

Article

Cross-Family Stem-Suffix Boundary Segmentation in Turkic and Uralic Languages: A ByT5 Study of Azerbaijani, Turkish, Hungarian, and Finnish

Kamran Ibiyev 

Doctoral School of Sciences and Technology, Pázmány Péter Catholic University, 1088 Budapest, Hungary; ibiyev.kamran@itk.ppke.hu

Abstract

Turkic and Uralic languages share almost everything except ancestry: both are agglutinative, both stack long suffix chains onto a stem, both use vowel harmony, and both put the verb last. That coincidence is what makes the pairing worth testing. If a neural model learns where to place a stem–suffix boundary in one family, does the skill carry over to the other, given that nothing but structure is shared? We fine-tuned ByT5-small, a byte-level encoder–decoder, on segmentation pairs built from UniMorph 4.0 and Universal Dependencies for Azerbaijani, Turkish, Hungarian, and Finnish. To our knowledge, no earlier segmentation benchmark covers these four languages together, and the SIGMORPHON 2022 shared task included only Hungarian among them. These targets contain exactly one boundary at most, a stem plus one undivided suffix block, so the supervised task is single-boundary segmentation, not full morpheme decomposition. On this task, in-domain exact match is high across a 4×4 train-on-one, test-on-another matrix run over three seeds, 0.91 to 0.98, against a copy-input baseline of 0.02 to 0.20. Cross-family transfer is not: Finnish-to-Azerbaijani reaches 0.640 against 0.208 the other way ($p < 10^{-66}$), and an exploratory reading of the twelve pairs suggests surface typology tracks this gap more closely than genealogy does, though one high-overlap pair dominates and the sample is too small to settle the question. The central finding comes from a multi-boundary stress test: on 800 words we hand-annotated from Wikipedia (model-correctness agreement was perfect, Cohen’s $\kappa = 1.00$), the models stall at two morphemes, and the 37% of words with three or more come out almost entirely wrong (one correct prediction in 897). Boundary-level scoring shows the shape of the failure: precision stays high, up to 0.97, while recall drops to 0.39–0.52, so the boundaries the models place are mostly right, they just place far too few. This ceiling is consistent with the two-part training targets rather than with the byte-level architecture, although only one architecture under one supervision regime was tested, so confirming the attribution will take retraining on fully decomposed targets. We release all code, data, and annotations to support that step.



Academic Editor: Paulo Quaresma

Received: 8 July 2026

Revised: 27 July 2026

Accepted: 27 July 2026

Published: 31 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: stem–suffix boundary segmentation; morpheme boundary detection; agglutinative languages; ByT5; byte-level models; cross-lingual transfer; low-resource NLP; language typology; Turkic languages; Uralic languages; natural language processing

1. Introduction

Splitting a word into its smallest meaningful units, the task known as morpheme segmentation, feeds a wide range of natural language processing applications: machine

translation, information retrieval, speech recognition, and language modeling all draw on it. The harder a language's morphology, the more it matters. In agglutinative languages a single orthographic word can carry what English would spread across an entire phrase or clause. Turkic and Uralic languages are textbook cases. They mark tense, case, possession, and other grammatical features with long suffix chains stacked onto a stem, and they impose vowel harmony, which constrains which suffixes can co-occur with which stems [1–3].

One clarification belongs at the very start, because it frames everything that follows. The supervision we can build from inflection tables is two-part by construction: a stem plus one undivided block holding all suffix material. The task our models are trained on is therefore single stem–suffix boundary prediction, not full morpheme decomposition, and the in-domain scores below measure exactly that. Full decomposition enters only through the out-of-domain evaluation, which acts as a multi-boundary stress test of what two-part supervision fails to teach. This distinction drives the central finding of the paper.

One especially demanding downstream application is morphological reinflection in context swapping a word in a sentence for the form its context calls for, as in the in-context (cloze) track of the CoNLL–SIGMORPHON reinflection shared task [4]. That task asks for more than boundary placement, and word-level segmentation, our subject here, is a tractable piece of it rather than a stand-in for it. Section 5.4 spells out the gap.

For all this typological similarity, the Turkic and Uralic families are genealogically unrelated. They were once lumped together under the discredited Ural–Altaic hypothesis; modern linguistic consensus keeps them separate [5]. What they share is structural, not lexical: agglutination, vowel harmony, suffix stacking, SOV word order, but no common vocabulary and no common morphological exponents. That combination makes them a natural experiment. It lets us ask whether character-level neural models carry boundary-placement behavior across genealogical boundaries when only the typological structure is shared.

Three observations motivate the study. The first is coverage. The SIGMORPHON 2022 shared task looked at nine languages, but omitted Turkish, Azerbaijani, and Finnish, three of the most morphologically productive languages anywhere. The second concerns the model. ByT5 was built for character-level work, yet most of its applications have been in NER, translation, and grapheme-to-phoneme conversion; how it fares on stem–suffix segmentation across several agglutinative families has never had a careful evaluation. The third is about transfer. Earlier work reports that intra-family transfer usually beats cross-family transfer on standard NLP tasks, but whether that gap holds at the sub-word level, where two unrelated families happen to share structural features, is still open.

This paper studies stem–suffix boundary segmentation across four agglutinative languages, Azerbaijani [aze], Turkish [tur], Hungarian [hun], and Finnish [fin], using one character-level architecture, ByT5-small. Section 3.3 details how the two-part targets are built. We offer two primary and two exploratory contributions. **(1, primary)** Our central result is a stress test. Evaluating the trained models on a new manually annotated out-of-domain dataset (Supplementary Materials), the first setting in these experiments where words demand more than one boundary, we document a recurring under-segmentation failure: predicted depth saturates near two segments whatever the true depth, boundary precision stays high while recall collapses, and accuracy on words with three or more morphemes is near zero. The pattern is uniform over four languages and three seeds, and it is what training on two-part targets would be expected to produce, an interpretation we defend, and bound, in Section 5.4. **(2, primary)** We build and release a four-language stem–suffix segmentation benchmark from UniMorph 4.0 and Universal Dependencies treebanks, to our knowledge the first segmentation benchmark to place these four languages together (the SIGMORPHON 2022 morpheme-segmentation data include Hungarian but none of

the other three), together with the 800-word annotated stress-test set. **(3, exploratory)** We measure cross-family transfer in a 4×4 train-on-X-test-on-Y matrix over three seeds, and show that boundary-placement transfer is strongly and significantly asymmetric, whereas an apparent intra-family asymmetry seen with one seed does not survive multi-seed evaluation, a methodological caution we foreground in the discussion. **(4, exploratory)** We report a correlation in which surface-typological overlap tracks transfer more closely than genealogical relatedness does. One high-overlap pair dominates this correlation and twelve pairs are far too few to establish a predictor, so we present it as an observation awaiting validation on a much larger language sample. All code is released, and the annotated out-of-domain dataset is available for research use under the Creative Commons Attribution 4.0 International license.

2. Related Work

2.1. Morpheme Segmentation as an NLP Task

Morpheme segmentation has a long history in computational linguistics. The field moved from rule-based finite-state systems to unsupervised statistical models, and lately to neural sequence-to-sequence approaches. Morfessor [6] brought in an unsupervised method built on the minimum description length principle, and it remains a strong baseline in low-resource contexts. Much of the recent progress traces back to the SIGMORPHON shared tasks. Several of these target morphological generation. The CoNLL-SIGMORPHON 2018 shared task on universal morphological inflection added a contextual (cloze) track that asked systems to inflect a word to fit its sentential context [4], and a companion 2022 task took on typologically diverse morphological inflection [7]. Among these formulations, in-context inflection comes closest to the real-world setting that motivates this work. The others, segmentation included, are more tractable proxies. The limitation they inherit from the UD/UniMorph labels they rest on is taken up in Section 5.4. The SIGMORPHON 2022 shared task on morpheme segmentation [8] covered nine languages, and the best system reached an average boundary-F1 of 97.29%. Hungarian was on the list; Turkish, Azerbaijani, and Finnish were not. One caveat on comparability: our primary metric is exact match, not boundary-F1, so our absolute scores do not line up directly with those shared-task figures.

Neural approaches usually frame the task as character-level sequence-to-sequence prediction [9,10] or character-level sequence labeling [11,12]. Byte-level alternatives like ByT5 [13] have started to draw attention, but a thorough benchmark spanning several agglutinative families is still absent.

2.2. Character-Level and Byte-Level Models

Character- and byte-level models sidestep the morpheme-boundary errors that subword tokenizers can introduce. ByT5 [13], a token-free variant of T5 that works directly on UTF-8 bytes, holds its own against subword baselines on multilingual tasks and copes better with noisy input. One theme keeps recurring in this literature: tokenization granularity interacts with morphological structure. That alone is a reason to look at byte-level models on agglutinative morphology specifically. Yet no published study has put ByT5 through a thorough evaluation for stem-suffix segmentation across multiple agglutinative families.

2.3. Cross-Lingual Transfer and Language Typology

Cross-lingual transfer learning is central to low-resource NLP. Bankula and Bankula [14] survey transfer across language families and report intra-family transfer above 0.70 against cross-family transfer that can drop below 0.50, with the gap widening for typologically distant pairs. The lang2vec library [15] supplies typological, geographical, and phylogenetic

vectors derived from URIEL, and the URIEL+ update [16] broadens the coverage. For all this infrastructure, no one has put a number on cross-family transfer at the morpheme-segmentation level, least of all between Turkic and Uralic.

2.4. Turkic and Uralic NLP

A growing body of work tackles Turkic and Uralic NLP, though usually one family at a time. The TurkicNLP toolkit [17] offers a unified multi-task pipeline for the Turkic family (tokenization, morphological analysis, part-of-speech tagging, dependency parsing, and named entity recognition), and Tukeyev and Rysbek [18] develop morphology-aware tokenization for Kazakh. Parallel Universal Dependencies treebanks have appeared recently for several Turkic languages, Azerbaijani and Turkish among them [19]. The Uralic side arrives at much the same conclusion from work on tokenization: morphology-aware segmentation lifts downstream performance for agglutinative low-resource languages [20]. Our study sits alongside these single-family efforts and goes one step further, placing the two families side by side on a shared task under a single architecture.

2.5. Multi-Agent Large Language Models as a Methodological Direction

A separate line of work coordinates several large language model agents on one problem. Hang et al. [21] study LLM teams that act as teammates in joint problem-solving, and Sultimov et al. [22] guide multi-agent coordination with episodic memory and cognitive task analysis. Neither system addresses segmentation, and we do not treat them as baselines or competitors for the ByT5 models studied here. Their relevance to this paper is methodological. Our annotation pipeline involves steps that such agent teams could plausibly support in future work: drafting candidate segmentations for human review, cross-checking a linguistic analysis against reference grammars, pre-sorting annotator disagreements before adjudication, and coordinating error analysis across four languages. We return to this possibility in Section 5.9.

2.6. The Gap This Paper Fills

Pieces of this space are covered by existing resources: SIGMORPHON 2022 supplies gold-segmented Hungarian data [8], CSE-based datasets and the TurkicNLP toolkit target Kazakh and other Turkic languages [17,18], and parallel UD treebanks add morphological annotation for several Turkic languages [19]. Against that background, three gaps remain, to our knowledge. (i) The most agglutinative Turkic and Uralic languages have never been benchmarked together for stem–suffix segmentation. (ii) ByT5 has not been put to a careful test on this task. (iii) Cross-family transfer between Turkic and Uralic has not been measured at the morpheme-boundary level. We close all three, and push the analysis further with an out-of-domain evaluation that the original design never anticipated.

3. Materials and Methods

3.1. Task Formulation

We cast stem–suffix segmentation as a sequence-to-sequence task. Given an input word w (a sequence of UTF-8 bytes), the model emits a segmented output made up of a stem s and zero or more suffixes separated by a delimiter,

$$f(w) \rightarrow s + \sigma_1 + \sigma_2 + \cdots + \sigma_n,$$

where ‘+’ marks the morpheme boundary. Bare forms come back unchanged, and in every case the stem and suffixes, once concatenated, must reconstruct the original input. In the training data n never exceeds one (Section 3.3): every target is either a bare form or a stem plus a single suffix block. The supervised task is therefore single stem–suffix boundary

prediction. We keep the general term segmentation for readability, flag wherever the single-boundary regime limits what a result can show, and use the out-of-domain evaluation (Section 4.4) as a multi-boundary stress test. Following T5’s multi-task convention, we prepend the task prefix segment: to each input word. Applying that prefix at inference time is not optional: drop it and the output degrades across the board.

3.2. Model Architecture

The model is ByT5-small [13], a 299.6M-parameter encoder–decoder transformer that runs directly on raw UTF-8 bytes. As ByT5’s design dictates, the encoder is far deeper than the decoder, 12 encoder layers against 4 decoder layers, with a hidden dimension of 1472, a feed-forward dimension of 3584, 6 attention heads, and a vocabulary of 384 tokens (256 byte values plus special tokens). Operating at the byte level avoids the wrong morpheme-boundary tokenization that subword vocabularies can impose, and that makes it a natural fit for morphologically rich languages.

3.3. Training Data

For each language we assemble a training set from two open-license sources. UniMorph 4.0 [23] contributes linguist-verified inflection tables for all four languages, and Universal Dependencies v2 [24] contributes gold (lemma, word_form) pairs from UD_Azerbaijani-TueCL, UD_Turkish-IMST, UD_Turkish-BOUN, UD_Hungarian-Szeged, and UD_Finnish-TDT.

A multi-case procedure converts inflection triples (lemma, inflected_form, features) into segmentation pairs. It handles simple suffixation, verb infinitive stripping (with language-specific suffix sets: -mək/-maq for Azerbaijani, -mek/-mak for Turkish, -ni for Hungarian, and the first-infinitive markers -da/-dä/-ta/-tä for Finnish), per-family consonant-mutation tables, and a longest-common-prefix fallback. Whatever the case, the procedure outputs a two-part target of the form stem + suffix, with everything after the stem concatenated into one suffix token, no matter its internal morpheme structure. This construction matters for reading the out-of-domain results in Section 4.4: the models see only two-part segmentations in training and never witness deeper decomposition.

We deduplicate the pairs and split them 85/10/5 into train/validation/test under a fixed random seed, which yields test sets of 350 (aze), 12,824 (tur), 46,079 (hun), and 2553 (fin) examples. Zero-boundary targets, bare forms that come back unchanged, make up 6.4% (Azerbaijani), 4.5% (Turkish), 2.5% (Hungarian), and 19.4% (Finnish) of the training pairs, and 4.0%, 4.8%, 2.4%, and 19.9% of the corresponding test sets. No split contains a target with more than one boundary. The split runs over (lemma, form) pairs rather than disjointly by lemma, so different inflected forms of one lemma can land in different partitions; as a result, 72–99% of test stems also turn up (under a different inflection) in training. Section 4.1 measures what this overlap does to generalization by reporting accuracy separately on the unseen-stem subset. The natural per-language sizes are wildly imbalanced, from 6992 to 921,558 pairs, so the combined-family models down- or up-sample each constituent language to 50,000 pairs before concatenation. That gives 100,000-pair Turkic and Uralic training sets and a 200,000-pair All-Four set. The quotas draw from each language’s training split, which means the Azerbaijani and Finnish pools, both short of 50,000 pairs, are up-sampled with replacement (some of their forms therefore recur in the combined sets), while Turkish and Hungarian, with far larger pools, are sub-sampled without replacement.

3.4. Cross-Family Transfer Measurement

We train one monolingual model per language and run each over every test set, producing a 4×4 matrix of exact-match scores; diagonal cells capture in-distribution

performance and off-diagonal cells capture transfer. We also train three combined models, Turkic (AZ+TR), Uralic (HU+FI), and All-Four (AZ+TR+HU+FI), and evaluate each on all four test sets to gauge multi-family interference.

3.5. Linguistic Distance Metrics

We correlate off-diagonal transfer exact match against three distance metrics. (a) Suffix overlap is the Jaccard similarity of curated orthographic suffix-string inventories, made up of 53 (aze), 64 (tur), 66 (hun), and 50 (fin) high-frequency nominal and verbal suffix forms, listed in Appendix B. Since the inventories are orthographic, morphologically equivalent suffixes that differ only in spelling (Azerbaijani *-lər* versus Turkish *-ler*, for example) do not match; the metric tracks surface string overlap, not abstract morpheme identity. (b) Vowel overlap is the Jaccard similarity of vowel inventories (Appendix B); Azerbaijani and Turkish get a shared Turkic vowel inventory, so their vowel overlap is 1.0 by construction. (c) An intra-family indicator is a binary value for same-family pairs. For each metric we report Pearson and Spearman correlations with the off-diagonal transfer exact match ($n = 12$ pairs). All inventories and the overlap functions ship with the code.

We do not report lang2vec syntactic, phonological, or genetic distances. The URIEL distance function was unavailable in our computing environment and returned null values for these metrics, so we could not compute them; the analysis is therefore limited to the three surface metrics above, a limitation we revisit in Section 5.8.

3.6. Out-of-Domain Evaluation Data

To probe generalization beyond the training distribution, and in particular to stress-test the models on naturally occurring words whose gold analysis contains more than one boundary, we gathered naturally occurring text from each language's Wikipedia edition with the wikipedia-api library. For each language we pulled the full text of one high-coverage article (Bakı for Azerbaijani, İstanbul for Turkish, Budapest for Hungarian, Helsinki for Finnish), broke it into sentences, and kept those of 40–250 characters. From those sentences we took word forms of at least five characters, deduplicated by lowercase form, shuffled under a fixed seed (seed 42), and kept the first 200 unique words per language, 800 words in all. The matching seed-42 monolingual model segmented each word, and native or near-native annotators corrected the predictions by hand (Section 4.5). Drawing on a single article per language narrows the scope, a point we take up in Section 5.8.

3.7. Hardware and Training Configuration

All training ran on a high-performance computing cluster with NVIDIA Tesla V100-SXM2-16GB GPUs (Volta architecture; NVIDIA Corporation, Santa Clara, CA, USA), the SLURM scheduler (SchedMD LLC, Lehi, UT, USA), and seven CPU cores per GPU.

ByT5 yields NaN loss under fp16 mixed precision: byte-level sequences run several times longer than subword sequences, and the internal activations overflow the fp16 range. The V100, for its part, has no bf16 support. Our training scripts spot the V100 and force fp32 even when fp16 or bf16 flags are passed, which stabilizes training at some cost in throughput. The total compute footprint is modest, on the order of tens of GPU-hours on V100 hardware. Table 1 lists the hyperparameters.

3.8. Experimental Design

3.8.1. Research Questions

The study takes on four linked questions. **Q1:** Can one character-level model reach accurate single-boundary (stem–suffix) segmentation across both families? **Q2:** How much boundary-placement behavior crosses between families that share typological features but no common ancestry? **Q3:** Which linguistic-distance metrics predict transfer success

at this level? **Q4:** Does multi-family training help the individual languages, or does interference drag performance down? A fifth question surfaced during analysis and is taken up in Section 4.4: how well do the trained models generalize to naturally occurring out-of-domain text, in particular to words whose gold analysis contains several boundaries, a multi-boundary stress test the training regime never prepares them for?

Table 1. Training hyperparameters. V100-specific settings.

Hyperparameter	Value
Optimizer	AdamW (weight_decay = 0.01)
Learning rate	1×10^{-4}
Per-device batch size	16 (V100 16 GB)
Gradient accumulation	4 (effective batch 64)
Precision	fp32 (V100 cannot do bf16; fp16 $\rightarrow$ NaN)
Max epochs	20 with early stopping (patience = 5)
Warmup ratio	0.1
Max input/target length	64/96 bytes
Gradient clipping	1.0 (max-norm)
Beam search	4 beams (inference only)
Random seeds	42, 123, 456 (monolingual); 42 (combined)

3.8.2. Models Trained

We train seven configurations. The four monolingual models (AZ, TR, HU, FI) each run with three random seeds (42, 123, 456). The three combined models (Turkic [AZ+TR], Uralic [HU+FI], and All-Four) run with a single seed (42).

3.8.3. Evaluation Metrics

Our primary metric is exact match: the share of predictions whose full segmented output matches the reference. We also compute stem accuracy (the first segment matches) and suffix accuracy (the suffix sequence matches), both reported alongside exact match in Section 4.1. For the out-of-domain stress test, correctness is the case-insensitive exact match between the model's output and the primary annotation pass. Both annotators returned the same binary correct/incorrect verdict on every item (Section 4.5), so this figure equals the annotators' unanimous judgment. Scoring against the stricter released gold is quantified in Section 4.5. On the stress test we additionally report position-level boundary precision, recall, and F1: each segmentation is mapped to the set of character offsets at which it places a boundary, true positives are offsets shared by prediction and reference, and precision, recall, and F1 are micro-averaged over words within each seed and then averaged over the three seeds. These boundary metrics apply to the out-of-domain set, the only setting where words with several boundaries occur; in-domain targets contain at most one boundary, so exact match and boundary detection nearly coincide there.

3.8.4. Baselines

We compare against a rule-based suffix stripper (greedy longest-match removal from a per-language curated inventory, deterministic) and Morfessor 2.0 [6] (unsupervised, trained on the same word frequencies). The All-Four combined model doubles as a multilingual point of comparison. Baselines and the neural model are scored on identical test splits. We also report a copy-input baseline that returns every word unchanged. Since word forms that need no segmentation are returned as they are, this baseline is correct on exactly the zero-boundary items, and it bounds how much exact match is available without placing a single boundary.

4. Results

4.1. In-Domain Performance

Within the UniMorph/UD distribution, ByT5-small segments all four languages accurately, posting exact-match scores of 0.972 (Azerbaijani), 0.978 (Turkish), 0.970 (Hungarian), and 0.914 (Finnish), each averaged over three seeds (Table 2). Across-seed variance is small but uneven, from ± 0.001 (Turkish) up to ± 0.020 (Finnish), which partly tracks test-set size, itself ranging from 350 items (Azerbaijani) to 46,079 (Hungarian). The model clears the unsupervised and rule-based baselines by wide margins on every language: the best baseline result, Morfessor on Hungarian at 0.711, still trails the matching neural score by 0.26, and elsewhere the gap tops 0.45 (Figure 1). The copy-input baseline scores 0.040 (Azerbaijani), 0.048 (Turkish), 0.024 (Hungarian), and 0.199 (Finnish), which is by construction the share of zero-boundary items in each test set (Section 3.3). Trivial copying therefore explains only a small part of the neural scores, one fifth at most, in the Finnish case. The baselines are admittedly weak, and every system here is judged on the same single-boundary targets; we therefore claim no state-of-the-art morpheme segmentation for these languages, and the absolute exact-match figures cannot be compared with the full-decomposition, boundary-F1 results of the SIGMORPHON 2022 systems (Section 2.1). Nor should the in-domain figures be read as performance on full morphological decomposition: they quantify single-boundary placement, and nothing in the in-domain data can test deeper analysis. One more caveat concerns what the in-domain figures actually measure. The stem-suffix targets come out of the automatic conversion in Section 3.3 applied to (linguist-verified) UniMorph and UD forms, so in-domain exact match reflects agreement with that conversion's boundary placement, not with an independently annotated gold standard. The out-of-domain evaluation (Section 4.4), whose targets were segmented from scratch and checked by native-speaker annotators (Section 4.5), is the setting that pits the trained models against human judgment directly. Boundary-level precision, recall, and F1 are reported there (Table 7), where words with several boundaries occur. The in-domain targets contain at most one boundary, so exact match already captures boundary detection in-domain. A stronger neural baseline trained on the same targets is left to future work.

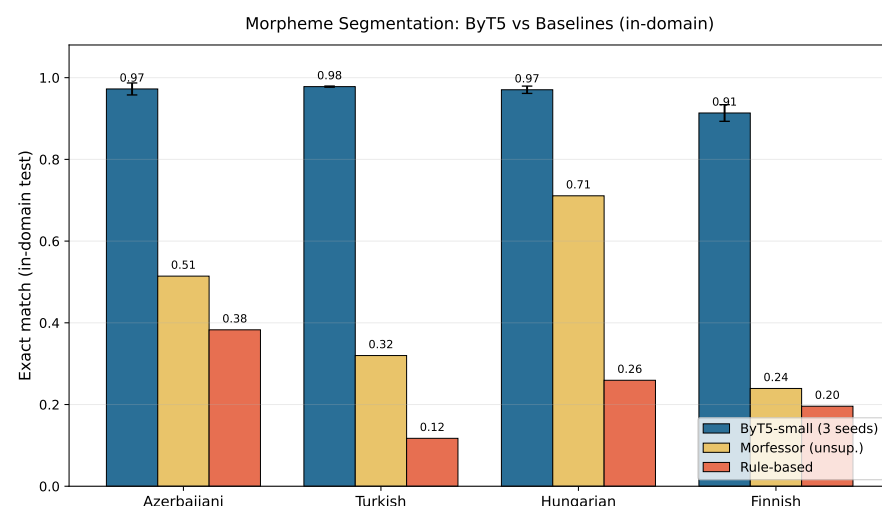


Figure 1. In-domain exact match by system and language: ByT5-small (3-seed mean) against the rule-based and unsupervised Morfessor baselines. The neural model leads on every language by a wide margin (the same values appear in Table 2).

Every training target is two-part (Section 3.3), and that shows up in the metrics. Stem accuracy and suffix accuracy track exact match almost exactly in-domain, staying within 0.006 of it for every language and matching to three decimals for Finnish

(0.914/0.914/0.914). Even in cross-lingual transfer the three barely separate (off-diagonal means: exact match 0.313, stem 0.336, suffix 0.320). We therefore report exact match as the primary metric throughout. Their near-coincidence follows directly from the two-part regime: with only one suffix boundary, there is just one place for the model to go wrong.

Table 2. In-domain exact match. ByT5 is the mean over seeds 42/123/456; baselines are single-run (rule-based deterministic, Morfessor near-deterministic). All systems evaluated on identical test splits ($n = 350/12,824/46,079/2553$). The copy-input baseline returns each word unchanged and is correct on exactly the zero-boundary items.

Model	Azerbaijani	Turkish	Hungarian	Finnish
Copy input	0.040	0.048	0.024	0.199
Rule-based	0.383	0.117	0.259	0.196
Morfessor 2.0	0.514	0.320	0.711	0.239
ByT5 (ours)	0.972	0.978	0.970	0.914

Generalization to unseen stems.

The split is over (lemma, form) pairs (Section 3.3), so most test stems also turn up in training under a different inflection. That overlap could inflate the in-domain figures through stem memorization. To isolate genuinely novel stems, we partition each test set by whether the stem occurs in training, score the two subsets separately, and average over the three seeds (Table 3). Accuracy on the unseen-stem subset stays high: 0.890 ± 0.012 (Turkish, $n = 462$), 0.873 ± 0.034 (Finnish, $n = 725$), and 0.787 ± 0.025 (Hungarian, $n = 280$), roughly 10–20 points below the matching seen-stem subset (Turkish 0.983, Finnish 0.942, Hungarian 0.987). Azerbaijani is the exception only for lack of data: its test set holds just eight unseen-stem items, too few to estimate reliably (mean 0.71, across-seed range 0.63–0.88). So stem overlap does inflate the headline figures. But unseen-stem accuracy is strong and stable, with low across-seed variance and a peak for Finnish, which also has the largest unseen subset. That pattern points to a model learning where boundaries fall rather than one memorizing training stems. We report unseen-stem exact match as the more conservative estimate of in-domain single-boundary generalization. A fully lemma-disjoint split, which would remove the residual overlap altogether, is planned as future work.

Table 3. In-domain exact match split by whether the test stem occurs in training (3-seed mean; $\pm$ std where the subset is large enough to report one). Unseen-stem accuracy is the more conservative estimate of in-domain generalization. The Azerbaijani seen-stem value is derived from the per-seed totals and the unseen-stem tallies; its unseen subset ($n = 8$) is too small for a reliable estimate, so the across-seed range is given instead of a standard deviation.

Language	Seen Stems		Unseen Stems	
	n	Exact Match	n	Exact Match
Azerbaijani	342	0.979	8	0.71 (range 0.63–0.88)
Turkish	12,362	0.983	462	0.890 ± 0.012
Hungarian	45,799	0.987	280	0.787 ± 0.025
Finnish	1828	0.942	725	0.873 ± 0.034

4.2. Cross-Family Transfer

Evaluating each monolingual model on every test set yields the transfer matrix in Table 4 and Figure 2. Transfer is far weaker than in-domain performance: across the twelve off-diagonal cells, mean exact match is 0.313. Two effects in this picture hold up statistically, and one commonly assumed effect is absent.

Table 4. Transfer matrix, exact match (3-seed mean $\pm$ std). The header arrows give the reading direction: rows list the training language (read down, $\downarrow$) and columns the test language (read across, $\rightarrow$), so each cell reports the model trained on its row language and evaluated on its column language. Diagonal cells (bold) are in-domain; off-diagonal cells are zero-shot transfer.

Train $\downarrow$ /Test $\rightarrow$	AZ	TR	HU	FI
AZ	0.972 $\pm$ 0.015	0.554 $\pm$ 0.088	0.250 $\pm$ 0.006	0.208 $\pm$ 0.014
TR	0.552 $\pm$ 0.085	0.978 $\pm$ 0.001	0.170 $\pm$ 0.023	0.222 $\pm$ 0.006
HU	0.125 $\pm$ 0.031	0.072 $\pm$ 0.006	0.970 $\pm$ 0.009	0.192 $\pm$ 0.004
FI	0.640 $\pm$ 0.010	0.420 $\pm$ 0.009	0.353 $\pm$ 0.043	0.914 $\pm$ 0.020

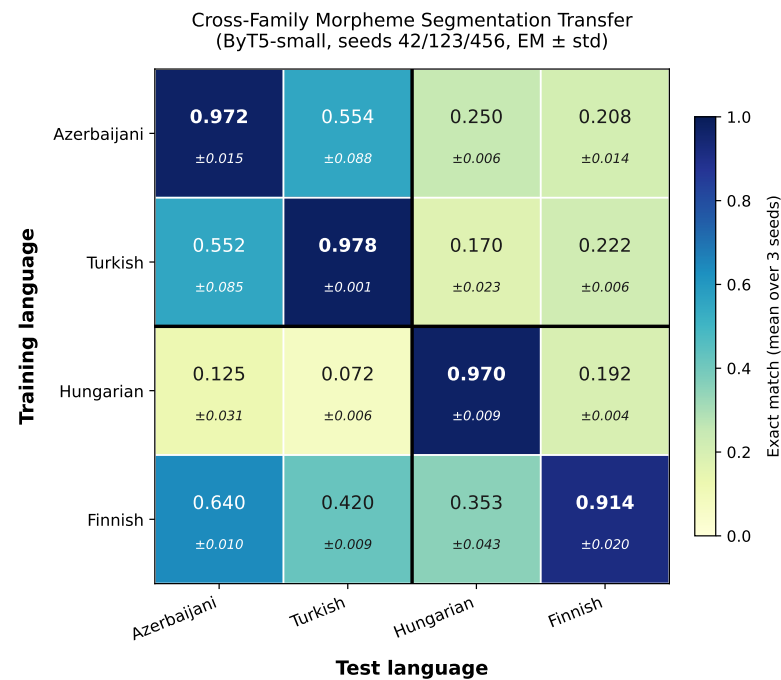


Figure 2. Cross-family transfer matrix (exact match, 3-seed mean). Black lines separate the Turkic (top-left) and Uralic (bottom-right) quadrants.

Asymmetry.

Cross-family transfer is strongly directional. Finnish to Azerbaijani reaches 0.640, while Azerbaijani to Finnish manages only 0.208 ($\Delta = 0.432$, 95% CI [0.379, 0.485], two-proportion $z = 17.3$, $p < 10^{-66}$). The same ordering holds for the Finnish–Turkish pair, 0.420 with Finnish as source against 0.222 in reverse ($\Delta = 0.198$, 95% CI [0.180, 0.217], $z = 18.8$, $p < 10^{-78}$). Finnish is consistently the stronger source; Hungarian is the weakest overall, averaging 0.130 onto the other three languages.

Intra-family symmetry.

Transfer between the two Turkic languages is symmetric. AZ to TR (0.554) and TR to AZ (0.552) are statistically indistinguishable ($\Delta = 0.001$, 95% CI [−0.052, 0.054], $z = 0.08$, $p = 0.94$). A single-seed analysis of the same models had suggested a sizeable Turkic asymmetry, but it did not survive averaging over three seeds, which we report as a caution against single-seed transfer claims. Differences were assessed with two-proportion z -tests. For comparisons on a shared test set (e.g., FI-to-AZ vs. TR-to-AZ) the unpaired test is conservative, so the reported p -value is an upper bound on the paired significance.

4.3. Typological Predictors of Transfer (Exploratory)

Across the twelve cross-language pairs, both surface-typological metrics correlate with transfer exact match (suffix overlap $r = 0.65$, $p = 0.02$; Spearman $\rho = 0.74$, $p = 0.006$; and vowel overlap $r = 0.66$, $p = 0.02$), whereas a binary intra-family indicator does not ($r = 0.39$, $p = 0.21$) (Table 5, Figure 3). Read at face value, this points to shared surface morphology, rather than shared ancestry, predicting transfer at the boundary-placement level, though the caveats below argue against a face-value reading.

These correlations come with heavy caveats, and we present them as exploratory rather than as an established predictor. Twelve pairs leave the analysis underpowered, and a single high-overlap pair dominates the surface metrics: the two Turkic languages (suffix overlap 0.33 and vowel overlap 1.0, against near-zero suffix overlap and lower vowel overlap for every other pair). Drop the Azerbaijani–Turkish pair and neither correlation survives (suffix $r = 0.57$, $p = 0.08$; vowel $r = 0.46$, $p = 0.18$; $n = 10$). The honest summary, then, is narrow: the pair with the highest surface overlap also shows the strongest mutual transfer, but four languages give too few pairs to establish surface similarity as a general predictor. Two further caveats apply. The vowel inventories of Azerbaijani and Turkish were defined identically (a shared Turkic inventory), so their vowel overlap is 1.0 by construction, and the suffix metric compares orthographic suffix-string inventories (Section 3.5), not abstract morpheme inventories.

Table 5. Correlation between distance metrics and off-diagonal transfer exact match ($n = 12$ pairs). Both surface metrics lose significance when the Azerbaijani–Turkish pair is removed (suffix $r = 0.57$, $p = 0.08$; vowel $r = 0.46$, $p = 0.18$; $n = 10$), so the correlations are exploratory. lang2vec metrics are not reported (see Section 3.5).

Distance Metric	Pearson r	p	Spearman ρ
Suffix-inventory overlap	+0.65	0.02	+0.74
Vowel-inventory overlap	+0.66	0.02	+0.60
Intra-family indicator	+0.39	0.21	+0.36

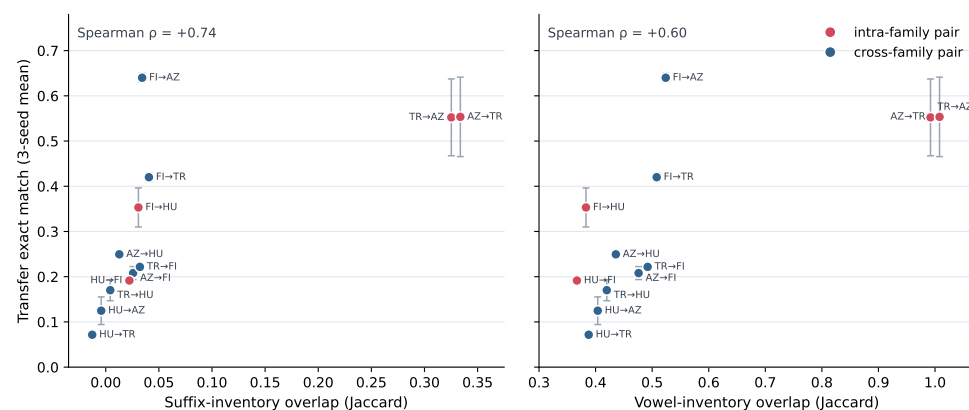


Figure 3. Transfer exact match against suffix-inventory overlap (left) and vowel-inventory overlap (right) for the twelve cross-language pairs, with per-cell standard-deviation error bars. Points are colored by intra- vs cross-family.

4.4. Out-of-Domain Evaluation as a Multi-Boundary Stress Test

We evaluated each monolingual model on 200 manually annotated Wikipedia words per language (800 total; Section 3.6). Exact match drops sharply from the in-domain figures: 0.468 (Azerbaijani), 0.428 (Turkish), 0.532 (Hungarian), and 0.353 (Finnish), each a three-seed mean, against in-domain scores of 0.91–0.98 (Figure 4). The copy-input baseline

reaches 0.310, 0.215, 0.260, and 0.140 respectively, close to the share of one-morpheme words in each sample. The models clear it everywhere, but the comparison shows how much of the out-of-domain exact match was available without segmenting anything. The domain gap (0.44–0.56) is significant in every language. In each case the in-domain score sits above the upper bound of the out-of-domain 95% bootstrap interval (10,000 resamples; Finnish, for instance, scores 0.914 in-domain against an out-of-domain interval of [0.290, 0.418]). Across-seed standard deviations of the out-of-domain scores stay small (0.006–0.020), which tells us the drop belongs to the training regime, not to any one model instance. Part of the gap reflects a shift in morphological depth, though, not genre alone. The in-domain test sets contain only one- and two-segment targets (no in-domain example has three or more segments; Section 3.3), whereas 37% of the out-of-domain words have three or more morphemes. The out-of-domain data is thus the first setting in which the models meet deep segmentation at all, and we return to this distinction when interpreting the result (Section 5.4).

The errors follow a clear pattern. Under-segmentation, where the model emits fewer morphemes than the gold analysis, accounts for 72–92% of all errors per language. Morphological depth governs the effect (Figure 5). Exact match stays high for one-morpheme words and drops to a lower but non-trivial level at two morphemes (0.52–0.86 across languages), then collapses once a word has three or more morphemes, where accuracy is effectively zero: one correct prediction out of 897 model–seed evaluations across the four languages (per-depth counts and accuracies in Appendix A). The mechanism is a ceiling on output length. Predicted segmentation depth saturates near two morphemes whatever the true depth of the input, reaching a mean of 2.0 even for five- and six-morpheme gold forms (Table 6). Since exact match demands every boundary, a two-morpheme output simply cannot match a five-morpheme target. And these deeply stacked words are not a fringe case: they make up 37% (299 of the 800 gold words) of the naturally occurring Wikipedia vocabulary in our sample.

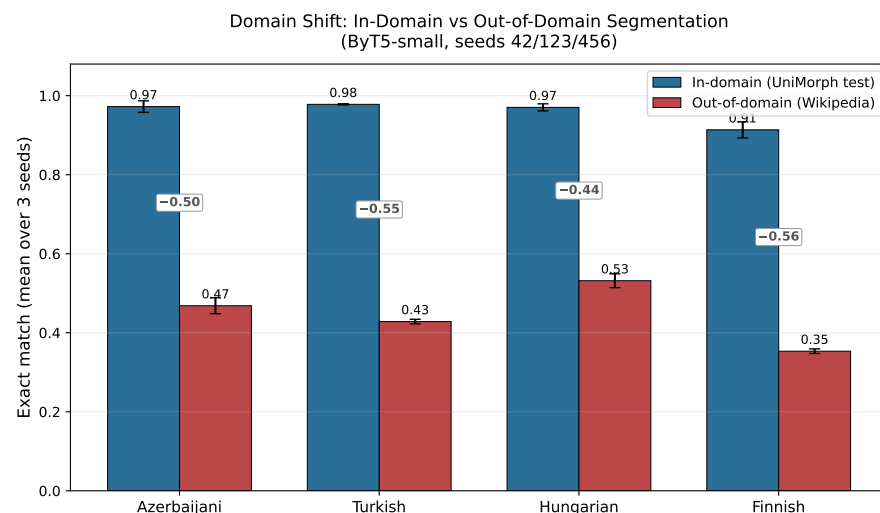


Figure 4. In-domain versus out-of-domain exact match per language (3-seed means; the per-pair gap is annotated).

Table 6. Mean predicted segmentation depth by gold morpheme count (mean over three seeds). Predicted depth saturates near two regardless of the true number of morphemes. A dash indicates no out-of-domain words at that gold depth for that language.

Gold Morphemes	AZ	TR	HU	FI
1	1.30	1.02	1.27	1.10
2	1.93	1.78	1.75	1.88
3	1.96	1.83	1.85	1.93
4	2.00	1.96	1.83	2.00
5	2.00	2.00	—	2.00
6	—	2.00	—	—

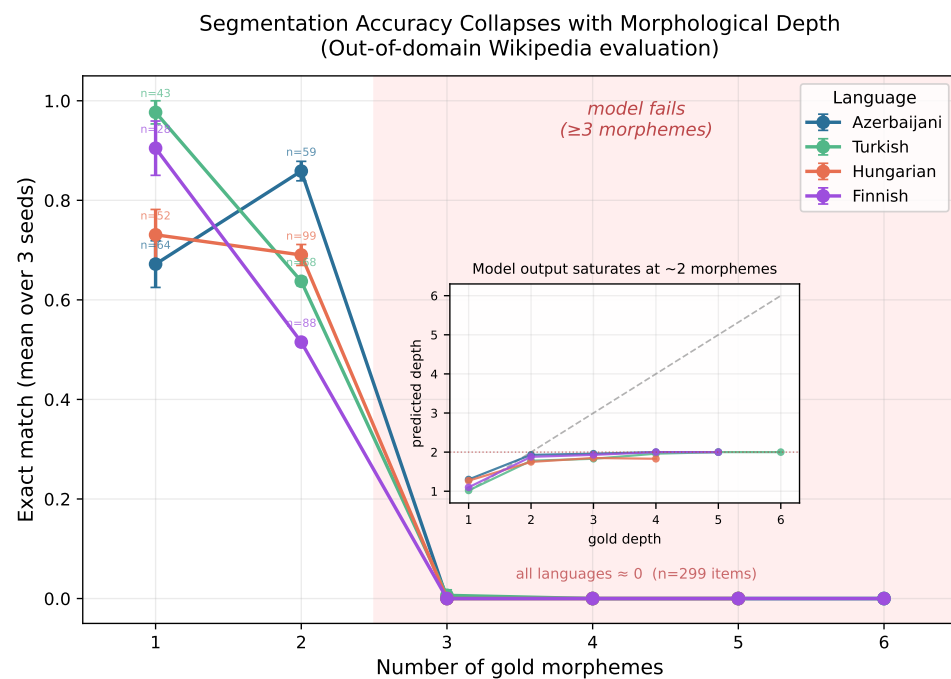


Figure 5. Out-of-domain exact match by gold morpheme count. Accuracy collapses at three morphemes for all four languages. Inset: mean predicted depth against gold depth, showing saturation near two morphemes.

Boundary-level scoring puts the same failure in metric form (Table 7). Precision is high for three of the four languages, 0.83–0.97, and reaches 0.97 for Turkish: when a model places a boundary, that boundary is almost always in the gold analysis. Recall tells the opposite story, 0.39–0.52 everywhere, because a model that stops at one boundary cannot recover the second and third boundaries that deeper words carry. Finnish is the outlier on precision as well (0.71), consistent with its harder in-domain task and its weaker out-of-domain exact match. The gap between precision and recall is the under-segmentation pattern itself: the models are conservative and mostly right about where to cut, they simply cut far too rarely. A morpheme-level variant of the same metrics, which compares the multisets of predicted and gold segments and needs no exclusions, gives the same ordering (F1 0.525, 0.496, 0.570, and 0.375 for Azerbaijani, Turkish, Hungarian, and Finnish).

Table 7. Boundary-level precision, recall, and F1 on the multi-boundary stress test (micro-averaged within each seed, then mean $\pm$ std over seeds 42/123/456), scored case-insensitively against the same reference as the exact-match figures. A total of 37 of the 800 items (1 Azerbaijani, 2 Turkish, 34 Finnish) are excluded because their gold analysis restores an underlying stem whose length differs from the surface form (for example Finnish *alussa* $\rightarrow$ *alku* + *ssa*), which leaves no well-defined surface boundary positions; Turkish proper-noun apostrophes are removed before computing positions.

Language	<i>n</i>	Precision	Recall	F1
Azerbaijani	199	0.834 $\pm$ 0.016	0.490 $\pm$ 0.007	0.617 $\pm$ 0.010
Turkish	198	0.970 $\pm$ 0.007	0.428 $\pm$ 0.004	0.594 $\pm$ 0.004
Hungarian	200	0.833 $\pm$ 0.022	0.517 $\pm$ 0.014	0.638 $\pm$ 0.014
Finnish	166	0.708 $\pm$ 0.018	0.386 $\pm$ 0.007	0.500 $\pm$ 0.008

4.5. Annotation Reliability

All 800 words were independently annotated by two native or near-native speakers per language (eight annotators in total). Each annotator was either a native speaker of the language or a near-native speaker who had studied or lived in the relevant country for several years. They contributed annotations voluntarily and consented to the release of those annotations for research use. On the binary judgment of whether the model's output was correct, the two annotators agreed on every single item, giving Cohen's $\kappa = 1.00$ across all languages and all morpheme-depth strata. On the full gold segmentation string, exact agreement averaged 92.4% (Azerbaijani 93.5%, Turkish 96.5%, Hungarian 84.0%, Finnish 95.5%). The two passes thus differed on the gold segmentation for 7.6% of items yet never disagreed on model correctness. That combination confirms the two annotation passes were independent, and it shows the model's errors read as categorical to native speakers even where the precise analysis allows legitimate variation. The 61 disagreements were adjudicated by the first author under a fixed segmentation policy (compounds decomposed; buffer consonants treated as separate morphemes), with a handful of transcription errors corrected along the way. Stated as a protocol: each annotator received their language's 200 words together with the seed-42 model's predicted segmentation for each word (Section 3.6), and corrected each prediction, working independently of the other annotator, into the full segmentation they judged right. A model output counted as correct exactly when it matched the annotation, compared case-insensitively. The finer analytic choices, above all how deeply to decompose Hungarian compounds, were where the two passes diverged, and those disagreements were settled by the predefined criteria just named rather than by case-by-case preference: the same two policy rules were applied uniformly to all 61 items. None of the disagreements touched the binary judgment, so the accuracies reported in this paper carry the annotators' unanimous verdicts and are computed against the primary annotation pass. Adjudication did, however, deepen the analysis of 63 of the 800 items in the released gold standard. Scoring the same predictions against that stricter reference yields lower exact-match figures: 0.463 (Azerbaijani), 0.412 (Turkish), 0.448 (Hungarian), and 0.347 (Finnish). We report the annotation-pass figures throughout because they reflect the unanimous binary judgment. The stricter figures reinforce the under-segmentation finding rather than soften it, since the adjudicated gold decomposes words more deeply than the models ever segment. Hungarian alone accounted for more than half of the disagreements (32 of 61), a reflection of how productive Hungarian nominal compounding is.

4.6. Multi-Family Training (Q4)

We trained three combined models (Turkic [AZ+TR], Uralic [HU+FI], and All-Four) and evaluated each on the four in-domain test sets (Table 8). On the languages a model

was trained on, exact match stays close to the matching monolingual baseline: across all trained-language cells the mean change is -0.012 . For the All-Four model the largest single drop is 0.025 (Turkish, 0.953 vs. 0.978), and its drop on Finnish is 0.016 (0.898 vs. 0.914). The biggest drop across every combined model is 0.026, for the Uralic model on Finnish (0.888 vs. 0.914). So the All-Four model, despite spanning two genealogically unrelated families, lands within 2.5 points of the monolingual models on every language. Azerbaijani even gains a little under joint Turkic training ($+0.017$), in line with helpful signal from the closely related and larger Turkish data. One byte-level model can therefore cover all four languages at negligible cost compared with per-language models. Cross-family transfer from the combined models echoes the monolingual asymmetry: the Uralic model transfers to Azerbaijani at 0.474, below the single Finnish model's 0.640, which suggests that blending the weaker Hungarian source with Finnish dilutes Finnish's unusually strong cross-family transfer. The combined models were trained with a single seed (42), and monolingual seed variance reaches ± 0.020 (Finnish), so the sub-0.03 differences fall within seed noise and should not be read as genuine interference.

Table 8. Combined-model exact match (single seed 42). Bold marks languages in the model's training set; the final row repeats the monolingual three-seed means for reference.

Model	AZ	TR	HU	FI
Turkic (AZ+TR)	0.989	0.957	0.345	0.246
Uralic (HU+FI)	0.474	0.208	0.955	0.888
All-Four	0.971	0.953	0.961	0.898
Monolingual (ref.)	0.972	0.978	0.970	0.914

5. Discussion

5.1. What the Transfer Results Suggest About Predictors

The clearest pattern in the transfer matrix is directional: Finnish is a strong source and a weak target, Hungarian is the weakest source, and family membership alone does not settle pair strength, since Finnish transfers to Azerbaijani (0.640) better than Azerbaijani's close Oghuz relative Turkish does (0.552).

Beyond this qualitative observation we explored whether surface-typological overlap predicts transfer quantitatively. Both suffix- and vowel-inventory overlap correlate with transfer ($r \approx 0.65$, $p < 0.05$) while a binary family indicator does not, but we stress that this is exploratory: with twelve pairs the correlation is underpowered and does not survive removing the single Turkic pair (Section 4.3). We therefore do not claim that surface similarity governs transfer. The most we can say is suggestive and directional (the one pair with high surface overlap also shows the strongest mutual transfer), and a properly powered test of the surface-similarity hypothesis would require many more language pairs. We flag this as a hypothesis for future work rather than an established predictor.

A further interpretive caution applies to all of the transfer results. Because the models are trained only to place a single stem-suffix boundary, the matrix measures transfer of boundary placement, not transfer of full morphological analysis; and boundary placement may be driven partly by surface cues such as stem-length and orthographic regularities rather than by morphological knowledge per se. It is worth being precise about what a zero-shot success shows. When a Finnish-trained model reproduces the automatically defined stem-suffix boundary of an Azerbaijani word, it shows that a string-level cutting pattern carried over. It does not show that abstract morphological rules or grammatical categories transferred. The models never see suffix identities or morphosyntactic features as explicit targets, so there is nothing in the supervision that would force a representation of case, number, or possession as such. What transfers is best described as an orthographic

boundary-placement pattern, and the distinction matters because the two readings support very different claims about multilingual representation. We therefore describe the transfer findings as boundary-placement transfer throughout, and read “Finnish is a strong source” as a statement about that sub-task. Distinguishing surface from genuinely morphological transfer (for instance by testing whether transfer correlates with stem-length distributions, by probing for morphosyntactic categories in the encoder, or by retraining on fully decomposed targets) is left to future work.

5.2. Hypotheses on Why Finnish Exports Well and Hungarian Poorly

The source-language ranking deserves a linguistic reading, and we offer one here explicitly as a set of hypotheses rather than conclusions; the present design cannot separate them. One candidate is the shape of Finnish case endings, which are string-distinctive and phonotactically regular (-ssa/-ssä, -sta/-stä, -lla/-llä, -lle, -ksi). A model trained on Finnish may learn a conservative, shape-driven cutting rule that ports to any language whose suffixes sit in a similar final-syllable window. The data offer a second candidate: the Finnish training set has by far the largest share of zero-boundary bare forms (19.4%; Section 3.3), which may teach the model when not to cut, and that habit would export as the high precision we in fact observe on the stress test. Consonant gradation points the same way. It makes the Finnish in-domain task the hardest of the four (0.914), yet it alters the stem rather than the boundary position, so the cutting pattern itself may stay regular and portable. On the Hungarian side, boundary position is more often lexically conditioned: linking vowels (-ok/-ek/-ak, -om/-em) and stem-final lengthening (a to á, e to é) tie the cut point to properties of the individual stem. Hungarian orthography also leans on digraphs (sz, gy, ny, cs) and a larger accented-vowel inventory, which makes its byte-level suffix contexts more language-specific, and its productive nominal compounding gives the notion of stem a wider range than in the other three languages, a fact that also drove most annotation disagreements (Section 4.5). The datasets themselves differ too, in provenance and in how often each conversion rule fired (Section 3.3), so part of the ranking may live in the data rather than the languages. Each of these is testable, for instance by ablating bare forms from the Finnish training set or by controlling suffix shape across sources, and we leave those probes to future work.

5.3. Asymmetry and the Value of Multi-Seed Evaluation

The directional asymmetry (Finnish strong as a source and weak as a target, Hungarian weakest of all as a source) holds steady across seeds. The intra-family Turkic case is instructive for a different reason: a single-seed reading of these same models suggested a substantial AZ–TR asymmetry, yet three-seed evaluation shows the two directions to be statistically identical ($p = 0.94$). Transfer asymmetries of the magnitude routinely reported in the literature can be artifacts of a single training run, and only multi-seed evaluation reliably separates signal from seed variance.

5.4. Under-Segmentation and the Two-Part Training Regime

The out-of-domain evaluation exposes a limitation that in-domain testing could not reveal. By construction, every training target takes the two-part form *stem + suffix*: the conversion from inflection tables concatenates all material following the stem into a single suffix token, regardless of how many morphemes it contains (Section 3.3). As a result, both the training data and the in-domain test sets consist entirely of one- and two-morpheme segmentations: across all four languages, no in-domain example has three or more segments. The models are therefore never trained on, nor in-domain evaluated on, deeper segmentation.

The Wikipedia evaluation is the first setting in which the models encounter words with three or more gold morphemes, which make up 37% of that vocabulary. On these words accuracy is effectively zero, and the predicted segmentation depth saturates near two regardless of the true depth (Table 6): the model reproduces the only pattern its supervision contained. The out-of-domain drop therefore reflects two compounded distribution shifts: a genre and vocabulary shift, and, more fundamentally, a shift into a morphological-depth regime entirely absent from training. We read this ceiling as a property of the inflection-table-derived supervision rather than of the byte-level architecture, and the evidence is consistent with that reading: the effect is uniform over four languages and three seeds, the depth saturation sits exactly where the two-part targets sit, and boundary precision stays high while recall collapses, which is the signature of a model that learned one cut well and never saw a second. Strictly, though, this study trains one architecture under one supervision regime, so the attribution remains a well-supported interpretation rather than a demonstrated cause. Settling it requires retraining the same architecture on fully decomposed targets (the SIGMORPHON 2022 corpora would serve), or comparing several architectures under identical two-part supervision. A model trained on fully decomposed segmentations should not face this ceiling, and testing that prediction is the clearest experiment this paper sets up.

This qualifies the common claim that byte-level models cope well with morphologically rich languages. They do cope, but only within the shallow segmentation regime they are trained on. On the deep suffix stacks that characterize natural agglutinative text, a model trained on two-part targets under-segments across the board.

A related conceptual point concerns what kinds of boundaries the stress test contains. The training pairs come from lemma to inflected-form relations, so the only boundary type the models ever see is inflectional, and even that is collapsed into a single block. The annotation policy for the gold standard, by contrast, decomposes compounds and treats linking elements such as buffer consonants as separate units, so the stress-test gold contains compound-internal, derivational, and linking-element boundaries on top of inflectional ones. Part of the out-of-domain gap is therefore a category shift on top of the depth shift: the models are asked for boundary types their supervision never defined. This bears directly on the per-language readings. Hungarian, where nominal compounding is most productive, drew most of the annotation disagreements (32 of 61) and the lowest full-string agreement (84.0%), and its stress-test gold is correspondingly rich in compound boundaries; its relatively high out-of-domain exact match (0.532) rides largely on the one- and two-morpheme items. Distinguishing performance by boundary type, rather than by depth alone, would sharpen this analysis and is a natural extension of the released annotations.

A second, orthogonal gap separates even full morpheme decomposition from the in-context reinflection task that ultimately motivates this work (Section 2.1). Correct segmentation identifies the morphemes present in an observed form, whereas reinflection must choose the surface form a suffix should take when a word is substituted into a new context, and that choice can depend on distinctions (such as the locative versus oblique function of a case suffix) that the UD/UniMorph feature bundles used to build our targets do not record. A model trained to reproduce annotation-derived segmentations therefore cannot, on its own, resolve the context-sensitive allomorphy that real reinflection requires. We regard this as a limitation of the supervision available to the whole task family rather than of our models specifically, and as further motivation for moving from type-level segmentation toward context-aware evaluation.

5.5. Bytes Are Not Characters

All four languages write in Latin-based scripts, but that does not make their byte sequences comparable in a simple way. The language-specific letters (Azerbaijani ə, ı, ş, ğ; Turkish ç, ö, ü; Hungarian á, é, ő, ű; Finnish ä, ö) occupy two UTF-8 bytes each, while the shared ASCII letters occupy one, so ByT5 does not process each visible character as one atomic unit, and the same suffix can differ in byte length across harmony variants (Finnish -ssa is three bytes, -ssä is four). One consequence is that cross-language transfer may partly ride on encoding-level material: many high-frequency suffixes begin with shared single-byte ASCII consonants (l, d, n, t, s), so a cutting pattern keyed to those bytes would carry across languages for reasons that have little to do with morphology. The suffix-overlap correlation of Section 4.3 is computed over orthographic strings and cannot separate this byte-sequence explanation from a morphological one. Another is that byte-level and character-level representations should not be treated as equivalent: a boundary one edit apart in characters can be two bytes apart in the model's input space, and the per-language mix of one- and two-byte letters differs. The boundary metrics in Table 7 are computed over characters, which is the annotation-level view. A byte-level error analysis, and a controlled test of whether transfer tracks shared byte n-grams, are open follow-ups.

5.6. What Single-Boundary Segmentation Is Useful for?

The introduction motivates segmentation through downstream applications, and the single-boundary regime obliges us to say which of those applications the present system can actually serve. A stem plus one undivided suffix block supports stemming-style uses: query and index normalization in information retrieval, lemma lookup, keyword and dictionary matching, deduplication of inflectional variants, and seeding or evaluating subword tokenizers. It does not supply what full decomposition provides. Morphological reinflection, feature-level tagging and parsing, interlinear glossing, and generation into an agglutinative language all need the individual morphemes, usually with their grammatical functions, and a system that returns *stem + rest* gives them neither the inner boundaries nor the labels. The out-of-domain results show the gap concretely: on the 37% of natural words with three or more morphemes, the analysis these applications would need is exactly the part the models fail to produce. Claims about downstream usefulness should therefore be read as claims about the stemming-style family of uses until a model trained on fully decomposed targets closes that gap.

5.7. Multi-Family Models

Training a single All-Four model costs at most 2.5 points of in-language exact match relative to per-language models (Turkish), and that difference is within seed noise. For deployment, one All-Four model is therefore an attractive alternative to four monolingual ones. The combined models do not, however, improve cross-family transfer; if anything, mixing a weak source (Hungarian) with a strong one (Finnish) dilutes the strong source's advantage.

5.8. Limitations

Several limitations apply. First, the four-language scope yields only twelve off-diagonal pairs, which caps the statistical power of the distance-transfer correlation; broader claims would need more language pairs. Second, *lang2vec* syntactic, phonological, and genetic distances could not be computed in our environment and are therefore absent from the correlation analysis; we report only the three surface metrics. Third, the combined-family models were trained with a single seed, so the multi-family interference estimates are less precise than the monolingual results. Fourth, the training data is dominated by

inflection tables and treebanks, which underrepresent derivational and compounding morphology, the very phenomena that drive the out-of-domain failures we report. Fifth, gold-segmentation disagreements were adjudicated by a single author under a fixed policy rather than by annotator consensus; the binary model-correctness judgment, on which annotators agreed perfectly, is unaffected, but the released gold segmentations reflect this adjudication choice. Sixth, the out-of-domain evaluation draws each language's 200 words from a single Wikipedia article (the city article for that language). A multi-article sampling was intended, but the first article's full text already supplied the target sentence count, so the collection terminated before reaching further articles. Consequently the out-of-domain vocabulary, while naturally occurring and unseen in training, is topically narrow (dominated by geography, history, and administrative terms), which limits the breadth of the out-of-domain morphology and means the absolute out-of-domain accuracies should be read as indicative rather than representative of each language as a whole. The morphological-depth finding (Section 5.4) survives this narrowing, however, since deep forms are abundant even within a single article (37% of the sampled words).

Seventh, and most fundamentally, the in-domain task is single-boundary segmentation: every training target is two-part by construction (Section 3.3), so the high in-domain exact match and the transfer matrix describe stem-suffix boundary placement, not full morpheme decomposition. The full-decomposition task is exercised only in the out-of-domain evaluation, where performance collapses. Eighth, the train/validation/test split is performed over (lemma, form) pairs rather than disjointly by lemma, so 72–99% of test stems also occur in training. We measure the effect directly (Section 4.1): unseen-stem exact match is roughly 10–20 points below the overall figure but remains high (0.79–0.89 for the three languages with sufficient unseen items), indicating that the headline numbers are inflated by stem overlap but that boundary placement does generalize to novel stems. A lemma-disjoint split would remove the inflation entirely and is planned as future work. Ninth, the learned baselines (a rule-based stripper and unsupervised Morfessor) are weak, and the copy-input baseline bounds only the trivial part of the task; we do not compare against established morphological analysers or strong neural segmenters for these languages, and in-domain we report exact match rather than the boundary-F1 used by the SIGMORPHON line, so our in-domain numbers are not directly comparable to that literature. Boundary-level metrics are reported for the stress test only (Table 7). Tenth, the study evaluates one architecture (ByT5-small) under one supervision regime (two-part targets), so statements tying the under-segmentation ceiling to the supervision are interpretations consistent with the evidence, not demonstrated causes. The controls that would settle them are listed in Sections 5.4 and 5.9. Finally, on research ethics, all four languages have substantial speaker communities and institutional support, so the work does not raise the equity concerns common to endangered-language NLP.

5.9. Future Directions

The natural next steps follow directly from these limitations. The most pressing is to train on gold-segmented data rather than inflection tables, which would test directly whether the under-segmentation ceiling is an artifact of supervision, together with a lemma-disjoint re-split of the in-domain data. A second is to extend the language set, both within Turkic and Uralic and beyond them, both to strengthen the correlation analysis and to put the surface-similarity hypothesis to a properly powered test. A third moves from words to sentences: sentence-level segmentation requires resolving morpheme boundaries jointly with word boundaries and so comes closer to the conditions of downstream applications, and connects directly to in-context reinflection [4], the application that most directly motivates this line of work and the one against which annotation-derived targets show the

discrepancy noted in Section 5.4. A fourth direction concerns the annotation pipeline itself. Building the stress-test gold involved drafting candidate segmentations, double annotation, and adjudication of disagreements, all by hand. Coordinated teams of large language model agents [21,22] could support these steps: generating candidate segmentations for human review, checking an analysis against reference grammars, pre-sorting annotator disagreements before adjudication, and running error analysis across languages in parallel. We see this line of work as methodological inspiration for scaling the annotation and analysis workflow, not as an alternative segmenter to benchmark against.

6. Conclusions

We presented a study of single-boundary stem–suffix segmentation across Azerbaijani, Turkish, Hungarian, and Finnish using one byte-level architecture, together with a multi-boundary stress test of what that training regime fails to teach. The stress test carries the central finding. On 800 hand-annotated Wikipedia words the models stall at two segments whatever the true depth and fail almost completely on the 37% of words with three or more morphemes. Boundary precision stays high (up to 0.97) while recall falls to 0.39–0.52. The most plausible reading ties this ceiling to the two-part training targets rather than the architecture, and retraining on fully decomposed targets is the experiment that would settle it. In-domain exact match is 0.91–0.98 (0.79–0.89 on unseen stems), transfer is strongly asymmetric with Finnish the best source, and the apparent Turkic asymmetry disappears under multi-seed evaluation. More tentatively, surface-typological overlap tracks transfer better than genealogy does, but that association is an exploratory observation that needs validation on a substantially larger language sample before it can count as a predictor. All code, data, and annotations are released.

Supplementary Materials: The supplementary data supporting this article are available at the DOI URL [<https://doi.org/10.5281/zenodo.21182942>], which archives a snapshot of the project’s GitHub repository, version 1.0.0 (tag v1.0.0), [<https://github.com/ibiyev/cross-family-morph>] (accessed on 26 July 2026), the tagged release used for this paper. The repository contains the data-preparation and training scripts, the curated suffix and vowel inventories, the derived stem–suffix segmentation datasets, the redistributable Universal Dependencies treebank files, the out-of-domain Wikipedia annotation pipeline and gold-standard files, all evaluation results and significance tests, and the scripts that reproduce every table and figure in this paper. See the LICENSE file in the repository for the mixed-license terms governing Hungarian-derived content.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. The study did not involve medical, clinical, or behavioral experimentation on human subjects or animals. The only human involvement was the voluntary expert annotation of publicly available word forms by the authors and collaborators, which does not constitute human subjects research under the host institution’s guidelines.

Informed Consent Statement: Informed consent was obtained from all annotators involved in the study. All annotators were informed of the research purpose and consented to the open release of their annotations under a Creative Commons Attribution 4.0 International license, permitting reuse, including for commercial purposes.

Data Availability Statement: All code, data-preparation scripts, SLURM submission scripts, and evaluation results are openly available at [<https://doi.org/10.5281/zenodo.21182942>], an archived snapshot of the project’s public GitHub repository. Content is released under mixed licenses, documented in full in the repository’s LICENSE file: the derived stem–suffix datasets for Azerbaijani, Turkish, and Finnish (including the Turkic combined set) are released under Creative Commons Attribution 4.0 International (CC BY 4.0), matching the licenses of their UniMorph 4.0 [23] (CC BY-SA 3.0) and Universal Dependencies sources [24] (UD_Azerbaijani-TueCL, UD_Turkish-BOUN,

UD_Turkish-IMST, UD_Finnish-TDT; all CC BY-SA 4.0). The Hungarian-derived datasets, the Uralic combined set, and the All-Four combined set contains content derived from UD_Hungarian-Szeged, which is licensed Creative Commons Attribution-NonCommercial-ShareAlike 3.0 (CC BY-NC-SA 3.0); these files, along with the redistributed raw Hungarian treebank, are therefore released under CC BY-NC-SA 3.0 rather than CC BY 4.0, restricting their reuse to non-commercial purposes. The morpheme-segmentation annotations created for the out-of-domain evaluation are released under CC BY 4.0; the underlying source words are drawn from Wikipedia and remain under its CC BY-SA license, to which we attribute and link rather than redistribute under altered terms. All random seeds are set explicitly (42, 123, 456 for the monolingual runs; 42 for the combined models).

Acknowledgments: I thank the volunteer native and near-native speakers who produced the out-of-domain annotations. I am also grateful to Gábor Prószték and Attila Novák (Pázmány Péter Catholic University) for their supervision and guidance throughout this work.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
UD	Universal Dependencies
OOD	Out-of-Domain
MDL	Minimum Description Length
SOV	Subject–Object–Verb (word order)
CI	Confidence Interval

Appendix A. Out-of-Domain Accuracy by Morpheme Depth

Table A1 reports the per-language out-of-domain exact match and word counts at each gold morpheme depth, underlying Figure 5. Accuracy is the mean over three seeds; the standard deviation across seeds is shown. Words with three or more gold morphemes (299 of 800, 37%) have effectively zero accuracy in every language.

Table A1. Out-of-domain exact match by gold morpheme depth (mean $\pm$ std over seeds 42/123/456), with the number of gold words at each depth.

Lang	Gold Depth	<i>n</i> Words	Exact Match
AZ	1	64	0.672 $\pm$ 0.047
AZ	2	59	0.859 $\pm$ 0.020
AZ	3	47	0.000 $\pm$ 0.000
AZ	4	24	0.000 $\pm$ 0.000
AZ	5–7	6	0.000 $\pm$ 0.000
TR	1	43	0.977 $\pm$ 0.023
TR	2	68	0.637 $\pm$ 0.008
TR	3	52	0.006 $\pm$ 0.011
TR	4	23	0.000 $\pm$ 0.000
TR	5–6	14	0.000 $\pm$ 0.000
HU	1	52	0.731 $\pm$ 0.051
HU	2	99	0.690 $\pm$ 0.021
HU	3	38	0.000 $\pm$ 0.000
HU	4–5	11	0.000 $\pm$ 0.000

Table A1. Cont.

Lang	Gold Depth	<i>n</i> Words	Exact Match
FI	1	28	0.905 ± 0.055
FI	2	88	0.515 ± 0.007
FI	3	51	0.000 ± 0.000
FI	4	22	0.000 ± 0.000
FI	5–7	11	0.000 ± 0.000

Appendix B. Linguistic Inventories Used for the Distance Metrics

The suffix-overlap and vowel-overlap metrics (Section 3.5) are computed as Jaccard similarities over the following inventories, which are also released with the code. The suffix inventories are sets of orthographic suffix strings; after de-duplication they contain 53 (Azerbaijani), 64 (Turkish), 66 (Hungarian), and 50 (Finnish) entries. The resulting pairwise overlaps are: Azerbaijani–Turkish suffix 0.33/vowel 1.00; Azerbaijani–Finnish 0.03/0.50; Turkish–Finnish 0.04/0.50; Hungarian–Finnish 0.03/0.38; Azerbaijani–Hungarian 0.00/0.41; Turkish–Hungarian 0.00/0.41.

Vowel inventories.

Azerbaijani and Turkish share a single Turkic inventory {e, i, ö, ü, ə, a, ı, o, u}; Hungarian {e, é, i, í, ö, ő, ü, ű, a, á, o, ó, u, ú}; Finnish {ä, ö, y, a, o, u, e, i}.

Suffix inventories.

The complete suffix-string lists for each language are provided in the released `utils/morphology.py`. They comprise high-frequency nominal case, possessive, and plural suffixes together with common verbal and copular endings; representative examples are Azerbaijani -lar/-lər, -da/-də, -nın/-nin; Turkish -ler/-lar, -de/-da, -nın/-nın; Hungarian -ban/-ben, -nak/-nek, -ok/-ek/-ak; and Finnish -ssa/-ssä, -lla/-llä, -sta/-stä.

References

1. Göksel, A.; Kerslake, C. *Turkish: A Comprehensive Grammar*; Routledge: London, UK, 2005.
2. Karlsson, F. *Finnish: An Essential Grammar*; Routledge: London, UK, 2008.
3. Rounds, C. *Hungarian: An Essential Grammar*; Routledge: London, UK, 2001.
4. Cotterell, R.; Kirov, C.; Sylak-Glassman, J.; Walther, G.; Vylomova, E.; McCarthy, A.D.; Kann, K.; Mielke, S.J.; Nicolai, G.; Silfverberg, M.; Yarowsky, D.; Eisner, J.; Hulden, M. The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 1–27.
5. Csátó, É.Á.; Johanson, L. *The Turkic Languages*; Routledge: London, UK, 1998.
6. Virpioja, S.; Smit, P.; Grönroos, S.-A.; Kurimo, M. *Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline*; Aalto University Publication Series SCIENCE + TECHNOLOGY 25/2013; Aalto University: Espoo, Finland, 2013.
7. Kodner, J.; Khalifa, S.; Batsuren, K.; Dolatian, H.; Cotterell, R.; Akkus, F.; Anastasopoulos, A.; Andrushko, T.; Arora, A.; Atanalov, N.; et al. SIGMORPHON–UniMorph 2022 Shared Task 0: Generalization and Typologically Diverse Morphological Inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 176–203.
8. Batsuren, K.; Bella, G.; Arora, A.; Martinovic, V.; Gorman, K.; Žabokrtský, Z.; Ganbold, A.; Dohnalová, Š.; Ševčíková, M.; Pelegrinová, K.; et al. The SIGMORPHON 2022 Shared Task on Morpheme Segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 103–116.
9. Kann, K.; Schütze, H. Single-Model Encoder-Decoder with Explicit Morphological Representation for Reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 555–560.

10. Wehrli, S.; Clematide, S.; Makarov, P. CLUZH at SIGMORPHON 2022 Shared Tasks on Morpheme Segmentation and Inflection Generation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 212–219.
11. Gırrbach, L. SIGMORPHON 2022 Shared Task on Morpheme Segmentation: Sequence Labelling for Word-Level Morpheme Segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 124–130.
12. Shao, Y. Cross-Lingual Word Segmentation and Morpheme Segmentation as Sequence Labelling. *arXiv* **2017**, arXiv:1709.03756.
13. Xue, L.; Barua, A.; Constant, N.; Constant, N.; Al-Rfou, R.; Narang, S.; Kale, M.; Roberts, A.; Raffel, C. ByT5: Towards a Token-Free Future with Pre-Trained Byte-to-Byte Models. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 291–306.
14. Bankula, A.; Bankula, P. Cross-Linguistic Transfer in Multilingual NLP: The Role of Language Families and Morphology. *arXiv* **2025**, arXiv:2505.13908.
15. Littell, P.; Mortensen, D.R.; Lin, K.; Kairis, K.; Turner, C.; Levin, L. URIEL and lang2vec: Representing Languages as Typological, Geographical, and Phylogenetic Vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 8–14.
16. Khan, A.; Shipton, M.; Anugraha, D.; Duan, K.; Hoang, P.H.; Khiu, E.; Doğruöz, A.S.; Lee, E.-S.A. URIEL+: Enhancing Linguistic Inclusion and Usability in a Typological and Multilingual Knowledge Base. In *Proceedings of the 31st International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 6937–6952.
17. Hakimov, S. TurkicNLP: An NLP Toolkit for Turkic Languages. *arXiv* **2026**, arXiv:2602.19174.
18. Tukeyev, U.; Rysbek, B. Morphology-Aware Segmentation and Tokenization for Turkic Languages: A CSE-Guided Framework (The Kazakh Case). *Information* **2026**, *17*, 128. <https://doi.org/10.3390/info17020128>.
19. Akhundjanova, A.; Akkurt, F.; Chontaeva, B.; Eslami, S.; Çöltekin, Ç. Parallel Universal Dependencies Treebanks for Turkic Languages. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 129–136.
20. Xu, N.; Kim, A. Tokenization and Morphological Fidelity in Uralic NLP: A Cross-Lingual Evaluation. *arXiv* **2026**, arXiv:2602.04241.
21. Hang, C.N.; Tan, C.W.; Chiu, D.M. LLM Teams: Harnessing Large Language Models as Multi-Agent Teammates for Joint Problem-Solving. In *Proceedings of the Thirteenth ACM Conference on Learning @ Scale (L@S '26)*; ACM: New York, NY, USA, 2026. <https://doi.org/10.1145/3774398.3811571>.
22. Sultimov, R.; Volkov, A.; Mitrovic, M.; Maximov, Y.V. LLM-Guided Multi-Agent Evacuation Coordination via Episodic Memory and Cognitive Task Analysis. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '26)*; International Foundation for Autonomous Agents and Multiagent Systems: Richland, SC, USA, 2026; pp. 4164–4166. <https://doi.org/10.65109/SEYR5220>.
23. Batsuren, K.; Goldman, O.; Khalifa, S.; Habash, N.; Kieraś, W.; Bella, G.; Leonard, B.; Nicolai, G.; Gorman, K.; Ate, Y.G.; et al. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*; European Language Resources Association (ELRA): Marseille, France, 2022; pp. 840–855.
24. Nivre, J.; de Marneffe, M.-C.; Ginter, F.; Hajič, J.; Manning, C.D.; Pyysalo, S.; Schuster, S.; Tyers, F.; Zeman, D.; et al. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*; European Language Resources Association (ELRA): Marseille, France, 2020; pp. 4034–4043.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.