DOI: **10.37943/25SCDM2312**

Yerlan Karabaliyev
PhD candidate, Clever System
y.karabaliyev@iitu.edu.kz, orcid.org/0009-0001-9465-3998
International Information Technology University, Kazakhstan
Kateryna Kolesnikova
Doctor of Technical Sciences, Professor
kkolesnikova@iitu.edu.kz, orcid.org/0000-0002-9160-5982
International Information Technology University, Kazakhstan

KAZMORPHLM: MORPHEME-AWARE LANGUAGE MODEL FOR KAZAKH AUTOMATIC SPEECH RECOGNITION

Abstract: This paper presents KazMorphLM, a morpheme-aware language model for automatic speech recognition (ASR) in the Kazakh language. Kazakh belongs to the Turkic family and is characterised by a highly agglutinative morphology, in which a single root can generate a large number of inflected forms through productive suffixation. This property causes severe data sparsity for conventional word-level language models and reduces recognition accuracy.

The proposed model introduces three main innovations. First, a rule-based morpheme segmenter uses an inventory of 230 suffixes across fourteen grammatical categories and includes phonological validation through vowel harmony and consonant assimilation rules. Second, a two-level interpolated n-gram architecture combines a 7-gram morpheme-level model with a 5-gram word-level model using an interpolation ratio of 0.6 to 0.4 and Witten–Bell smoothing. Third, a four-channel rescoring mechanism integrates acoustic confidence, word-level and morpheme-level language-model probabilities, and a vowel-harmony consistency score.

KazMorphLM was integrated into a hybrid ASR pipeline combining NVIDIA FastConformer and Meta MMS-1B acoustic models. On the FLEURS test set, the system achieves a word error rate of 6.86%, a 14.6% relative improvement over word-level KenLM rescoring. The results indicate that higher-order morpheme modelling is essential for agglutinative languages and that corpus quality outweighs corpus size. The approach is applicable to other morphologically rich Turkic languages.

Keywords: morpheme language model; Kazakh speech recognition; agglutinative morphology; vowel harmony; morpheme segmentation; n-gram interpolation; ASR rescoring; Turkic languages; low-resource ASR.

Introduction

Automatic Speech Recognition (ASR) systems have achieved near-human performance for high-resource languages such as English, yet under-resourced languages with complex morphological systems remain a significant challenge [1, 2]. The Kazakh language, spoken by over 13 million people as the state language of Kazakhstan, exemplifies these challenges due to its agglutinative morphology, productive suffixation, and strict vowel harmony constraints [3, 4].

A fundamental limitation of existing Kazakh ASR systems is their reliance on word-level language models. In agglutinative languages like Kazakh, a single nominal root such as "кітап" (book) can generate over 100 distinct inflected forms through combinations of plural, possessive, and case suffixes: "кітаптар" (books), "кітаптарым" (my books), "кітаптарымда" (in my books), "кітаптарымдағы" (the one in my books). Word-level n-gram models treat each of these as an independent vocabulary entry, leading to severe data sparsity and poor generalization [4, 5].

Previous work on Kazakh ASR has reported WER ranging from 21.9% to 56.87% [6, 7, 23]. While recent pre-trained models such as Meta MMS [8] and NVIDIA FastConformer [9] have improved acoustic modeling through transfer learning, the language modeling component remains

Received: 12.02.2026 Accepted: 25.02.2026 Published: 30.03.2026

a critical bottleneck. Morpheme-level language models have shown promise for other agglutinative languages, including Turkish [8] and Finnish [9, 10]. Recent work on Finnish and Estonian ASR has further confirmed the value of morpheme-level modelling for agglutinative languages [25, 26]. However, no such model has been developed for Kazakh ASR.

This paper makes the following contributions:

• KazMorphLM - a morpheme-aware language model for Kazakh featuring a rule-based segmenter with 230 suffixes across 14 grammatical categories, incorporating vowel harmony and consonant assimilation rules specific to Kazakh phonology;

• A two-level interpolated n-gram architecture combining 7-gram morpheme and 5-gram word models with Witten-Bell smoothing, optimized for agglutinative structure;

• A four-channel rescoring mechanism that integrates acoustic, lexical, morphological, and phonological (vowel harmony) scores;

• Experimental validation achieving 6.86% WER on FLEURS, with ablation studies demonstrating the critical role of morpheme-level modeling and corpus curation.

Morphological Challenges of the Kazakh Language

Kazakh belongs to the Kipchak branch of the Turkic language family and is characterized by agglutinative morphology, where grammatical information is encoded through sequences of suffixes appended to a root [11, 12]. Understanding these morphological properties is essential for designing an effective language model.

1. Agglutinative Suffixation

Kazakh employs extensive suffixation to express grammatical categories, including number, possession, case, tense, person, voice, mood, and negation. A single word can carry up to 5–6 suffixes. For example, the verb form "жазғандарымыз" (those that we wrote) decomposes as:

$$жаз + ған + дар + ымыз \rightarrow write + PAST.PTCP + PL + POSS.1PL \ . \qquad (1)$$

This productive morphology means the number of possible word forms grows combinatorially. A word-level language model must observe each combination independently, while a morpheme-level model can learn compositional suffix patterns from far fewer examples.

2. Vowel Harmony

Kazakh enforces strict vowel harmony, where all vowels within a word must belong to either the back (hard) class or the front (soft) class. This constraint determines the allomorphic form of each suffix, as shown in Table 1:

Table 1. Kazakh Vowel Harmony Classes

Vowel Class	Vowels	Example
Back (жуан)	а, о, у, ы	қала+лар (cities)
Front (жіңішке)	ә, ө, ү, і	көл+дер (lakes)

The vowel harmony score of a word is computed as:

$$H(w) = max(|V\_back|, |V\_front|) / (|V\_back| + |V\_front|) \ , \qquad (2)$$

where V_back and V_front are the sets of back and front vowels in word w. The neutral vowel "е" can occur with either class. A well-formed Kazakh word has H(w) = 1.0; violations indicate a recognition error, providing a signal for ASR rescoring.

3. Consonant Assimilation

The initial consonant of a suffix undergoes assimilation based on the final phoneme of the preceding morpheme, as summarised in Table 2:

Table 2. Consonant Assimilation: Locative Suffix

Stem-Final Class	Suffix Form	Example
Voiceless (к, п, т, с)	-та/-те	мектеп+те (at school)

Voiced (б, д, ж, з)	-да/-де	қала+да (in city)
Nasal (м, н, ң)	-нда/-нде	маман+нда (at specialist)

The segmenter validates all allomorphic variants against the phonological context. Kazakh consonants are classified into voiceless (к, қ, п, с, т, ф, х, ц, ч, ш, щ), voiced (б, в, г, ғ, д, ж, з), sonorants (л, м, н, ң, р, й), and nasals (м, н, ң) as a subgroup.

KazMorphLM Architecture

KazMorphLM consists of three components: (1) a morpheme segmenter that decomposes Kazakh words into stems and suffixes; (2) a two-level interpolated n-gram language model; and (3) a four-channel rescoring mechanism for ASR hypothesis selection.

1. Morpheme Segmenter

The segmenter employs a greedy right-to-left matching algorithm with vowel harmony validation. Given an input word, the algorithm iteratively matches the longest known suffix, validates vowel harmony compatibility, and recurses until no further suffixes can be extracted or the stem reaches the minimum length of 3 characters. The maximum recursion depth is 5 suffixes per word.

The suffix inventory comprises 230 morphemes organized into 14 grammatical categories, enumerated in Table 3:

Table 3. Suffix Inventory: Grammatical Categories

Category	**Count**	**Examples (back/front)**	**Function**
Plural	3	-лар/-лер, -дар/-дер, -тар/-тер	Number marking
Case	20	-ның/-нің (GEN), -ға/-ге (DAT), -да/-де (LOC)	7 grammatical cases
Possessive	13	-ым/-ім (1SG), -ың/-ің (2SG), -сы/-сі (3SG)	Ownership
Denominal	17	-лы/-лі, -сыз/-сіз, -шы/-ші	Noun→Adj/Noun
Deverbal	14	-ған/-ген, -ушы/-уші, -ғыш/-гіш	Verb→Noun/Adj
Tense	11	-ды/-ді (PST), -ады/-еді (AOR)	Temporal
Person	9	-мын/-мін (1SG), -сың/-сің (2SG)	Agreement
Voice	10	-ыл/-іл (PASS), -тыр/-тір (CAUS)	Pass/Caus/Refl
Mood	5	-са/-се (COND), -айын/-ейін (OPT)	Modality
Negation	3	-ма/-ме, -ба/-бе, -па/-пе	Verbal negation
Question	3	-ма/-ме, -ба/-бе, -па/-пе	Interrogative
Comparative	2	-рақ/-рек	Degree
Diminutive	2	-шық/-шік	Size/Affection
Total	**230**		**14 categories**

Each suffix entry stores back-vowel and front-vowel allomorphs, grammatical category, and optional special forms for post-voiceless, post-nasal, and post-vocalic contexts. The inventory is pre-sorted by length (longest first) for greedy matching.

Representative segmentation outputs are shown in Table 4:

Table 4. Example Morpheme Segmentations

Input Word	Segmentation	Gloss
балаларға	бала + лар + ға	child + PL + DAT
мектептерде	мектеп + тер + де	school + PL + LOC
жазғандарымыз	жаз + ған + дар + ымыз	write + PTCP + PL + POSS.1PL
оқушылардың	оқу + шы + лар + дың	study + AGENT + PL + GEN

2. Two-Level Interpolated N-gram Model

KazMorphLM combines two n-gram language models at different granularity levels. The morpheme-level model operates on a token sequence where each word is replaced by its morpheme sequence with word boundary markers ⟨w⟩:

$$кітаптарымда \rightarrow кітап\ ⟨w⟩\ тар\ ⟨w⟩\ ым\ ⟨w⟩\ да\ ⟨w⟩\ . \quad (3)$$

The morpheme model uses 7-gram order. This high order is essential for Kazakh: with an average of 2.3–2.5 morphemes per word plus boundary tokens, a 7-gram context spans approximately 2–3 words, capturing both intra-word suffix patterns and inter-word dependencies. The word-level model uses 5-gram order on unsegmented tokens.

Both models employ Witten-Bell smoothing:

$$P\_WB(w|c) = C(c,w) / (C(c) + T(c)) \ \ if\ C(c,w) > 0\ , \quad (4)$$

$$P\_WB(w|c) = T(c) / (C(c) + T(c)) \cdot P\_WB(w|c')\ \ otherwise\ , \quad (5)$$

where C(c,w) is the n-gram count, C(c) is the total count of tokens following context c, T(c) is the number of unique types following c, and c′ is the backoff context. Witten-Bell was chosen over Kneser-Ney because morpheme types follow a more uniform frequency distribution than words, making type-based discounting more appropriate [10, 22, 27].

The combined score is computed through linear interpolation:

$$P\_combined(w) = \lambda \cdot \hat{P}\_morph(w) + (1 - \lambda) \cdot P\_word(w),\ \ \lambda = 0.6\ . \quad (6)$$

The morpheme-level score is normalized to per-word granularity: $\hat{P}\_morph(w_1..w_n) = P\_morph(m_1..m_k) \cdot (n/k)$, where n is word count, and k is morpheme token count, preventing the morpheme model from being penalized for producing more tokens per word.

3. Four-Channel Rescoring

KazMorphLM rescores ASR hypotheses using the four scoring channels listed in Table 5:

$$S(h) = \alpha \cdot S\_acoustic(h) + \beta \cdot S\_word(h) + \gamma \cdot S\_morph(h) + \delta \cdot S\_harmony(h)\ . \quad (7)$$

Table 5. Rescoring Channels

Channel	Source	Range	Description
S_acoustic	ASR model	[0, 1]	CTC decoder confidence
S_word	KenLM 5-gram	$\log_{10} \rightarrow [0,1]$	Word-level LM, min-max normalized

S_morph	KazMorphLM	$\log_{10} \rightarrow [0,1]$	Morpheme+word interpolated score
S_harmony	Harmony scorer	[0, 1]	Sentence vowel harmony score

The harmony channel exploits a Kazakh-specific phonological constraint: ASR errors in suffixes frequently produce vowel harmony violations (e.g., back-vowel suffix on a front-vowel stem), which generic language models cannot detect. Grid search yielded optimal weights: $\alpha \in [0.05, 0.1]$, $\beta \in [0.0, 0.2]$, $\gamma \in [0.6, 0.8]$, $\delta \in [0.05, 0.2]$. The results indicate that γ has the largest contribution to the rescoring process, suggesting that morpheme-level language modeling is a key factor in improving recognition accuracy.

ASR Pipeline Integration

KazMorphLM is integrated as the final rescoring stage in a hybrid ASR pipeline: Audio → FastConformer-CTC → MMS-1B → ROVER fusion → KenLM → KazMorphLM → Text.

FastConformer-CTC [7] and MMS-1B [6] serve as complementary acoustic models. MMS-1B uses adapter layers (~138K trainable parameters, 0.01% of 1B total), enabling fine-tuning on consumer GPUs (RTX 4060 Ti, 8 GB VRAM) with fp16 and gradient checkpointing. Both models are fine-tuned on 559 hours of Kazakh speech: KSD (552h) [13, 28] + FLEURS (7h) [14, 29]. ROVER [13] aligns and merges hypotheses through word-level voting, producing a richer candidate set for language model rescoring [15, 16].

Methods and Materials

1. Data

The training data used in this study combines two Kazakh speech corpora, as summarised in Table 6.

Table 6. Training Data

Dataset	Samples	Duration	Description
KSD (OpenSLR-140)	203,480	552 hours	Natural Kazakh speech
FLEURS (Google)	4,425	~7 hours	Read speech, diverse speakers
Combined	**207,905**	**~559 hours**	**Training set**

Samples >15s are filtered (GPU memory), yielding ~202,016 training samples with 90/10 train/validation split. Test set: 100 samples from FLEURS Kazakh test split.

2. Language Model Corpus

KazMorphLM is trained on 181,000 curated Kazakh sentences from Kazakh Wikipedia and Leipzig Kazakh corpus. Preprocessing: lowercasing, punctuation removal, length filtering (3–100 words), Kazakh character validation (ә, ғ, қ, ң, ө, ұ, ү, і, h), deduplication. Average morphemes per word: 2.3–2.5.

3. Model Parameters

The final configuration of KazMorphLM, including language model orders, smoothing, and interpolation weights, is reported in Table 7.

Table 7. KazMorphLM Configuration

Parameter	Value
Morpheme LM order	7-gram
Word LM order	5-gram
Smoothing	Witten-Bell
Interpolation λ	0.6 (morph) / 0.4 (word)
Training corpus	181,000 sentences
Suffix inventory	230 morphemes, 14 categories
Model size	~810 MB

Results

1. Pipeline Stage Evaluation

The word error rate (WER) achieved at each stage of the hybrid ASR pipeline is reported in Table 8.

Table 8. WER at Each Pipeline Stage (100 FLEURS Test Samples)

Pipeline Stage	WER	Relative Δ vs KenLM
FastConformer-CTC	15.28%	—
MMS-1B fine-tuned	8.03%	—
ROVER (FC + MMS)	12.44%	—
+ KenLM 5-gram	8.09%	baseline
+ KazMorphLM	**6.86%**	**14.6%**

A key result is the reduction from 8.09% (KenLM) to 6.86% (KazMorphLM) - a 14.6% relative WER reduction achieved solely by replacing word-level rescoring with morpheme-aware rescoring.

2. Ablation Study: Corpus Composition

The effect of different training-corpus compositions on the final WER is summarised in Table 9.

Table 9. Effect of Corpus Composition on WER

Corpus	Size	WER
Core (Wiki + kazakh_corpus)	**181K sents**	**6.86%**
+ KazakhTTS transcripts	288K sents	7.31%
+ Leipzig news/wiki crawls	1.3M sents	> 8%

Adding KazakhTTS data (136K sentences) degraded WER to 7.31% (+6.6% relative) due to domain mismatch. Adding Leipzig corpora (1.3M sentences) caused further degradation, requiring partly reduction of morpheme n-gram order from 7 to 5 (memory constraint: 7-gram on 1.3M needs ~18+ GB RAM).

3. Ablation Study: N-gram Order

Table 10 reports the effect of varying the morpheme-level n-gram order on WER.

Table 10. Effect of Morpheme N-gram Order

Morph Order	Word Order	WER
7-gram	**5-gram**	**6.86%**
5-gram	5-gram	> 7%

Reducing morpheme order from 7 to 5 on the same corpus caused regression. With ~2.4 morphemes/word + boundary tokens, a 5-gram context spans only ~1.5 words - insufficient for cross-word dependencies. The 7-gram spans ~2-3 words.

4. Rescoring Weight Analysis

Table 11 presents the optimal weight ranges obtained through grid search for the four rescoring channels.

Table 11. Optimal Rescoring Weight Ranges

Weight	Optimal Range	Interpretation
α (acoustic)	0.05 - 0.10	Low: already reflected in hypotheses
β (word LM)	0.0 - 0.20	Low: subsumed by morpheme LM
γ (morph LM)	0.60 - 0.80	Dominant: primary rescoring signal

δ (harmony)	0.05 -0.20	Supplementary phonological check

5. Comparison with Literature

A comparison of the proposed approach with previously reported Kazakh ASR results is provided in Table 12.

Table 12. Comparison with Reported Kazakh ASR Results

Study	Method	Data	WER
Google Research, 2021	Multilingual Model	>100h	21.9%
Baevski et al., 2020	wav2vec 2.0	~10h	23.7%
Yessenbayev et al., 2021	Google STT API	~50h	34.5%
Mamyrbayev et al., 2022	E2E Transfer Learning	~100h	~20%
This work	**Pipeline + KazMorphLM**	**559h**	**6.86%**

KazMorphLM achieves a WER of 6.86%, corresponding to an estimated 68.7% relative improvement over the best previously reported result (21.9%). The ablation study further demonstrates a 14.6% improvement over word-level rescoring alone.

Discussion

The results suggest that morpheme-level language modeling plays an important role for ASR in agglutinative languages. Several key insights emerge.

Why morphemes outperform words. Word-level models face a combinatorial explosion in agglutinative languages [17, 18]. The root "бала" (child) with 3 plural, 13 possessive, and 20 case forms generates hundreds of word forms. A morpheme-level model shares n-grams across all forms: "бала + лар + ымыз + да" shares morpheme context with "кітап + тар + ымыз + да" at all positions except the stem.

The role of vowel harmony scoring. The harmony channel provides an orthogonal signal detecting phonological well-formedness. ASR suffix errors frequently produce vowel harmony violations that the harmony score detects even when the LM assigns a reasonable probability. This language-specific feature is unavailable to generic models.

Witten-Bell vs Kneser-Ney. Morpheme types follow a more uniform frequency distribution than words (most suffixes occur with moderate frequency), making type-based discounting (Witten-Bell) more appropriate than count-based discounting (Kneser-Ney), which is designed for the highly skewed Zipfian distribution of word frequencies.

Corpus quality principle. The finding that 181K curated sentences outperform 1.3M mixed-domain sentences has practical implications: for low-resource languages, investing in corpus curation yields greater returns than accumulating more data [19, 20].

Applicability to Turkic languages. The morphological principles underlying KazMorphLM - agglutinative suffixation, vowel harmony, consonant assimilation are shared across the Turkic family (Turkish, Uzbek, Kyrgyz, Turkmen, Azerbaijani) [21, 22]. The architecture can be adapted by replacing the suffix inventory and harmony rules while retaining the two-level interpolated n-gram framework and four-channel rescoring.

Limitations. The evaluation was conducted on a relatively small test set consisting of 100 samples from the FLEURS dataset. While the results are indicative, they may not fully reflect generalization performance across broader domains and speaker variability. Future work will include evaluation on larger and more diverse test sets to ensure robustness.

Conclusion

This paper presented KazMorphLM, a morpheme-aware language model for Kazakh ASR that addresses the challenges of agglutinative morphology. By combining a rule-based segmenter with a two-level interpolated n-gram architecture and a four-channel rescoring mechanism incorporating vowel harmony, the proposed approach achieves a WER of 6.86%, corresponding to a 14.6% relative improvement over word-level KenLM rescoring.

The results indicate that sufficiently high-order morpheme n-grams (7-gram) are required to capture dependencies spanning multiple words, while corpus quality and domain relevance have a stronger impact on performance than dataset size. In addition, incorporating vowel harmony provides a complementary phonological signal that enhances rescoring effectiveness.

Future work will focus on extending the approach using neural morpheme-level language models, exploring unsupervised segmentation methods such as Morfessor [24], and adapting the framework to other Turkic languages.

References

[1] Khassanov, Y., Mussakhojayeva, S., Mirzakhmetov, A., Adiyev, A., Nurpeiissov, M., & Varol, H. A. (2021). A crowdsourced open-source Kazakh speech corpus and initial speech recognition baseline. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL) (pp. 697–706). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.58

[2] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

[3] Mussakhojayeva, S., Khassanov, Y., & Varol, H. A. (2021). A study of multilingual end-to-end speech recognition for Kazakh, Russian, and English. In Speech and Computer (SPECOM 2021). Lecture Notes in Computer Science, vol. 12997, pp. 448–459. Springer. https://doi.org/10.1007/978-3-030-87802-3_41

[4] Mussakhojayeva, S., Dauletbek, Y., Yeshpanov, R., & Varol, H. A. (2023). Multilingual speech recognition for Turkic languages. Information, 14(2), 74. https://doi.org/10.3390/info14020074

[5] Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2022). Automatic speech recognition for under-resourced languages: A survey. Speech Communication, 140, 79–104. https://doi.org/10.1016/j.specom.2022.04.002

[6] Karabaliyev, Y., & Kolesnikova, K. (2024). Kazakh speech and recognition methods: Error analysis and improvement prospects. Scientific Journal of Astana IT University, 20, 62–75. https://doi.org/10.37943/20HKZC2614

[7] Mamyrbayev, O., Alimhan, K., Oralbekova, D., Bekarystankyzy, A., & Zhumazhanov, B. (2022). Identifying the influence of transfer learning method in developing an end-to-end automatic speech recognition system with a low data level. Eastern-European Journal of Enterprise Technologies, 1(9(115)), 84–92. https://doi.org/10.15587/1729-4061.2022.252801

[8] Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., et al. (2024). Scaling speech technology to 1,000+ languages. Journal of Machine Learning Research, 25(97), 1–52. https://jmlr.org/papers/v25/23-1318.html

[9] Rekesh, D., Koluguri, N. R., Kriman, S., Majumdar, S., Noroozi, V., Huang, H., et al. (2023). Fast Conformer with linearly scalable attention for efficient speech recognition. In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2023) (pp. 1–8). IEEE. https://doi.org/10.1109/ASRU57964.2023.10389717

[10] Smit, P., Virpioja, S., Grönroos, S.-A., & Kurimo, M. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. In Proceedings of the Demonstrations at the 14th

Conference of the European Chapter of the Association for Computational Linguistics (EACL) (pp. 21–24). Association for Computational Linguistics. https://aclanthology.org/E14-2006/

[11] Johanson, L. (2021). Turkic. Cambridge Language Surveys. Cambridge University Press. https://doi.org/10.1017/9781139016704

[12] Toleu, A., Tolegen, G., & Makazhanov, A. (2021). Character-aware neural morphological disambiguation for Kazakh. Cognitive Computation, 13(6), 1480–1490. https://doi.org/10.1007/s12559-021-09926-6

[13] Xu, H., Povey, D., Mangu, L., & Zhu, J. (2021). Minimum Bayes risk decoding and system combination based on a recursion for edit distance. Computer Speech & Language, 65, 101147. https://doi.org/10.1016/j.csl.2020.101147

[14] Bérard, A., Calapodescu, I., Dymetman, M., Roux, C., Meunier, J.-L., & Nikoulina, V. (2021). Efficient inference for multilingual neural machine translation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 8563–8583). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.674

[15] Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., et al. (2021). VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 993–1003). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.80

[16] Ren, Z., Yolwas, N., Slamu, W., Cao, R., & Wang, H. (2022). Improving hybrid CTC/Attention architecture for agglutinative language speech recognition. Sensors, 22(19), 7319. https://doi.org/10.3390/s22197319

[17] Varjokallio, M., Virpioja, S., & Kurimo, M. (2021). Morphologically motivated word classes for very large vocabulary speech recognition of Finnish and Estonian. Computer Speech & Language, 66, 101141. https://doi.org/10.1016/j.csl.2020.101141

[18] Longpre, S., Yauney, G., Reif, E., Lee, K., Roberts, A., Zoph, B., et al. (2024). A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, and toxicity. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (pp. 3245–3276). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.179

[19] Çöltekin, Ç., Doğruöz, A. S., & Çetinoğlu, Ö. (2023). Resources for Turkish natural language processing: A critical survey. Language Resources and Evaluation, 57(1), 449–488. https://doi.org/10.1007/s10579-022-09605-4

[20] Ruokolainen, T., Kohonen, O., Virpioja, S., & Kurimo, M. (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 (CoNLL) (pp. 29–37). Association for Computational Linguistics. https://aclanthology.org/W13-3504/

[21] Mamyrbayev, O., Oralbekova, D., Kydyrbekova, A., Turdalykyzy, T., & Bekarystankyzy, A. (2021). End-to-end model based on RNN-T for Kazakh speech recognition. In Proceedings of the 3rd International Conference on Computer Communication and the Internet (ICCCI) (pp. 163–167). IEEE. https://doi.org/10.1109/ICCCI51764.2021.9486811

[22] Mamyrbayev, O., Kydyrbekova, A., Alimhan, K., Oralbekova, D., Zhumazhanov, B., & Nuranbayeva, B. (2021). Development of security systems using DNN and i & x-vector classifiers. Eastern-European Journal of Enterprise Technologies, 4(9(112)), 32–45. https://doi.org/10.15587/1729-4061.2021.239186

[23] Orken, M., Dina, O., Keylan, A., Tolganay, T., & Mohamed, O. (2022). A study of transformer-based end-to-end speech recognition system for Kazakh language. Scientific Reports, 12, 8337. https://doi.org/10.1038/s41598-022-12260-y

[24] Peters, B., & Martins, A. F. T. (2022). Beyond characters: Subword-level morpheme segmentation. In Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology (pp. 131–138). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sigmorphon-1.14

[25] Enarvi, S., Smit, P., Virpioja, S., & Kurimo, M. (2017). Automatic speech recognition with very large conversational Finnish and Estonian vocabularies. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(11), 2085–2097. https://doi.org/10.1109/TASLP.2017.2743344

[26] Kurimo, M., Enarvi, S., Tilk, O., Varjokallio, M., Mansikkaniemi, A., & Alumäe, T. (2017). Modeling under-resourced languages for speech recognition. Language Resources and Evaluation, 51(4), 961–987. https://doi.org/10.1007/s10579-016-9336-9

[27] Singh, M., Virpioja, S., Smit, P., & Kurimo, M. (2019). Subword RNNLM approximations for out-of-vocabulary keyword search. In Proceedings of INTERSPEECH 2019 (pp. 4235–4239). ISCA. https://doi.org/10.21437/Interspeech.2019-1329

[28] Mussakhojayeva, S., Khassanov, Y., & Varol, H. A. (2022). KSC2: An industrial-scale open-source Kazakh speech corpus. In Proceedings of INTERSPEECH 2022 (pp. 1367–1371). ISCA. https://doi.org/10.21437/Interspeech.2022-421

[29] Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., et al. (2023). FLEURS: Few-shot learning evaluation of universal representations of speech. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT) (pp. 798–805). IEEE. https://doi.org/10.1109/SLT54892.2023.10023141