Evaluating Morphological Plausibility of Subword Tokenization via Statistical Alignment with Morpho-Syntactic Features

Anonymous ACL submission

Abstract

We present a novel metric for the evaluation of morphological plausibility of subword segmentation. Unlike the typically used morpheme boundary or retrieval F-score, which requires gold segmentation data that is either unavailable or of inconsistent quality across many languages, our approach utilizes morpho-syntactic features. These are available in resources such as Universal Dependencies or UniMorph for a much wider range of languages. The metric works by probabilistically aligning subwords with morphological features through an IBM Model 1. Our experiments show that the metric correlates well with traditional morpheme boundary recall while being more broadly applicable across languages with different morphological systems.

1 Introduction

Subword tokenization is a fundamental preprocessing step in modern NLP systems. When evaluating tokenizers, researchers consider both extrinsic metrics (downstream task performance) and intrinsic properties, including morphological plausibility—how well tokenization aligns with morphological segmentation (Uzan et al., 2024; Libovický and Helcl, 2024; Arnett and Bergen, 2025)—alongside statistical measures like compression ratio and vocabulary coverage (Limisiewicz et al., 2023; Zouhar et al., 2023; Schmidt et al., 2024).

Evaluating morphological plausibility faces challenges due to unavailable or inconsistent gold standard segmentation data, which biases experiments toward high-resource languages. Even among these languages, cross-dataset inconsistencies exist. For example, the German word *Wolkenkratzer* ("skyscraper") is segmented as *Wolke* + *n* + *kratz* + *er* in German-CELEX (Gulikers et al., 1995) but left unsegmented in German MorphyNet (Batsuren et al., 2021), raising concerns about evaluation validity.

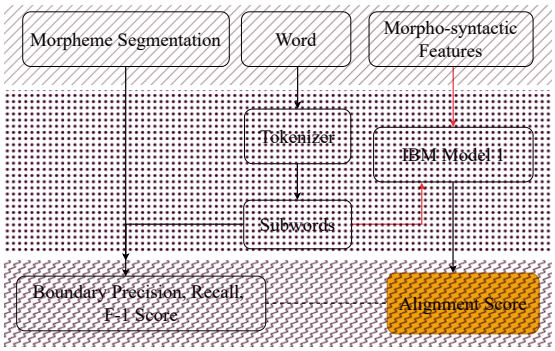

Figure 1: The workflow subword tokenization evaluation for morphological plausibility: Instead of comparing subwords with morpheme segmentation, we align morpho-syntactic features with subwods using IBM Model 1 to compute our proposed alignment score.

Additionally, some tokenizers employ non-concatenative approaches, using latent variables (Samuel and Øvrelid, 2023), incomplete text coverage (Hofmann et al., 2022), or additional tags for casing and diacritics (Popel et al., 2022; Forsythe, 2023). For these approaches, traditional morpheme boundary metrics are not well-defined.

We address these limitations with a novel metric that evaluates morphological plausibility without gold segmentation data, instead utilizing morpho-syntactic features from resources like UniMorph (Batsuren et al., 2022b), available for 169 languages. These features enable cross-lingual comparison through a language-independent schema, such as the Dutch word *adviseren* ("to advise") tagged as `V;SBJV;PRS;PL`.

Our method uses IBM Model 1 (Brown et al., 1993) to align morphological features with subword tokens, which is applicable to both concatenative and non-concatenative tokenization schemes. We extract alignment probabilities between subword tokens and morpho-syntactic features and aggregate them into a single measure. The metric shows a strong correlation with traditional boundary re-

Word	Segmentation	Morpho-syntactic Feature
rýžový	rýžlovlý	ADJ;ACC;MASC;INAN;SG
bázeň	bázleň	N;ACC;SG;FEM
projet	proljelt	V;V.PTCP;MASC;PASS;SG

Table 1: The curated dataset by combining the segmentation from Universal Segmentations and morpho-syntactic features from UniMorph.

call across diverse languages. The alignment-based approach rests on the principle that morphemes and morphological features are inherently linked. Well-segmented subwords should capture units that consistently express particular grammatical functions, leading to strong feature-subword alignments. Poor segmentation produces arbitrary character sequences that align weakly with any specific features, resulting in lower scores.

We conceptualize the workflow as illustrated in Figure 1. The data curation block deals with collecting datasets with morpheme segmentation and morpho-syntactic features mapped to the word forms. We train tokenizers and the IBM Model 1 using the created data structure (Table 1) and propose an alignment score that quantifies the morphological plausibility. The datasets along with the experimental codes[1] are also released.

2 Related Work

Most frequently used subword tokenizers are trained with a statistical heuristic, such as greedily shortening the training corpus (Sennrich et al., 2016) or minimizing negative log likelihood of the training data in a unigram model (Kudo, 2018). The properties of the resulting tokenizer depend not only on the algorithm itself but also, to a large extent, on data preprocessing and the languages in the training data mix.

Intrinsic evaluation of tokenizers includes information-theoretical properties, such as compression ratio or Renyi efficiency (Zouhar et al., 2023). In multilingual setups, the evaluation can include vocabulary allocation for different languages (Limisiewicz et al., 2023) or literal and semantic token overlap between languages (Hämmerl et al., 2025).

Morphological qualities of word segmentation are usually evaluated either via morpheme precision and recall in more linguistic contexts (Batsuren et al., 2022a) or via morpheme boundary precision and recall in the context of subword segmentation (Uzan et al., 2024; Libovický and Helcl, 2024).

IBM Models for word alignment in statistical machine translation were previously shown to be able to discover the relationship between morphemes and morpho-syntactic features (Stephen et al., 2024) and can be used for unsupervised extraction of morphological categories for morphemes. These results indicate that the alignment probabilities might be a good indicator of the morphological quality of subword segmentation.

3 The Alignment Score

Our metric uses IBM Model 1 (Brown et al., 1993) to establish probabilistic alignments between subword tokens and morphological features. It operates as an expectation-maximization algorithm that learns translation probabilities between source and target elements by iteratively maximizing the likelihood of observing the target given the source, without requiring any initial alignment. In our context, it discovers the probability distribution $P(f \mid s)$ between subword s tokens and morphological features f by treating each subword-feature pair as a potential alignment, then converging toward alignments that best explain the observed co-occurrences in the data.

The morphological plausibility score for tokenizer T

$$\frac{1}{|W|} \sum_{w \in W} \frac{1}{|S_w|} \sum_{s \in S_w} \mathrm{agg}_{f \in F_w} P(f \mid s) \quad (1)$$

where W is our corpus, F_w are features for word w, S_w are subwords of w, agg is a function that aggregates the probability scores for a single subword. We implement different aggregation functions: maximum, minimum, sum of the probabilities, sum of their logarithms, and mean.

To eliminate noisy alignments, we only consider probabilities $P(f \mid s)$ over a certain threshold, which is a hyperparameter of our method.

4 Experiments

We validate the proposed metric by measuring the Spearman correlation of the proposed metric and the established way of measuring morphological plausibility via morpheme boundary precision, recall, and F_1 score.

[1] https://anonymous.4open.science/r/morph-tok-eval-C27B/

Type	Aggregate	Boundary Precision										Boundary Recall									
		cs	de	en	fi	hr	hu	hy	kn	nl	sk	cs	de	en	fi	hr	hu	hy	kn	nl	sk
Joint	Sum	.06	-.28	-.67	-.14	-.13	.71	.32	-.65	-.24	.30	.94	.84	.67	.78	.74	-.04	.73	.91	.80	.71
	Log	.27	.72	.79	-.15	-.35	.56	-.08	-.78	.81	.16	-.67	-.84	-.78	-.69	-.29	-.38	.49	.99	-.95	.86
	Mean	.05	-.35	-.70	-.17	-.13	.60	.32	-.65	-.24	-.32	.94	.87	.71	.82	.74	.11	.73	.91	.80	.70
	Min	.04	-.36	-.70	-.15	-.13	.42	.32	-.65	-.24	-.32	.94	.88	.71	.95	.74	.30	.73	.91	.80	.70
	Max	.05	-.35	-.70	-.13	-.13	.61	.32	-.65	-.24	-.33	.94	.87	.71	.79	.74	.11	.73	.91	.80	.70
Split	Sum	-.03	-.31	-.70	-.15	.02	.60	.15	-.67	-.17	-.35	.98	.87	.71	.82	.60	.11	.81	.95	.72	.71
	Log	.21	.60	.79	.09	-.55	.65	-.29	-.71	.15	.44	-.96	-.88	-.78	-.84	-.27	-.63	-.64	.91	-.60	-.84
	Mean	-.04	-.34	-.72	-.11	-.16	.58	.12	-.67	-.16	-.34	.98	.88	.72	.91	.81	.15	.83	.96	.72	.72
	Min	-.05	-.36	-.72	-.02	-.00	.43	-.05	-.63	-.14	-.33	.98	.89	.72	.94	.82	.27	.88	.90	.70	.73
	Max	-.04	-.35	-.71	-.10	-.19	.60	.14	-.66	-.16	-.34	.98	.88	.71	.80	.76	.11	.82	.97	.71	72

Table 2: The Spearman correlation of our metric scores using different aggregates over full and split tags across languages. The threshold of the alignment probabilities obtained through IBM Model 1 is 0.01.

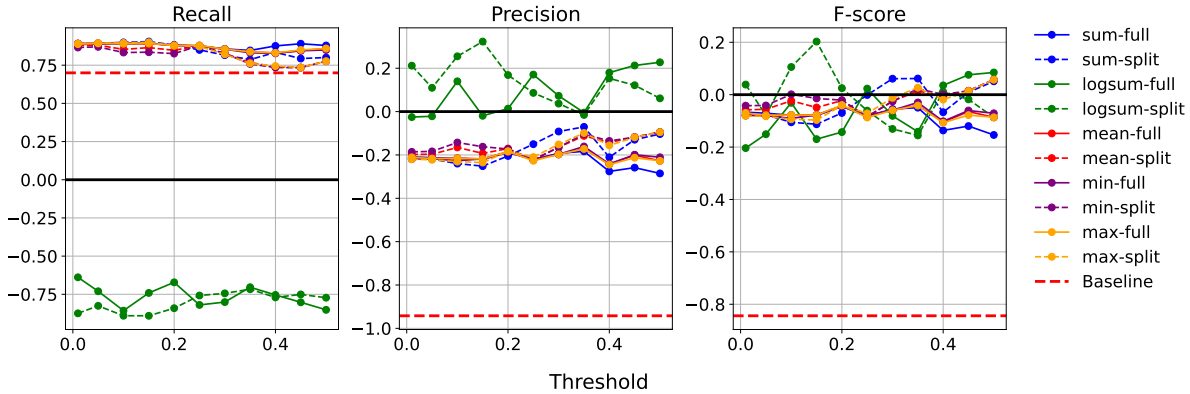

Figure 2: The Spearman correlation of our metric scores with boundary recall, precision, and F_1 score for Finnish. Plots for other languages are in the Appendix.

4.1 Linguistic Resources

We use segmentation resources from Universal Segmentations (Žabokrtský et al., 2022), a data resource that houses harmonized datasets for morphological segmentation. The data resources used are Armenian-MorphyNet (hy), Finnish-MorphyNet (fi), Kannada-KCIS (kn), English-CELEX (en), German-CELEX (de), Dutch-CELEX (nl), Czech-DeriNet (cs), Serbo-Croatian-MorphyNet (hr), and Hungarian-MorphyNet (hu). Additionally, we use the Slovak (sk) dataset by Ološtiak et al. (2015).

We use UniMorph (Batsuren et al., 2022b) for extracting the morpho-syntactic feature tags for the word forms.

We test two approaches to feature representation: "Joint" tags containing complete morphological information, and "Split" tags where composite features are separated into atomic units (e.g., 'V' for verb and 'SG' for singular are considered separate units). Even though considering the morpho-syntactic features jointly makes more sense linguistically, treating them independently provides a richer training signal for estimating probabilities with the IBM model.

The data for training the tokenizers and the metric was created by mapping the word forms along with their morphological segmentation with UniMorph feature tags.

4.2 Subword Tokenizers

We evaluate several standard subword tokenizers: Byte-Pair-Encoding (Sennrich et al., 2016), WordPiece (Schuster and Nakajima, 2012), and the Unigram model (Kudo, 2018). We train the tokenizers using 1M sentences from the CC100 corpus (Wenzek et al., 2020) for the respective languages with vocabulary sizes of 2k, 4k, 8k, 16k, 24k, 32k, 40k, 48k, 56k, 64k, 72k, and 80k. Additionally, we add character segmentation and gold morphological segmentation to the correlation study.

We segment the UniMorph vocabulary using the tokenizers and run the IBM model for 10 epochs, which is enough for convergence. We evaluate the metric with 11 thresholds between 0.01 and 0.5.

4.3 Results

We measure the correlation between the proposed morphological plausibility metric and traditional boundary-based metrics across 10 languages with different morphological structures. Table 2 presents the Spearman correlation coefficients between our alignment-based scores and boundary precision/recall measures.

Correlation with Boundary Metrics. Our alignment-based metric correlates strongly with traditional boundary recall measures across the evaluated languages. Most languages show high positive correlations (> 0.70) when using Sum, Mean, Min, and Max aggregation functions (See Appendix A). Czech (cs) exhibits particularly strong correlations, reaching 0.94-0.98 for boundary recall. Correlations with boundary precision vary more, with several negative correlations observed, except for Hungarian (hu), suggesting our metric aligns more closely with the recall aspect of morphological segmentation, rather than precision, which measures over-segmentation for almost all the languages.

Impact of Feature Representation. Split representation generally produces stronger correlations with boundary recall, particularly for morphologically complex languages like Finnish (fi) and Armenian (hy). For Finnish (compare Table 2 and Figure 2), the correlation increases from 0.79 (Joint-Mean) to 0.91 (Split-Mean). Although, for the Hungarian dataset we find stronger correlations with boundary precision.

Aggregation Function Analysis. Sum, Mean, Min, and Max aggregation functions show similar performance patterns, with strong positive correlations with boundary recall. The Log aggregation function often yields inverse correlations (e.g., -0.96 for Czech, -0.88 for German). The Mean aggregation function performs consistently across languages, with an average correlation of 0.86 with boundary recall when using split features.

Cross-lingual Observations. Performance varies across language families. The three Germanic languages (German, English, and Dutch) exhibit consistent correlation patterns, while Finnish and Kannada show some of the strongest correlations with the Min aggregation function (0.94 and 0.90 respectively). Czech and Slovak correlate well with boundary recall but show more variable relationships with precision.

5 Discussion

The metric's inherent tendency to penalize over-segmentation can be seen through the weak correlation with the boundary precision scores. However, we observe consistent patterns of correlations across languages, which provides a strong signal for the metric's cross-lingual viability. The metric is also able to capture the distinct morphology of languages well. Split representations have higher correlations with boundary recall (especially Mean and Max functions) for Armenian, Czech, Finnish, Kannada, Serbo-Croatian, and Slovak, illustrating that the metric is sensitive towards agglutination and allomorphy, Hungarian being the only exception. This is additionally supported by almost identical results for Joint and Split representations for English and German, where we find a weaker fusional morphology.

The performance of the metric is constrained by the granularity of morpheme segmentation and the alignment between morphemes and morpho-syntactic features. Since most morphemes align to features on a continuous rather than binary scale (even with thresholds), feature assignment lacks clear boundaries, reducing precision.

6 Conclusion

In this paper, we propose a new metric to assess the morphological plausibility of subword segmentation, addressing limitations of traditional evaluation metrics. We use datasets where the morpho-syntactic features of the word forms are also mapped, along with their morpheme segmentation. The morpho-syntactic features are taken from UniMorph, and the gold morpheme segmentation is accessed from Universal Segmentations. We train our tokenizers with varying vocabulary sizes using Byte-Pair-Encoding, WordPiece, and the Unigram model. The resulting subword tokens per word form are, respectively, aligned with the morpho-syntactic features using the IBM Model 1. The resulting alignment probabilities are aggregated in multiple ways to obtain a statistically robust array of results. These results are correlated with the traditional boundary precision, recall, and F_1 score obtained by comparing the tokenizer outputs with the gold morpheme segmentation data in hand.

Our proposed metric correlates strongly with boundary recall across languages with varied morphological systems, making it a capable contender for testing morphological plausibility.

Limitations

The current stream of experiments is correlated with the existing gold segmentation data, which could potentially carry some inherent flaws. We do not run any checks whatsoever to quantify the accuracy of the gold segmentation data.

References

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022a. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, and 76 others. 2022b. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Alasdair Forsythe. 2023. Tokenmonster. `https://github.com/alasdairforsythe/tokenmonster`.

Léon Gulikers, Gilbert Rattink, and Richard Piepenbrock. 1995. German linguistic guide. *The CELEX Lexical Database (CD-ROM). Linguistic Data Consortium, Philadelphia, PA*.

Katharina Hämmerl, Tomasz Limisiewicz, Jindřich Libovický, and Alexander Fraser. 2025. Beyond literal token overlap: Token alignability for multilinguality. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 756–767, Albuquerque, New Mexico. Association for Computational Linguistics.

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393, Dublin, Ireland. Association for Computational Linguistics.

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.

Jindřich Libovický and Jindřich Helcl. 2024. Lexically grounded subword segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7420, Miami, Florida, USA. Association for Computational Linguistics.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681, Toronto, Canada. Association for Computational Linguistics.

Martin Ološtiak, Ján Genči, and Soňa Rešovská. 2015. *Retrográdny morfematický slovník slovenčiny*. Filozofická fakulta Prešovskej univerzity v Prešove.

Martin Popel, Jindřich Libovický, and Jindřich Helcl. 2022. CUNI systems for the WMT 22 Czech-Ukrainian translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 352–357, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

David Samuel and Lilja Øvrelid. 2023. Tokenization with factorized subword encoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14143–14161, Toronto, Canada. Association for Computational Linguistics.

Craig W Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. 2024. Tokenization is more than compression. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 678–702, Miami, Florida, USA. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152. IEEE.

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.

Abishek Stephen, Vojtěch John, and Zdeněk Žabokrtský. 2024. Unsupervised extraction of morphological categories for morphemes. In *Text, Speech, and Dialogue*, pages 239–251, Cham. Springer Nature Switzerland.

Omri Uzan, Craig W. Schmidt, Chris Tanner, and Yuval Pinter. 2024. Greed is all you need: An evaluation of tokenizer inference methods. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 813–822, Bangkok, Thailand. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Zdeněk Žabokrtský, Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Magda Ševčíková, and Jonáš Vidra. 2022. Towards universal segmentations: UniSegments 1.0. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1137–1149, Marseille, France. European Language Resources Association.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

A Appendix: Detailed Results

Figure 3 shows the correlation of the alignment score with boundary recall for different thresholds for 9 languages, Figure 4 shows the correlation with boundary recall, and Figure 5 shows the correlation with F_1 score. We present the results for Hungarian in Figure 6.

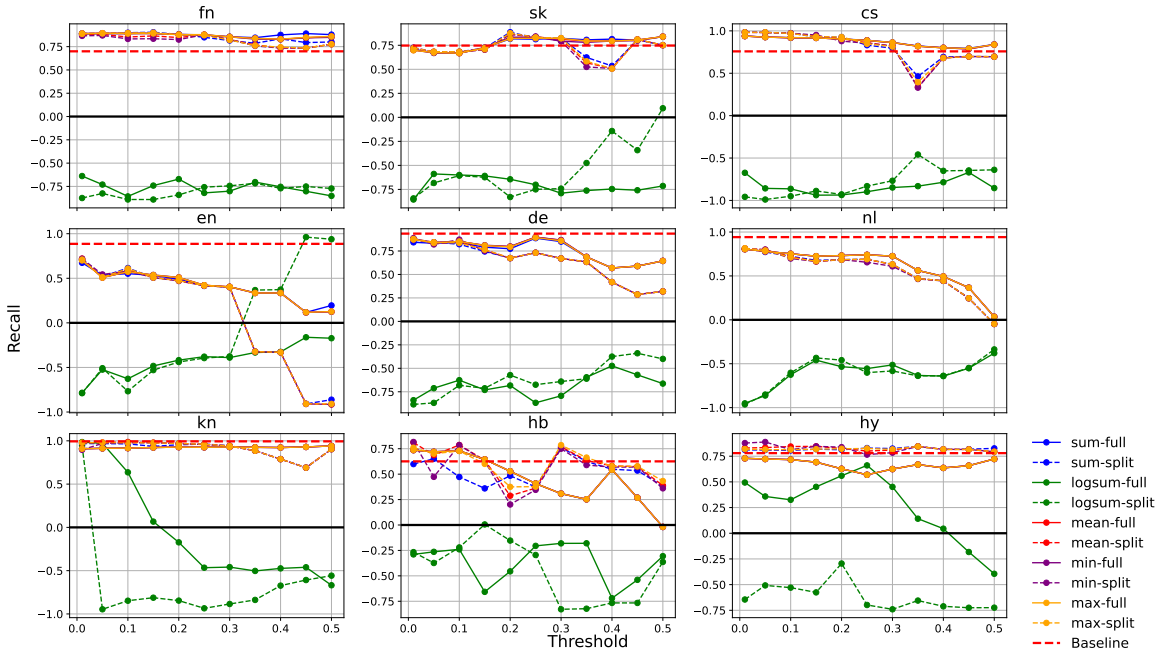

Figure 3: The correlation of the alignment-based score with boundary recall for all languages.

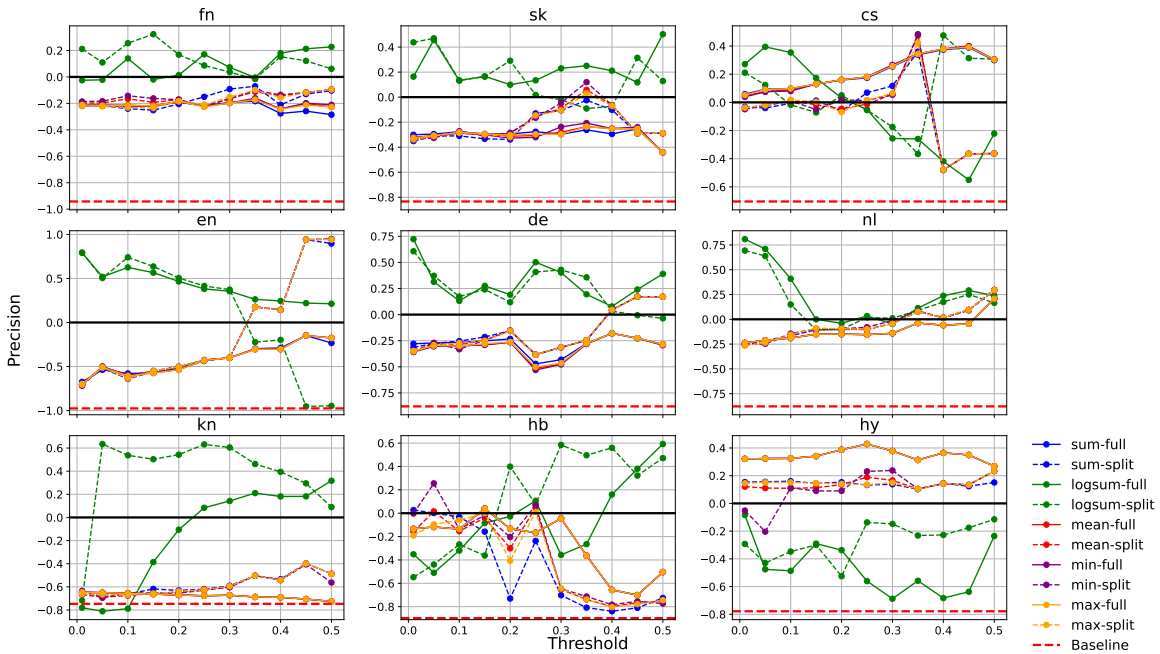

Figure 4: The correlation of the alignment-based score with boundary precision for all languages.

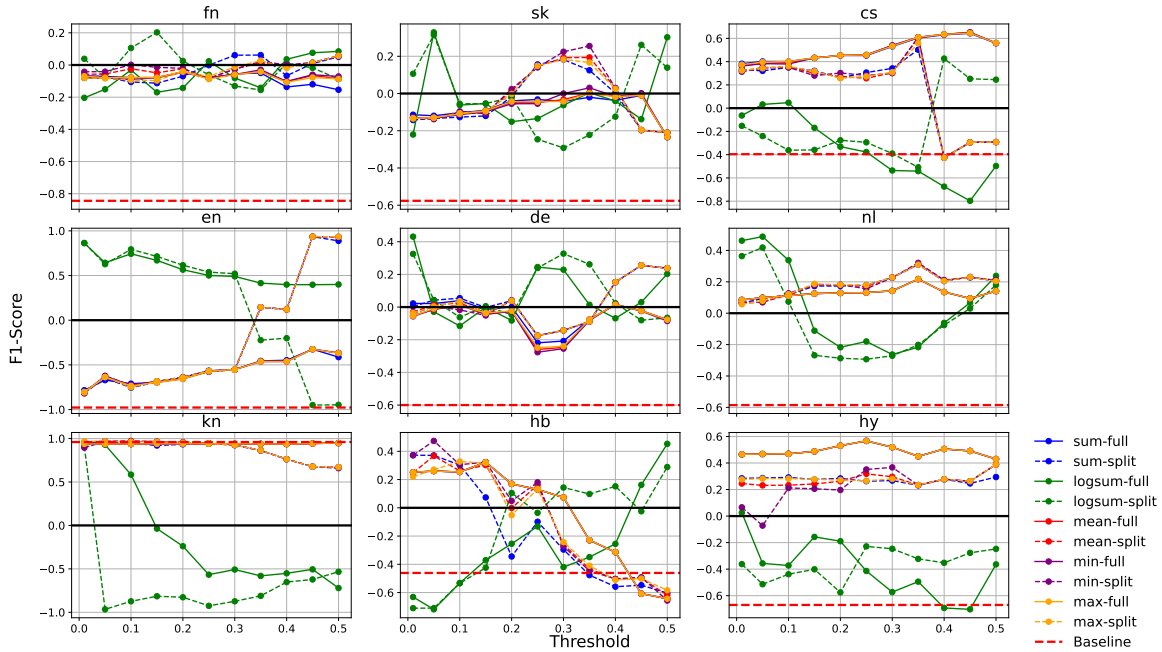

Figure 5: The correlation of the alignment-based score with F_1 score for all languages.

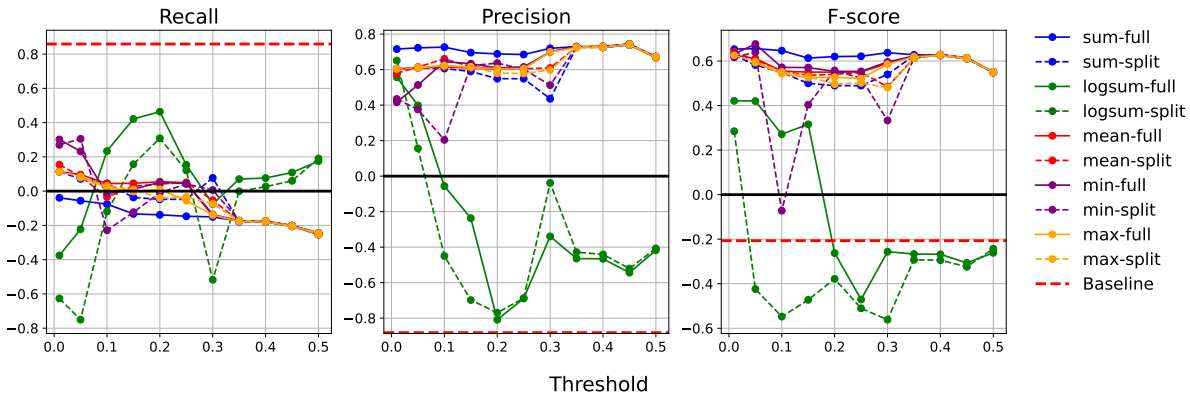

Figure 6: The Spearman correlation of our metric scores with boundary recall, precision, and F_1 score for Hungarian.