwegian	7.0766	2.8401
Swedish	7.5166	3.3496

Table 3: Mean word length and standard deviation for each language in the CANINE embeddings.

A.4 Full Results

Table 4 presents the LOLO results (F1 scores) for all probing tasks and all languages.

Feature	*no*	*da*	*sv*	*fa*	*fi*	*is*
global_type_consonant	0.88	0.81	0.95	0.83	0.86	0.82
global_type_vowel	0.87	0.81	0.91	0.81	0.89	0.86
global_type_diacritic	0.87	0.56	0.80	0.96	0.78	0.86
global_type_suprasegmental	0.82	0.92	0.71	0.86	0.77	0.81
consonant_voicing_voiced	0.57	0.64	0.59	0.43	0.43	0.67
consonant_voicing_voiceless	0.57	0.65	0.66	0.39	0.43	0.50
consonant_place_alveolar	0.46	0.74	0.63	0.58	0.69	0.66
consonant_place_bilabial	0.83	0.83	0.82	0.85	0.86	0.85
consonant_place_labio-dental	0.89	0.90	0.87	0.88	0.94	0.87
consonant_place_palatal	0.92	0.96	0.96	0.95	0.95	0.89
consonant_place_velar	0.86	0.92	0.85	0.90	0.86	0.90
consonant_manner_approximant	0.86	0.86	0.96	0.91	0.87	0.95
consonant_manner_nasal	0.72	0.78	0.83	0.69	0.73	0.84
consonant_manner_non-sibilant-fricative	0.89	0.75	0.85	0.86	0.94	0.73
consonant_manner_plosive	0.64	0.57	0.62	0.52	0.67	0.54
vowel_height_close	0.63	0.77	0.89	0.82	0.48	0.84
vowel_height_close-mid	0.81	0.72	0.74	0.90	0.92	0.83
vowel_backness_front	0.51	0.54	0.41	0.37	0.33	0.53
vowel_backness_back	0.71	0.65	0.61	0.71	0.33	0.81
vowel_roundness_rounded	0.80	0.71	0.67	0.83	0.67	0.77
vowel_roundness_unrounded	0.80	0.72	0.70	0.81	0.67	0.83

Table 4: LOLO results (F1 scores, weighted) for all probing tasks and all languages.