Bootstrapping NLP for Sakha: Named Entity Recognition and Sentiment Analysis in an Extremely Low-Resource Setting

Mariia Everstova, Nikolai Efimov, Valerio Basile
University of Turin
eve.mm49@gmail.com, enifimov@gmail.com, valerio.basile@unito.it

Abstract

We present the first systematic study of core NLP tasks for Sakha (Yakut), a low-resource Turkic language with approximately 450,000 speakers in northeastern Siberia. We introduce two manually annotated datasets: a 690-sentence NER corpus (921 entities: PER, LOC, ORG) and an 798-sentence sentiment corpus (positive, negative, neutral). Using mBERT and RuBERT in controlled 2×2 experiments, we report a twofold effect: on the one hand, it improves performance when base unknown-token rates exceed approximately 10% (RuBERT: +9.4 F1); on the other hand, it leads to worse performance otherwise (mBERT: –6.1 F1), despite improving tokenization in both cases. Cross-domain transfer (news vs forums) reveals severe asymmetry: formal-to-informal training achieves 47% accuracy while the reverse yields only 26%—a 21-point gap demonstrating that domain composition dominates model architecture choice in low-resource settings. Neutral-boundary detection is the primary bottleneck, with 89% of disagreements clustering around subjective/objective distinctions rather than polarity confusions. With fewer than 1,000 samples per task, we establish first benchmarks for Sakha NER (53.5 F1) and sentiment analysis (54% accuracy).

Keywords: Corpus (Creation, Annotation, etc.), Less-Resourced/Endangered Languages, Named Entity Recognition, Opinion Mining/Sentiment Analysis

1. Introduction

Natural Language Processing has made remarkable progress in recent years, yet its benefits remain concentrated among approximately 20 high-resource languages (Joshi et al., 2020). However, the vast majority of the world's 7,000+ languages lack even basic computational resources: annotated datasets, pre-trained models, or morphological analyzers necessary for fundamental NLP tasks. This disparity reflects not merely a data gap but a structural inequality regarding who benefits from language technology.

Sakha, the endonym for the language often referred to by the exonym Yakut, is a Turkic language spoken by approximately 450,000 people in northeastern Siberia, and exemplifies these challenges. Despite its official status in the Sakha Republic (Yakutia), its active presence in the media, and its literary tradition, Sakha remains severely underrepresented in the NLP infrastructure. Before this work, no publicly available human-annotated datasets existed for core tasks such as named entity recognition (NER) or sentiment analysis.

This paper presents the first systematic investigation of Sakha NLP through two complementary tasks: Named Entity Recognition and Sentiment Analysis. We make four primary contributions:

(1) Two annotated datasets establishing baselines: A 690-sentence NER corpus with 921 manually annotated entities (PER, LOC, ORG) achieving inter-annotator agreement of F1=0.897 (Cohen's κ=0.96), and a 798-sentence sentiment corpus spanning formal journalism and informal forum discussion with moderate agreement (Cohen's κ=0.54, weighted κ-quadratic=0.55) reflecting genuine linguistic ambiguity.

(2) Empirical evidence on vocabulary extension: Through controlled experiments comparing base and extended BERT vocabularies, we establish a practical decision criterion: vocabulary extension improves performance when base unknown-token rates exceed approximately 10% (RuBERT: 12.25% $\rightarrow$ +9.4 F1) but degrades performance when base coverage is adequate (mBERT: 3.52% $\rightarrow$ –6.1 F1).

(3) Cross-domain generalization analysis: Bidirectional sentiment transfer experiments reveal severe asymmetry: training on formal news and testing on informal forums achieves 47% accuracy, while the reverse yields only 26% — a 21 percentage point gap demonstrating that domain composition matters more than model architecture choice in low-resource settings.

(4) Linguistic documentation: A systematic analysis of annotation challenges shows that for both tasks, the bulk of disagreements in sentiment annotation (89%) stem from detecting truly neutral cases. These difficulties are driven by morphological complexity, code-switching, orthographic variation, and cultural semantics unique to Sakha.

2. Related Work

2.1. Low-Resource NLP and Turkic Languages

Joshi et al. (2020) categorize the world's languages into six resource classes, revealing that 88% fall into Class 0 (virtually no digital resources). Among

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3295–3303
11–16 May 2026.

Turkic languages, Turkish dominates NLP research with established benchmarks (Yeniterzi, 2011). Recent work on Kazakh — another agglutinative Turkic language — provides a directly relevant context. Yeshpanov and Varol (2022) introduced KazNERD (112,702 sentences, 136,333 entities), achieving F1=0.97 for NER and KazSAnDRA (180,064 reviews), achieving F1=0.81 for binary sentiment but collapsing to F1=0.39 for 5-class classification (2024). Their findings on agglutinative morphology, code-switching, and class imbalance parallel our Sakha results despite 200× more training data, suggesting these challenges are intrinsic to Turkic language structure rather than merely data scarcity.

2.2. Named Entity Recognition

NER has evolved from rule-based systems (Grishman and Sundheim, 1996) through statistical approaches (CRFs) (Lafferty et al., 2001) to neural architectures (Lample et al., 2016). Transformer-based models such as BERT (Devlin et al., 2019) now achieve 92-94% F1 on English benchmarks (CoNLL-2003), but performance degrades substantially in low-resource settings. Adelani et al. (2021) report F1 scores of 0.79-0.89 across 10 African languages with comparable annotation schemes, while transfer learning effectiveness varies by typological distance (Pires et al., 2019) and training set size (Rahimi et al., 2019).

For morphologically rich languages, vocabulary coverage becomes critical: Artetxe et al. (2020) demonstrate that extending pre-trained vocabularies can improve downstream performance, though conditions under which extension helps versus harms remain underspecified—a question our work addresses empirically for Sakha.

2.3. Sentiment Analysis

Sentiment analysis faces distinct challenges in low-resource settings. Abdul-Mageed et al. (2012) report that Arabic subjectivity detection (distinguishing opinion from fact) is more complicated than polarity classification (65.6% vs 82.4% accuracy), a pattern we observe for Sakha. Cross-domain transfer studies show asymmetric generalization: Vilares et al. (2014) find that training on formal text transfers better to informal contexts than vice versa, attributed to the “grammatical completeness“ of formal registers.

2.4. Annotation Challenges

Morphologically rich languages introduce systematic annotation difficulties. For agglutinative Turkic languages specifically: sentiment-bearing suffixes create boundary ambiguity, negation scopes over entire morphological words non-locally, and code-switching complicates sentiment attribution when matrix and embedded languages provide conflicting cues (Myers-Scotton, 1993). Our work documents these patterns systematically for Sakha, providing empirical grounding for challenges often mentioned anecdotally in low-resource NLP literature.

3. The Sakha Language Context

Sakha (ISO 639-3: sah) is a Turkic language spoken by approximately 450,000 people primarily in the Sakha Republic (Yakutia), Russian Federation. Typologically, it shares core features with other Turkic languages — agglutinative morphology, SOV word order, vowel harmony — but diverges in key respects relevant to NLP:

Morphological complexity: Extensive suffixation is illustrated in (1) and case marking in (2), both of which create tokenization challenges when entity boundaries must include inflected forms.

(1) *балык-сыт-тар-быт*
balık-sıt-tar-bıt
fish-NMLZ-PL-1PL.POSS

‘our fishermen’

(2) *Дьокуускай-га*
Dokuuskay-ga
Yakutsk-DAT

‘to Yakutsk’

Orthographic variation: Sakha uses five Cyrillic letters absent from Russian keyboards (ҥ, ө, ҕ, ү, ҥ). Informal digital writing systematically substitutes these: ҥ→ь/c/english h, (киҥи→киси), ө→о/e/8, (сөх→сох), ҕ→г/х/5 (дьылҕа→дьыл5а). Analysis of our forum data reveals 62% of sentences contain letter substitutions — not random errors but stable sociolinguistic markers of informality.

Code-switching: Sakha-Russian bilingualism is near-universal in urban contexts. Technical vocabulary, intensifiers, and entire clauses mix languages freely, complicating both entity and sentiment attribution.

Cultural semantics: Sentiment expressions resist direct translation. The Sakha word *соһуччу sohuchchu* ’surprising/unexpected’ leans negative by default — implying uncomfortable disruption — unlike sentiment-neutral English/Russian equivalents.

These properties create annotation challenges distinct from well-studied languages and motivate our focus on documenting linguistic complexity alongside computational baselines.

4. Datasets and Annotation

4.1. NER Corpus Construction

Data source: All text derives from the February 2018 Sakha Wikipedia dump (LINDAT/CLARIAH-CZ). While this introduces genre bias toward encyclopedic content, Wikipedia's high name density makes it efficient for NER annotation under severe resource constraints.

Annotation scheme: Flat BIO tagging with three entity types: PER (person names including given names, patronymics, surnames), LOC (geopolitical entities, geographic features, infrastructure), and ORG (organizations with institutional permanence—governmental bodies, educational institutions, cultural organizations). Critical design decision: morphological suffixes are preserved within entity spans (*Дьокуускайга Dokuuskayga* 'to-Yakutsk' fully labeled as LOC) rather than excluded, reflecting alignment with subword tokenizer boundaries and ecological validity of inflected name usage.

Sampling strategy: Three-phase pipeline combining random sampling, active learning, and targeted class balancing achieved 3× annotation efficiency over uniform sampling — equivalent coverage under uniform random sampling would have required annotating approximately 2,300 sentences, of which roughly 1,600 would be entity-free:

- Phase 1: 500 random sentences (30.6% entity-bearing) → 329 entities
- Phase 2: 100 sentences via uncertainty sampling — using Maximal Marginal Relevance (MMR, λ=0.7), balancing uncertainty and diversity(100% entity-bearing) → 325 entities
- Phase 3: 100 ORG-targeted sentences (96% entity-bearing) → 267 entities

This approach reduced annotation burden by 70% compared to uniform sampling while maintaining high IAA (κ=0.96). Final corpus: 690 sentences (349 entity-bearing, 341 entity-free), 921 entities: 266 PER (28.9%), 495 LOC (53.7%), 160 ORG (17.4%).

Annotation procedure: Primary annotation was carried out by a single native speaker (co-author Mariia Everstova), a graduate in digital humanities with a BA in philology. Two native speakers then independently annotated 100 Phase-2 sentences for inter-annotator agreement assessment, achieving entity-level F1=0.897 (exact span + label matches) and Cohen's κ=0.96. Disagreement analysis reveals systematic patterns: boundary ambiguities with grammatical suffixes (27% of errors), ORG vs. LOC confusion for geopolitical institutions (18%), and proper name vs. demonym distinctions (10%). Notably, PER achieves F1=0.983 while ORG achieves only F1=0.812, predicting downstream model performance patterns.

Example: In "Прокопий иккистээн Лев Толстойдуун Яснай Полянаҕа тиийэн көрсүҥэр." (Prokopiy visited Lev Tolstoy at Yasnaya Polyana), annotators must identify person names and locations with overt case morphology: Прокопий (PER), Лев Толстой*дуун* (PER, genitive suffix -duun included), and Яснай Поляна*ҕа* (LOC, dative suffix -gha included). Inflected forms are preserved as complete entity spans, illustrating the morphological challenges documented in our 27% suffix-boundary disagreements.

4.2. Sentiment Analysis Corpus

Data sources: To test how robust our models are across different registers, we designed a dual-domain dataset instead of simply chasing large numbers. On one side, we collected 497 sentences from the Sakha-language newspaper Kyym, drawn from sections such as Politics, Economy, Society, and Culture—each sentence averaging about 11.2 tokens. On the other side, we turned to the informal world of the internet: 301 sentences pulled from the archived forum at forum.ykt.ru. These average only 7.9 tokens per sentence—yet consciously capture real-world messiness: 62% of them include orthographic substitutions or non-standard spellings, typical of casual online discourse.

The rationale for including both domains in our language resource is to create a more diverse corpus and therefore a more challenging cross-domain benchmark.

Label distribution: Negative: 174 (22%), Neutral: 434 (54%), Positive: 190 (24%). The 54% neutral dominance is typical of real-world sentiment data and necessitates evaluation metrics beyond accuracy.

Annotation procedure: Sentence-level three-class classification was carried out by a single native speaker, co-author Everstova — this choice reflects practical constraints: Sakha-literate individuals with linguistic training required for sentiment annotation are geographically concentrated in Yakutsk, and no funding was available for annotator compensation. A native independent annotator without formal linguistic training then conducted a targeted inter-annotator agreement (IAA) assessment on 150 stratified sentences (about 19% of the corpus). The training asymmetry between annotators is a limitation on the IAA interpretation of the sentiment data: the moderate κ=0.54 may also reflect differences in annotation background.

Disagreement analysis: Unlike NER's near-perfect agreement, sentiment annotation reveals genuine linguistic ambiguity: 89% (33/37) disagreements involve neutral boundaries rather than

positive-negative confusions, splitting between positive↔neutral (49%) and negative↔neutral (40%). Primary sources include: orthographic variation (18%), code-switching attribution conflicts (3%), context-dependent affection terms requiring pragmatic inference (49% of positive-neutral disagreements), and dialectal proverbs carrying non-transparent sentiment (3%). This pattern suggests determining whether sentiment is expressed at all is fundamentally more complex than identifying entity boundaries.

Example: The forum post "*олус учугэй Доллулар бааргыт доооо*" (Are you here, the best Dollul residents?) illustrates orthographic-affective coupling. One annotator marked positive (praise via *учугэй*='best' + excitement from tripled vowels *доооо*), another marked neutral (simple greeting question). The deliberate vowel lengthening signals prosodic intensity, but whether this constitutes evaluative sentiment or merely emphatic informality remains ambiguous.

5. Methodology

5.1. Vocabulary Extension Strategy

To compensate for the limited representation of Sakha in existing pretrained vocabularies (with UNK rates of 3.52% for mBERT and 12.25% for RuBERT), we extended both models by introducing Sakha-specific subwords. Our technique is inspired by vocabulary extensions for domain adaptation (Tai et al., 2020; Hong et al., 2021), however we apply it to mitigate the extra token splitting that a non-Sakha model would face if applied directly to Sakha.

Token selection: We began by training a Sakha WordPiece tokenizer on the 2018 Sakha Wikipedia dump. This produced 14,577 candidate tokens. After removing whitespace, punctuation, control symbols, and duplicates already present in the base vocabularies, we obtained 11,112 new tokens for mBERT (expanding its vocabulary from 119,547 to 130,659 entries) and 11,364 for RuBERT (120,138 → 131,502).

Embedding initialization: For embedding initialization, we adopted a two-stage approach. When a new token could be decomposed into known subwords, we initialized its embedding by taking a length-weighted average of the subword vectors (Chau et al., 2020). Tokens that did not decompose—typically those containing Sakha-specific Cyrillic characters (ө, ү, ҥ, ҕ, ҥ)—were instead initialized by sampling from embeddings that included similar characters. In practice, 90.5% of new mBERT tokens were initialized via averaging, compared with only 69.0% for RuBERT, highlighting the more substantial morphological overlap between Sakha and the multilingual model The resulting extended embeddings showed a lower standard deviation ($\sigma = 0.0154$ vs. $\sigma = 0.0463$), suggesting a more compact initialization space.

Metric	RuBERT		mBERT	
	Base	Ext	Base	Ext
Avg pieces/word	2.37	1.39	3.21	1.46
Single-token (%)	36.0	73.6	17.2	71.6
UNK rate (%)	12.25	0.00	3.52	0.00

Table 1: Tokenization quality on 20,000 Sakha words.

Tokenization improvement: As Table 1 shows, these extensions dramatically reduced token fragmentation. Average pieces per word dropped from 3.21 to 1.46 in mBERT and from 2.37 to 1.39 in RuBERT, while single-token coverage more than doubled in both models. However, as the experiments later reveal, smoother tokenization alone does not guarantee better downstream performance.

5.2. Experimental Design

Model selection: We compared four model variants in a controlled 2×2 setup. The first pair involved mBERT (base and extended), which provides multilingual coverage across 104 languages including Turkish and Tatar, thereby offering useful Turkic priors. The second pair used RuBERT (base and extended), a Russian-specific model that benefits from geographical proximity and frequent code-switching with Sakha. All models shared the BERT-base architecture (12 layers, 768 dimensions, ≈110 M parameters) and identical training procedures, ensuring that any observed differences stemmed solely from vocabulary extension.

NER configuration: For the NER experiments, we used a BERT encoder followed by dropout (0.1) and a linear classifier mapping 768 to 7 entity classes. Training used the AdamW optimizer (lr = 3e-5) with a cosine schedule, 10% warmup, and up to 8 epochs with early stopping (patience = 3). The dataset was split into 483 training, 103 development, and 104 test sentences, stratified by entity density and with ORG entities oversampled (2×). Evaluation employed entity-level F1 via seqeval.

Sentiment configuration: For sentiment classification, we applied the same architecture but trained with lr = 1e-5 and a sequence length of 210 tokens. We ran two evaluation protocols: (1) pooled five-fold cross-validation stratified by label and domain; and (2) cross-domain transfer (Kyym → Forum, Forum → Kyym). We used inverse-frequency class weighting and reported weighted F1 as the primary metric, complemented by macro F1 and per-

Model	F1	PER	LOC	ORG
RuBERT base	0.335	0.361	0.374	0.232
RuBERT ext	0.429	0.474	0.434	0.361
Δ	**+0.094**	**+0.113**	**+0.060**	**+0.129**
mBERT base	**0.535**	0.590	0.560	0.423
mBERT ext	0.474	0.486	0.489	0.415
Δ	**−0.061**	**−0.104**	**−0.071**	**−0.008**

Table 2: NER test set performance showing opposite vocabulary extension effects. RuBERT improves +9.4 F1 while mBERT degrades −6.1 F1.

class scores.

Reproducibility: To ensure reproducibility, all experiments used a global random seed (42), deterministic computation settings, and dataset fingerprints for deduplication. We did not perform any hyperparameter tuning; all configurations were kept constant to preserve strict comparability between base and extended models.

6. Results

6.1. Named Entity Recognition

Table 2 presents test set performance for all four NER variants.

Key findings:

(1) Vocabulary extension effects are model-dependent and opposite in direction. RuBERT improved +9.4 F1 overall (0.335→0.429), while mBERT degraded −6.1 F1 (0.535→0.474). This 15.5-point divergence contradicts the assumption that better tokenization universally benefits downstream performance.

(2) Vocabulary extension effects depend on base model coverage. Extension improved RuBERT performance (+9.4 F1) when base unknown-token rates were high (12.25%), but degraded mBERT (−6.1 F1) with adequate base coverage (3.52%). This mixed pattern indicates the need for a practical threshold: vocabulary extension addresses genuine coverage gaps when base UNK rates exceed approximately 10%, but it introduces training instability when base coverage is adequate. The divergent outcomes provide a practical decision criterion: measure base tokenizer UNK rates on target-language text before extending vocabularies, and prioritize extension only when coverage gaps are demonstrable.

(3) ORG remains the bottleneck across all models. Despite the use of several optimization techniques (2× oversampling, active learning for ORG-rich sentences, vocabulary extension), ORG achieved F1=0.232-0.423 compared to 0.361-0.590 for PER and 0.374-0.560 for LOC. Only 160 training instances, distributed across highly diverse

Model	Acc	W-F1	M-F1	SD
mBERT	0.539	0.520	0.454	±0.052
RuBERT	0.494	0.492	0.450	±0.044
Δ	**+4.5**	**+2.8**	**+0.4**	—

Table 3: Pooled 5-fold CV results (mean). mBERT outperforms RuBERT by 4.5pp despite Russian proximity.

Direction	Accuracy	Weighted F1
Kyym → Forum	0.468	0.433
Forum → Kyym	0.256	0.249
Asymmetry	**21.3pp**	**18.4pp**

Table 4: Bidirectional cross-domain transfer (mBERT).

naming patterns (government bodies, universities, media outlets, cultural institutions), proved insufficient for generalization.

(4) All models substantially underperform with respect to human agreement. The best model (mBERT base, F1=0.535) reaches only 59.6% of inter-annotator agreement (F1=0.897). The 36-point gap is larger than typically observed in high-resource NER (10-15 points for English), suggesting ultra-low-resource constraints constitute a fundamental upper bound.

6.2. Sentiment Analysis

Table 3 shows pooled cross-validation results.

Key findings:

(1) mBERT's multilingual exposure outweighs Russian proximity. Despite Sakha-Russian bilingualism and geographic proximity, mBERT consistently outperformed RuBERT (4/5 folds). Hypothesis: mBERT's training on Turkish, Tatar, and Azerbaijani provides more relevant Turkic structural priors than RuBERT's Russian fusional morphology.

(2) Neutral-boundary detection is the bottleneck. Per-class analysis shows neutral F1=0.64 (mBERT) vs 0.37-0.38 for polarized classes. This mirrors the 89% neutral-boundary disagreement rate in IAA, suggesting the fundamental challenge is detecting *whether* sentiment is expressed, not *which* sentiment.

(3) Domain composition dominates model choice. Table 4 reveals severe cross-domain transfer asymmetry.

Training on 497 formal sentences and testing on 301 informal sentences achieved 47% accuracy, while the reverse yielded only 26%—a 21-point gap exceeding the 4.5-point model difference. This asymmetry reflects: 65% more training data (497 vs 301 sentences), 42% longer sentences provid-

ing richer context (11.2 vs 7.9 tokens), orthographic consistency in formal text (vs 62% substitutions in forum), and formal grammar's fuller inventory that transfers to informal subsets while informal-specific markers (emoji, vowel lengthening) don't generalize.

Practical implication: When annotation budgets permit only 500-1000 samples, prioritizing domain consistency over diversity and favoring formal registers yields better generalization.

7. Discussion

7.1. Cross-Task Insights

Both tasks reveal converging evidence on low-resource NLP challenges:

(1) Neutral/objective boundary detection is more complex than polarity/entity-type classification. NER achieves near-perfect PER agreement (F1=0.983) but struggles with ORG vs. LOC semantic overlaps (F1=0.812). Sentiment shows 89% of disagreements at neutral boundaries, not positive-negative confusions. This pattern generalizes: Arabic subjectivity detection (opinion vs. fact) is harder than polarity classification (Abdul-Mageed et al., 2012), and Kazakh score classification collapsed (F1=0.39) despite 140K samples (Yeshpanov and Varol, 2024).

(2) Sample size alone cannot overcome class imbalance and semantic ambiguity. Targeted interventions (ORG oversampling, active learning, vocabulary extension) failed to improve ORG performance with 160 training instances substantially. Similarly, the sentiment's positive class remained unstable across folds despite class weighting. This suggests minimum viable dataset sizes: 300+ instances for stable category learning, likely 500-1000 for robust generalization to rare patterns.

(3) Domain composition matters more than model architecture when data is scarce. The 21pp sentiment cross-domain asymmetry exceeds the 9.4pp vocabulary extension benefit and 4.5pp model difference, indicating annotation strategy choices dominate algorithmic choices in ultra-low-resource regimes.

7.2. The Tokenization-Performance Paradox

Despite mBERT achieving superior tokenization (1.46 vs 1.39 avg pieces/word post-extension), its NER performance degraded while RuBERT's improved. This decoupling reveals vocabulary extension as conditionally beneficial: it helps when base model has genuine coverage gaps (RuBERT UNK=12.25%), frequent entities fragment catastrophically, and new embeddings can be initialized from linguistically similar base tokens; it harms when base coverage is adequate (mBERT UNK=3.52%), new embeddings are poorly initialized (low standard deviation indicating compressed initialization), and training data is insufficient to update 11K new parameters meaningfully. The 10% UNK threshold provides a practical heuristic: extend vocabulary when the base tokenizer demonstrably fails on the target language.

7.3. Practical Guidelines

For languages with <1000 samples, our findings suggest a series of guidelines: i) prioritize domain consistency over heterogeneity, favor formal registers for better transfer, use multi-phase sampling for efficiency, and invest in dual annotation for ambiguous cases. ii) Measure base tokenizer UNK rates before vocabulary extension (only extend if >10%). iii) Multilingual models may outperform geographically-motivated monolingual models for typologically distant languages. iv) Report per-class metrics and cross-domain evaluation alongside human agreement baselines.

Finally, we note an important issue: mBERT and RuBERT differ in pre-training data and language coverage. We present UNK rate as correlated to the observed outcome, not a confirmed mechanism.

7.4. Linguistic Documentation

Systematic analysis reveals orthographic-affective coupling (87.5% of lengthened vowels carry polarity), code-switching attribution conflicts, and cultural semantics with cross-linguistic non-equivalence—patterns generalizing to other Turkic languages (Yeshpanov and Varol, 2024; Kurt et al., 2019).

The moderate sentiment IAA (κ=0.54) reflects genuine linguistic ambiguity rather than unclear guidelines. Two lines of evidence support this: First, disagreements cluster systematically at neutral boundaries (89%) rather than distributing randomly across class pairs, splitting evenly between positive↔neutral (49%) and negative↔neutral (40%) with only 3% positive↔negative confusions. Second, weighted and unweighted κ are nearly identical (Δ=0.01), ruling out systematic ordinal confusion between adjacent categories. This pattern aligns with findings in Arabic (Abdul-Mageed et al., 2012) and Kazakh (Yeshpanov and Varol, 2024), confirming neutral-boundary detection is fundamentally more complex across morphologically rich languages.

7.5. Comparison with Related Work

Our NER results (best F1=0.535) fall between preliminary Turkic NER attempts and well-resourced

languages. Turkish NER achieves 85-90% F1 with mature resources (Yeniterzi, 2011), while Kazakh reaches 97% with 112,702 sentences (Yeshpanov et al., 2022). We establish 54% as the practical floor for transformer-based NER with <1000 samples. The 36-point gap to human agreement (0.897) exceeds typical English gaps (10-15 points), confirming ultra-low-resource constraints are fundamental.

Sentiment results (54% accuracy, 0.52 weighted F1) align with Arabic subjectivity detection (65.6%) when accounting for task differences (Abdul-Mageed et al., 2012). Yeshpanov and Varol's Kazakh binary sentiment (F1=0.81 with 167,961 samples) vs. 5-class collapse (F1=0.39 with 140,126 samples) parallels our finding that neutral-boundary detection is the bottleneck: their model defaulted to majority class (5-star reviews), just as ours struggles at subjective/objective boundaries (Yeshpanov and Varol, 2024).

The 21pp cross-domain asymmetry exceeds Vilares et al.'s 16pp Spanish result (Alonso Pardo et al., 2014), likely due to compounding effects: a smaller sample size (497 vs 301), 62% orthographic inconsistency in informal data, and shorter sentences (11.2 vs 7.9 tokens), which reduce contextual cues.

8. Conclusion

We present the first systematic study of Named Entity Recognition and Sentiment Analysis for Sakha, a low-resource Turkic language. Through two manually annotated datasets totaling 1,488 sentences, we establish baseline performance and document linguistic challenges arising from Sakha's agglutinative morphology and code-switching patterns.

Our key empirical findings challenge common assumptions: (1) vocabulary extension helps when base unknown-token rates exceed approximately 10%—below this threshold, it degrades performance despite improving tokenization; (2) domain composition is more impactful than model architecture in ultra-low-resource settings—the 21-point cross-domain asymmetry exceeds model differences; (3) neutral-boundary detection (subjective vs. objective) is more complex than polarity/entity-type classification—89% of sentiment disagreements cluster at this boundary.

With fewer than 1,000 training samples per task, we establish first NER and sentiment benchmarks for Sakha (53.5% F1; 54% accuracy). However, this remains below high-resource language performance (85-95%) and practical utility thresholds. Closing this gap requires not just more data but also Sakha-specific infrastructure: morphological analyzers handling informal orthography, bilingual sentiment lexicons, and potentially dedicated pretrained models.

Beyond computational baselines, our systematic documentation of annotation challenges—such as orthographic variation as a sentiment signal (87.5% of lengthened vowels carry polarity), code-switching attribution conflicts, and cultural semantics with cross-linguistic non-equivalence—provides linguistic insights valuable for future Sakha NLP and comparable low-resource efforts. The moderate IAA on sentiment (κ=0.54, weighted κ=0.55) reflects genuine ambiguity that both humans and models struggle with, rather than inadequate guidelines.

For researchers facing similar ultra-low-resource constraints, our methodology provides a replicable pathway: multi-phase sampling achieving 3× annotation efficiency, preservation of orthographic variation rather than normalization, cross-domain evaluation to assess generalization, and systematic documentation of annotation challenges as empirical findings.

Sakha now has its first NLP baselines. The path from proof-of-concept to practical utility remains long, but the foundation is established.

Data and Code Availability

Datasets, annotation guidelines, and experimental code are available at: https://github.com/enifimov81k/sakha-nlp-2026

Limitations and Future Work

Corpus size (690 NER, 798 sentiment sentences) falls below transformer fine-tuning thresholds, likely underestimating achievable performance. Single-annotator primary annotation (with 15-19% dual coverage) cannot rule out consistent bias. No hyperparameter tuning, classical ML baselines, or domain-adaptive pretraining were performed to isolate vocabulary extension effects. Domain coverage is limited to Wikipedia and two text registers, with unknown generalization to dialects (Northern, Vilyuy) or other genres (speech, literature, social media). The temporal span (2015-2024) suggests that language evolution may degrade future performance. All experiments used a single random seed (42), preventing variance estimation across runs. Immediate priorities include expanding to 2,000+ samples with complete dual annotation and building bilingual sentiment lexicons exploiting code-switching patterns.

Ethical Considerations

All data was collected from publicly accessible sources (Sakha Wikipedia, archived forum posts,

and newspaper articles) in compliance with applicable platform terms of service and copyright regulations.

References

Muhammad Abdul-Mageed, Sandra Kuebler, and Mona Diab. 2012. SAMAR: A system for subjectivity and sentiment analysis of Arabic social media. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 19–28, Jeju, Korea. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane Mboup, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima Diop, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 1116–1131.

Miguel Alonso Pardo, David Vilares, and Carlos Gómez-Rodríguez. 2014. A syntactic approach for opinion mining on spanish reviews. *Natural Language Engineering*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Jimin Hong, TaeHee Kim, Hyesu Lim, and Jaegul Choo. 2021. AVocaDo: Strategy for adapting vocabulary to downstream domain. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4692–4700, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Fatih Kurt, Kezban Dilek Kisa, and Pinar Karagoz. 2019. Investigating the effect of segmentation methods on neural model based sentiment analysis on informal short texts in turkish. *CoRR*, abs/1902.06635.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Carol Myers-Scotton. 1993. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford University Press.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. Association for Computational Linguistics.

Reyyan Yeniterzi. 2011. Exploiting morphology in Turkish named entity recognition system. In *Proceedings of the ACL 2011 Student Session*, pages 105–110, Portland, OR, USA. Association for Computational Linguistics.

Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. KazNERD: Kazakh named entity recognition dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 417–426, Marseille, France. European Language Resources Association.

Rustem Yeshpanov and Huseyin Atakan Varol. 2024. KazSAnDRA: Kazakh sentiment analysis dataset of reviews and attitudes. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9657–9667, Torino, Italia. ELRA and ICCL.