http://dx.doi.org/10.14500/aro.12641

A Novel Central Kurdish Part-of-Speech Corpus and Deep Tagging Model Evaluation

Haneen Al-Raghefy[1†], Halgurd S. Maghdid[2], and Akar H. Taher[1]

[1]Department of Software Engineering, Faculty of Engineering, Koya University, Koya, Kurdistan Region – F.R. Iraq

[2]Department of Engineering Research Center, Deanship of R&D Center, Koya University, Koya, Kurdistan Region – F.R. Iraq

***Abstract*—For many low-resource languages, including the central Kurdish language (CKL), building effective natural language processing (NLP) tools has been a challenge. This is due to the lack of annotated text. Without a large corpus that specifies how words function grammatically, it is difficult to perform basic tasks such as part-of-speech (POS) tagging, which is the building block of many language technologies. To address this issue, this study presents the first comprehensive POS-tagged corpus for CKL. This dataset consists of 108,680 words manually tagged with 86 tags. Unlike simpler tagging schemes, the 86 tags account for the complexity of Kurdish grammar and allow a single word to have multiple valid tags, reflecting the language's natural ambiguity. Using this resource, this study benchmarks a range of deep models, including neural networks such as bidirectional long short-term memory (BiLSTM). To address the ambiguity challenge, this paper introduces a new method, adaptive tag cycling within the BiLSTM that trains the model to consider all possible tags. The most advanced model in this study, an ensemble of neural sequence taggers, achieves 92.3% accuracy with stop-words retained and 89.5% with stop-words removed on broad grammatical categories (main tags). On the full fine-grained tagset (detailed tags), the same model attains 79.0% accuracy with stop-words and 76.2% without stop-words. Therefore, this study provides two key contributions: (i) a new dataset that supports future Kurdish NLP research, and (ii) a strong performance benchmark for CKL POS tagging.**

***Index Terms*—Annotated dataset, Central Kurdish language, Deep learning, Long short-term memory, Part-of-speech.**

I. Introduction

Part-of-speech (POS) tagging is a fundamental task in natural language processing (NLP), enabling grammatical analysis and supporting higher-level applications such as parsing, information extraction, spell checking, text generation systems, and machine translation (MT) (Daelemans, 2011; Qiu, et al., 2020; Jurafsky and Martin, 2023). While POS tagging has been widely studied for high-resource languages, it remains a challenging task in low-resource languages due to the scarcity of annotated corpora and variability in orthography and complex morphology.

For high-resource languages such as English and Spanish, the availability of large annotated corpora and pretrained models has enabled POS taggers to achieve consistently high performance (Pota, et al., 2019). However, these advances have not transferred equally to low-resource languages.

Recent advancements in generative artificial intelligence (AI) have made it possible for models to produce high-quality text, speech, and images with just a single pipeline trained on data from the web at scale, eliminating the need for resource-intensive ones. Yet, Kurdish has not shared equally in these developments; there is much needed to be done in this AI era within and among its dialects. Large pretrained models have limited coverage of Kurdish because (i) relatively little Kurdish text is available online and (ii) existing materials frequently show orthographic and grammatical inconsistencies, reflecting the absence of a single standardized codification.

In this vein, the Central Kurdish Language (CKL), also known as Kurdish Sorani Language, spoken by millions in Iraq and Iran, exhibits rich inflectional morphology in which a single surface form can correspond to multiple grammatical categories depending on context (Naserzade, et al., 2022). Additional orthographic inconsistencies, such as variations in representing certain consonants, increase further ambiguity (Ahmadi and Masoud, 2020). These characteristics make the POS tagging for CKL challenging and limit the direct transfer of the developed models to other languages.

Previous POS tagging efforts for CKL have been constrained by small datasets or coarse-grained tagsets (e.g., 38 tags), which restrict the ability to capture fine morphological distinctions (Maulud, Jacksi and Ali, 2023a). Furthermore, the last developed comprehensive standardized tagset (97 tags) for CKL. (Sabr, et al., 2025) has not been digitally used to date. Modern neural architectures (e.g., bidirectional long-short-term-memory [BiLSTM] and conditional random field [CRF]) have not yet been

ARO-The Scientific Journal of Koya University
Vol. XIV, No.1 (2026), Article ID: ARO.12641. 11 pages
DOI: 10.14500/aro.12641
Received: 25 September 2025; Accepted: 11 February 2026
Regular research paper: Published: 20 June 2026
†Corresponding author's e-mail: haneen.hayder@koyauniversity.org

systematically benchmarked for CKL in combination with subword tokenization and multi-tag annotation strategies.

Therefore, this study addresses these challenges by introducing a 108,680 tokens manually annotated via an academic-domain corpus from Koya University with an 86-tags POS tagset, most of which are taken from (Sabr, et al., 2025) and designed for morphological and syntactic coverage. This is followed by developing seven progressively enhanced architectures that are evaluated by Hidden Markov model (HMM), BiLSTM, BiLSTM with adaptive tag cycling, BiLSTM + Ensemble, BiLSTM + CRF, BiLSTM + adaptive tag cycling + CRF, and BiLSTM + adaptive tag cycling + CRF + ensemble.

All models are implemented in two variants, the full 86-tags detailed (fine-grained) scheme and a collapsed 10-category main tagset. By systematically comparing these architectures, this research aims to quantify the incremental benefits of adaptive cycling in multi-tag annotation, CRF-based sequence modeling, and ensemble learning, while also demonstrating the value of a standardized, fine-grained tagging tagset for the CKL.

II. Related Work

Due to limited annotated corpora and also due to the dialectal variation of CKL in different regions, there are many challenges for low-resource NLP. The multi-dialectal nature of the Kurdish language also complicates the generalization of a model across geographical regions. Despite these challenges, several approaches, such as hybrid statistical-rule-based systems and cross-lingual transfer, have been explored in recent studies, with initial corpus-driven studies demonstrating these in practice.

An early study with a corpus-driven approach addressed the tasks of classification and parsing. (Malmasi, 2016) developed one of the first CKL datasets and applied support vector machines with n-gram features to classify subdialects, reporting a classification accuracy of 96.7%. For Kurmanji, (Gökırmak and Tyers, 2017) built the first dependency treebank under the Universal Dependencies framework (10k tokens); this resource is not sufficient for reliable parsing.

Lexical and text-processing tools have, however, also been developed (Salavati and Ahmadi, 2018). The research presented Peyv, a lemmatizer and Rênûs, a spell-checker for CKL, which is trained on the 18M-word Pewan corpus. These tools were promising but limited due to lexicon coverage. (Hassani, 2022) created a POS-tagged CKL lexicon via transferring annotations from Persian; however, inconsistencies and domain adaptation issues remained. Elsewhere, larger corpus projects attempted to broaden the data landscape.

The *Bianet* Turkish–Kurdish–English dataset (Ataman, 2018) and Awta corpus (Amini, et al., 2021) provided initial resources for Kurdish MT, though their size remains small relative to modern neural MT requirements (Ahmadi and Masoud, 2020) highlighted tokenization and morphological segmentation as critical for improving Kurdish–English MT performance. Beyond MT, research has also explored POS tagging directly.

POS tagging research has been limited, but impressive. The DASTAN corpus (74k words), reported at 96% accuracy, used a hybrid HMM and rule-based tagger (Maulud, Jacksi and Ali, 2023a). But their tagset was limited to 38 tags, with little morph-syntactic granularity. Similarly, (Azzat, Jacksi and Ali, 2024) applied a similar HMM-based method for the Badini dialect and constructed a POS-tagged Badini corpus (51k words) for ontology development. Applied an HMM-based method on the Badini dialect and built a POS-tagged Badini corpus (51k words) for ontology building. These studies have shown that Kurdish POS tagging is possible, highlighting issues such as small corpora, inventory, and evaluation benchmarks. Taken together, these studies have demonstrated important advances but also enduring gaps. Recent progress in Kurdish NLP is based on transformer models, such as for sentiment analysis and named entity recognition. Such examples include a bidirectional encoder representation from the transformer (BERT) model used for the sentiment classification of Central Kurdish (Awlla, Veisi and Abdullah, 2025) and the fine-tuning of RoBERTa in low-resource NER settings (Abdullah, et al., 2024). These applications are concerned with fewer label categories and higher semantic levels than POS tagging, which this work specifically considers through the lens of token-level sequence labeling across 86 tags, where transformer models typically need orders of magnitude larger annotated corpora than that available here (108,680 tokens). Thus, the mBERT baseline is included only for completeness, with the main experiments investigating recurrent architectures that maintain a distinct advantage in terms of effective parameters for low-resource morphologically rich languages under limited supervision. As a result, large-scale annotated corpora and strong tagging models are still missing for CKL. Table I compares this corpus with recent Kurdish corpora.

To address these gaps, this study makes two main contributions. First, it introduces 108,680 tokens, manually annotated CKL corpus using an 86-tag fine-grained POS scheme, which provides substantially greater morpho-

TABLE I
Comparison of POS Tagging Latest Studies for the Kurdish Corpus

Study	Dialect	Corpus size	Tagset	Method	Reported accuracy (%)
(Maulud, Jacksi and Ali, 2023a)– DASTAN	Sorani	74k words	38 tags	HMM+Rule-based	96
(Azzat, Jacksi and Ali, 2024)– UOZBDN	Badini	51k words	~30 tags	HMM+Rule-based, Ontology extraction	95
This study (2025)	CKL	108,680 words	86 tags	BiLSTM+CRF+Cycling+Ensemble	92.3 flexible accuracy (Main tagset with stop-words)

POS: Part-of-speech, HMM: Hidden Markov model, CKL: Central Kurdish language, BiLSTM: Bidirectional long short-term memory, CRF: Conditional random field

syntactic coverage. Second, it systematically evaluates a progression of architectures from traditional HMM to modern neural systems (BiLSTM, BiLSTM + CRF), including adaptive tag cycling and ensemble approaches.

III. Methodology

This study develops a CKL corpus to provide a reliable resource for POS tagging in a low-resource setting. This section describes the dataset and procedures used in this study.

A. Corpus Creation and Preprocessing

The corpus contains 108,680 manually annotated tokens extracted from 20 academic articles published in the Koya University Journal of Humanities and Social Sciences. To prepare the texts for annotation, a multi-stage preprocessing pipeline is implemented as shown in Fig. 1. The main steps included:

1. Normalization and standardization
   All texts are converted to UTF-8, and common script inconsistencies are corrected (e.g., Arabic ك → ک Kurdish, ی→ي). Punctuation and spacing are standardized, and inconsistent encodings are fixed along with custom rules (Ahmadi, 2020b).
2. Abbreviation handling
   The errors made by tokenization models frequently happen in academic abbreviations (Abdulrahman and Hassani, 2020).
3. Sentence segmentation and tokenization
   The sentences are tokenized using a hybrid approach (Ahmadi, 2020a). Involving the Kurdish Language Processing Toolkit and regex heuristics. Special attention is paid to not split within abbreviations, URLs, or decimal numbers extending. To assess the reliability of the tokenization process, a manually inspected subset of the corpus is evaluated. A total of 5,250 tokens is reviewed, and 167 tokenization errors are identified, corresponding to an error rate of 3.18% and an overall tokenization accuracy of 96.82%, where the tokenization accuracy is calculated as the complement of the tokenization error rate.
4. Language identification
   Since academic texts may include Arabic, Persian, or English passages, sentences are filtered using character-based heuristics (Veisi, MohammadAmini, and Hosseini, 2019).
5. Stop-words filtering
   Stop words are identified using a nearly 240-word stop-word list proposed by (Mustafa and Rashid, 2018). Given that stop words are syntactically informative and should be handled by POS taggers, the main models are evaluated both with and without stop-word filtering. The stop-word-filtered setting is considered to analyze its effect on feature sparsity and corpus efficiency, as discussed in (Maulud, Jacksi and Ali, 2023b).

Fig. 1 illustrates the overall pipeline from document collection through preprocessing, normalization, annotation, and final corpus storage.

After preprocessing, trained linguists from the Kurdish Language Department and the Malay Gawra Center at Koya University conducted manual annotation. Each token may receive up to three parallel tags to capture morpho-syntactic richness: A primary syntactic role (Tag1), a secondary morphological property (Tag2), and a context-sensitive finer

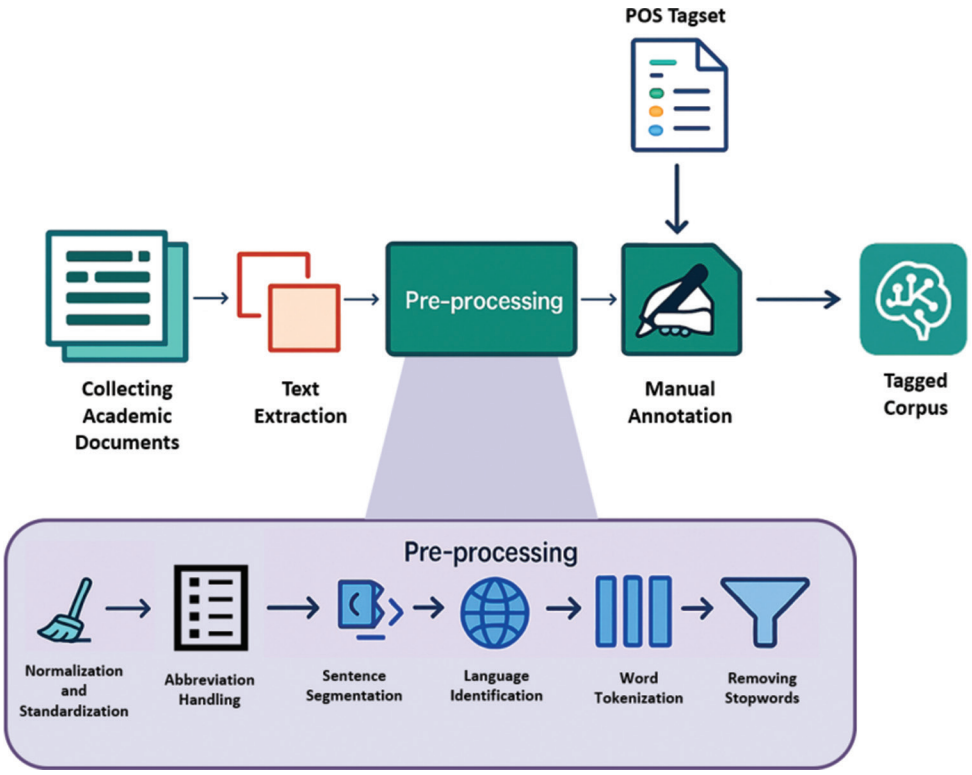

Fig. 1. The process of creating the central Kurdish language annotated corpus.

tag (Tag3). For example, ڕستە (“sentence”) can be annotated by three compact tags, a general noun (N-GEN), a known noun (N-KNOWN), and a singular noun (N-SG). Standard non-lexical tokens (NUM, PUNCT, EMAIL, and URL) are auto-labelled during pre-processing and supervised corrected when ambiguous. This hybrid process yields a corpus that is both reasonably consistent in structure and rich in language.

B. Corpus Annotation

To capture the morpho-syntactic richness of CKL, a fine-grained POS tagset of 86-tags is designed in consultation with Kurdish linguists. The 86-tags scheme used in this study represents a corpus-attested subset of the broader 97-tag standard proposed by Sabr, et al. (2025). Some tags from the original scheme did not occur in any token and, therefore, were not instantiated in the final dataset. This scheme expands the 38-tag systems used in previous CKL corpora (Maulud, Jacksi, and Ali, 2023a) and introduces detailed subcategories. The structure of the tagset appears in Table II.

The annotated corpus is formatted in a table with columns for sentence, token (word), Tag1, Tag2, and Tag3. Annotation follows a two-pass method: the initial pass by the main annotator and a validation pass by a linguist team. Conflicts are resolved during adjudication meetings based on guidelines derived from an 86-tag scheme.

Inter-annotator agreement is estimated to measure annotation quality and consistency. Agreement is first evaluated at the main POS category level (main tags) using Cohen’s κ, adjusted for chance agreement. Following standard practice in large-scale corpus annotation, a representative subset of 5,250 tokens (≈5% of the corpus) is chosen to be a representative range of sentence lengths and POS categories and a proportion of morphologically ambiguous tokens. An agreement rate of 84.62% is observed, with Cohen’s κ value as 74.5%, showing substantial agreement (Landis and Koch, 1977).

Implementation-wise, annotators use structured spreadsheets for consistency, as shown in Fig. 2. Any personal information has been anonymized.

A summary of key corpus statistics is presented in Table III. Fig. 3 illustrates the distribution of main POS categories in the CKL corpus, showing dominance of nouns and punctuation, followed by verbs, adjectives, and other classes. Fig. 4 shows the distribution of multi-tag annotations in the corpus, showing that nearly one-third of tokens carry more than one valid tag.

C. ML Models

All models in this study are trained and evaluated on the CKL annotated corpus. The traditional ML, such as HMM, is trained on the annotated corpus. HMMs estimate the probability of tag sequences through transition probabilities ($P(tag_i|tag_{i-1})$) and emission probabilities ($P(word|tag_i)$). With decoding performed using the Viterbi algorithm (Jurafsky and Martin, 2023). As a standard HMM assigns a single hidden state per token, it cannot directly model multi-

TABLE II
CENTRAL KURDISH POS TAGSET

Category	Abbreviation	Category	Abbreviation
Noun	N-S	Adverb	ADV-TIME
	N-EXT		ADV-PLACE
	N-COMP		ADV-S
	N-PROP		ADV-EXT
	N-GEN		ADV-COMP
	N-KNOWN		ADV-EMPH
	N-UNK	Number	NUM-CARD
	N-SG		NUM-EXT
	N-PL		NUM-COMP
	N-M		NUM-ORD
	N-F		NUM-FRAC
	N-NEUT		NUM-BASE
	N-DUAL	Verb	V-SIM
	N-MAT		V-EXT
	N-ABS		V-COMP
Pronoun	PR-1P		V-CONN
	PR-2P		V-PST
	PR-DEM		V-NPST
	PR-INT		V-REQ
	PR-IND		V-CAUS
	PR-REF		V-PASS
	PR-POSS		V-NEG
Adjective	ADJ-QUAL		V-POS
	ADJ-COL		V-TR
	ADJ-DESC		V-INTR
	ADJ-IND		V-INCOMP
	ADJ-GRAD		V-COMPLETE
	ADJ-ATTR	Gerund	GER-A
	ADJ-SUPL		GER-T
	ADJ-COMP		GER-D
	ADJ-QUAN		GER-W
	ADJ-DEM		GER-Y
	ADJ-INT		GER-S
	ADJ-PART		GER-EXT
	ADJ-USE		GER-COMP
	ADJ-S	Particle	PART-LEX
	ADJ-EXT		PART-CONN
	ADJ-COMPOUND		PART-DET
Adverb	ADV-M	UNKNOWN	UNK
	ADV-D	Punctuation	PUNCT
	ADV-R	Abbreviation	ABBR
	ADV-INT	Email	Email
	ADV-CAUSE	Uniform resource locator	URL

POS: Part-of-speech

TABLE III
SUMMARY STATISTICS OF THE CKL CORPUS

Measure	Value
Total tokens	108,680
Total sentences	3,612
Unique word types	20,717
Type–token ratio	0.19
Avg. sentence length	30 tokens (median 23)
Max. sentence length	377 tokens
Avg. word length	5.3 characters
Multi-tagged tokens	30.3% (2 tags: 18.7%, 3 tags: 11.5%)
Hapax legomena	53.8% of vocabulary

CKL: Central Kurdish language

 http://dx.doi.org/10.14500/aro.12641

Sentence	Word	Tag#1	Tag#2	Tag#3
کۆتبەندکردنی بەرهەمهێنان	کۆتبەندکردنی	GER-D - چاوگی دالی	N-COMP - ناوی لێکدراو	GER-COMP - چاوگی لێکدراو
کۆتبەندکردنی بەرهەمهێنان	بەرهەمهێنان	GER-A - چاوگی ئەلفی	N-COMP - ناوی لێکدراو	GER-COMP - چاوگی لێکدراو
کۆتبەندکردنی بەرهەمهێنان		N-S - ناوی سادە	N-PROP - ناوی تایبەتی	N-F - ناوی مێ
کۆتبەندکردنی بەرهەمهێنان		N-S - ناوی سادە	N-PROP - ناوی تایبەتی	N-M - ناوی نێر
کۆتبەندکردنی بەرهەمهێنان		N-S - ناوی سادە	N-PROP - ناوی تایبەتی	N-M - ناوی نێر
کۆتبەندکردنی بەرهەمهێنان		EMAIL		
کۆتبەندکردنی بەرهەمهێنان		NUM	NUM-CARD - ژمارەی بنچی	
کۆتبەندکردنی بەرهەمهێنان	سەربەرشتیار	N-EXT - ناوی دارژاو	N-DUAL - ناوی دولایەن	N-SG - ناوی تاك
کۆتبەندکردنی بەرهەمهێنان :		PUNCT		
کۆتبەندکردنی بەرهەمهێنان	د.	ABBR		
		N-S - ناوی سادە	N-PROP - ناوی تایبەتی	N-M - ناوی نێر
		N-S - ناوی سادە	N-PROP - ناوی تایبەتی	N-M - ناوی نێر
بەشی زمانی – کۆڵێژی – زانکۆی	بەشی	N-S - ناوی سادە	N-SG - ناوی تاك	N-UNK - ناوی نەناسراو
	زمانی	N-S - ناوی سادە	N-SG - ناوی تاك	N-UNK - ناوی نەناسراو
بوخته: ئەم توێژینەوەیە لەژێر ناونیشانی		N-EXT - ناوی دارژاو	N-ABS - ناوی واتایی	N-SG - ناوی تاك
(کۆتبەندکردنی بەرهەمهێنان)ە کە زۆرجار _		PUNCT		
بەپێی قۆناغی لێکۆڵینەوەزمانییەکان دو و	کۆڵێژی	N-S - ناوی سادە	N-SG - ناوی تاك	N-UNK - ناوی نەناسراو
زاراوەی جیاوازی بۆبەکاردێت: 1		N-S - ناوی سادە	N-SG - ناوی تاك	N-GEN- ناوی گشتی
-blocking، کە لە کوردیش زۆرجار زاراوەی _		PUNCT		
بلۆککردنی بۆبەکاردێنن، کە بە واتای	زانکۆی	N-COMP - ناوی لێکدراو	N-SG - ناوی تاك	N-UNK - ناوی نەناسراو
کۆتبەندکردن دێت، 2 -هەندێ جاریش هەبە		N-S - ناوی سادە	N-PROP - ناوی تایبەتی	N-SG - ناوی تاك
زاراوەی competition، یان	بوختە	N-EXT - ناوی دارژاو	N-UNK - ناوی نەناسراو	N-SG - ناوی تاك
بەرانبەر (المنافسه) عەرەبی دەوسێ. هەردو و :		PUNCT		

Fig. 2. Manual annotation sheet sample.

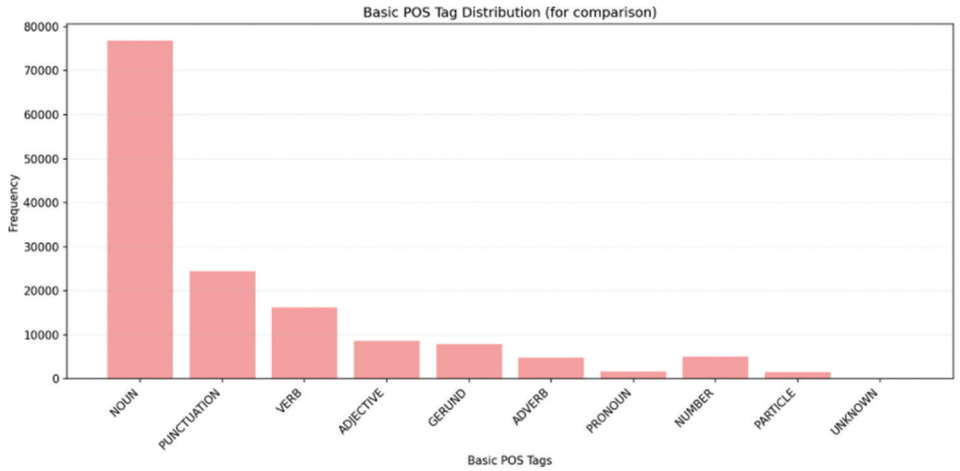

Fig. 3. The distribution of main part-of-speech categories in the central Kurdish language corpus.

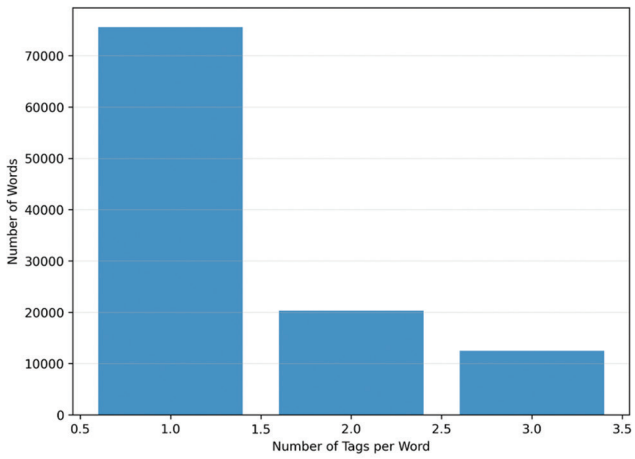

Fig. 4. The distribution of multi-tag annotations in the central Kurdish language corpus.

label annotations; therefore, the Tag1–Tag3 annotations are incorporated via training-data expansion, where each sentence is replicated using Tag1, Tag2, and Tag3 (when available) as alternative single-label sequences.

While HMMs cannot capture long-range dependencies or subword features, they provide a strong baseline for comparison with neural architectures. To capture longer-range dependencies, the first neural model used is a BiLSTM network (Hochreiter and Schmidhuber, 1997), which encodes tokens in both forward and backward directions to capture long-range contextual dependencies. The architecture comprises several key components: (a) Tokenization employs byte pair encoding with a 5,000 subword vocabulary; (b) the embeddings utilize pretrained fastText Central Kurdish vectors (cc.ckb.300.vec), represented as 300-dimensional vectors (Grave, et al., 2018); (c) subword embeddings are aggregated through mean pooling and linear projection; (d) a three-layer BiLSTM encoder captures bidirectional context with a hidden size of 512 and a dropout rate of 0.4; (e) adaptive tag cycling alternates between three tags during training to mitigate label bias; (f) training is executed using

AdamW optimization, with gradient clipping (max norm = 1.0), and cosine annealing learning-rate scheduling.

To further refine the consistency of sequences, a CRF output layer (Lafferty, McCallum, and Pereira, 2001) is added on top of the BiLSTM. Models the joint of tag sequences, scoring transitions, which leads to coherent predictions and discourages inconsistent tags (Huang, Xu and Yu, 2015). With the aim of leveraging diversity among models and datasets, this method represents all valid annotations at training time, leading to greater robustness and generalization than single-model ML.

IV. Experimental Design and Scenarios

All experiments use stratified 5-fold cross-validation, with balanced tag distributions and the inclusion of rare tags in the training set. Results are reported as mean ± standard deviation of metrics across the five folds. The measures to prevent data leakage through tokenization, training the tokenizer only on training data, and not sharing sentences between folds, as well as only assigning rare tags to the training set. The total corpus comprises 108,680 tokens, allocated 80% for training (86,944 tokens) and 20% for testing (21,736 tokens), applied consistently across each cross-validation fold.

Oversampling is used to deal with the problem of class imbalance. For HMM experiments, sentences with tags that occur fewer than 10 times are ×5 copied in the training. For BiLSTM and BiLSTM+CRF experiments, sentences that contain tags with <25 occurrences are oversampled ×4. After preprocessing and normalization, the neural models are optimized using AdamW with weight decay, cosine annealing learning-rate scheduling, and gradient clipping (maximum norm = 1.0). The best epoch for each fold is selected based on validation flexible accuracy; therefore, a prediction is considered correct if it matches any of the annotated tags (Tag1, Tag2, Tag3). For robustness, results are averaged over folds, and 95% confidence intervals are computed. In practice, the neural models in PyTorch are implemented and decoded with the provided torchcrf library (Nguyen and Salazar, 2019). Training is performed on an NVIDIA RTX 4090 GPU, while HMM is run on a central processing unit. The software stack included Python, NumPy, pandas, scikit-learn (for evaluation metrics), and optuna for hyperparameter optimization. The selected settings are embedding dimension of 300, a BiLSTM hidden dimension of 512, three BiLSTM layers, dropout of 0.4, learning rate of 0.001, batch size of 32, and 150 epochs. To systematically evaluate the effects of model architecture, a set of 14 experimental scenarios is designed. These scenarios span three major architecture families:

- HMM: A statistical sequence model trained at the word level, serving as a traditional ML
- BiLSTM: A neural encoder with softmax classification (Abdullah, et al., 2025), optionally combined with adaptive tag cycling, ensemble learning, and embedding normalization
- BiLSTM + CRF: The same BiLSTM encoder extended with a linear-chain CRF decoder for structured sequence prediction.

Beyond these configurations, the impact of tagset granularity is evaluated by comparing the full 86-tags inventory with a main tag categories version of the tagset, consisting of the main grammatical categories (Noun, Pronoun, Verb, Adjective, Adverb, Number, Particle, Gerund, Unknown, and Punctuation). This reduced tagset, referred to as main categories, helps to analyze whether performance improves when morpho-syntactic detail is abstracted into broader classes. Both the full and main categories tagsets are used across the experimental scenarios.

All scenarios used the same dataset, preprocessing pipeline, and stratified 5-fold cross-validation described previously. The only differences are in model architecture, training configuration, and tag granularity.

To analyze the impact of stop-word retention on tagging performance, the main models are evaluated with and without stop words, allowing for transparency regarding potential accuracy inflation. This study trains two representative models: the HMM as a statistical baseline and the BiLSTM + CRF + Cycling + Ensemble model as the strongest neural architecture. Two granularities, an 86-tags scheme and a main categories tagset, are tested across both models, using consistent data splits, training protocols, and evaluation measures.

This methodology outlines the fundamentals of corpus construction, model architectures, and varied experimental scenarios, leading to results presented across fourteen scenarios. It emphasizes the influence of model architecture, CRF decoding, ensemble learning, tag cycling, and tag granularity on performance.

V. Results and Discussion

In the absence of explicit indication, the experiments reported in this section are conducted on the stop-word-filtered version of the corpus. This setting is adopted to control feature sparsity and reduce the dominance of high-frequency function words in model training and evaluation. To address concerns regarding realism and comparability, additional evaluations of the main models with stop words retained are reported separately to contextualize the results and to illustrate the effect of stop-word handling on POS tagging performance.

A. Overall Model Performance

Table IV presents the flexible accuracy results over fourteen experimental settings, where a prediction is considered successful if it coincides with at least one valid annotation assigned to a token (Tag1–Tag3). The HMM baseline model is tooled to a flexible accuracy of 86.0% on the main POS categories, and 65.0% for the detailed tagset. This sharp decrease emphasizes the lack of expressive power of statistical sequence models for applying to fine-grained, morphologically complex tagsets.

 http://dx.doi.org/10.14500/aro.12641

TABLE IV
SUMMARY OF MODELS' RESULTS

Scenario	Tag granularity	Flexible accuracy (%)
1 (HMM)	Main	86.0
2 (HMM)	Detailed	65.0
3 (BiLSTM)	Detailed	73.0
4 (BiLSTM-cycling)	Detailed	75.6
5 (BiLSTM-CRF)	Detailed	74.0
6 (BiLSTM-ensemble)	Detailed	75.1
7 (BiLSTM-cycling-CRF)	Detailed	75.3
8 (BiLSTM-cycling-ensemble-CRF)	Detailed	76.2
9 (BiLSTM)	Main	87.6
10 (BiLSTM-cycling)	Main	88.0
11 (BiLSTM-CRF)	Main	88.0
12 (BiLSTM-ensemble)	Main	88.2
13 (BiLSTM-cycling-CRF)	Main	88.1
14 (BiLSTM-cycling-ensemble-CRF)	Main	89.5

HMM: Hidden Markov model, BiLSTM: Bidirectional long short-term memory, CRF: Conditional random field

Performance of neural architectures is consistently higher than the HMM baseline in all setups. While on the detailed tagset, BiLSTM-based models obtain flexible accuracy in the range from 73.0% to 76.2%, whereas performance on main categories is higher up to 89.5% for the best-case configuration. These results confirm the effectiveness of contextual neural representations in handling ambiguity and sparsity in Central Kurdish POS tagging.

When compared with prior work, such as the DASTAN corpus, the reported accuracy appears lower. DASTAN achieved approximately 96% accuracy using a hybrid HMM and rule-based system on a smaller 38-tag scheme with about 74k tokens. This difference is expected given the substantially more challenging setting adopted in the present study, which employs a purely statistical HMM and a fine-grained 86-tag scheme that results in fewer training instances per class. Consequently, although absolute accuracy is lower, the experimental setting is more demanding and linguistically expressive. To better understand the sources of these gains, the next subsection analyzes the contribution of individual architectural components.

In addition to recurrent and statistical models, a transformer-based baseline (mBERT) is also fine-tuned and evaluated under the same experimental protocol. On the main POS categories tagset, mBERT achieved a flexible accuracy of 77.0%. Although competitive, this result remains below that of the best BiLSTM–CRF configurations, suggesting that under the available data scale and tag granularity, recurrent architectures with structured decoding are better suited to Central Kurdish POS tagging.

B. Contribution of Model Components (Ablation Study)

To better understand the contribution of individual architectural components, an ablation study is conducted by incrementally adding adaptive tag cycling, CRF decoding, and ensemble learning to a BiLSTM baseline. The results are presented in Table V.

Introducing adaptive tag cycling leads to a clear improvement over the baseline, which illustrates that exposing a model to all available annotation layers (Tag1–Tag3) when training introduces richer supervision information. Including a CRF decoder brings modest, but reliable improvement by imposing sequence-level constraints and penalizing invalid tag transitions. Ensemble learning also enhances performance through variance reduction as well as the combination of diverse predictions made by multiple independently trained models. In addition to architectural components, the effect of data imbalance is also determined to affect overall results. With regard to the BiLSTM models, oversampling is performed for sentences with rare tags (<25 samples ×4) after some small experiments that found this level yielded a reasonable accuracy improvement, averaging about +1.2% points. While for HMM model (<10 samples, ×5) shows the best results with an accuracy improvement averaging about +0.5% points. These results for models with detailed tags usage, while for main tags models, the effect of the oversampling is much smaller and often negligible.

TABLE V
ABLATION ANALYSIS OF MODEL COMPONENTS

Model variant	Cycling	CRF	Ensemble	Detailed tagset accuracy (%)	Main tagset accuracy (%)
BiLSTM	✗	✗	✗	73.0	87.6
+ Cycling	✓	✗	✗	75.6	88.0
+ CRF	✗	✓	✗	74.0	88.0
+ Ensemble	✗	✗	✓	75.1	88.2
+ Cycling+CRF	✓	✓	✗	75.3	88.1
+ Cycling+CRF+ Ensemble	✓	✓	✓	76.2	89.5

CRF: Conditional random field. ✓ indicates that the component is included in the model variant, while ✗ indicates that the component is not included.

The full BiLSTM-cycling-ensemble-CRF model performs best, with 76.2% flexible accuracy for the detailed tagset and 89.5% for the main categories, benefiting from each of these components. To the best of our knowledge, this study is among the first to demonstrate the benefit of combining multiple independently trained BiLSTM-based models for POS tagging in Central Kurdish, reducing variance and capturing complementary tagging decisions.

C. Evaluation Strategy, Stability, and Stop-word Effects

To provide a more comprehensive evaluation beyond flexible accuracy, multiple complementary metrics are reported. In addition to flexible accuracy, which accounts for annotation ambiguity by accepting any of the valid POS labels (Tag1–Tag3), strict accuracy based solely on the primary annotation (Tag1) is also included for comparability with prior work. Furthermore, weighted and macro F1 scores are reported to better capture performance under class imbalance.

Table VI presents the results of the baselines in terms of both tag granularities. Given the highly unbalanced distribution of POS categories, in particular for the 86-tag scheme, weighted F1 is used as the dominant F-measure, which measures tagging quality across all classes while taking class frequency into consideration. For instance, some rare system tags, such as EMAIL, occur only 4 times in the entire corpus, whereas frequent categories, such as common

TABLE VI
PERFORMANCE OF THE MAIN MODELS

Model	Tag granularity	Flexible accuracy (%)	Strict accuracy (%)	Weighted F1 (%)	Macro F1 (%)
HMM	Detailed	65.0	58.0	58.0	34.7
HMM	Main	86.0	86.0	85.0	69.0
BiLSTM-cycling-ensemble-CRF	Detailed	76.2	61.4	68.0	57.0
BiLSTM-cycling-ensemble-CRF	Main	89.5	89.2	89.3	78.0

HMM: Hidden Markov model, BiLSTM: Bidirectional long short-term memory, CRF: Conditional random field

noun tags, occur thousands of times (e.g., N-S occurs 12,115 times and N-GEN occurs 13,306 times). Under such conditions, macro F1 is strongly affected by rare classes, as each category contributes equally regardless of its frequency, leading to lower macro-averaged scores. Nevertheless, macro F1 is additionally reported to highlight per-class performance trends and to explicitly reflect the impact of rare and underrepresented categories.

Across five-fold cross-validation, the BiLSTM-cycling-ensemble-CRF model achieved a flexible accuracy of 89.5 ± 0.2% STD on the main POS categories and 76.2 ± 0.5% STD on the detailed tagset. The low standard deviation across folds indicates stable and consistent performance, suggesting that the reported gains are not driven by a particular data split but generalize well across the corpus.

To contextualize the main results and assess robustness, the effect of stop-word handling is examined by re-evaluating a representative statistical baseline (HMM) and the strongest neural configuration (BiLSTM-cycling-ensemble-CRF) on a version of the corpus in which stop-words are retained. As can be observed from Table VII, stop-word inclusion consistently increases the flexible accuracy for both models and both tag granularities. This is an expected behavior, since stop-words are often high-frequency function words and relatively low in tag ambiguity.

D. Effect of Tag Granularity

Besides the evaluation method, tag granularity also shows its influence on performance. Flexible accuracy is significantly improved, reaching up to 89.5%, as the original 86 fine-grained tags are thus grouped into main categories. In comparison, using the full detailed tagset, the performance is maximized at 76.2%. This discrepancy implies that learning becomes substantially easier when label sparsity is relaxed, especially in a low-resource scenario.

However, the detailed tagset is linguistically useful since it reflects subtle morphological and functional contrasts that are collapsed into main tags. Under this more challenging setting, a relaxed accuracy of 76.2% with the 86-tag scheme is competitive with CKL results published in earlier works on corpora with comparable size. In general, these results show both the penalty of larger tag granularity and the ability of neural models to handle this complexity well.

E. Error Analysis and Class-Level Behavior

To gain insight into systematic errors, an error analysis based on per-class F1 scores, as shown in Table VIII and the row-normalized confusion matrix presented in Fig. 5,

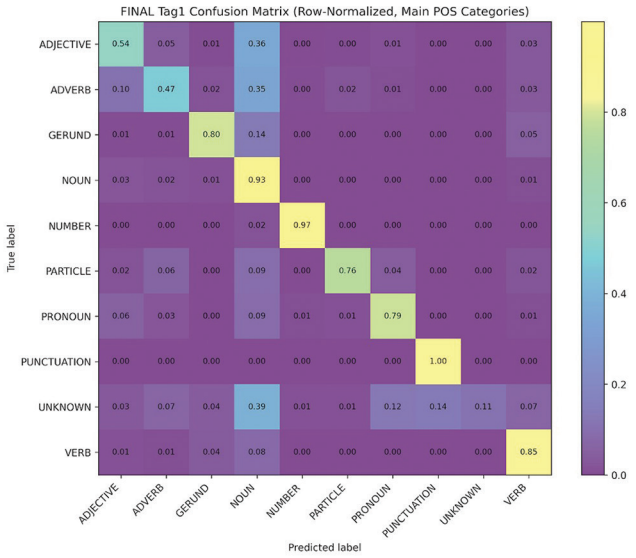

Fig. 5. The row-normalized confusion matrix for the bidirectional long short-term memory-cycling-ensemble-conditional random field model using main part-of-speech categories

is analyzed for the best-performing model, the BiLSTM-cycling-ensemble-CRF model, using the main tags. High-frequency and statistically stable tags, such as NOUN VERB NUMBER PUNCTUATION, yield largely robust performances across models, given the clear syntactic roles and negligible context ambiguity.

On the other hand, ADJECTIVE and ADVERB have noticeably lower F1 compared to some of the top-scoring classes, despite having relatively high token counts (tag distributions presented in Fig. 3). The confusion matrix shows a high rate of erroneous classification between such categories and NOUN. This is largely because they are almost completely functionally superimposed in CKL, where identical surface forms work as adjectives or adverbs based on simple context and therefore constantly get mixed up with one another and with NOUN.

The less frequent GERUND category, on the other hand, shows a relatively good performance as the morphological and syntactic constraints are clearer, so that there is much less competition for labels. The class UNKNOWN has the lowest and the highest variance, with most instances incorrectly classified as NOUN or PUNCTUATION. Such a result is relatively predictable given that UNKNOWN constitutes a small and motley collection of tokens, which are symbols and foreign strings, instead of grouping under an appropriate concept grammar.

 http://dx.doi.org/10.14500/aro.12641

TABLE VII
Effect of Stop-words Inclusion on POS Tagging Performance

Model	Tag granularity	Stop-words	Flexible accuracy (%)	Strict accuracy (%)
HMM	Detailed	Removed	65.0	58.0
HMM	Detailed	Kept	74.0	70.0
HMM	Main	Removed	86.0	86.0
HMM	Main	Kept	88.0	88.0
BiLSTM-cycling-ensemble-CRF	Detailed	Removed	76.2	61.4
BiLSTM-cycling-ensemble-CRF	Detailed	Kept	79.0	69.0
BiLSTM-cycling-Ensemble-CRF	Main	Removed	89.5	89.2
BiLSTM-cycling-ensemble-CRF	Main	Kept	92.3	92.1

POS: Part-of-speech, HMM: Hidden Markov model, BiLSTM: Bidirectional long short-term memory

TABLE VIII
Per-class F1 Scores (Mean±Standard Over Five-fold Cross-validation) for the BiLSTM-Cycling-Ensemble-CRF Model using Main POS Categories

POS class	Mean F1 (%)±standard
NOUN	91.5±0.2
VERB	86.5±0.7
ADJECTIVE	58.1±0.7
ADVERB	52.0±0.8
PRONOUN	80.0±1.9
NUMBER	97.3±0.4
PARTICLE	77.1±1.1
GERUND	80.5±1.1
PUNCTUATION	99.9±0.0
UNKNOWN	21.7±13.2

CRF: Conditional random field, BiLSTM: Bidirectional long short-term memory, POS: Part-of-speech

In addition to class-level error, model performance is compared at varying levels of ambiguity in annotation as presented in Table IX. Products with 1, 2, or 3 appropriate tags are treated separately. Flexible accuracy remains relatively constant with the level of ambiguity, even as strict accuracy drops dramatically. This observation is the most convincing evidence that, in many cases, the model produces a linguistically acceptable tag even if it does not correspond to the primary annotator's preference, showing the impact of single-label evaluation and the need for ambiguity-aware metrics.

F. Practical Implications

In summary, his work has a great impact on practical applications in Kurdish NLP, including grammar and spell corrections and MT. The presented work provides a reusable, annotated corpus and trained models that allow to shorten development cycles and increase the reliability of the Kurdish language technologies. From a research perspective, the 86-tag annotated corpus will benefit the Kurdish NLP community by offering a more detailed benchmark compared to existing resources, enabling the exploration of advanced modeling approaches and advancing research on morphologically rich, low-resource languages.

TABLE IX
Performance by Ambiguity Level

Ambiguity level	Flexible accuracy (%)	Strict accuracy (%)
1 annotation	76.4	76.4
2 annotations	75.0	58.1
3 annotations	76.0	39.5

VI. Conclusion and Future Work

This study presents the first large-scale, fine-grained POS-tagged corpus for CKL consisting of 108,680 tokens manually annotated with an 86-tag scheme. Using this resource, a systematic evaluation of statistical and neural sequence-labeling models is conducted, including HMM, BiLSTM, BiLSTM–CRF, and ensemble-cycling architectures, under a flexible evaluation framework that accounts for annotation ambiguity.

Beyond its performance, this work emphasizes the importance of ambiguity-aware evaluation for morphologically rich and low-resource languages, where it is demonstrated that flexible accuracy provides better results than strict single-label evaluation.

From these experiments, three key findings emerge:

1. Neural architectures substantially outperform HMMs, though HMMs remain competitive for collapsed tag categories
2. Adaptive tag cycling plays a crucial role, often providing larger gains than CRF decoding by exploiting the multi-annotation structure of the corpus
3. Ensemble-cycling with CRF achieves the strongest performance, with flexible accuracy reaching 89.5% on main categories and 76.2% on the full 86-tags scheme — the highest reported so far for CKL POS tagging at this level of granularity.

When compared to prior work such as the DASTAN corpus, however, performance appears lower (96% on a 38-tag set). This difference reflects the much greater difficulty of the present setting, more tags and fewer tokens per tag. While earlier CKL studies incorporated rule-based components, these systems rely on handcrafted rules tailored to smaller tagsets and specific datasets that are not publicly available. Because of this, and given CKL's rich and

complex morphology, the generalizability of such rule-based approaches to the 86-tags scheme cannot be confirmed. Thus, even a score of 76.2% on the detailed scheme is a significant milestone for Kurdish NLP.

Despite these gains, challenges remain for inherently ambiguous categories such as ADJECTIVE and ADVERB, as well as for the heterogeneous UNKNOWN class, indicating that residual errors are driven more by linguistic overlap than by data sparsity alone.

Future work will focus on extending the corpus beyond academic writing to include other domains, increasing the dataset size to better support detailed tagging, and expanding the annotation work for more Kurdish dialects such as Kurmanji and Badini. A multilingual BERT baseline is evaluated in the current work; recurrent architectures are found to work best at this scale of data, and future experimentation with bigger datasets should consider transformer models.

In summary, this study will pave the way for these extensions. By making available a well-annotated corpus as well as strong baselines, this work provides a foundation for the next stage of Kurdish NLP research, paving the way towards POS tagging model improvement through additional trained/annotated tokens, parsing, MT, and linguistic resource development.

References

Abdullah, A.A., Hasan, S., Toufiq, D., Maghdid, H.S., Rashid, T.A., Farho, P., Sabr, S., Taher, A.H., Sabir, D., Veisi, H., and Asaad, A.T., 2024. *NER- RoBERTa: Fine-Tuning RoBERTa for Named Entity Recognition (NER) within Low-resource Languages*. arXiv [Preprint].

Abdullah, A.A., Mohammed, N.S., Khanzadi, M., Asaad, S.M., Abdul, Z.K., and Maghdid, H.S., 2025. In-depth analysis on machine learning approaches: Techniques, applications, and trends. *ARO-The Scientific Journal Of Koya University*, 13(1), pp.190-202.

Abdulrahman, R.O., and Hassani, H., 2020. *Using Punkt for Sentence Segmentation in Non-Latin Scripts: Experiments on Kurdish (Sorani) Texts.* arXiv [Preprint].

Ahmadi, S., 2020a. A Tokenization System for the Kurdish Language. In: Zampieri, M., Nakov, P., Ljubešić, N., Tiedemann, J., and Scherrer, Y., Eds. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*. International Committee on Computational Linguistics (ICCL), Barcelona, Spain, pp.114-127. Available from: https://aclanthology.org/2020.vardial-1.11 [Last accessed on 2025 May 03].

Ahmadi, S., 2020b. KLPT-Kurdish language processing toolkit. In: Park, E.L., Hagiwara, M., Milajevs, D., Liu, N.F., Chauhan, G., and Tan, L., Eds. *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. Association for Computational Linguistics, Pennsylvania, pp.72-84.

Ahmadi, S., and Masoud, M., 2020. Towards Machine Translation for the Kurdish Language. In: Karakanta, A., Ojha, A.K., Liu, C.H., Abbott, J., Ortega, J., Washington, J., Oco, N., Lakew, S.M., Pirinen, T.A., Malykh, V., Logacheva, V., and Zhao, X., Eds. *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*. Association for Computational Linguistics, Suzhou, China, pp.87-98.

Amini, Z., Mohammadamini, M., Hosseini, H., Mansouri, M., and Jaffet, D., 2021. *Central Kurdish Machine Translation: First Large Scale Parallel Corpus and Experiments*. arXiv [Preprint]. Available from: https://hal.science/hal-03263105 [Last accessed on 2025 May 03].

Ataman, D., 2018. *Bianet: A Parallel News Corpus in Turkish, Kurdish and English*. arXiv [Preprint].

Awlla, K.M., Veisi, H., and Abdullah, A.A., 2025. Sentiment analysis in low-resource contexts: BERT's impact on Central Kurdish. *Language Resources and Evaluation*, 59(3), pp.2213-2243.

Azzat, M., Jacksi, K., and Ali, I., 2024. A Hybrid Approach to ontology construction for the Badini Kurdish language. *Information*, 15(9), p.578.

Daelemans, W., 2011. POS tagging. In: Sammut, C., and Webb, G.I., Eds. *Encyclopedia of Machine Learning*. Springer, US, Boston, MA, pp.776-779.

Gökırmak, M., and Tyers, F.M., 2017. A Dependency Treebank for Kurmanji Kurdish. In: Montemagni, S., and Nivre, J., Eds. *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*. Linköping University Electronic Press, Pisa, Italy, pp.64-72. Available from: https://aclanthology.org/W17-6509 [Last accessed on 2025 May 03].

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T., 2018. Learning Word Vectors for 157 Languages. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). Available from: https://fasttext.cc/docs/en/crawl-vectors.html [Lase accessed on 2025 Sep 16].

Hassani, H., 2022. *Part of Speech Tagging (POST) of a Low-resource Language using another Language (Developing a POS-Tagged Lexicon for Kurdish (Sorani) using a Tagged Persian (Farsi) Corpus*. arXiv [Preprint].

Hochreiter, S., and Schmidhuber, J., 1997. Long short-term memory. *Neural Computation*, 9(8), pp.1735-1780.

Huang, Z., Xu, W., and Yu, K., 2015. *Bidirectional LSTM-CRF Models for Sequence Tagging.* arXiv [Preprint].

Jurafsky, D., and Martin, J.H., 2023. *Speech and Language Processing*. 3rd ed. Stanford University, California.

Lafferty, J.D., McCallum, A., and Pereira, F.C.N., 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data In: *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, pp.282-289.

Landis, J.R., and Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), p.159.

Malmasi, S., 2016. Subdialectal Differences in Sorani Kurdish. In: Nakov, P., Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., and Malmasiet, S., Eds. *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*. The COLING 2016 Organizing Committee, Osaka, Japan, pp.89-96. Available from: https://aclanthology.org/W16-4812 [Last accessed on 2025 May 03].

Maulud, D., Jacksi, K., and Ali, I., 2023a. A hybrid part-of-speech tagger with annotated Kurdish corpus: Advancements in POS tagging. *Digital Scholarship in the Humanities*, 38(4), pp.1604-1612.

Maulud, D., Jacksi, K., and Ali, I., 2023b. Towards a complete Kurdish NLP Pipeline: Challenges and opportunities. *Jurnal Informatika*, 17, pp.1-17.

Mustafa, A.M., and Rashid, T.A., 2018. Kurdish stemmer pre-processing steps for improving information retrieval. *Journal of Information Science*, 44(1), pp.15-27.

Naserzade, M., Mahmudi, A., Veisi, H., Hosseini, H., and MohammadAmini, M., 2022. *CKMorph: A Comprehensive Morphological Analyzer for Central Kurdish*. arXiv [Preprint].

Nguyen, T.Q., and Salazar, J., 2019. *Transformers without Tears: Improving the Normalization of Self-Attention.* arXiv [Preprint].

Pota, M., Marulli, F., Esposito, M., De Pietro, G., and Fujita, H., 2019. Multilingual POS tagging by a composite deep architecture based on character-level features and on-the-fly enriched word embeddings. *Knowledge-Based Systems*, 164, pp.309-323.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X., 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10), pp.1872-1897.

Sabr, S.S., Sabr Mustafa, N., Omar, T.S., Rasool, S.H., Omer, N.A., Hamad, D.S., Abdulhameed Shams, H., Kareem, O.M., Noori, R.A., Abdullah, K.A., Mohammad, M,A., Al-Raghefy, H., Asaad, S.M., Mohammed, S.J.,… & Maghdid, H.S., 2025. *A Comprehensive Part-of-Speech Tagging to Standardize Central-Kurdish Language: A Research Guide for Kurdish Natural Language Processing Task*s. arXiv [Preprint].

Salavati, S., and Ahmadi, S., 2018. *Building a Lemmatizer and a Spell-checker for Sorani Kurdish*. arxiv [Preprint].

Veisi, H., MohammadAmini, M., and Hosseini, H., 2019. Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus. *Digital Scholarship in the Humanities,* 35, pp.176-193.