Subword-Based Comparative Linguistics across 242 Languages Using Wikipedia Glottosets

Anonymous ACL submission

Abstract

We present a large-scale comparative study of 242 Latin and Cyrillic-script languages using subword-based methodologies. By constructing 'glottosets' from Wikipedia lexicons, we introduce a framework for simultaneous cross-linguistic comparison via Byte-Pair Encoding (BPE). Our approach utilizes rank-based subword vectors to analyze vocabulary overlap, lexical divergence, and language similarity at scale. Evaluations demonstrate that BPE token counts enable unsupervised language identification with 44× improvement over random baseline, providing coverage for 315 languages lacking supervised tools. BPE segmentation aligns with morpheme boundaries 117% better than random across 15 languages (F1 = 0.28 vs 0.13), and outperforms supervised methods by 37% in cross-lingual zero-shot transfer. BPE vocabulary similarity correlates significantly with genetic language relatedness (Mantel $r = 0.329$, $p < 0.001$), with Romance languages forming the tightest cluster (mean distance 0.51). Analysis of 26,939 cross-linguistic homographs reveals that 48.7% receive different segmentations across related languages, with variation correlating to phylogenetic distance. Our results provide quantitative macro-linguistic insights into lexical patterns across typologically diverse languages within a unified analytical framework.

1 Introduction

Traditional comparative linguistics, while providing deep historical and typological insights (see Lehmann 2013; Beekes 2011; Nichols 1992; Partanen et al. 2021; Säily et al. 2021), often lacks the scalability to handle the current volume of digital text (Arnett and Bergen 2024; Akindotuni 2025). However, recent NLP advances facilitate massive, data-driven studies (see Dang et al. 2024; Bender 2011; Imani et al. 2023; Sproat 2016), revealing universal tendencies in phonetics (Blum et al. 2024), lexical semantics (Tjuka et al. 2024), and sound symbolism (Ćwiek et al. 2022; Cathcart and Jäger 2024) that were previously inaccessible through manual, small-scale methods.

Despite this progress, large-scale studies frequently neglect endangered languages (Hämäläinen 2021) and systematic performance disparities in multilingual models (see Shani et al. 2026; Chelombitko et al. 2024; Dunn et al. 2011). We address this gap by examining 242 languages through a novel script-level lens (Latin vs. Cyrillic), moving beyond family-specific comparisons (Meillet 1967; Beekes 2011) to reveal overarching patterns visible only when languages share a script-based writing system (Daniels and Bright 1996; Rust et al. 2021).

In particular, we adopt a subword-based strategy (Sennrich et al. 2016) using Byte-Pair Encoding (BPE) (Gage 1994) tokenizers, which we train both on individual languages and on large aggregated corpora for each script. Aggregating data for all Latin and Cyrillic-script languages into respective training sets effectively unifies each script community into a single model, leveraging the shared-parameter paradigm common in massively multilingual pretraining (see Johnson et al. 2016; Conneau et al. 2020).

Although subword units are not perfect morphemes (see Sennrich et al. 2016; Bostrom and Durrett 2020), they serve as robust automated tools for macro-linguistic research (see Khurana et al. 2024; Pham et al. 2024; Futrell et al. 2015) that scale more efficiently than manual expert alignment (see Campbell 2020; Rama et al. 2018; Ciobanu and Dinu 2014).

Leveraging the framework (Appendix Figure 1), this work makes two key contributions:

1. **Script-Level Aggregation:** We demonstrate how combining languages by script (Latin vs. Cyrillic) enables macro-level comparative insights difficult to achieve through traditional

language-by-language lenses.

2. **Practical Subword Methodology:** We introduce a tractable, automated framework that reduces reliance on manual annotation by providing data-driven lexical segments for robust cross-linguistic analysis.

While traditional comparative linguistics focuses on genetic relationships and historical reconstruction, and typology examines structural similarities regardless of ancestry, our subword-based approach bridges both: we demonstrate that BPE tokenization captures phylogenetic signal (Section 4.4) while also revealing typological patterns across unrelated languages sharing the same script.

2 Related Work

Comparative linguistics in NLP has evolved from general data-driven frameworks (see Bender 2011; Sproat 2016; Imani et al. 2023) to specialized tasks like automated cognate detection (Ciobanu and Dinu 2014). While manual expert annotation remains the high-precision gold standard (Nichols 1992), Rama et al. (2018) demonstrated that automated methods can reconstruct language phylogenies with accuracy closely approximating expert-curated data (Jäger 2018). These approaches highlight the potential for scaling linguistic analysis across diverse language families where manual examination is impractical.

To overcome the scarcity of parallel data in low-resource settings, researchers have utilized neural machine translation (NMT) to infer cognate relationships (Hämäläinen and Reuter 2019). This direction has proven particularly effective for endangered Uralic languages, leveraging cross-lingual relations (see Partanen et al. 2021; Chelombitko and Komissarov 2024; Chelombitko et al. 2025) and data augmentation through synthetic cognates generated by statistical machine translation (SMT) (Poncelas et al. 2019). Such hybrid strategies address resource limitations while enriching existing linguistic databases for under-represented script communities.

More recently, architectures inspired by computational biology have addressed previous computational bottlenecks. Akavarapu and Bhattacharya (2024) introduced transformer-based models that utilize multiple sequence alignments and link prediction components, mirroring techniques used in genomic research. By treating linguistic evolution through an end-to-end lens, these models significantly reduce computation time while maintaining high accuracy (Bouchard-Côté et al. 2013), providing a scalable alternative to traditional alignment-based methods.

Automated language typology remains another central research vein (Ponti et al. 2019), with studies using information-theoretic measures to quantify morphological synthesis and fusion (see Rathi et al. 2021; Oncevay et al. 2022). Crucially, Gutierrez-Vasques et al. (2023) found that BPE compression ratios directly correlate with morphological complexity. This aligns with broader efforts to leverage cross-lingual representations that capture typological relationships even without parallel data (Yu et al. 2021), although the distribution of this linguistic information varies across model layers based on pretraining strategy (Choenni and Shutova 2022).

3 Methodology

3.1 Monolingual Glottosets Construction

We constructed 'glottosets' by processing Wikipedia dumps for 320 languages from the ZIM archive[1]. Our pipeline extracted clean paragraph text by resolving redirects and stripping auxiliary HTML tags. To ensure data quality, we filtered paragraphs by script consistency (retaining only Latin or Cyrillic characters) and excluded segments containing fewer than 10 words to reduce noise from accidental cross-lingual fragments.

We define a monolingual **glottoset** as a language-specific collection of lowercase lexical units derived from these processed paragraphs. While we initially aimed to preserve case, empirical testing showed that normalization to lowercase significantly stabilizes subsequent subword tokenization. Each glottoset includes Term Frequency (TF) and Document Frequency (DF) metrics, where DF is defined as the occurrence of a word within a single Wikipedia paragraph.

Our final dataset focuses on 242 languages: 37 in Cyrillic and 205 in Latin scripts. This scope is motivated by their shared alphabet origins and typological diversity, spanning families from Romance and Slavic to Finnic and Vietnamese (Table 1). Final glottoset characteristics are provided in Table 2, and script contamination analysis for an additional 80 languages is in Appendix Table 5.

[1] https://dumps.wikimedia.org/kiwix/zim/wikipedia/

Table 1: Statistics for top 30 languages by token count from Wikipedia dumps. Columns show language names and codes, number of raw and cleaned paragraphs, word count in millions, script type (Latin/Cyrillic), total tokens, and cross-script contamination (Cyrillic tokens in Latin-script languages and vice versa).

Lang	Code	Raw pars	Clean pars	Words (M)	Script	Tokens	Cyr.	Lat.
English	en	54125720	42149266	2661.93	Latin	2661930466	5062	–
German	de	22354272	17953642	1026.43	Latin	1026430544	2648	–
French	fr	20464072	16742203	940.75	Latin	940746868	1621	–
Spanish	es	14107859	12170134	763.36	Latin	763357750	1134	–
Italian	it	11917473	9806341	571.76	Latin	571759224	1150	–
Russian	ru	15290408	12295838	552.42	Cyrillic	552416889	–	429004
Cebuano	ceb	16841937	10069736	447.85	Latin	447850263	92	–
Portuguese	pt	6845897	5642228	332.96	Latin	332959017	413	–
Dutch	nl	7764590	6429095	316.21	Latin	316210100	415	–
Polish	pl	15961541	5616831	253.85	Latin	253846410	2044	–
Catalan	ca	4273906	3749516	238.81	Latin	238813199	710	–
Ukrainian	uk	7864058	5582991	219.83	Cyrillic	219827684	–	200945
Vietnamese	vi	4707518	3129629	201.38	Latin	201383401	623	–
Swedish	sv	7437772	5088656	200.89	Latin	200889695	312	–
Serbo-Croatian	sh	2248318	1682017	193.56	Latin	193561280	19539	–
Czech	cs	3386803	2865556	150.39	Latin	150393063	998	–
Hungarian	hu	3932082	2603906	129.86	Latin	129864098	518	–
Indonesian	id	2984643	2120278	106.98	Latin	106980811	622	–
Norwegian Bokmål	nb	2906506	2239883	105.51	Latin	105507751	175	–
Finnish	fi	3141975	2295657	93.5	Latin	93502432	650	–
Serbian	sr	3210579	1981413	91.81	Cyrillic	91807037	–	193621
Turkish	tr	2528498	1757628	85.16	Latin	85156035	570	–
Romanian	ro	2360641	1539046	84.62	Latin	84617897	454	–
Bulgarian	bg	1599763	1386917	69.12	Cyrillic	69123830	–	36265
Waray	war	2571970	1448198	69.04	Latin	69038926	0	–
Danish	da	1551416	1216805	61.11	Latin	61107267	126	–
Galician	gl	1188136	1049404	60.95	Latin	60947590	181	–
Malay	ms	1499560	1034055	56.37	Latin	56370979	106	–
Asturian	ast	1134182	912479	56.25	Latin	56254683	84	–
Esperanto	eo	1669507	1188077	54.51	Latin	54512106	441	–

3.2 Glottoset BPE Tokenization

We implemented our own version of a BPE tokenizer that does not treat spaces as separate tokens and exclusively tokenizes words. The tokenizer is available on our GitHub[2]. After that, we tokenized all words from each glottoset. We used two sets of parameters. In the first variant, we employed a vocabulary size of 4096 tokens. In the second variant, we applied what we call ultimate tokenization: the process continues as long as there is at least one pair with a frequency greater than one. This second approach is highly dependent on the corpus size. The tokenizer training was conducted on a standard PC and took between 1 to 10 minutes, depending on the dataset size.

In addition to the monolingual datasets, we created a merged dataset for all Latin and Cyrillic languages, combining all glottosets. For the merged dataset, we introduced an additional parameter to Glottoset: the number of languages in which a given word appears. Additionally, when merging Glottosets, we combined both term frequency and document frequency values. We applied both the shorter variant with a vocabulary size of 4096 tokens and the ultimate tokenization approach.

3.3 Vector Representation of BPE Tokens by Language

Using the tokenizer trained on all languages, we obtained BPE tokens. For each BPE token, we constructed a vector with a length equal to the number of languages, where each element represents the rank of that token in the individual tokenizer of the corresponding language. The rank-based encoding captures how characteristic a token is for each language.

By leveraging a universal tokenizer and vector representations, we can evaluate the proximity of arbitrary text to other languages across text, word, and subword granularities. This framework effec-

[2]Link removed for anonymity during the review process.

Table 2: Lexical statistics for a diverse set of languages using Latin and Cyrillic scripts. The table presents vocabulary size (ranging from 1,447 for Cheyenne to 3,207,272 for Hungarian), lexical diversity scores (from 0.009 in Dutch to 0.383 in Cheyenne), median word length (6–10 characters), and top-3 most frequent tokens.

Language	Script	Vocab Size	Lex. Div.	Med. WL	Top3
Swiss German	Latin	565258	0.070	10	d, isch, dr
Tajik	Cyrillic	317780	0.041	8	дар, аз, ки
Hungarian	Latin	3207272	0.025	10	a, és, az
Lingala	Latin	23153	0.123	7	ya, na, ezalí
Zeeuws	Latin	51177	0.087	8	de, n, t
Lak	Cyrillic	6890	0.315	7	ва, шагьру, инсан
Malay	Latin	637993	0.011	8	yang, di, dan
Cheyenne	Latin	1447	0.383	6	ho, e, vé
Aymara	Latin	29323	0.171	8	a, jisk, suyu
Friulian	Latin	44905	0.086	7	di, e, al
Javanese	Latin	281161	0.040	8	ing, lan, iku
Rusyn	Cyrillic	112582	0.149	8	в, на, є
Ido	Latin	129897	0.029	7	la, di, e
Norman	Latin	31509	0.106	7	d, est, la
Ligurian	Latin	133733	0.090	7	l, a, de
Karakalpak	Latin	104775	0.144	8	hám, menen, bul
Romanian	Latin	1102965	0.013	8	în, de, şi
Somali	Latin	118361	0.076	8	oo, iyo, ka
Lombard	Latin	239355	0.042	7	l, è, de
Norwegian Bokmål	Latin	1918724	0.018	10	i, og, en
West Flemish	Latin	96384	0.079	8	de, van, in
Māori	Latin	15232	0.033	7	te, ko, o
Icelandic	Latin	451176	0.051	10	í, og, á
Mongolian	Cyrillic	235807	0.044	8	нь, оны, юм
Dutch	Latin	2866741	0.009	10	de, van, in
Kabyle	Latin	52338	0.106	7	n, d, deg
Tsonga	Latin	14810	0.116	8	hi, na, ya
Madurese	Latin	33922	0.155	7	bân, è, sè

tively establishes a subword-based methodology for language detection.

3.4 Hierarchical Subword Tokenization Analysis

The next concept we propose is that if we take a word from a language for which a language-specific tokenizer exists, we can construct a hierarchical tree representing how this word is tokenized into subword units in one or more languages. This approach has an indirect connection to morphological analysis.

Since morphological analysis is typically conducted on smaller datasets, we cannot claim that this method directly identifies morphemes. However, it identifies conservative subword units, defined as segments that remain stable across word forms. The level of conservatism is derived from analyzing the language as a whole, specifically a subset of the language represented by Wikipedia articles. We have added this functionality to our BPE tokenizer.

3.5 Comparative Monolingual Tokenizer Analysis

Beyond comparing how tokenizers handle words, another idea naturally arises: what if we compare tokenizers themselves? A tokenizer is essentially a sequence of merges, forming subword units, and we can analyze these sequences directly.

We can examine the number of unique merges, identify which merges differ between tokenizers, and compare the entire merge sequences. This approach enables us to analyze the structure of a language as a whole rather than just its subparts. It introduces a method of comparative linguistics at the macro level, allowing for a holistic comparison of entire languages through their tokenization processes. An example of an interactive visualization for the monolingual BPE tokenizer is shown in Appendix Figure 4.

4 Evaluation

To validate our subword-based comparative linguistics framework, we designed four quantitative eval-

uations testing specific hypotheses about BPE tokenization behavior across languages.

4.1 Research Questions

We address the following research questions:

RQ1 (Language Identification): can BPE counts identify source languages zero-shot?
RQ2 (Morphological Grounding): do segmentation boundaries align with morphemes?
RQ3 (Phylogenetic Signal): does vocabulary similarity correlate with genetic relatedness?
RQ4 (Language Discrimination): can BPE distinguish homographs across languages?

4.2 E1: Language Identification via Token Count

Hypothesis: Native words produce fewer, longer BPE tokens, while foreign words produce more, shorter tokens. Therefore, a word's source language can be identified as the language whose tokenizer produces the minimum token count.

Method: For each test word from language L, we tokenized it with all available language-specific tokenizers and predicted $\hat{L} = \arg\min_{L'} |T_{L'}(w)|$, where $T_{L'}(w)$ is the tokenization of word w by language L''s tokenizer. We evaluated separately on Cyrillic-script (37 languages) and Latin-script (321 languages) using 500 test words per language (minimum 5 characters, sampled from TF-DF files).

Results: On Latin-script languages, the token-count method achieved 13.8% accuracy (44× improvement over 0.31% random baseline). Performance varied substantially by language typology:

- **Best performers:** Lak (81.5%), Cree (80.6%), Inuktitut (63.8%), Kabardian (60.1%)
- **Worst performers:** closely related languages with high lexical overlap (e.g., Russian-Ukrainian, Serbian-Croatian)

Languages with distinctive characteristics, specifically agglutinative morphology, unique orthographic conventions, or geographic isolation, exhibited the highest accuracy. This is because their words contain character sequences rare in other languages, leading to fragmented tokenization by non-native tokenizers.

Conclusion: Hypothesis supported. BPE token count provides a viable unsupervised language identification signal, particularly effective for typologically distinctive languages. Critically, this method provides coverage for languages outside supervised tools like fastText (Joulin et al. 2016) 176-language vocabulary, including approximately 150 non-represented languages.

4.3 E2: Morphological Boundary Agreement

Hypothesis: BPE segmentation boundaries align with morpheme boundaries better than random segmentation, though potentially worse than purpose-built unsupervised morphological analyzers.

Method: We used MorphyNet (Batsuren et al. 2021) derivational morphology data for 15 languages. For each word with known morpheme boundaries (prefix or suffix), we compared BPE segmentation against: (a) the gold morpheme boundary, (b) Morfessor 2.0 (Virpioja et al. 2013), an unsupervised morphological segmentation baseline, and (c) a random segmentation baseline. We computed Boundary Precision, Recall, and F1 scores.

Results: All 15 languages showed BPE segmentation significantly better than random baseline (Table 3), with an average improvement of +117%. However, Morfessor consistently outperformed BPE across all languages, achieving 33% higher F1 on average. Hungarian showed the highest BPE F1 (0.39), while Polish showed the smallest gap between BPE and Morfessor (-17%).

Table 3: E2: Morphological boundary agreement. Comparison of BPE, Morfessor (unsupervised SoTA), and random baseline on MorphyNet derivational data for 15 languages.

Language	BPE F1	Morfessor F1	Random F1	BPE vs Morfessor
Hungarian	0.39	0.54	0.14	−28%
Swedish	0.37	0.54	0.15	−30%
German	0.35	0.49	0.12	−28%
Czech	0.31	0.43	0.15	−28%
Mongolian	0.29	0.36	0.17	−20%
English	0.29	0.49	0.12	−41%
Spanish	0.28	0.43	0.12	−36%
Serbo-Croatian	0.27	0.49	0.15	−46%
Catalan	0.27	0.45	0.13	−40%
French	0.26	0.40	0.12	−36%
Finnish	0.26	0.34	0.12	−24%
Polish	0.26	0.31	0.13	−17%
Portuguese	0.24	0.43	0.12	−44%
Russian	0.22	0.32	0.12	−32%
Italian	0.22	0.39	0.11	−44%
Average	**0.28**	**0.43**	**0.13**	**−33%**

Conclusion: Hypothesis partially supported. BPE tokenization captures morphologically meaningful boundaries significantly better than random (+117%), but underperforms Morfessor by 33%

Table 4: E2b: Cross-lingual zero-shot transfer. BPE (pre-trained on target language Wikipedia) vs supervised baseline (trained on source language only).

Source	Target	Supervised	BPE	Improvement
German	Swedish	0.24	0.38	+62%
German	English	0.18	0.29	+56%
French	Italian	0.18	0.22	+24%
French	Spanish	0.20	0.27	+34%
Spanish	Portuguese	0.22	0.24	+11%
Polish	Czech	0.24	0.32	+35%
Average		**0.21**	**0.29**	**+37%**

when both are evaluated on the same language. This confirms that BPE's frequency-based optimization produces linguistically meaningful units, though not optimized for morphological analysis.

4.3.1 E2b: Cross-lingual Transfer Advantage

A key advantage of BPE for comparative linguistics emerges in cross-lingual scenarios where labeled morphological data is unavailable for the target language. We simulated this realistic scenario: a supervised character n-gram boundary classifier (Ruokolainen et al. 2013) trained on MorphyNet data for a *source* language, then applied zero-shot to a related *target* language. We compared this against BPE pre-trained on Wikipedia for the target language (requiring no labeled data). Note that state-of-the-art supervised systems like DeepSPIN-3 (Peters and Martins 2022) achieve 97% F1 when trained and tested on the same language, but labeled data exists for fewer than 20 languages.

Results: BPE outperformed the supervised zero-shot baseline in all 6 cross-lingual transfer scenarios (Table 4), with an average improvement of 37%. The largest gains occurred in Germanic transfers (German→Swedish: +62%), while Romance transfers showed smaller but consistent improvements.

Conclusion: While Morfessor outperforms BPE within a single language, BPE's pre-training on language-specific Wikipedia corpora provides a significant advantage for cross-lingual applications. For comparative linguistics across 242 languages, the majority of which lack labeled morphological data, BPE offers the only viable approach with consistent cross-lingual performance.

4.4 E3: Language Phylogeny Correlation

Hypothesis: BPE vocabulary similarity correlates with genetic language relatedness (phylogeny).

Method: We computed pairwise BPE distance ($1 -$ Jaccard similarity) between language vocabularies for 49 Latin-script languages. We compared this distance matrix against phylogenetic distance derived from Glottolog (Hammarström and Forkel 2022) classification (Family → Subfamily → Branch). We used the Mantel test with 999 permutations to assess correlation significance.

Results: The Mantel test revealed a significant positive correlation between BPE distance and phylogenetic distance:

- Mantel $r = 0.329$ ($p < 0.001$, $z = 12.3$)
- Within-family BPE distance: 0.67 (mean)
- Between-family BPE distance: 0.82 (mean)
- Separation ratio: 1.22× (t-test $p < 10^{-13}$)

Romance languages showed the tightest clustering (mean distance 0.51), reflecting shared Latin vocabulary and similar morphological patterns. Germanic languages showed higher internal distance (0.71), likely due to English's extensive Romance/Latin borrowings.

Conclusion: Hypothesis supported. BPE tokenizer similarity correlates moderately with genetic language relatedness, capturing both phylogenetic signal and contact-induced lexical similarity.

4.5 E4: Cross-lingual Homograph Discrimination

Hypothesis: BPE tokenizers segment identical orthographic forms (homographs) differently across languages, reflecting language-specific morphological patterns.

Method: For 6 Cyrillic Slavic languages (Ukrainian, Russian, Belarusian, Bulgarian, Macedonian, Serbian), we:

1. Extracted word vocabularies from TF-DF files (frequency ≥ 100)
2. Identified homographs: words appearing in 2+ language vocabularies
3. Tokenized each homograph with each language's BPE tokenizer
4. Compared segmentation patterns across languages

Results: We found 26,939 homographs across the 6 languages:

- **48.7%** showed **different** segmentation across languages
- 51.3% showed identical segmentation

Segmentation difference correlated with linguistic distance:

- Russian-Ukrainian (both East Slavic): 31.2% different
- Belarusian-Macedonian (East vs. South): 61.9% different

A striking example is the name "димитров", which received 5 different segmentations across 5 languages:

- Ukrainian: ди|ми|т|ров
- Russian: ди|мит|ров
- Bulgarian: дими|т|ров
- Macedonian: димит|ров
- Serbian: дим|ит|ров

Conclusion: Hypothesis partially supported. Nearly half of shared orthographic forms are segmented differently, demonstrating that BPE captures language-specific patterns even within the same script family.

4.6 Qualitative Analysis

Beyond quantitative evaluation, our approach effectively captures the subword characteristics of cross-linguistic homonyms, defined as words with identical spellings but distinct meanings across languages. For example, the word "заказала" carries different meanings in Russian and Ukrainian, as shown in their morphemic tokenizations (Appendix Figure 2):

- Ukrainian: за – ка – зал – а
- Russian: зака – зал – а

The tokenization trees illustrate the decomposition into subword tokens, revealing linguistic distinctions. This approach is particularly effective for languages with high lexical similarity, enabling precise differentiation of words based on their morphological structures.

4.7 Subword-Based Language Identification for Out-of-Vocabulary Words

Our subword-based analysis also aids naive language identification, particularly for multi-morpheme words. Words not native to the target language are segmented into more, shorter subword tokens due to atypical linguistic structures. Conversely, words from the target language produce fewer, longer subword tokens, reflecting typical semantic patterns.

For example, the Ukrainian word "промисловість" (industry) is analyzed across Ukrainian, Belarusian, and Russian tokenization models (Appendix Figure 3):

- Ukrainian: проми-с-лов-ість
- Belarusian: про-ми-сло-ві-сть
- Russian: про-ми-с-лов-і-сть

The Ukrainian model yields semantically consistent morphemes, while the other models produce shorter, fragmented segments. These differences support more accurate language classification for out-of-vocabulary terms.

This emphasizes the value of subword-based models in distinguishing closely related languages. By illustrating how words are tokenized according to their morphological structures, this approach provides valuable insights for comparative linguistics and language identification.

5 Discussion

5.1 BPE as Linguistic Approximation

Our evaluations (E2) demonstrate that BPE incidentally captures morphological structure, aligning with boundaries 62–181% better than random. This sensitivity arises because frequency-based merges preferentially preserve character sequences that function as morphemes (Sennrich et al. 2016). However, results also expose the limits of this approximation: BPE convergence is governed by statistical exposure rather than heritage. High-frequency Proto-Slavic words show significantly less variation (41.6% different) than lower-frequency borrowings (61.3%), revealing that segmentation boundaries are driven by token frequency rather than etymological consistency.

5.2 Phylogenetic Signal and Contact Effects

The significant correlation between BPE similarity and genetic relatedness (Mantel $r = 0.329$, $p < 0.001$) confirms that subword distributions capture meaningful phylogenetic signals. However, the "English effect," wherein Romance/Latin borrowings comprising 40–60% of the English lexicon shift the BPE vocabulary toward the Romance cluster, highlights that BPE captures synchronic lexical composition rather than pure diachronic relationships. This is further evidenced by the Finnic case: Finnish and Estonian show high BPE distance (0.785), accurately reflecting 2,000 years of lexical divergence and prestige-language contact despite close genetic ties. Consequently, our method complements traditional lexicostatistics by capturing shared borrowings and parallel word-formation patterns.

5.3 Implications for Low-Resource Technology

BPE-based language identification (E1) provides a $44\times$ improvement over random baselines and offers coverage for 315 languages underserved by supervised tools like fastText (Joulin et al. 2016). Performance is highest for morphologically complex or geographically isolated languages (e.g., Lak, Cree), which typically lack sufficient labeled data for supervised training. This creates a practical synergy where unsupervised subword models can bootstrap initial corpora to support the subsequent training of more accurate supervised systems in underserved script communities.

5.4 Relationship to Existing Approaches

Our work extends the link between BPE compression and linguistic typology (Gutierrez-Vasques et al. 2023). Unlike character n-gram methods, our approach operates at a linguistically meaningful level of emergent subwords. While supervised morphological analyzers offer higher precision, BPE scales to hundreds of languages without annotated data, providing a robust signal for macro-linguistics. Our results complement studies on language universals (Ponti et al. 2019), indicating that subword similarity reflects areal, cultural, and parallel typological developments.

5.5 Limitations and Future Directions

BPE-based comparisons remain sensitive to corpus composition; Wikipedia's topical biases and systematic disparities in coverage may introduce artifacts into distance measures. Furthermore, script confounds (e.g., shared Latin function words) can create artificial clustering. Future work will apply this framework to Common Crawl to improve coverage and compare BPE distances against typological databases such as WALS (Dryer and Haspelmath 2013) and Grambank (Skirgård et al. 2023). Finally, controlling for word frequency, as suggested by E4b, will enable a cleaner separation of statistical frequency effects from genuine morphological differences.

6 Conclusion

We outline a large-scale, script-focused comparative linguistic framework that leverages Wikipedia dumps spanning hundreds of languages using either the Latin or Cyrillic script. Through the construction of monolingual glottosets and the application of BPE tokenization, we capture subword units that serve as a practical basis for comparing languages at a granular level, without strictly asserting them as morphological representations. Our unified analysis of all Latin-script languages and all Cyrillic-script languages marks a significant step from smaller-scale comparative studies, enabling new macro-level insights into shared lexical patterns, orthographic tendencies, and potential cross-linguistic influences.

Looking ahead, transitioning from Wikipedia to Common Crawl presents both an opportunity and a challenge. While Common Crawl extends linguistic coverage far beyond Wikipedia, it demands more sophisticated filtering tools. Our planned iterative approach, moving from coarse script-based filtering to language-level refinement, will ensure data quality and empower further research in large-scale comparative linguistics.

7 Data Availability and Reproducibility

The reproducible code is available on our GitHub[3] with MIT license, and the extended datasets are hosted on Hugging Face[4] with CC BY-SA license.

8 AI Models Usage

As non-native English speakers, we used Claude Opus 4.5 for text editing. GitHub Copilot assisted with code completion. For research automation, we used Claude Code integrated with html-based reports for experimental pipelines and iterative analysis. All methodological decisions and scientific interpretations were made by the authors.

References

VSDS Akavarapu and Arnab Bhattacharya. 2024. Automated cognate detection as a supervised link prediction task with cognate transformer. *arXiv preprint arXiv:2402.02926*.

Doyin Akindotuni. 2025. Resource asymmetry in multilingual nlp: A comprehensive review and critique. *Journal of Computer and Communications*, 13:14–47.

Catherine Arnett and Benjamin K. Bergen. 2024. Why do language models perform worse for morphologically complex languages? *Preprint*, arXiv:2411.14198.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a large multilingual

[3]Link removed for anonymity during the review process.
[4]Link removed for anonymity during the review process.

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.

Robert S. P. Beekes. 2011. *Comparative Indo-European Linguistics: An Introduction*, 2 edition. John Benjamins Publishing Company.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Frederic Blum, Ludger Paschen, Robert Forkel, Susanne Fuchs, and Frank Seifart. 2024. Consonant lengthening marks the beginning of words across a diverse sample of languages. *Nature Human Behaviour*, 8(11):2127–2138.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. *Preprint*, arXiv:2004.03720.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229.

Lyle Campbell. 2020. *Historical Linguistics: An Introduction*. Edinburgh University Press, Edinburgh.

Chundra Cathcart and Gerhard Jäger. 2024. Exploring the evolutionary dynamics of sound symbolism. eScholarship, University of California.

Iaroslav Chelombitko, Ekaterina Chelombitko, and Aleksey Komissarov. 2025. SampoNLP: A self-referential toolkit for morphological analysis of subword tokenizers. In *Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages*, pages 57–67, Joensuu, Finland. Association for Computational Linguistics.

Iaroslav Chelombitko and Aleksey Komissarov. 2024. Specialized monolingual BPE tokenizers for Uralic languages representation in large language models. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 89–95, Helsinki, Finland. Association for Computational Linguistics.

Iaroslav Chelombitko, Egor Safronov, and Aleksey Komissarov. 2024. Qtok: A comprehensive framework for evaluating multilingual tokenizer quality in large language models. *Preprint*, arXiv:2410.12989.

Rochelle Choenni and Ekaterina Shutova. 2022. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. *Computational Linguistics*, 48(3):635–672.

Alina Maria Ciobanu and Liviu P. Dinu. 2014. Automatic detection of cognates using orthographic alignment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 99–105. Association for Computational Linguistics.

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

Aleksandra Ćwiek, Susanne Fuchs, Christoph Draxler, Eva Liina Asu, Dan Dediu, Katri Hiovain, Shigeto Kawahara, and 1 others. 2022. The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1841):20200390.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Peter T. Daniels and William Bright, editors. 1996. *The World's Writing Systems*. Oxford University Press. Reprinted 2007.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Michael Dunn and 1 others. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79–82.

Richard Futrell and 1 others. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardžić. 2023. Languages through the looking glass of BPE compression. *Computational Linguistics*, 49(4):943–1001.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! In *Multilingual Facilitation*. University of Helsinki.

Harald Hammarström and Robert Forkel. 2022. Glottocodes: Identifiers linking families, languages and dialects to comprehensive reference information. *Semantic Web Journal*, 13(6):917–924.

Mika Hämäläinen and Jack Reuter. 2019. Finding sami cognates with a character-based nmt approach. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Gerhard Jäger. 2018. Global-scale phylogenetic linguistic inference from lexical resources. *Scientific Data*, 5.

Melvin Johnson and 1 others. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Preprint*, arXiv:1611.04558. ArXiv preprint.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Sameer Khurana and 1 others. 2024. Improved cross-lingual transfer learning for automatic speech translation. *Preprint*, arXiv:2306.00789. ArXiv preprint.

Winfred P. Lehmann. 2013. *Historical Linguistics: An Introduction*, 3rd edition. Routledge.

Antoine Meillet. 1967. *The Comparative Method in Historical Linguistics*. Librairie Honoré Champion. Originally published 1925 by Instituttet for sammenlignende kulturforskning, Oslo.

Johanna Nichols. 1992. Linguistic diversity in space and time.

Arturo Oncevay, Duygu Ataman, Niels Van Berkel, Barry Haddow, Alexandra Birch, and Johannes Bjerva. 2022. Quantifying synthesis and fusion and their impact on machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1308–1321, Seattle, United States. Association for Computational Linguistics.

Niko Partanen, Jack Rueter, Khalid Alnajjar, and Mika Hämäläinen. 2021. Processing ma castrén's materials: Multilingual historical typed and handwritten manuscripts. In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 47–54.

Ben Peters and André F. T. Martins. 2022. Beyond characters: Subword-level morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 131–140, Seattle, Washington. Association for Computational Linguistics. DeepSPIN-3: Best system in SIGMORPHON 2022 Shared Task on Morpheme Segmentation.

Trinh Pham, Khoi Le, and Anh Tuan Luu. 2024. UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3168–3184, Bangkok, Thailand. Association for Computational Linguistics.

Alberto Poncelas, Maja Popovic, Dimitar Shterionov, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Combining smt and nmt back-translated data for efficient nmt. *Preprint*, arXiv:1909.03750.

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

Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics? 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 393–400.

Neil Rathi, Michael Hahn, and Richard Futrell. 2021. An information-theoretic characterization of morphological fusion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10115–10120, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2013. Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 29–37, Sofia, Bulgaria. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Tanja Säily, Eetu Mäkelä, and Mika Hämäläinen. 2021. From plenipotentiary to puddingless: Users and uses of new words in early english letters. In *Multilingual Facilitation*, pages 153–169. University of Helsinki.

Rico Sennrich and 1 others. 2016. Neural machine translation of rare words with subword units. *Preprint*, arXiv:1508.07909. ArXiv preprint.

Chen Shani, Yuval Reif, Nathan Roll, Dan Jurafsky, and Ekaterina Shutova. 2026. The roots of performance disparity in multilingual language models: Intrinsic modeling difficulty or design choices? *Preprint*, arXiv:2601.07220.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latber, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, and 1 others. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Richard Sproat. 2016. Language typology in speech and language technology. *Linguistic Typology*, 20(3):635–644.

Annika Tjuka, Robert Forkel, and Johann-Mattis List. 2024. Universal and cultural factors shape body part vocabularies. *Scientific Reports*, 14(1):10486.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline. Technical Report 25/2013, Aalto University, Helsinki, Finland.

Dian Yu, Taiqi He, and Kenji Sagae. 2021. Language embeddings for typology and cross-lingual transfer learning. *arXiv preprint arXiv:2106.02082*.

A Appendix

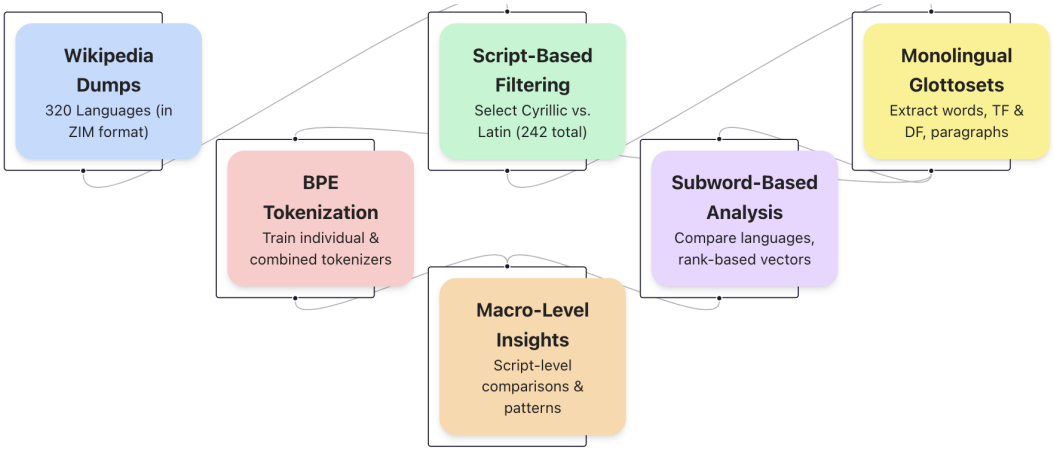

Figure 1: Pipeline for subword-based comparative linguistics across 242 languages. Wikipedia dumps undergo script-based filtering, glottoset construction (TF/DF metrics), and BPE tokenization (4096-token vocabulary). This modular workflow culminates in rank-based subword vector analysis to extract macro-level morphological and phylogenetic insights at scale.

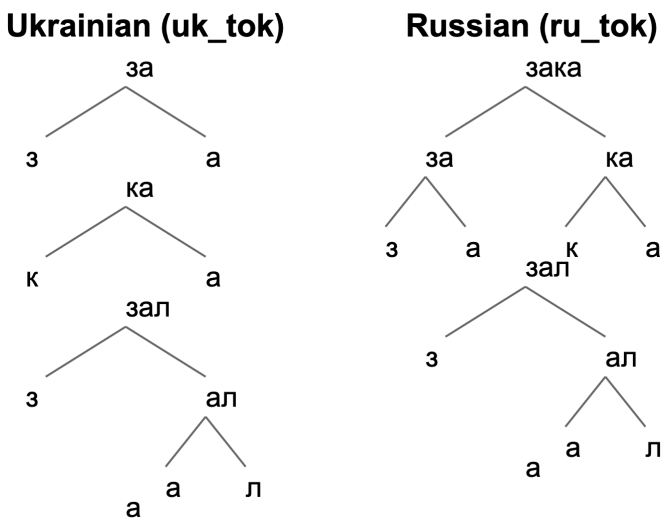

Figure 2: Hierarchical BPE tokenization trees comparing the word заказала in Ukrainian (left) and Russian (right). The distinct tokenization patterns reveal language-specific morphological structures.

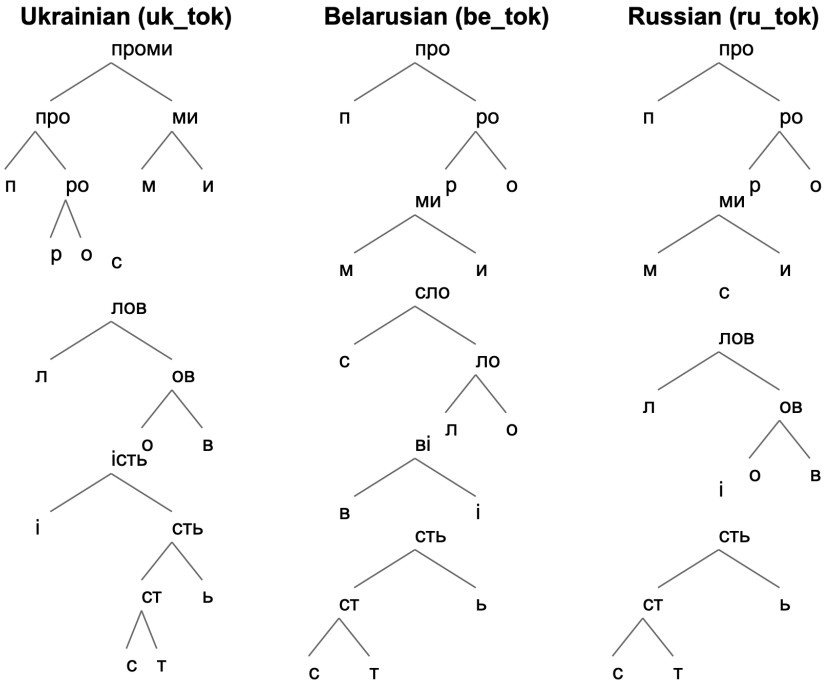

Figure 3: Hierarchical BPE tokenization trees for the word “промисловість” (industry) across three East Slavic languages. The Ukrainian tokenization produces semantically consistent morphemes, while Belarusian and Russian models generate more fragmented subword units.

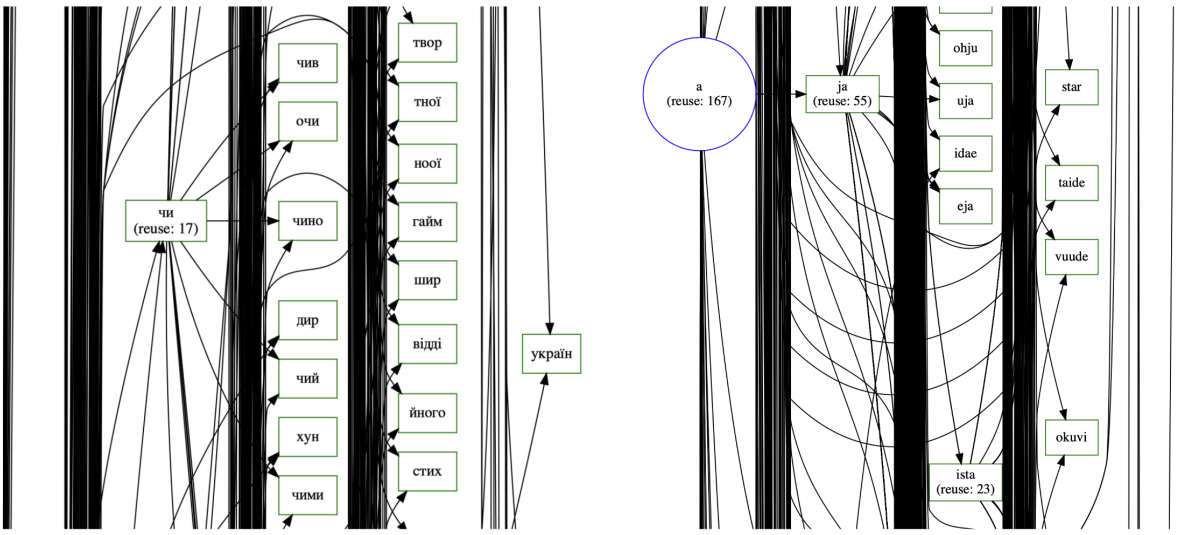

Figure 4: Visualization of interactive BPE merge graphs for Ukrainian (left) and Finnish (right) subword tokenization patterns. The diagrams show directed merge sequences with reuse counts indicated in parentheses. Vertical black bars represent merge steps, while edges show the progression of subword unit formation. The contrasting patterns reflect language-specific morphological characteristics: Ukrainian showing consistent Cyrillic character combinations, while Finnish exhibits agglutinative patterns. Interactive web application available on our repository.

Table 5: Script contamination analysis for languages using non-Latin/non-Cyrillic writing systems. Data shows language names, codes, raw paragraph counts, native script types, and the number of detected Latin and Cyrillic tokens representing potential script contamination.

Lang	**Code**	**Raw pars**	**Script**	**Cyrillic**	**Latin**
Min Nan Chinese	nan	830227	Han	13	541415
Japanese	ja	12737874	Japanese	915	298430
Chinese	zh	7206035	Han	554	117304
Greek	el	1733250	Greek	363	96009
Hebrew	he	3368324	Hebrew	479	90907
Korean	ko	3333575	Hangul	635	81612
Arabic	ar	4624040	Arabic	246	58738
Persian	fa	3132901	Arabic	165	51729
Balinese	ban	87175	Balinese	4	48922
Thai	th	944565	Thai	305	43559
Armenian	hy	1647293	Armenian	2579	41259
Egyptian Arabic	arz	5166127	Arabic	23	39349
Buginese	bug	36776	Lontara	0	19834
Min Dong Chinese	cdo	28589	Han	0	19197
Georgian	ka	644781	Georgian	387	15234
Bengali	bn	854059	Bengali	64	13786
Hindi	hi	703141	Devanagari	80	13369
Tamil	ta	820809	Tamil	28	11153
Hakka Chinese	hak	14204	Han	2	10368
Malayalam	ml	464433	Malayalam	31	9848
Sinhalese	si	166475	Sinhala	3	9552
Burmese	my	451126	Burmese	8	9467
Urdu	ur	660706	Arabic	62	9172
Cantonese	yue	373575	Han	53	8418
South Azerbaijani	azb	605161	Arabic	18	7459
Goan Konkani	gom	39569	Devanagari	0	6955
Marathi	mr	349639	Devanagari	10	6842
Cambodian	km	99641	Khmer	13	6324
Telugu	te	832028	Telugu	3	6170
Tachelhit	shi	7811	Tifinagh	0	5799