Part-of-Speech Tagging for Northern Kurdish

Peshmerge Morad[1] **Sina Ahmadi**[2] **Lorenzo Gatti**[3]
[1,3]University of Twente [2]University of Zurich
[1]p.morad@hotmail.com [2]sina.ahmadi@uzh.ch [3]l.gatti@utwente.nl

Abstract

In the growing domain of natural language processing, low-resourced languages like Northern Kurdish remain largely unexplored due to the lack of resources needed to be part of this growth. In particular, the tasks of part-of-speech tagging and tokenization for Northern Kurdish are still insufficiently addressed. In this study, we aim to bridge this gap by evaluating a range of statistical, neural, and fine-tuned-based models specifically tailored for Northern Kurdish. Leveraging limited but valuable datasets, including the Universal Dependency Kurmanji treebank and a novel manually annotated and tokenized gold-standard dataset consisting of 136 sentences ($2,937$ tokens). We evaluate several POS tagging models and report that the fine-tuned transformer-based model outperforms others, achieving an accuracy of 0.87 and a macro-averaged F1 score of 0.77. Data and models are publicly available under an open license at https://github.com/peshmerge/northern-kurdish-pos-tagging

Keywords: Part-of-Speech tagging, morphosyntactic analysis, Northern Kurdish, low-resource NLP

1. Introduction

Automatic part-of-speech (POS) tagging or grammatical tagging is the process of assigning POS tags to each word/token in a given text. POS tagging is essentially a disambiguation task because words naturally are ambiguous and can have more than one correct tag depending on the context and their position in the sentence. POS tagging serves many purposes in natural language processing (NLP) applications, and it is traditionally considered a building block for other tasks such as named entity recognition (Ma and Liu, 2021), information extraction (Luan et al., 2017), spelling correction (Nagata et al., 2018), text classification (Pranckevičius and Marcinkevičius, 2016), natural language generation (Li et al., 2019), and machine translation (Hlaing et al., 2022).

Just as part-of-speech tagging serves as a precursor for tasks like syntactic parsing, tokenization is a crucial task in NLP and a prerequisite for POS tagging. Tokenization is segmenting the input text into smaller, distinct units termed *tokens*. These tokens can encompass compound words, single words, sub-words, symbols, or other significant elements. At its most fundamental level, tokenization separates tokens using whitespace as a delimiter (Mitkov, 2022, p. 549).

Unlike high-resourced languages (HRLs) like English and French, for which POS tagging and tokenization have been extensively addressed, low-resourced languages (LRLs) like Kurdish lack sufficient tools and resources (Ahmadi, 2020a). Although Northern Kurdish is included in Universal Dependencies (UD) (Nivre et al., 2020) (using the 'Kurmanji' label since version 2.1) based on Gökırmak and Tyers (2017)'s treebank, hence serving as a benchmark, achieving high-accuracy POS tagging for LRLs may require a greater emphasis on linguistic insights as observed in other languages (Manning, 2011). Our literature review indicates that there is room for effective and open-source contributions to Kurdish POS tagging.

In this paper, we report on the progress we have made in addressing the task of POS tagging for Northern Kurdish. More specifically, we revisit the UD Kurmanji treebank (Gökırmak and Tyers, 2017) by reannotating tokens that belong to specific word classes and introducing a different annotation scheme with more fine-grained linguistic features of Northern Kurdish. Secondly, we create a manually tokenized and annotated gold-standard dataset for Northern Kurdish with a total of 136 sentences and $2,937$ tokens. To that end, we deploy an annotation scheme different from that of UD Kurmanji that aims for a more fine-grained representation of linguistic features of Northern Kurdish, notably noun phrases containing *Izafe* (also spelled *Ezafe*) acting as a relativizer and linker. Thirdly, we evaluate the effect of different POS techniques along with the annotation schemes. Finally, we implement different POS tagging models and introduce a state-of-the-art transformer-based POS tagger for Northern Kurdish.

The rest of the paper is organized as follows. In section 2, we provide an overview of the Kurdish language and its dialects, focusing on Northern Kurdish. Section 3 presents a comprehensive review of related work and state-of-the-art studies on POS tagging for LRLs in general, with a specific focus on Northern Kurdish. We then detail the annotation schemes for the training and testing datasets in section 4. In section 5, we discuss the process of collecting and annotating testing data, as well as augmenting the training data. Additionally, we provide a detailed explanation of the tokenization and POS tagging methods. Subse-

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@LREC-COLING-2024, pages 70–80. May 25, 2024. ©European Language Resources Association: CC BY-NC 4.0

quently, section 6 presents our evaluation results, accompanied by an in-depth analysis. Finally, our conclusions are presented in section 7.

2. Kurdish Language

The Kurdish language belongs to the Northwestern Iranic branch within the Indo-European languages family, spoken by more than 30 million people. The Kurdish language (ISO 639-3 code kur) is divided into many dialects (with corresponding ISO 639-3 languages codes): Northern Kurdish or Kurmanji (kmr), Central Kurdish or Sorani (ckb), Southern Kurdish (sdh), and Laki (ldk) and is closely related to Zaza-Gorani languages (Ahmadi et al., 2019). Northern Kurdish is widely spoken in Syria and Turkey but also in the Kurdistan Region of Iraq, Iran, Armenia and among the Kurdish diaspora. It is written using Kurdified Latin-based and Arabic-based scripts. The Latin-based script is widely known as the Hawar alphabet introduced by Jeladet Ali Bedirkhan in 1932.

Northern Kurdish has a subject–object–verb word order and specifies grammatical gender (feminine and masculine). The noun in its absolute state and without any suffixes represents the generic and definite senses of the noun, and it marks four cases, namely nominative, oblique, *Izafe*, and vocative. In addition, it has a split-ergative alignment in the past tense with transitive verbs. Furthermore, the passive voice (conjugated in all persons, moods, and tenses) is constructed using the verb *hatin* 'to come' and *dan* 'to give' plus the infinitive.

Both the oblique and the *Izafe* case (construct case) are essential in Northern Kurdish for indicating the roles of the nouns and the pronouns in a sentence. Nouns, proper nouns, personal pronouns, and demonstrative adjectives, in both cases, undergo a form change as in "*komputera min*" (my computer) where 'a' is an *Izafe* linking '*komputera*' (computer) to '*min*' (my). They are either completely altered, or the case markers are added to the end of the noun and proper nouns. Those markers, shown in Table 1, are unstressed markers that reveal the gender and number of nouns. In this study, our introduced annotation scheme, discussed in Section 5.1, particularly revolves around addressing and segmenting the oblique and *Izafe* case markers in our datasets.

Nonetheless, *Izafe* case markers differ from oblique case markers in the fact that they can also appear as separate particles serving the same purpose within definite nouns; this phenomenon is referred to as *construct extender* (Thackston, 2006) because it allows extending the *Izafe* case by adding adjectives or nouns to the first *Izafe* case.

	OBLIQUE		IZAFE	
	Definite	Indefinite	Definite	Indefinite
SG. F.	*-ê*	*-ekê*	*-a*	*-eke/-eka*
SG. M.	*-î*	*-ekî*	*-ê*	*-ekî*
PL	*-an*	*-inan*	*-ên*	*-ine*

Table 1: Case markers based on the number, gender, and definiteness of the noun in Northern Kurdish. If the noun ends in a vowel, the case markers will be preceded by a *-y*.

3. Related Work

The task of POS tagging has been addressed using various methods. Rule-based techniques (Brill, 1992; Karlsson, 1990) were the first methods applied. Decision Trees have also been employed for the task (Schmid, 1994). Furthermore, hidden Markov models (HMMs) and conditional random fields (CRFs) have been widely used and proved to be effective for this task (Schmid and Laws, 2008; Pradhan and Yajnik, 2023; Yousif, 2019; Stratos et al., 2016; Silfverberg et al., 2014).

Additionally, deep learning based approaches like recurrent neural networks and (Bi)LSTMs have shown to be powerful in capturing temporal dependencies when performing POS tagging (Wang et al., 2015; Qi et al., 2020; Horsmann and Zesch, 2017). Those are often combined with other techniques such as convolutional neural networks, HMMs, and CRFs (Shao et al., 2017; Plank et al., 2016; Ma and Hovy, 2016; Maimaiti et al., 2017).

In recent years, the rise of transformer-based architectures introduced by Vaswani et al. (2017) has led to the development of large language models (LLMs) such as GPT2 (Radford et al., 2019), BERT (Kenton and Toutanova, 2019) and RoBERTa (Liu et al., 2019). These models have greatly influenced NLP in various fields. However, despite being trained on multiple languages, they don't always perform better than single-language models, especially in less-resourced languages, for tasks like POS tagging (Conneau et al., 2020). Nonetheless, they can adapt and improve their performance when fine-tuned (Maimaiti et al., 2021).

For Kurdish, Walther et al. (2010) presents the first dedicated work on POS tagging for Northern Kurdish, where a morphological lexicon (KurLex) and a POS tagger were created. The authors report an 85.7% precision, however on a small annotated corpus of 13 sentences. Although Gökırmak and Tyers (2017)'s treebank for Northern Kurdish is available on UD and has been used in various consecutive studies in multilingual training setups as in Qi et al., 2020 (BiLSTM) and Nguyen et al., 2021 (transformer-based fine-tuning) inter alia, there is still no tool or fine-grained dataset in-

dicating the existing gap in the literature (Ahmadi, 2020a).

4. Annotation Schemes

4.1. UD Kurmanji Scheme

The UD Kurmanji treebank (Gökırmak and Tyers, 2017) is a treebank for Northern Kurdish that contains morpho-syntactic information such as POS tags and some morphological features. The data in the treebank is drawn from fiction and encyclopedic data in roughly equal measure. It consists of the Kurdish translation of *The Adventure of the Speckled Band* story and sentences from the Northern Kurdish Wikipedia. UD Kurmanji contains $10,189$ tokens and has been annotated following the UD annotation scheme (Nivre et al., 2020), meaning it does not allow multi-word expressions, and it instructs to undone contractions. In addition, the case markers, shown in Table 1, within nouns are not segmented. Moreover, the construct extenders in the treebank are tagged as ADP. For example, the noun phrase '*Beşa Felsefeyê*' (department of philosophy) is tagged as NOUN and NOUN, respectively, while having *Izafe* and oblique case markers in both nouns.

4.2. Our Scheme

We propose a different, fine-grained annotation scheme taking into account all case and indefinite noun markers. In addition, we address multi-word prepositions such as 'lê' (from, analogous to *au/aux* in French), adverbs, and compound verbs and tag them as single tokens. It is worth mentioning that the UD annotation scheme (Nivre et al., 2020) serves as a basis for our scheme.

Case Markers and Determiners One of the main differences between our scheme and the UD Kurmanji scheme is how we segment the nouns and their attached indefinite, oblique, and *Izafe* case markers. We use the POS tags from the UD tagset (Petrov et al., 2012). While we use DET for indefinite and oblique case makers, we introduce a new POS tag named IZAFE for the *Izafe* case markers. For example, the noun phrase *Beşa Felsefeyê* (department of philosophy) is split into four tokens *Beş*, *a*, *Felsefe yê* and respectively tagged as NOUN, IZAFE, NOUN and DET.

Multi-word Expressions In UD Kurmanji, the tag X is assigned to nouns that are part of the compound verbs; in our case, we tag those nouns either as a NOUN or all together with the verbs they belong to as a multiword expression VERB. For instance, in UD Kurmanji, the compound verb '*pêşkêş dikin*' (presenting) is split into two tokens: *pêşkêş* and *dikin* and tagged X and VERB, respectively. Within our annotation scheme, we tag it as VERB.

Regarding compound prepositions, we annotate the compound preposition '*li ser*' (on/upon) as ADP, while in UD Kurmanji, it is separated into two tokens '*li*' (in/at) and '*ser*' (onto) where both are tagged as ADP. In addition, compound adverbs such as '*bi tenê*' (only) are also separated into two tokens '*bi*' (with) and '*tenê*' (alone), both are annotated as ADP. However, we treat it as a multi-word token, and we annotate it as ADV.

Moreover, the verb to be in Northern Kurdish '*bûn*' (to be) is always annotated as AUX in UD Kurmanji treebank, while we tag it as a VERB unless it appears as a light verb. In addition, the particles *-ê* and *dê* are used for forming the future tense in Northern Kurdish and are tagged as AUX in UD Kurmanji. However, we tag those particles as PART because they are not auxiliary verbs.

Furthermore, the tokens '*jî*' (also/too) and '*her*' (every) are annotated as PART and either as DET or ADV in the UD Kurmanji, respectively. We annotate the former as ADV and the latter as PRON.

5. Methodology

5.1. Data Collection and Annotation

We collect 136 ($2,937$ tokens) sentences written in Northern Kurdish from multiple news websites. The first 100 sentences are taken from the unannotated Pewan corpus (Esmaili et al., 2013). The remaining 36 sentences are taken from three Kurdish news websites, mainly Kurdistan24[1], Xwebûn[2], and Hawar News[3]. We annotated those sentences according to our annotation scheme introduced in section 4.2. We call this collection the "gold-standard dataset", and we use it as a test set to evaluate our POS tagging models. Figure 1 demonstrate the statistics of this dataset.

Similar to the UD Kurmanji treebank, for each given sentence in our gold-standard dataset, we provide: 1) the raw (untokenized) sentence where tokens are delimited by whitespaces and the case markers are not split-off, and 2) a list of tokens with corresponding POS tags where the case markers are segmented and annotated.
The availability of the untokenized sentence, along with the list of the tokens, enables us to evaluate various tokenization methods. The untokenized sentence can be fed to any tokenizer, and its output can be compared against the list of tokens we already have, which we consider as gold tokens.

[1] https://www.kurdistan24.net/kmr
[2] https://xwebun1.org
[3] https://hawarnews.com/kr

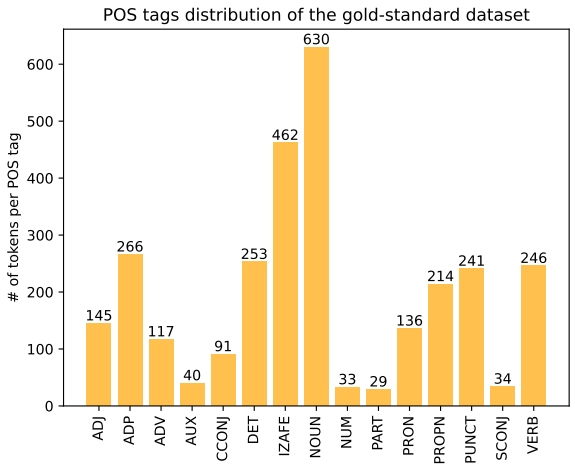


Figure 1: Number of tokens per POS tags in our gold-standard dataset

5.2. Data Augmentation

We augment the UD Kurmanji treebank by splitting the case and indefinite markers from the tokens they are attached to. Thus introducing new tokens. For example, we split '*hevalekî*' (a male friend) into three separate tokens, each with its corresponding POS tag: *heval* as NOUN, *ek* (indefinite noun marker) as DET, and finally *î* as IZAFE. In addition, we re-tag independent *Izafe* markers (construct extender) as IZAFE instead of ADP. Finally, we reverse the splitting of the contracted prepositions (*jê, lê, pê, tê*) in the treebank.

Our approach for augmenting the UD Kurmanji treebank bears a close resemblance to the research described by (Seddah et al., 2023). The authors made significant steps in addressing tokenization issues to ensure consistency in the NArabizi treebank annotations (Farah et al., 2020), the user-generated content variety of Arabic Algerian, which is known for its frequent usage of code-switching. For instance, they carefully segmented specific classes of words, such as determiners in noun phrases.

As a result of this augmentation step, the number of tokens increased in the treebank ($12,233$ tokens). We refer to this augmented version as UD Kurmanji augmented, while we refer to the version with its initial annotation scheme as UD Kurmanji original.

5.3. Tokenization

In addition to the KLPT tokenizer, Ahmadi, 2020b provided multiple neural tokenization models trained (unsupervised) on Northern Kurdish raw corpora. We use three of those models: Unigram (Kudo, 2018), Byte-Pair Encoding (BPE) (Sennrich et al., 2016), and wordPiece (Schuster and Nakajima, 2012) tokenizers.

Moreover, we use the NLTK tokenizer and a manual tokenization method. The manual tokenization, as the name suggests, is the process of manually tokenizing any given text. This method is mostly performed in pairs with the task of manually annotating tokens with the corresponding POS tags. Despite being very time-consuming, it is considered to have the best outcome because it is done by humans with good linguistic knowledge of the language. Therefore, the manually tokenized text can be considered the ground truth that can be used for evaluating other automatic tokenization methods.

5.4. POS Tagging

The task of POS tagging can be seen as a multi-class classification task where a model is trained on annotated data to enable it to classify each token in any given sequence of tokens. There are multiple approaches to tackle the task of POS tagging. Generally, those approaches can be grouped into four categories: rule-based, statistical, neural-based, and transformer-based fine-tuned (Jurafsky and Martin, 2009; Kanakaraddi and Nandyal, 2018).

Except for the work of Walther et al., 2010, there has been no dedicated work for the task of POS tagging for Northern Kurdish. Therefore, we propose seven supervised POS tagging models. The goal is to cover POS methods as much as possible to establish a baseline method and to examine the effectiveness of those methods. Those methods will be explained in the following subsections.

It is worth mentioning that we train all POS tagging models independently, once on the UD Kurmanji original and once on the UD Kurmanji augmented. We take this approach because we want to assess the impact of the annotation scheme on the models' performance. Hence, the labels (augmented) and (original) within the models' names indicate the dataset used for training the model, either UD Kurmanji augmented or UD Kurmanji original.

5.4.1. Statistical-based Models

Our first model is a Unigram model from the NLTK Python package (Bird et al., 2009). This model assigns tags based on word frequency observed during training. It uses conditional frequency distributions to calculate the most likely tag for each given token. The model may encounter unfamiliar words in linguistically resource-limited settings like ours (*out-of-vocabulary*). Therefore, we specify the default POS tag as NOUN when it fails to determine a POS tag for a token. This is a common practice when establishing a baseline, and it is motivated by Bird et al. (2009).

In addition, we create HMM (Huang et al., 2001) and CRF (based on CRFsuite library (Okazaki, 2007)) models using the implementation available in the NLTK Python package.
Finally, we create an ExtraTrees POS model using the implementation from Scikit-learn (Pedregosa et al., 2011).

5.4.2. Neural-based Models

Our first neural-based model is the Averaged Perceptron POS tagging model, similar to the Extra Trees model, which has the notion of feature engineering. However, here we do not define our own set of features, we use the standard features set defined by the NLTK Python package since we use their implementation[4].

In addition, we use the Flair Python package (Akbik et al., 2019) to create a BiLSTM model using a configurable BiLSTM architecture as originally proposed by Huang et al. (2015). For this model, we use pre-trained sub-word fastText embeddings (Grave et al., 2018) specifically pre-trained on Northern Kurdish data. FastText enables us to generate embeddings from character-level n-grams, thereby being better at capturing morphological nuances.

5.4.3. Transformer-based Fine-tuned Models

In contrast to the previous models, where each model was trained from scratch for our task, we fine-tune the pretrained multilingual XLM-RoBERTa model (Conneau et al., 2020) on the UD Kurmanji original and UD Kurmanji augmented. We utilize the 'base' version of XLM-RoBERTa because of its lower computational requirements, making it easier to fine-tune. The fine-tuning is performed using Trankit (Nguyen et al., 2021), which offers a relatively fast and straightforward approach for fine-tuning LLMs like XLM-RoBERTa, thanks to the utilization of Adapters (Pfeiffer et al., 2020). We refer to the fine-tuned POS model as **N**orthern **K**urdish **XLM**-**R**oBERTa **(NK-XLMR)**.

6. Experiments

6.1. Tokenization Performance

We distinguish between two types of tokenization evaluation: 1) intrinsic evaluation and 2) extrinsic evaluation.
Within the intrinsic evaluation, we want to evaluate the quality of the tokenization system in isolation from the later stages, POS tagging in our case. The intrinsic evaluation directly measures the tokenization system's capabilities by comparing it to similar systems. We follow the same approach of (Ahmadi, 2020b) by performing tokenization evaluation using the Bilingual Evaluation Understudy Score (BLEU).

Table 2 shows the BLEU scores of the tokenization methods we used in this study using the gold-standard dataset as testing data. We see that the BLEU scores for the KLPT tokenizer are the highest, outperforming other tokenizers by a great margin. In contrast to other tokenizers, the KLPT tokenizer is characterized by its extensive knowledge of Northern Kurdish, enabling it to correctly recognize case markers and handle multi-word expressions like compound verbs and compound prepositions.

Tokenizer	BLEU-1	BLEU-2	BLEU-3	BLEU-4
KLPT	0.73	0.65	0.59	0.53
unigram	0.54	0.44	0.36	0.29
NLTK	0.50	0.41	0.33	0.25
BPE	0.50	0.39	0.31	0.24
wordPiece	0.45	0.36	0.28	0.21

Table 2: BLEU scores for all tokenization methods on the gold-standard dataset.

Within the extrinsic evaluation, we evaluate the tokenization system by measuring its impact on our whole NLP pipeline. In our case, the tokenization system's quality greatly affects the POS tagger's performance. Therefore, the tokenization correctness can also be determined by examining the F1 and accuracy scores of the POS tagger presented in section 6.2.

6.2. POS Tagging Performance

We present the evaluation results (accuracy and macro-averaged F1 score) of all POS tagging models. In order to make the comparison clearer, we divide the results based on the used training data (UD Kurmanji original and augmented). While table 4 provides a detailed comparison of all models trained on the UD Kurmanji augmented, table 3 demonstrates the results of the same POS model but trained on UD Kurmanji original.

By comparing the results in both tables and regardless of the tokenization method, we observe a performance increase among the models. This increase is the highest within the manual tokenization method and the lowest within the wordPiece tokenization method. This confirms the importance and the impact of the data augmentation we did on the UD Kurmanji original treebank for the task of POS tagging. In addition, it stipulates the impact the performance of the tokenization method has on

[4]This implementation is based on Matthew Honnibal's implementation: https://explosion.ai/blog/part-of-speech-pos-tagger-in-python

Model / Tokenizer	manual		KLPT		NLTK		unigram		BPE		wordPiece	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Baseline (Unigram)	0.4	0.51	0.35	0.41	0.37	0.32	0.36	0.34	0.36	0.32	0.34	0.31
HMM	0.37	0.46	0.33	0.36	0.35	0.32	0.34	0.33	0.34	0.32	0.34	0.31
ExtraTrees	0.41	0.52	0.37	0.42	0.38	0.33	0.37	0.34	0.38	0.33	0.34	0.32
AveragedPerceptron	0.44	0.54	0.37	0.42	0.40	0.36	0.38	0.37	0.39	0.35	0.36	0.33
BiLSTM	0.42	0.51	0.40	0.41	0.45	0.35	0.43	0.36	0.44	0.34	**0.42**	0.33
CRF	0.46	0.54	0.41	0.44	0.42	0.36	0.40	0.37	0.40	0.35	0.35	0.33
NK-XLMR	**0.57**	**0.62**	**0.46**	**0.47**	**0.47**	**0.38**	**0.45**	**0.39**	**0.45**	**0.37**	0.40	**0.35**

Table 3: The macro-averaged F1 scores and accuracy (Acc) of the POS tagging models trained on the UD Kurmanji original and evaluated on our gold-standard dataset.

Model / Tokenizer	manual		KLPT		unigram		NLTK		BPE		wordPiece	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Baseline (Unigram)	0.59	0.73	0.47	0.52	0.41	0.37	0.4	0.32	0.4	0.33	0.37	0.33
HMM	0.62	0.77	0.48	0.53	0.4	0.37	0.41	0.33	0.4	0.34	0.38	0.33
ExtraTrees	0.61	0.79	0.49	0.56	0.43	0.4	0.41	0.36	0.41	0.36	0.37	0.34
AveragedPerceptron	0.68	0.83	**0.57**	0.57	0.47	0.41	0.49	0.37	0.45	0.37	0.40	0.35
BiLSTM	0.72	0.83	0.52	0.57	0.45	0.40	0.43	0.36	0.43	0.37	0.41	0.34
CRF	0.74	0.84	0.55	0.59	0.48	0.42	0.48	0.39	0.46	0.38	0.42	0.35
NK-XLMR	**0.77**	**0.87**	0.56	**0.59**	**0.49**	**0.43**	**0.51**	**0.39**	**0.47**	**0.39**	**0.44**	**0.36**

Table 4: The macro-averaged F1 scores and accuracy (Acc) of the POS tagging models, trained on the UD Kurmanji augmented and evaluated on our gold-standard dataset.

POS tagging for Northern Kurdish. While this performance increase is in part due to the different annotation scheme, which is explained in section 4.2, the introduction of this richer scheme improved the performance of the POS models on specific POS tags other than `IZAFE` and `DET`. A detailed analysis of this improvement is reported in section 6.3.

Further observation reveals that within the context of the training on UD Kurmanji augmented, both the BiLSTM and AveragedPerceptron models exhibit identical accuracy scores, although their macro-averaged F1 scores diverge slightly but remain comparable. Conversely, when utilizing the UD Kurmanji original, a similar trend of identical accuracy emerges between the AveragedPerceptron and the CRF models. Additionally, it is notable that the HMM model falls behind, even when compared to the baseline.

Moreover, the NK-XLMR model is our best model as it outperforms all other models. This was an expected performance, and it is in line with our finding in section 3 where we showed how LLMs achieve state-of-the-art results for multiple NLP tasks, including POS tagging.

However, comparing the scores of NK-XLMR and CRF models in Table 4, we observe very close performance between the two. The difference is very small, 0.03 for the macro-averaged F1 and the accuracy scores. This is a notable result, especially with regard to the computational resources required for fine-tuning XLM-RoBERTa and for training the CRF model from scratch for the task of POS tagging. Based on our experiments in this study, fine-tuning XLM-RoBERTa for POS tagging took notably longer than training the CRF for the same task.

6.3. Analysis

The presented results in the Tables 4 and 3 unambiguously demonstrate two trends in our results. First, training the POS models on the UD Kurmanji augmented undeniably results in higher accuracy and F1 scores when compared with the outcomes of POS models trained on the UD Kurmanji original. Second, the performance of POS models tends to decline as we transition away from the manual tokenization method. The further we move, the less knowledge of the linguistic characteristics of Northern Kurdish the tokenizers have. To further analyze this, we present two confusion matrices in Figures 3a and 3b demonstrating the performance of the NK-XLMR(augmented) and NK-XLMR(original).

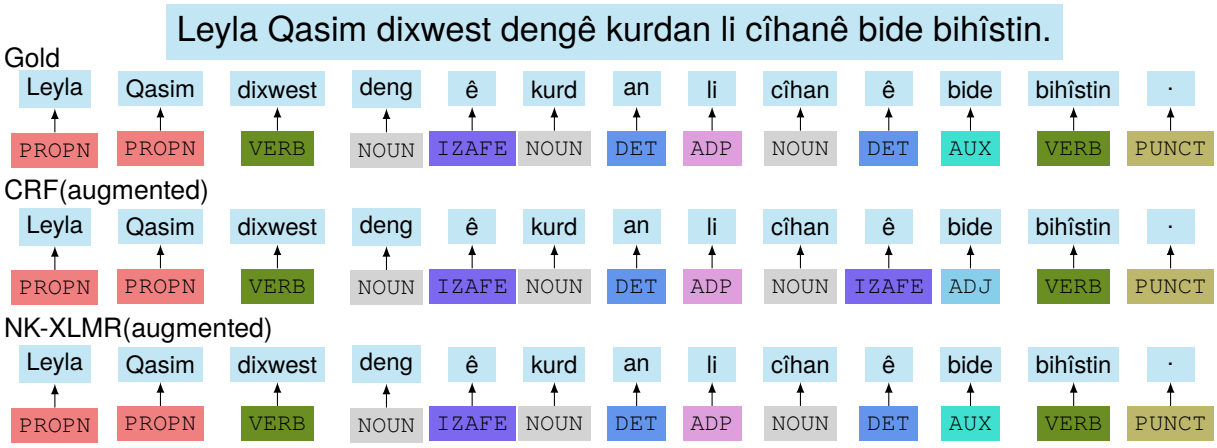
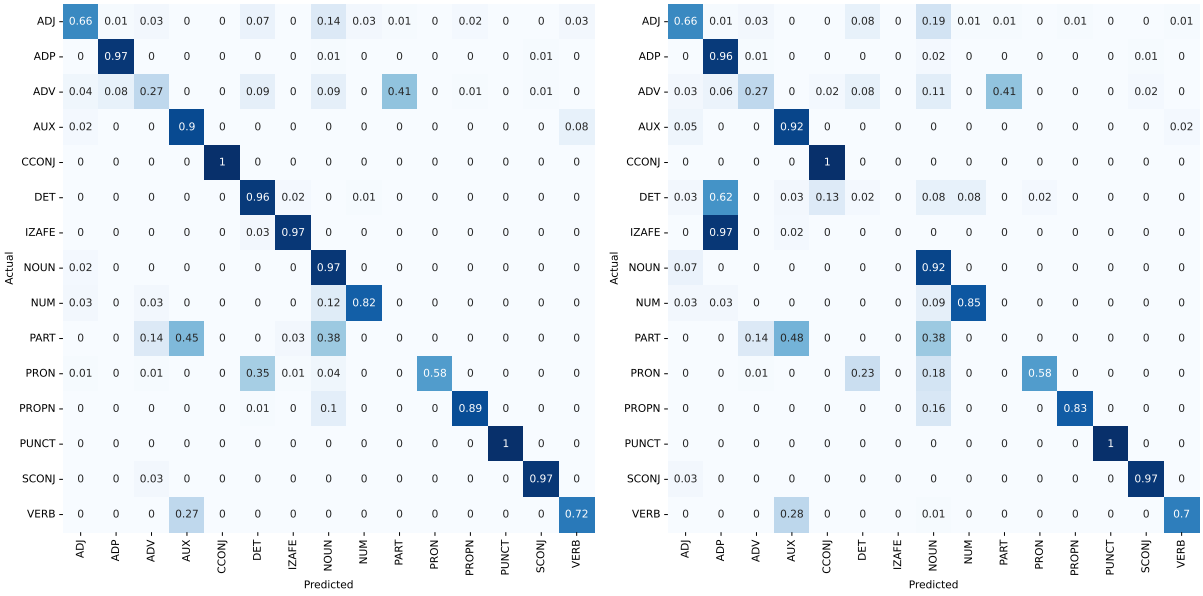


Figure 2: Outputs of the CRF and NK-XLMR compared to the gold annotations for a sentence from the gold-standard dataset (Translation: 'Leyla Qasim wanted to make the Kurdish voice heard in the world.')



(a) NK-XLMR (augmented)

(b) NK-XLMR (original)

Figure 3: Confusion matrices of NK-XLMR (augmented) and NK-XLMR (original) models. Although both models exhibit inadequacy in handling the PART and ADV tags, NOUN and PROPN benefit from data augmentation.

The UD Kurmanji augmented is characterized by the enhancements we have introduced and discussed in detail in section 5.2. The data augmentation affected tokens from the following POS tags: NOUN, PROPN, DET, and ADP, which are important elements in the *Izafe* and oblique cases in Northern Kurdish.

By comparing the confusion matrices, we observe that NOUN and PROPN benefit the most from the data augmentation, demonstrating 0.05 and 0.06 accuracy improvement, respectively, and the ADP and VERB to a lesser extent. In addition, we see that the tags DET and IZAFE enjoy huge improvement when trained on the UD Kurmanji augmented. However, we cannot consider it reliable since the IZAFE tag was not present in the UD Kurmanji original.

Nevertheless, it is evident that the NK-XLMR (original and augmented) exhibits a notable inadequacy in handling the PART and ADV tags. Examined outputs of NK-XLMR(augmented) and the error rates presented in section 6.3 and section 6.3 also verify this inadequacy. The tag PART has an error rate of 1.0, which means the model completely fails in recognizing tokens belonging to this tag correctly. We argue that this can be attributed to a misalignment in the annotation schemes between the UD Kurmanji and ours rather than a lim-

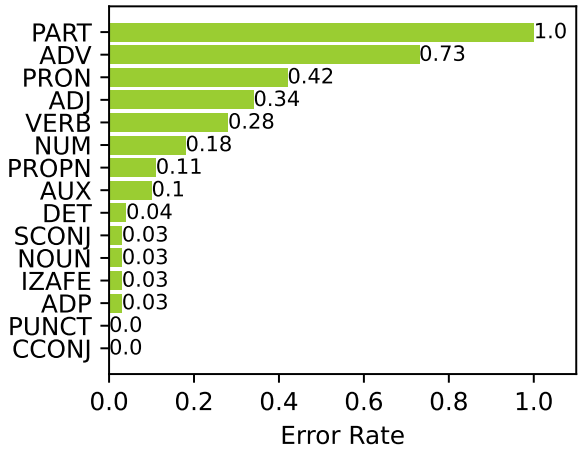


Figure 4: Error rates of NK-XLMR(augmented)

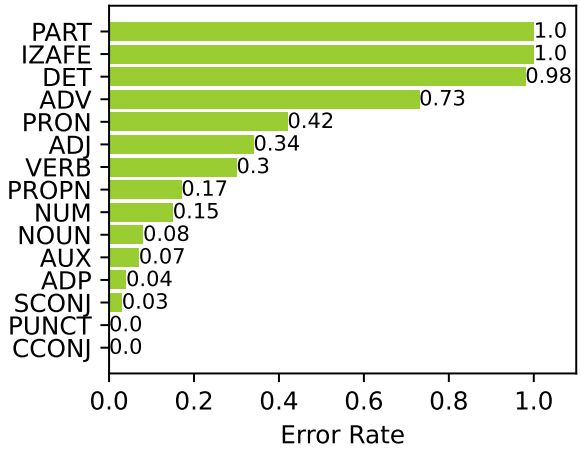


Figure 5: Error rates of NK-XLMR(original).

itation within the model itself.

Additionally, in section 6.3, we see that IZAFE and DET also have error rates of 1.0 and 0.98. This happens due to the fact that NK-XLMR(original) has no knowledge of the *Izafe* and oblique case markers and, therefore, fails to perform POS tagging correctly when evaluated on the gold-standard dataset where those markers are explicitly represented.

Regarding the second trend, the most straightforward reason for this is the fact that the tokenization methods are generating, in most cases, either fewer or more tokens than the ground truth. This can be attributed to the linguistic knowledge the tokenizer has about Northern Kurdish, such as the *Izafe*, oblique case markers, and multi-word expressions.

7. Conclusions and Discussion

The main objective of this study was to address the task of POS tagging for Northern Kurdish by utilizing the currently available resources. On the one hand, our multifaceted approach for this study enabled us to establish a baseline POS tagger for Northern Kurdish using the Unigram(augmented) model with an accuracy of 0.73 and a macro-averaged F1 score of 0.59 evaluated on the gold-standard dataset. On the other hand, the CRF(augmented) model achieves the second-best performance with 0.84 and 0.74 for accuracy and macro-averaged F1 score, making it the best-performing model among statistical POS tagging models. In addition, the CRF model stands out because of its quick training time.

The transformer-based NK-XLMR (augmented) outperforms all other models with an accuracy of 0.87 and a macro-averaged F1 score of 0.77., thus setting a new state-of-the-art performance for the task of POS tagging in Northern Kurdish. Our results are particularly robust compared to the work of Walther et al., 2010, where their POS tagger for Northern Kurdish was evaluated on only 13 sentences. This comparison underscores the reliability of our findings, considering the granularity of linguistic features in our gold-standard dataset and the larger number of test sentences (136 sentences) we used for evaluation.

Moreover, we further explored the impact of tokenization methods on POS tagging accuracy by comparing their outcomes against the gold standard tokens in our dataset. While encountering difficulties with certain linguistic nuances, the KLPT tokenizer demonstrated notable proficiency in capturing Northern Kurdish linguistic traits.

Finally, we successfully demonstrated the effect of the various linguistic features of Northern Kurdish, such as the *Izafe* and oblique case markers and contracted prepositions on the task by evaluating both variants of the models (original and augmented). Our POS tagging models trained on the UD Kurmanji augmented showed improvements on NOUN, PROPN, VERB, and ADP POS tags.

Limitations While this study has made several contributions to the field of Kurdish NLP, several limitations should be noted. Firstly, we did not target the task of syntactic parsing. Secondly, we did not explore the employment of LLMs or POS models from other closely related languages like Persian or dialects like Central Kurdish. Furthermore, we did not examine the impact of our POS tagging models and annotation schemes on other downstream tasks like named entity recognition, sentiment analysis , or parsing.

References

Sina Ahmadi. 2020a. KLPT–Kurdish language processing toolkit. In *Proceedings of second work-*

shop for NLP open source software (NLP-OSS), pages 72–84.

Sina Ahmadi. 2020b. A tokenization system for the Kurdish language. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 114–127.

Sina Ahmadi, Hossein Hassani, and John P McCrae. 2019. Towards electronic lexicography for the kurdish language. In *Proceedings of the sixth biennial conference on electronic lexicography (eLex)*. eLex 2019.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Eric Brill. 1992. A simple rule-based part of speech tagger. Technical report, Pennsylvania Univ Philadelphia Dept of Computer and Information Science.

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

Kyumars Sheykh Esmaili, Donya Eliassi, Shahin Salavati, Purya Aliabadi, Asrin Mohammadi, Somayeh Yosefi, and Shownem Hakimi. 2013. Building a test collection for Sorani Kurdish. In *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, pages 1–7. IEEE.

Djamé Farah, Seddah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Ortiz Suarez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 1139–1150.

Édouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Memduh Gökırmak and Francis M. Tyers. 2017. A dependency treebank for Kurmanji Kurdish. In *Proceedings of the Fourth International Conference on Dependency Linguistics (DepLing, 2017)*, pages 64–73.

Zar Zar Hlaing, Ye Kyaw Thu, Thepchai Supnithi, and Ponrudee Netisopakul. 2022. Improving neural machine translation with POS-tag features for low-resource language pairs. *Heliyon*, 8(8).

Tobias Horsmann and Torsten Zesch. 2017. Do LSTMs really work so well for PoS tagging? – a replication study. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 727–736, Copenhagen, Denmark. Association for Computational Linguistics.

Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, and Raj Reddy. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st edition. Prentice Hall PTR, USA.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.

Suvarna G Kanakaraddi and Suvarna S Nandyal. 2018. Survey on parts of speech tagger techniques. In *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*, pages 1–6. IEEE.

Fred Karlsson. 1990. Constraint grammar as a framework for parsing running text. In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the*

56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 66–75.

Hailiang Li, YC Adele, Yang Liu, Du Tang, Zhibin Lei, and Wenye Li. 2019. An augmented transformer architecture for natural language generation tasks. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 1–7. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. *arXiv preprint arXiv:1708.06075*.

Liwen Ma and Weifeng Liu. 2021. An enhanced method for entity trigger named entity recognition based on pos tag embedding. In *2021 IEEE 7th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, pages 89–93.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Maihemuti Maimaiti, Aishan Wumaier, Kahaerjiang Abiderexiti, and Tuergen Yibulayin. 2017. Bidirectional long short-term memory network with a conditional random field layer for Uyghur part-of-speech tagging. *Information*, 8(4):157.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, Zegao Pan, and Maosong Sun. 2021. Improving data augmentation for low-resource NMT guided by POS-tagging and paraphrase embedding. *Transactions on Asian and Low-Resource Language Information Processing*, 20(6):1–21.

Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *International conference on intelligent text processing and computational linguistics*, pages 171–189. Springer.

Ruslan Mitkov. 2022. *The Oxford handbook of computational linguistics*. Oxford University Press.

Ryo Nagata, Tomoya Mizumoto, Yuta Kikuchi, Yoshifumi Kawasaki, and Kotaro Funakoshi. 2018. A POS tagging model adapted to learner English. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 39–48.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs). *URL http://www. chokkan. org/software/crfsuite*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 2089–2096. European Language Resources Association (ELRA).

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging

with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418.

Ashish Pradhan and Archit Yajnik. 2023. Parts-of-speech tagging of Nepali texts with Bidirectional LSTM, Conditional Random Fields and HMM. pages 1–17. Springer.

Tomas Pranckevičius and Virginijus Marcinkevičius. 2016. Application of logistic regression with part-of-the-speech tagging for multi-class text classification. In *2016 IEEE 4th workshop on advances in information, electronic and electrical engineering (AIEEE)*, pages 1–5. IEEE.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 777–784.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Arij Seddah, Djamé Riabi, and Mahamdi Menel. 2023. Enriching the NArabizi treebank: A multifaceted approach to supporting an underresourced language. In *The 17th Linguistic Annotation Workshop (LAW-XVII)@ ACL 2023*, page 266.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics (ACL).

Yan Shao, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2017. Character-based Joint Segmentation and POS Tagging for Chinese using Bidirectional RNN-CRF. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 173–183.

Miikka Silfverberg, Teemu Ruokolainen, Krister Lindén, and Mikko Kurimo. 2014. Part-of-speech tagging using conditional random fields: Exploiting sub-label dependencies for improved accuracy. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 259–264.

Karl Stratos, Michael Collins, and Daniel Hsu. 2016. Unsupervised part-of-speech tagging with anchor hidden markov models. *Transactions of the Association for Computational Linguistics*, 4:245–257.

Wheeler M Thackston. 2006. *Kurmanji Kurdish:A Reference Grammar with Selected Readings*. Harvard University.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Géraldine Walther, Benoît Sagot, and Karën Fort. 2010. Fast development of basic NLP tools: Towards a lexicon and a POS tagger for Kurmanji Kurdish. In *International conference on lexis and grammar*.

Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. 2015. Part-of-speech tagging with bidirectional long short-term memory recurrent neural network. *arXiv preprint arXiv:1510.06168*.

Jabar Yousif. 2019. Hidden Markov Model tagger for applications based Arabic text: A review. *Journal of Computation and Applied Sciences IJOCAAS*, 7(1).