

Entropic properties of languages in phonetic versus orthographic alphabets

Carla Olabe  and Christian Bentz 

Language Science and Technology
Saarland University, Germany

carla.olabe-rodriguez@uni-saarland.de, chris@christianbentz.de

Abstract

Computational linguistics predominantly operates on written text. However, writing represents spoken language only indirectly. Whether phonetic and orthographic representations share fundamental information-theoretic properties is an open question. We present a cross-linguistic analysis of Shannon entropy across modalities using the DoReCo corpus featuring 42 languages of 25 families after preprocessing. By computing continuous entropy trajectories at orthographic word, subword (BPE), and character levels (UTF-8/SAMPA), we reveal a systematic modality convergence: orthographic and phonetic encodings yield consistently small entropy differences across all three granularity levels. This general pattern, however, masks notable cross-linguistic variation—some languages, such as Evenki, converge tightly across all levels, while others show modality gaps at specific granularities—for instance, English at the character level, Bora at the word level, and Sanzhi Dargwa at the subword level. In sum, Shannon entropy tends toward modality-invariance as a general tendency across the Latin-script languages examined, while leaving room for typologically conditioned divergence. Whether this pattern extends to non-alphabetic writing systems remains an open question.

 [properties-spoken-written-language](#)

1 Introduction

Language exists in multiple modalities, yet computational linguistics predominantly focuses on written text. This bias extends to fundamental assumptions: tokenization algorithms, language models, and information-theoretic analyses often treat orthographic representations as primary. Systematic comparisons of information-theoretic properties across spoken and written modalities remain scarce (Coupé et al., 2019; Petrini et al., 2023), particu-

larly at the token level where these properties can be directly measured.

We ask the following: Do orthographic and phonetic encodings exhibit equivalent information-theoretic profiles? Using the DoReCo corpus in Cross-Linguistic Data Format (CLDF)—providing orthographic and phonetic transcriptions for 42 diverse languages across the world—we compute Shannon entropy trajectories at three granularities: *words* (orthographic words vs. phone words), *subwords* (BPE-tokenized), and *characters* (UTF-8 characters vs. SAMPA phones).

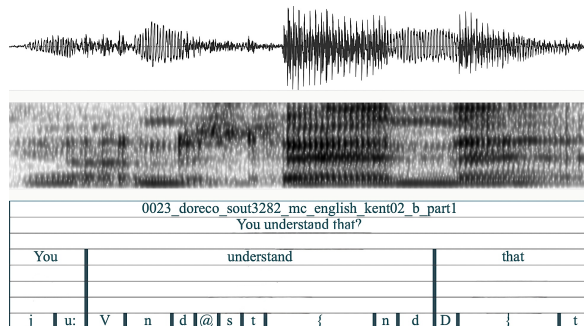


Figure 1: Praat TextGrid annotation of an English (Southern England) utterance (“You understand that?”) from the DoReCo corpus (Seifart et al., 2024), showing the waveform, spectrogram, orthographic transcription, and SAMPA phone tier (Boersma and Weenink, 2026).

Our analysis reveals two complementary patterns. First, a general modality convergence: orthographic and phonetic entropy trajectories remain close across all three granularity levels, with the smallest average gap at the subword level. Second, systematic cross-linguistic variation in the size and location of this gap: languages such as Evenki (even1259) and Jahai (jeha1242) show close convergence throughout, while languages such as Southern English (sout3282), Bora (bora1263), and Sanzhi Dargwa (sanz1248) exhibit larger modality gaps at specific granularities. The typological properties of languages and scripts shape

DoReCo Languages Overview by Family

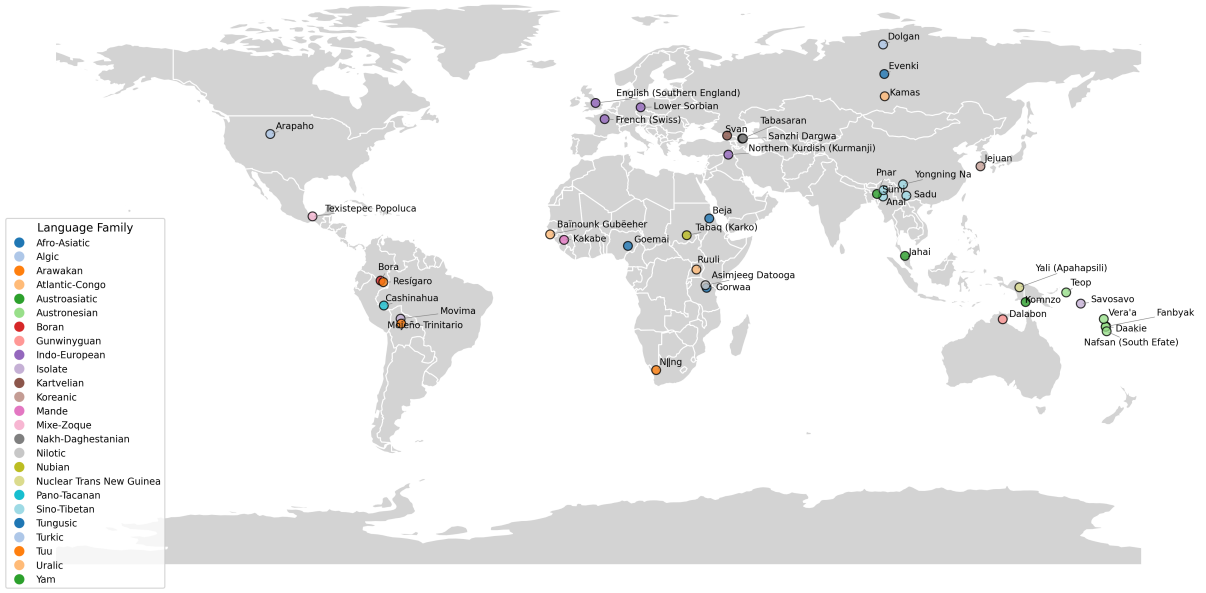


Figure 2: Geographic distribution of the languages in our DoReCo sample, coloured by language family. The sample spans diverse genealogical and typological groups.

how strongly and at which granularity level modalities diverge.

2 Data and Methods

2.1 The DoReCo Corpus

We use the DoReCo (Documentary Reusable Corpus) dataset (Seifart et al., 2022) in its CLDF representation version 1.2.1 (Seifart et al., 2024), a collection of naturalistic speech recordings with orthographic and phonetic transcriptions for 51 typologically diverse languages (of which 42 were retained after preprocessing in this study). The corpus covers languages from overall 25 families according to the Glottolog classification (Hammarström et al., 2026), including Indo-European, Niger-Congo, Austronesian, Sino-Tibetan, Turkic, Trans-New Guinea, among others. This also includes two language isolates: Savosavo (savo1255) and Movima (movi1243). Each recording comes with word-level orthographic transcriptions and phone-level transcriptions with precise timestamps, providing the cross-modal data required for information-theoretic comparisons. Figure 1 shows an example utterance from the English (Southern England) recording (“*You understand that?*”), illustrating the alignment between the needed orthographic and SAMPA phone tiers that we study in this paper, as provided in DoReCo. Several DoReCo languages feature recordings with multiple speakers. Words

from all speakers are included in the token streams, ordered by sequential word ID within each recording session. We access the corpus through its CLDF representation, which provides structured tables for words and phones with associated metadata. Per-language word and phone token counts vary considerably across the sample, reflecting differences in recording duration and speaker participation across DoReCo datasets. Full corpus size statistics for all 42 languages, including word token counts, word type counts, phone token counts, and phone type counts, are provided in Table 4 in the Appendix.

2.2 Encoding levels and preprocessing

We analyze each language at three levels of granularity, further divided into both orthographic and phonetic encodings:

Word level. Orthographic words as written in the corpus transcriptions versus phone words (where the full phone sequence of each word is treated as a single token).

Subword level. Byte-Pair Encoding (BPE) (Sennrich et al., 2016) is applied separately to orthographic and phonetic streams. We train language-specific BPE tokenizers with vocabulary size 200, following Gutierrez-Vasques et al. (2021). The sensitivity of our results to this choice is assessed in Section 3.

Modality	Encoding	Granularity	Word 1	Word 2	Word 3
<i>English (Southern England) (Schiborr, 2022)</i>					
written	UTF-8	words	you	understand	that
written	UTF-8	chars	y o u	u n d e r s t a n d	t h a t
written	UTF-8	subwords	Y ou	u n d e r s t a n d	t h a t
spoken	SAMPA	words	ju:	Vnd@st{nd	D{t
spoken	SAMPA	chars	j u:	V n d @ s t { n d	D { t
spoken	SAMPA	subwords	ju:	Vn d @ s t { n d	D { t
<i>Bora (Seifart, 2022a)</i>					
written	UTF-8	words	éée	Tsíñé	wákiméí
written	UTF-8	chars	é é e	T s í í ñ é	w á k i m é í
written	UTF-8	subwords	é é e	T s í í ñ é	w á k i m é í
spoken	SAMPA	words	E:E	ts_hi:JE	k_wak_himEi
spoken	SAMPA	chars	E: E	ts_h i: J E	k_w a k_h i m E i
spoken	SAMPA	subwords	E: E	ts_hi: JE	k_wa k_h i m E i
<i>Evenki (Kazakevich and Klyachko, 2022)</i>					
written	UTF-8	words	on	bi	ohot'ičāw
written	UTF-8	chars	o n	b i	o h o t' i č ā w
written	UTF-8	subwords	on	bi	o h o t' i č ā w
spoken	SAMPA	words	on	bi	ohot_jitSa:w
spoken	SAMPA	chars	o n	b i	o h o t_j i t S a: w
spoken	SAMPA	subwords	on	bi	o h o t_j i t S a: w
<i>Sanzhi Dargwa (Forker and Schiborr, 2022)</i>					
written	UTF-8	words	xuc?anu	xuraib	ducic:ew
written	UTF-8	chars	x u c ? a n u	x u r a i b	d u s i c : e w
written	UTF-8	subwords	x u c ? a n u	x u r a i b	d u s i c : e w
spoken	SAMPA	words	xuts_>anu	xuraib	duhit:sew
spoken	SAMPA	chars	x u t s _ > a n u	x u r a i b	d u s i t : s e w
spoken	SAMPA	subwords	x u t s _ > a n u	x u r a i b	d u s i t : s e w

Table 1: The six combinations of modality and granularity applied to example words from four typologically diverse DoReCo languages: English (Southern England), Bora (Boran language spoken in the western Amazon rainforest), Evenki (Tungusic language of Siberia), and Sanzhi Dargwa (Nakh-Daghestanian language of the Caucasus). UTF-8 rows reflect orthographic representations; SAMPA rows reflect X-SAMPA phonetic transcriptions. Subword boundaries (BPE) are learned from the full corpus for each language and independent of modality.

Character level. Individual UTF-8 characters in orthographic text versus individual SAMPA phones in phonetic transcriptions.

Table 1 illustrates the six encoding levels applied to example words drawn from four typologically diverse DoReCo languages. The table also highlights cases where orthographic and phonetic tokenizations deviate from a one-to-one mapping. In Bora, tonal distinctions marked by accent diacritics in the Latin script (e.g., í vs. i) are absent in the SAMPA transcription, where only vowel length is preserved—resulting in a systematically reduced phonetic type inventory. In Sanzhi Dargwa, ejective consonants are represented by a single raised glottal stop character in orthography (e.g., cʔ) but

require a two-character sequence in X-SAMPA (e.g., ts_>), producing longer phonetic tokens and driving divergence at the subword level. These encoding asymmetries directly motivate the cross-linguistic variation in modality gaps reported in Section 3.

Our preprocessing pipeline ensures that orthographic and phonetic streams are filtered and ordered consistently through three steps. First, we remove DoReCo annotation markers (e.g., <<tag>content>), pause tokens (<p:>, <p>, <sil>), and placeholder markers (****, [INT]) uniformly across all encoding levels. Second, we sort all tokens by word ID sequence number, which reflects the original corpus order and ensures each speaker’s contribution forms a contiguous block.

Words lacking phone transcriptions in DoReCo are excluded from all encoding streams to preserve cross-modal comparability.

We process tokens as continuous streams rather than fixed-size chunks, preserving the natural sequential structure of discourse and avoiding data loss from discarded remainder tokens; entropy trajectories thus reflect the change in information content over the full corpus in its original order.

For BPE training on phones, we concatenate phones without whitespace, allowing the algorithm to learn natural subword boundaries unconstrained by pre-existing segmentation. This approach enables diacritics and length markers to merge with base phones according to co-occurrence patterns. Importantly, BPE tokenization is applied to the same word ID-ordered sequences, ensuring that subword encodings maintain the same ordering as word-level and character level encodings.

2.3 Entropy estimation

We estimate Shannon entropy (Shannon, 1948) in bits based on the standard formula (Cover and Thomas, 2006, p. 14):

$$H(\mathcal{X}) = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \quad (1)$$

where $\mathcal{X}$ is the “alphabet”, i.e. the set of unique types, and $p(x)$ is the probability of a type x . In our study, types can be characters/phones, character subwords/phone subwords, and orthographic words/phone words. For approximating their probability we use the maximum likelihood or plug-in estimator:

$$\hat{p}(x_i) = \frac{f_i}{\sum_{i=1}^{|\mathcal{X}|} f_i}, \quad (2)$$

where f_i is the raw frequency count of a given type x_i . The raw frequency is normalized by the sum of all frequency counts for types in the “alphabet” to yield the relative frequency as a probability estimate.

More advanced entropy estimators apply smoothing techniques to reduce estimation bias (Hausser and Strimmer, 2009; Arora et al., 2022), and language models can estimate entropy by exploiting sequential context (Takahashi and Tanaka-Ishii, 2018; Cotterell et al., 2018; Mielke et al., 2019; Shannon, 1951). However, both approaches yield estimates highly correlated with the simpler plug-in estimator (Bentz et al., 2017; Koplenig et al., 2025). Since our objective is to compare entropy across

granularities and encodings rather than estimate an absolute “true” entropy, the plug-in estimator already captures the properties we need, without the overhead of more complex models.

Rather than computing entropy only at sequence endpoints, we calculate continuous entropy trajectories: At position i , we compute entropy based on all tokens from position 1 to i , tracking both the growing type inventory and shifting frequency distributions. This reveals the dynamics of entropy stabilisation.

2.4 Data & Code

The data and code for replicating these analyses is given in the provided GitHub repository. For the permanent version relevant for this paper see Zenodo.¹

3 Results

3.1 Small entropy differences across modalities

Entropy growth at the character and subword levels typically slows down rapidly within the first 500–1000 tokens, likely reflecting the smaller inventory sizes at these granularities. At the word level, trajectories continue rising more noticeably.

To see this, Figure 3 presents entropy trajectories for all six combinations of modality and granularity across our DoReCo language sample. We observe a consistently small modality gap in the smoothed trajectories across all three granularity levels, with some cross-linguistic variation in both the magnitude and location of this gap for particular languages.

At the word level, entropies for both orthographic words and phone words reach approx. 7–10 bits at the final token position. The two smoothed trajectories follow closely parallel paths—orthographic words and phone words differ by only a small margin on average—suggesting that written and spoken modalities encode similar amounts of information at the word level.

At the subword level, BPE tokenization reduces entropy to approximately 5–6 bits for both modalities. This demonstrates that compression through learned subword units operates similarly regardless of whether the input is graphemic or phonetic.

At the character level, UTF-8 characters and SAMPA phones both fall approx. in the 4–5 bit

¹Code and data available at <https://doi.org/10.5281/zenodo.22133437>

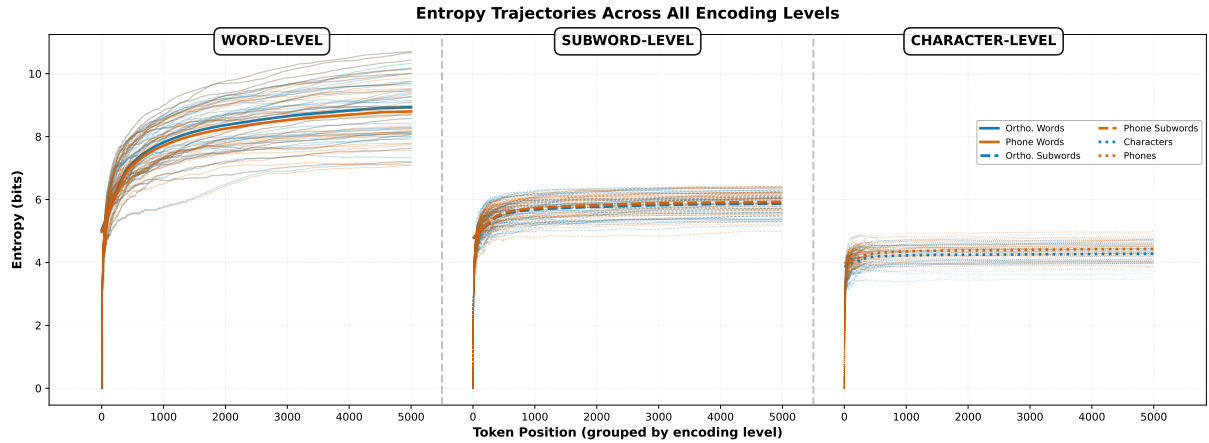


Figure 3: Shannon entropy trajectories across word-level, subword-level, and character level encodings. Each thin line represents a single language; bold lines show LOESS-smoothed cross-linguistic trends.

Gran.	Enc.	Mean H	SD	ΔH
word	written	9.347	1.015	0.152
	spoken	9.195	1.054	
subword (100 merges)	written	5.207	0.257	0.153
	spoken	5.360	0.240	
subword (200 merges)	written	5.955	0.361	0.022
	spoken	5.977	0.358	
subword (400 merges)	written	6.549	0.437	0.017
	spoken	6.532	0.423	
character	written	4.314	0.360	0.139
	spoken	4.453	0.289	

Table 2: Mean Shannon entropy (H , in bits), cross-linguistic standard deviation (SD), and modality gap ($\Delta H = |\text{mean written } H - \text{mean spoken } H|$) across 42 DoReCo languages, for each combination of granularity and encoding. Entropy is taken at the final token position of each language’s trajectory; SD is the sample standard deviation across languages. For the subword level, results are shown for three BPE vocabulary sizes (100, 200, and 400 merges); 200 is the default setting used elsewhere in the paper.

range. This is the finest granularity we examine, where differences in inventory size between different Latin scripts and phone inventories are expected to produce divergent entropy values.

Table 2 reports the mean entropies, cross-linguistic standard deviations, and modality gaps (ΔH , the absolute difference between mean written and mean spoken entropy) at each encoding level, computed as the entropy value at the final token position of each trajectory, averaged across all 42 languages. For the subword level, we additionally report results across three BPE vocabulary sizes (100, 200, and 400 merges) to assess the sensitivity of our entropy estimates to this parameter; unless stated otherwise, the main text refers to the 200-merge setting. These values are lower bounds, as

language entropy keeps growing—though at decreasing rates. Two patterns stand out.

First, the modality gap is consistently small across all three granularity levels: written and spoken encodings differ by 0.152 bits at the word level (9.347 vs. 9.195), by only 0.022 bits at the subword level (5.955 vs. 5.977) at 200 merges, and by 0.139 bits at the character level (4.314 vs. 4.453). Hence, on average, the subword level shows the smallest gap. We additionally tested vocabulary sizes of 100 and 400 merges to assess this sensitivity. As shown in Table 2, the mean entropy values increase with vocabulary size, as expected from a larger merge budget, but the modality gap remains consistently small across all three settings (0.153 bits at 100 merges, 0.022 bits at 200 merges, and 0.017 bits at 400 merges). This confirms that the modality convergence observed at the subword level is robust to the choice of BPE vocabulary size.

Second, cross-linguistic standard deviation (SD) drops sharply from the word level to the subword and character levels, indicating that finer granularities behave more consistently across typologically diverse languages.

These patterns hold regardless of the varying corpus sizes across our sample (Table 4). Figure 5, 6 and 7 in the Appendix plot entropy trajectories for all six encoding levels for each language individually, using a 5,000-token window that matches the size of the smallest corpora in our sample (Dalabon, 4,128 word tokens; Tabasaran, 5,059 word tokens); this window was chosen to allow direct cross-linguistic comparison over a common, data-supported range rather than as an arbitrary truncation. Within this range, character- and subword-

Entropy Trajectories: All Encoding Levels per Language

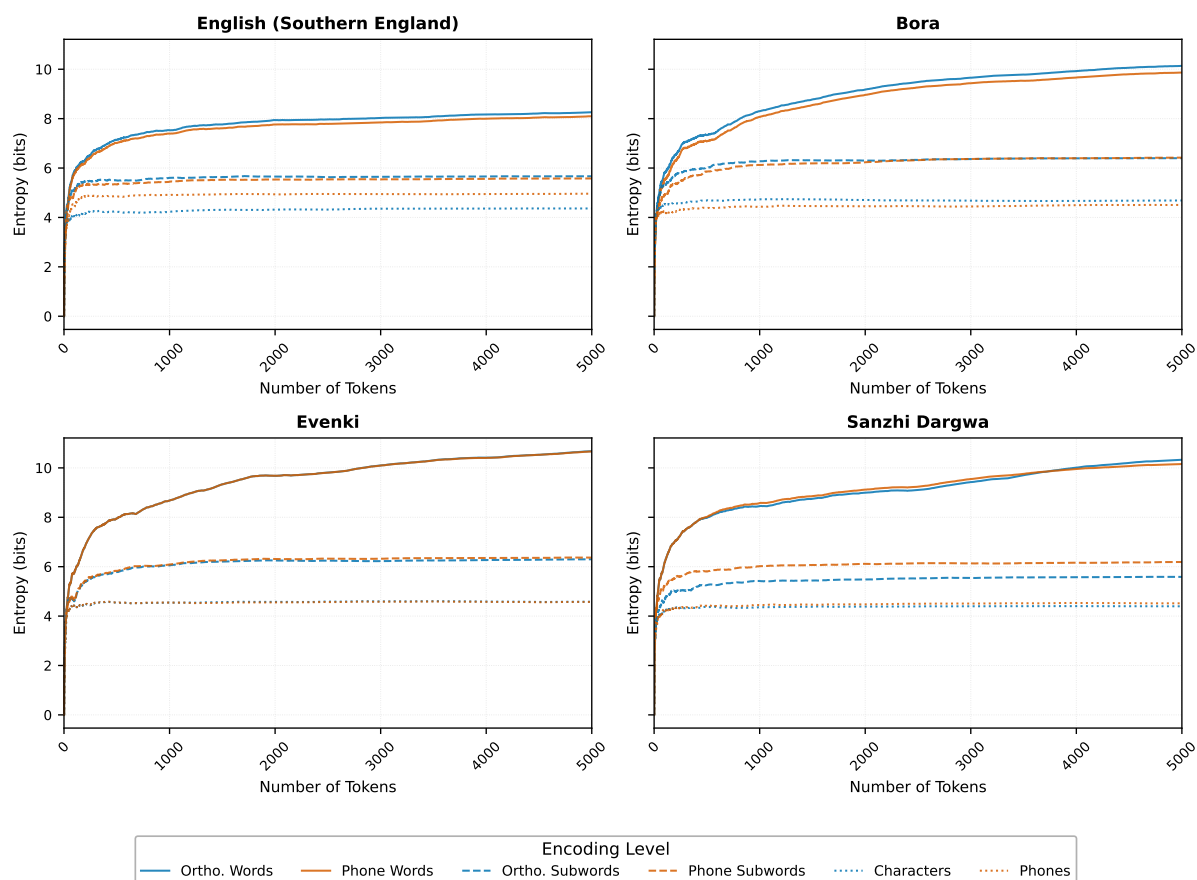


Figure 4: Entropy trajectories for all six encoding levels across four representative languages: English (Southern England), Bora, Evenki, and Sanzhi Dargwa. Each panel shows one language with all encodings overlaid, illustrating both the general tendency toward modality convergence and cross-linguistic variation in the size and location of the modality gap or convergence.

level trajectories plateau within the first 500–1,000 tokens for every language in the sample. Word-level entropy follows the same qualitative growth dynamic across languages—an initial steep rise as new word types are encountered, followed by a decelerating increase as recurring words come to dominate the stream.

3.2 Cross-linguistic entropy differences

While the patterns in Figure 3 demonstrate a general modality invariance of entropy, zooming in on individual languages reveals interesting variation in both entropy magnitudes and entropy differences between modalities. Figure 4 presents entropy trajectories for the subset of four languages also given in Table 1.

Evenki shows the closest entropy similarity across modalities, with trajectory pairs nearly identical at all three granularities. Evenki is a Tungusic

language spoken in Siberia. Since the late 1930s, it is by default written with a Cyrillic alphabet (Bulatova and Grenoble, 1999, p. 6). However, grammars of Evenki (cf. Bulatova and Grenoble, 1999) might use a Latin transliteration of the spoken language, and this is also used by the DoReCo corpus at the word tier (Kazakevich and Klyachko, 2022), as in the example below:

Evenki (even1259)²

- (1) tx: он би оһотичав
 wd: on bi ohot'ičāw
 ph: o n b i o h o t _ j i t S a: w
 ft: “The way I hunted.”

The orthographic text (tx) is written in Cyrillic alphabet; our analyses use the word (wd) and phone (ph) tiers, since the text tier is not tokenized in

²File: doreco_even1259_2007_Mutoray_Dmitriev_L01

DoReCo.

English (Schiborr, 2022), on the other hand, displays similar entropy at the word and subword levels but shows a visible gap at the character level, where SAMPA phones yield higher entropy than UTF-8 characters. English orthography is less precise about distinguishing between different sounds than SAMPA phonetic transcriptions. For instance, the UTF-8 character string <ea> is variably mapped onto [I@] as in <years> vs. [jI@z], [e] as in <meadow> vs. [med@U], [i:] as in <team> vs. [ti:m], etc. This leads to more information in the phone stream, i.e. higher entropy. However, interestingly, this difference is neutralized at the subword level due to early merging of frequently co-occurring characters such as <ea>.

Bora (Seifart, 2022a) also features modality gaps, especially at the word and the character granularity, while at the subword level the entropies are near indistinguishable after approx. 2000 tokens. Here, the gaps are due to higher entropy in the written modality compared to the spoken modality.

Bora (bora1263)³

- (2) tx: éée Tsííñé wákiméí
wd: éée Tsííñé wákiméí
ph: E: E ts_h i: J E k_w a k_h i m E i
ft: “Another job to do.”

Bora is a tonal language of the western Amazon rainforest. It features two tones which are indicated in the Latin script used to write the language by an acute accent, e.g., <á> for high tone, and the lack of an accent, e.g., <a> for low tone (Thiesen and Weber, 2012, p. 27), as in (2). While tone differences are indicated in the Latin script (tx and wd), e.g. <í> vs. <i>, these are dropped in the phonetic transcription—only vowel length is preserved by usage of the colon as a length marker, e.g., [i:]. Hence, the tone information is systematically lost in the phonetic transcription, and the entropy decreases compared to the Latin script.

Sanzhi Dargwa (Forker and Schiborr, 2022) shows yet another pattern: words and characters have similar entropies, but its subword encodings yield a substantially larger modality gap, with phone subwords having higher entropy than orthographic subwords. Sanzhi Dargwa is a Nakh-Dagestanian language of the Eastern Caucasus. It has a large inventory of more than 50 consonant phonemes (cf. Forker, 2020, p. 2). The respective

phones and phonemes have to be represented by relatively long strings of characters in X-SAMPA transcription and in orthographic writing respectively. For instance, ejectives are represented by a raised glottal stop sign <ʔ> in orthography (wd tier) but by a combination of underscore and larger than sign [_>] in X-SAMPA (ph tier), compare <qʔ> vs. [q_>], <kʔ> vs. [k_>], <cʔ> vs. [ts_>], etc. These encoding differences seem to drive the modality gap at the subword level.

4 Discussion

Our results reveal small modality gaps – spoken vs. written – across orthographic and phonetic encodings at different granularities: characters, subwords, words. These are average trends across 42 typologically diverse languages examined. Despite small gaps on average – the entropy differences vary cross-linguistically, pointing toward typological properties as a modulating factor.

4.1 Implications for subword tokenization

Mielke et al. (2021) review the landscape of tokenization approaches from word to character level and conclude that no single tokenization strategy is optimal for all applications. At least for the question of how subword tokenization fares with regards to different modalities—spoken vs. written—our results show that on average the differences are very small.

Byte-Pair Encoding (Sennrich et al., 2016), now a standard tokenization approach in multilingual NLP pipelines (Mielke et al., 2021; Gezmu and Nürnberger, 2023), iteratively merges the most frequent adjacent token pairs to build subword vocabularies. BPE-tokenized orthographic and phonetic sequences stabilize at similar entropy values (5–6 bits) and display the smallest modality gap. This proves that the algorithm discovers comparable compression strategies regardless of whether the input is graphemic or phonetic—though as the Sanzhi Dargwa case illustrates—there can still be divergence also at the subword level for particular languages and granularities.

The small modality gap at the subword-level is particularly interesting given recent findings on how BPE interacts with morphological typology. Gutierrez-Vasques et al. (2021, 2023) demonstrate that BPE exhibits a “turning point”—at a relatively small number of merges (around 200–350)—where cross-linguistic entropy values converge to their

³File: doreco_bora1263_aakityeJP1

most similar distributions. Their analysis of 47 typologically diverse languages shows that before this turning point, subword-level complexity is uncorrelated with word-level morphological complexity, but after approximately 200 merges the two begin to align as BPE captures increasingly morpheme-like units. Our subword analysis falls within the range of this turning point, and we observe a corresponding cross-modal similarity on average: orthographic and phonetic encodings yield similar entropy values ($\approx 5\text{--}6$ bits) within each language at this compression level.

Other recent work explores information-theoretic approaches to improve tokenization itself (Zouhar et al., 2023; Hu et al., 2025), for instance using pointwise mutual information and boundary entropy to guide BPE toward more linguistically motivated segmentation (Hu et al., 2025). Our findings suggest a complementary perspective: the 5–6 bits we observe at the subword level may reflect a natural compression point that emerges from frequency patterns alone, largely independent of modality—though this tendency is stronger for some languages than others.

For multilingual NLP, these findings have concrete implications. Our findings highlight that orthographically trained BPE vocabularies have analogous compression properties to those trained on phonetic input for many languages. This has practical relevance for speech-and-text models, where the granularity mismatch between subword text tokens and phone-level speech units poses a known challenge for cross-modal transfer (Chou et al., 2023). Subword tokenization can help bridge this gap by bringing the two modalities closer in terms of information density. More broadly, our results illustrate that entropy trajectories offer a principled evaluation criterion for tokenization quality: if a tokenizer produces substantially different entropy values across orthographic and phonetic representations of the same language, this asymmetry signals that the tokenizer captures modality-specific surface patterns rather than underlying linguistic structure—a concern directly relevant to multilingual speech-text models that train on both modalities across many languages (Bapna et al., 2022). Conversely, tokenizers that achieve similar entropic properties across modalities may transfer more robustly in multilingual and cross-modal settings.

5 Implications for Language Modelling

In parallel to the question which languages are hard to language model (Cotterell et al., 2018; Mielke et al., 2019), we can also ask: *which modality is hard to language model?* Our entropic results point in the direction that spoken and written language are equally hard to language model on average. However, this conclusion is limited to the Latin-script languages in our sample; writing systems that encode less phonological information—such as logographic or consonantal scripts—may yield substantially larger modality gaps and thus different implications for language modelling difficulty. Entropy divergence between the two modalities studied is then mainly driven by differences in encodings (e.g., Bora tones being encoded in Latin script but not in X-SAMPA phones) or tokenization (e.g., divergence at the subword level for Sanzhi Dargwa). This is in line with Rust et al. (2021), who find that specialized monolingual tokenizers are a main factor driving performance in language modelling and downstream tasks.

6 Theoretical Implications

The consistent entropy increments across granularity levels—from approximately 4–5 bits at the character level to 5–6 bits at the subword level to 8–10 bits at the word level—reflect a structured layering of information across granularities that is consistent across both modalities and typologically diverse languages. Each encoding level captures independently quantifiable distributional structure: phones or characters encode basic segments, subword units capture phonological and morphological patterns, and orthographic words represent lexical units.

Preserving natural sequential order in preprocessing ensures that entropy trajectories reflect the genuine distributional structure of each language rather than artifacts of arbitrary segmentation, suggesting hierarchical information organisation is a property of linguistic content itself—not an artifact of how that content is written or transcribed.

The tendencies reported here are compatible with the broader hypothesis that information density is subject to cross-linguistic efficiency constraints (Piantadosi et al., 2011; Petrini et al., 2023), and that entropic differences between languages can be reduced to a minimum given a change in granularity of tokenization (Coupé et al., 2019; Gutierrez-Vasques et al., 2021).

7 Conclusion

We presented a cross-linguistic analysis of Shannon entropy across orthographic and phonetic encodings using the DoReCo corpus in CLDF representation, spanning 42 typologically diverse languages. Our central finding is that written and spoken modalities exhibit a consistently small modality gap across all three granularity levels examined: orthographic words, BPE subwords, and characters. This pattern is accompanied by a sharp decrease in cross-linguistic standard deviation from the word level to finer granularities, suggesting that information density becomes increasingly consistent across languages as the unit of analysis becomes smaller.

Beyond the cross-linguistic average, the case study on four languages reveals meaningful variation in both the size and location of the modality gap. Evenki shows a tight fit across all granularities, while English exhibits a modality gap at the character/phone level, Bora exhibits a gap at the word level, and Sanzhi Dargwa exhibits an unexpectedly large gap at the subword level. These cross-linguistic disagreements show how typological properties—reflected in the relationship between orthographic and phonetic representations—modulate where and how strongly entropic properties of modalities diverge.

In the context of multilingual NLP, these findings illustrate that information-theoretic measures computed on orthographic text generalise more robustly to spoken language representations than previously assumed. Entropy trajectories offer a principled basis for evaluating tokenization quality across modalities. Future work should investigate whether these entropy patterns extend to non-alphabetic writing systems, larger corpora, tokenization algorithms beyond BPE, and a broader range of typologically diverse languages to better characterise the conditions under which modality convergence holds.

Limitations

- All the languages in this analysis are represented in Latin scripts at the word tier. Other writing systems and scripts such as Chinese Hanzi (morphosyllabaries), Devanagari (syllabaries), or Abjads (consonantal alphabets) can yield larger modality gaps.
- The DoReCo corpus consists of naturalistic speech with orthographic transcriptions produced by trained annotators. The orthographic

representations therefore reflect transcription conventions that may not be fully comparable across languages.

- Our BPE tokenizers are trained on small, language-specific corpora. We use a vocabulary size of 200 merges; a sensitivity analysis with vocabulary sizes of 100 and 400 merges confirms that the modality convergence finding is robust to this parameter choice (see Table 2).
- The opposition between spoken vs. written can be interpreted as a modality difference (speech vs. orthography) or as a register/genre difference. We here focus on the technical aspect of the modality difference. We do not control for genre, speaker, or recording conditions, all of which may introduce variability in entropy estimates independent of modality.
- Our analysis is limited to Shannon entropy estimated via ML for unigrams; other information-theoretic measures such as conditional entropy or mutual information could reveal additional cross-modal differences not captured here.
- Several DoReCo languages contain recordings with multiple speakers; words from all concurrent speakers are included in the token streams. Since we compute unigram Shannon entropy over cumulative token frequencies, this does not affect the converged entropy values reported—though it means that word-level trajectories run across multiple speakers rather than a single discourse.

Acknowledgments

Funded by the European Union (ERC, EVINE, 101117111). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Aryaman Arora, Clara Meister, and Ryan Cotterell. 2022. [Estimating the entropy of linguistic distributions](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 175–195.

- Mathieu Avanzi, Marie-José Béguelin, Gilles Corminboeuf, Federica Diémoz, and Laure Anne Johnsen. 2022. [French \(Swiss\) DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Jocelyn Aznar. 2022. [Nisvai DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. [mSLAM: Massively multilingual joint pre-training for speech and text](#). *Preprint*, arXiv:2202.01374.
- Hauke Bartels and Marcin Szczepański. 2022. [Lower Sorbian DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i Cancho. 2017. [The entropy of words — Learnability and expressivity across more than 1000 languages](#). *Entropy*, 19(6):1–32.
- Paul Boersma and David Weenink. 2026. [Praat: doing phonetics by computer \[computer program\]](#). Accessed February 28, 2026.
- Natalia Bogomolova, Dmitry Ganenkov, and Nils Norman Schiborr. 2022. [Tabasaran DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Nadezhda Bulatova and Lenore Grenoble. 1999. *Evenki*. Lincom Europa, München.
- Niclas Burenhult. 2022. [Jahai DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Ju-Chieh Chou, Chung-Ming Chien, Wei-Ning Hsu, Karen Livescu, Arun Babu, Alexis Conneau, Alexei Baevski, and Michael Auli. 2023. [Toward Joint Language Modeling for Speech Units and Text](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6582–6593, Singapore. Association for Computational Linguistics.
- Alexander Yao Cobbinah. 2022. [Bainounk Gubeeher DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. [Are All Languages Equally Hard to Language-Model?](#) In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics.
- Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. [Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche](#). *Science Advances*, 5(9):eaaw2594.
- Thomas M. Cover and Joy A. Thomas. 2006. *Elements of information theory*. John Wiley & Sons, New Jersey.
- Andrew Cowell. 2022. [Arapaho DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Chris Lasse Däbritz, Nina Kudryakova, Eugénie Stapert, and Alexandre Arkhipov. 2022. [Dolgan DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Christian Döhler. 2022. [Komnzo DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Diana Forker. 2020. *A grammar of Sanzhi Dargwa*, volume 2 of *Languages of the Caucasus*. Language Science Press, Berlin.
- Diana Forker and Nils Norman Schiborr. 2022. [Sanzhi Dargwa DoReCo dataset](#). In Frank Seifart, Ludger

- Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Michael Franjieh. 2022. [Fanbyak DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Alexandro Garcia-Laguia. 2022. [Northern Alta DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Andargachew Mekonnen Gezmu and Andreas Nürnberger. 2023. [Morpheme-Based Neural Machine Translation Models for Low-Resource Fusion Languages](#). *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(9).
- Sonja Gipper and Jeremías Ballivián Torrico. 2022. [Yurakaré DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Jost Gippert. 2022. [Svan DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Richard Griscom. 2022. [Asimjeeg Datooga DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Valentin Gusev, Tiina Klooster, Beáta Wagner-Nagy, and Alexandre Arkhipov. 2022. [Kamas DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardžić. 2023. [Languages through the looking glass of BPE compression](#). *Computational Linguistics*, 49(4):943–1001.
- Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardžić. 2021. [From characters to words: the turning point of BPE merges](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468, Online. Association for Computational Linguistics.
- Tom Güldemann, Martina Ernszt, Sven Siegmund, and Alena Witzlack-Makarevich. 2022. [Nlɪŋg DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Geoff Haig, Maria Vollmer, and Hanna Thiele. 2022. [Northern Kurdish \(Kurmanji\) DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2026. [Glottolog 5.2](#).
- Iren Hartmann. 2022. [Hoocək DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Andrew Harvey. 2022. [Gorwaa DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Katharina Haude. 2022. [Movima DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Jean Hausser and Korbinian Strimmer. 2009. [Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks](#). *Journal of Machine Learning Research*, 10(7).
- Birgit Hellwig. 2022. [Goemai DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

- Birgit Hellwig, Gertrud Schneider-Blum, and Khaleel Bakheet Khaleel Ismail. 2022. [Tabaq \(Karko\) DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Yifan Hu, Frank Liang, Dachuan Zhao, Jonathan Geuter, Varshini Reddy, Craig W. Schmidt, and Chris Tanner. 2025. [Entropy-Driven Pre-Tokenization for Byte-Pair Encoding](#). *Preprint*, arXiv:2506.15889.
- Olga Kazakevich and Elena Klyachko. 2022. [Evenki DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Soung-U Kim. 2022. [Jejuan DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Alexander Koplenig, Sascha Wolfer, Jan Oliver Rü diger, and Peter Meyer. 2025. [Human languages trade off complexity against efficiency](#). *PLOS Complex Systems*, 2(2):e0000032.
- Manfred Krifka. 2022. [Daakie DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Alexis Michaud. 2022. [Yongning Na DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, and 1 others. 2021. [Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP](#). *arXiv preprint arXiv:2112.10508*.
- Sabrina J Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. [What kind of language is hard to language-model?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989.
- Ulrike Mosel. 2022. [Teop DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Carmel O’Shannessy. 2022a. [Light Warlpiri DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Carmel O’Shannessy. 2022b. [Warlpiri DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Pavel Ozerov. 2022. [Anal DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Sonia Petrini, Antoni Casas-i Muñoz, Jordi Cluet-i Martinell, Mengxue Wang, Chris Bentz, and Ramon Ferrer-i Cancho. 2023. [Direct and indirect evidence of compression of word lengths. Zipf’s law of abbreviation revisited](#). *Glottometrics*, 54:58–87.
- Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. [Word lengths are optimized for efficient communication](#). *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.
- Maïa Ponsonnet. 2022. [Dalabon DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Juan Diego Quesada, Stavros Skopeteas, Carolina Pasamonik, Carolin Brokmann, and Florian Fischer. 2022. [Cabécar DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Sabine Reiter. 2022. [Cashinahua DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage

- (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Sonja Riesberg. 2022. [Yali \(Apahapsili\) DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Hiram Ring. 2022. [Pnar DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Françoise Rose. 2022. [Mojeño Trinitario DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. [How good is your tokenizer? on the monolingual performance of multilingual language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.
- Nils Norman Schiborr. 2022. [English \(Southern England\) DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Stefan Schnell. 2022. [Vera’a DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Frank Seifart. 2022a. [Bora DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Frank Seifart. 2022b. [Resígaro DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Frank Seifart, Ludger Paschen, and Matthew Stave. 2022. [Language documentation reference corpus \(DoReCo\) 1.2](#).
- Frank Seifart, Ludger Paschen, Matthew Stave, Robert Forkel, and Frederic Blum. 2024. [CLDF dataset derived from the DoReCo core corpus \(v1.2.1\)](#). Data set.
- 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.
- C. E. Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27(3):379–423.
- C. E. Shannon. 1951. [Prediction and entropy of printed english](#). *Bell System Technical Journal*, 30(1):50–64.
- Stavros Skopeteas. 2022. [Yucatec Maya DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Stavros Skopeteas, Violeta Moisi, Nutsa Tsetereli, Johanna Lorenz, and Stefanie Schröter. 2022. [Urum DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Shuntaro Takahashi and Kumiko Tanaka-Ishii. 2018. [Cross entropy of neural language models at infinity — a new bound of the entropy rate](#). *Entropy*, 20(11):839.
- Amos Teo. 2022. [Sümi DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Nick Thieberger. 2022. [Nafsan \(South Efate\) DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

- Wesley Thiesen and David Weber. 2012. *A grammar of Bora: with special attention to tone*. SIL International.
- Martine Vanhove. 2022. [Beja DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Alexandra Vydrina. 2022. [Kakabe DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Claudia Wegener. 2022. [Savosavo DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Søren Wichmann. 2022. [Texistepec Popoluca DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Alena Witzlack-Makarevich, Saudah Namyalo, Anatol Kiriggwajjo, and Zarina Molochieva. 2022. [Ruuli DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- Xianming Xu and Bibo Bai. 2022. [Sadu DoReCo dataset](#). In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.
- 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.
- languages present in the DoReCo corpus were excluded due to insufficient token counts after preprocessing: Cabécar ([Quesada et al., 2022](#)), Hoocak ([Hartmann, 2022](#)), Light Warlpiri ([O’Shannessy, 2022a](#)), Nisvai ([Aznar, 2022](#)), Northern Alta ([Garcia-Laguia, 2022](#)), Urum ([Skopeteas et al., 2022](#)), Warlpiri ([O’Shannessy, 2022b](#)), Yucatec Maya ([Skopeteas, 2022](#)), and Yurakaré ([Gipper and Ballivián Torrico, 2022](#)).

A Appendix

Table 3 lists the 42 DoReCo language datasets included in the analysis, with their citations. Nine

Language	Family	Glottocode	Reference
Anal	Sino-Tibetan	anal1239	(Ozerov, 2022)
Arapaho	Algic	arap1274	(Cowell, 2022)
Asimjeeg Datooga	Nilotic	tsim1256	(Griscom, 2022)
Bainounk Gubäeher	Atlantic-Congo	bain1259	(Cobbinah, 2022)
Beja	Afro-Asiatic	beja1238	(Vanhove, 2022)
Bora	Boran	bora1263	(Seifart, 2022a)
Cashinahua	Pano-Tacanan	cash1254	(Reiter, 2022)
Daakie	Austronesian	port1286	(Krifka, 2022)
Dalabon	Gunwinyguan	ngal1292	(Ponsonnet, 2022)
Dolgan	Turkic	dolg1241	(Däbritz et al., 2022)
English (Southern England)	Indo-European	sout3282	(Schiborr, 2022)
Evenki	Tungusic	even1259	(Kazakevich and Klyachko, 2022)
Fanbyak	Austronesian	orko1234	(Franjeh, 2022)
French (Swiss)	Indo-European	stan1290	(Avanzi et al., 2022)
Goemai	Afro-Asiatic	goem1240	(Hellwig, 2022)
Gorwaa	Afro-Asiatic	goro1270	(Harvey, 2022)
Jahai	Austroasiatic	jeha1242	(Burenhult, 2022)
Jejuan	Koreanic	jeju1234	(Kim, 2022)
Kakabe	Mande	kaka1265	(Vydrina, 2022)
Kamas	Uralic	kama1351	(Gusev et al., 2022)
Komnzo	Yam	komn1238	(Döhler, 2022)
Lower Sorbian	Indo-European	lowe1385	(Bartels and Szczepański, 2022)
Mojeño Trinitario	Arawakan	trin1278	(Rose, 2022)
Movima	Isolate	movi1243	(Haude, 2022)
Nlɪŋg	Tuu	nngg1234	(Güldemann et al., 2022)
Nafsan (South Efate)	Austronesian	sout2856	(Thieberger, 2022)
Northern Kurdish (Kurmanji)	Indo-European	nort2641	(Haig et al., 2022)
Pnar	Austroasiatic	pnar1238	(Ring, 2022)
Resígaro	Arawakan	resi1247	(Seifart, 2022b)
Ruuli	Atlantic-Congo	ruul1235	(Witzlack-Makarevich et al., 2022)
Sadu	Sino-Tibetan	sadu1234	(Xu and Bai, 2022)
Sanzhi Dargwa	Nakh-Daghestanian	sanz1248	(Forker and Schiborr, 2022)
Savosavo	Isolate	savo1255	(Wegener, 2022)
Sümi	Sino-Tibetan	sumi1235	(Teo, 2022)
Svan	Kartvelian	svan1243	(Gippert, 2022)
Tabaq (Karko)	Nubian	kark1256	(Hellwig et al., 2022)
Tabasaran	Nakh-Daghestanian	taba1259	(Bogomolova et al., 2022)
Teop	Austronesian	teop1238	(Mosel, 2022)
Texistepec Popoluca	Mixe-Zoque	texi1237	(Wichmann, 2022)
Vera'a	Austronesian	vera1241	(Schnell, 2022)
Yali (Apahapsili)	Nuclear Trans New Guinea	apah1238	(Riesberg, 2022)
Yongning Na	Sino-Tibetan	yong1270	(Michaud, 2022)

Table 3: The 42 DoReCo language datasets included in the analysis, sorted alphabetically, with language family, Glottocode, and dataset citations. Languages with no genealogical affiliation are marked as Isolate.

Language	Word tokens	Word types	Phone tokens	Phone types
Anal	14,438	4,178	50,132	48
Arapaho	10,444	5,044	41,540	27
Asimjeeg Datooga	8,784	3,383	40,352	40
Bainounk Gubëeher	12,550	4,588	58,125	43
Beja	15,466	7,243	77,800	54
Bora	29,370	12,110	50,461	42
Cashinahua	10,050	3,490	49,177	28
Daakie	12,039	1,118	43,156	38
Dalabon	4,128	1,773	24,617	32
Dolgan	18,928	7,339	45,858	67
English (Southern England)	35,296	4,040	26,624	45
Evenki	8,378	4,617	54,391	41
Fanbyak	19,612	2,367	35,609	39
French (Swiss)	13,339	1,852	39,229	38
Goemai	26,225	2,153	27,613	70
Gorwaa	19,992	4,800	43,051	44
Jahai	7,954	1,508	31,817	34
Jejuan	9,367	3,622	34,538	37
Kakabe	46,656	5,869	27,775	36
Kamas	40,336	11,271	49,782	55
Komnzo	33,875	7,332	52,599	41
Lower Sorbian	10,926	2,285	38,262	49
Mojeño Trinitario	18,396	5,226	44,409	32
Movima	10,249	2,591	52,618	32
Nlɪŋg	27,903	3,173	25,770	88
Nafsan (South Efate)	29,833	5,117	42,879	25
Northern Kurdish (Kurmanji)	9,658	2,445	38,026	35
Pnar	20,527	3,326	27,425	44
Resígaro	8,084	3,361	42,057	37
Ruuli	8,747	4,033	50,420	34
Sadu	11,760	1,098	28,055	49
Sanzhi Dargwa	15,232	6,397	27,378	56
Savosavo	12,081	2,000	48,268	25
Sümi	14,458	2,075	32,601	40
Svan	10,460	4,594	50,376	44
Tabaq (Karko)	10,053	4,070	40,185	50
Tabasaran	5,059	1,856	24,148	58
Teop	12,149	1,269	43,735	27
Texistepec Popoluca	8,479	1,902	40,491	56
Vera'a	17,795	2,021	45,162	24
Yali (Apahapsili)	15,365	2,742	33,418	25
Yongning Na	7,260	2,331	31,854	60

Table 4: Corpus size statistics for the 42 DoReCo languages included in the analysis, after preprocessing. Word and phone counts reflect tokens remaining after removal of pause markers, annotation tags, and words lacking phone transcriptions.

Entropy Trajectories: All Encoding Levels per Language (1/3)

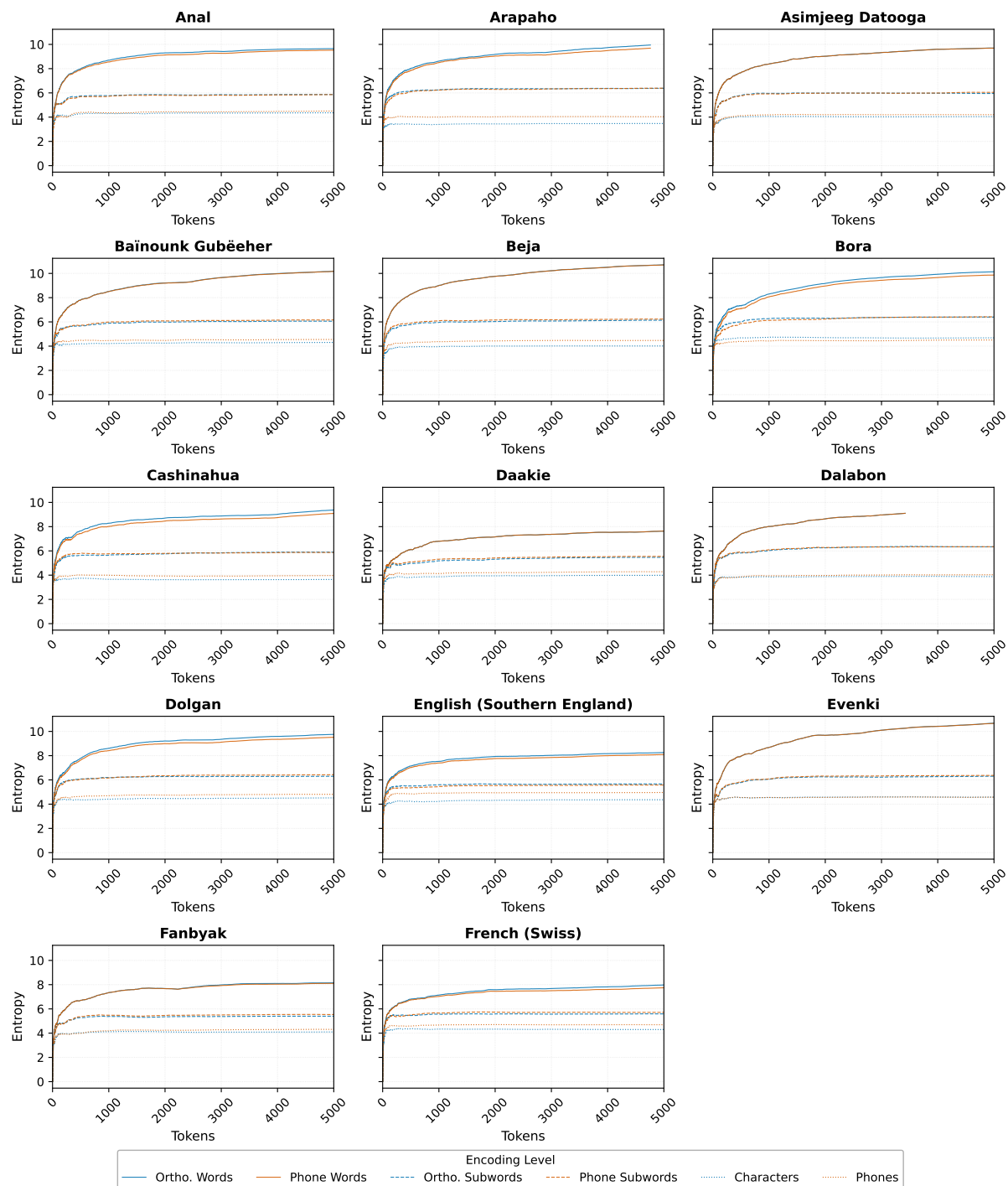


Figure 5: Entropy trajectories for all six encoding levels (orthographic words, phone words, orthographic subwords, phone subwords, characters, and phones), faceted by language, over the first 5,000 tokens (1/3).

Entropy Trajectories: All Encoding Levels per Language (2/3)

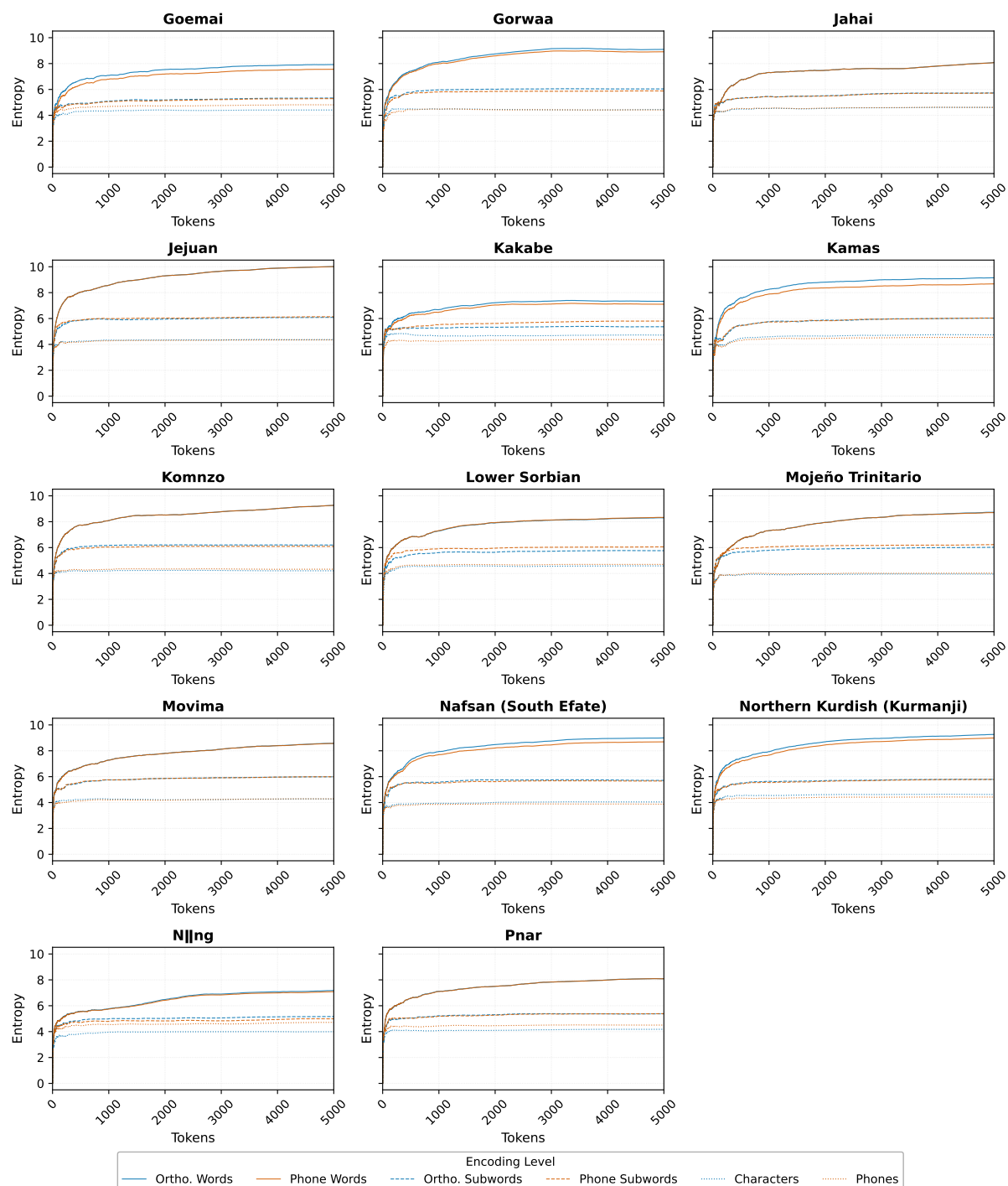


Figure 6: Entropy trajectories for all six encoding levels (orthographic words, phone words, orthographic subwords, phone subwords, characters, and phones), faceted by language, over the first 5,000 tokens (2/3, continued from Figure 5).

Entropy Trajectories: All Encoding Levels per Language (3/3)

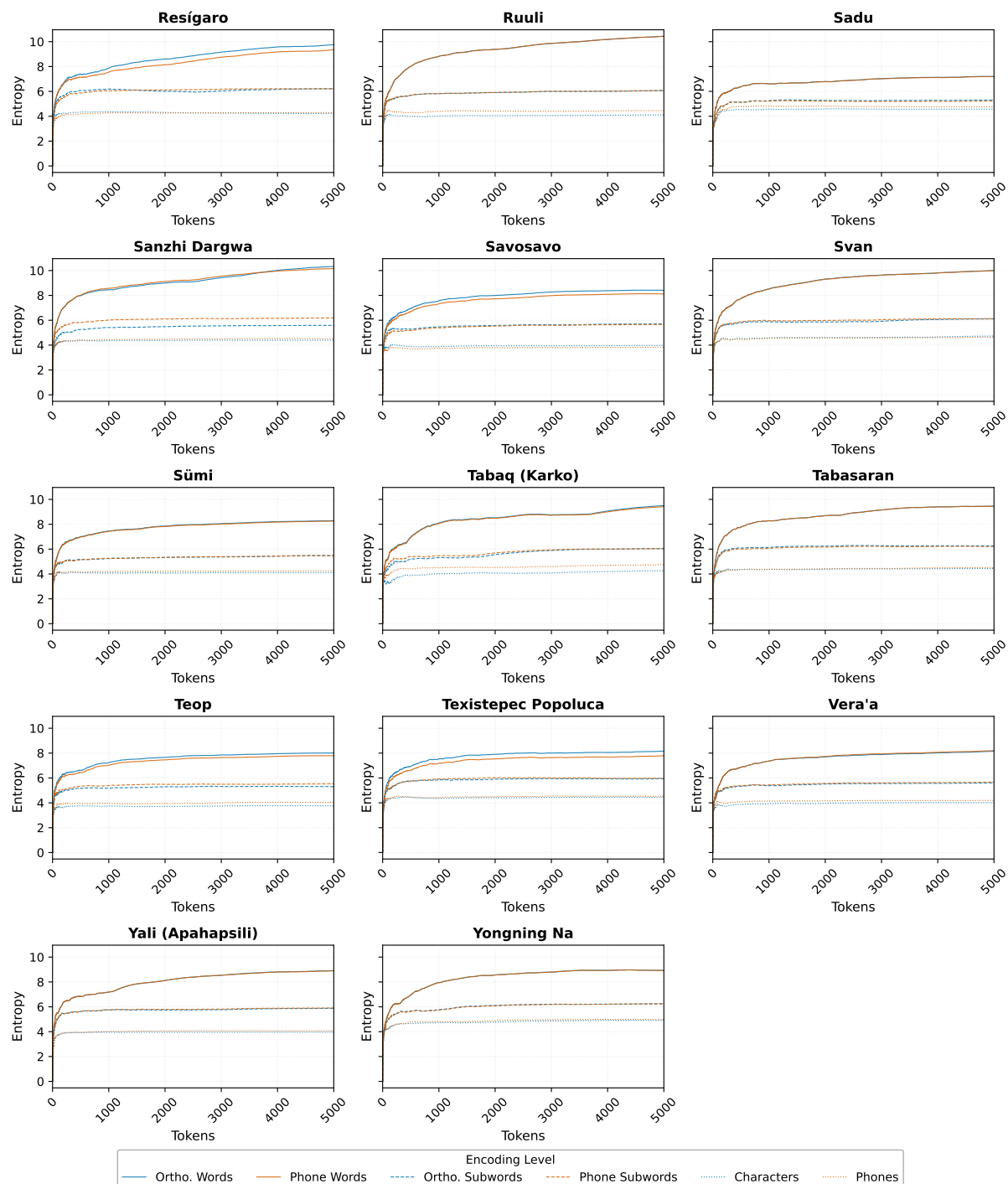


Figure 7: Entropy trajectories for all six encoding levels (orthographic words, phone words, orthographic subwords, phone subwords, characters, and phones), faceted by language, over the first 5,000 tokens (3/3, continued from Figure 6).