A Practical Tool to Help Automate Interlinear Glossing: a Study on Mukrī Kurdish

Hiwa Asadpour*[1], Shu Okabe*[2,3], Alexander Fraser[2,3,4]
[1]Goethe University Frankfurt
[2]School of Computation, Information and Technology, Technische Universität München (TUM)
[3]Munich Center for Machine Learning
[4]Munich Data Science Institute
* Equal contribution
Correspondence: asadpour@lingua.uni-frankfurt.de, shu.okabe@tum.de

Abstract

Interlinear gloss generation aims to predict linguistic annotations (gloss) for a sentence in a language that is usually under ongoing documentation. Such output is a first draft for the linguist to work with and should reduce the manual workload. This article studies a simple glossing pipeline based on a Conditional Random Field and applies it to a small fieldwork corpus in Mukrī Kurdish, a variety of Central Kurdish. We mainly focus on making the tool as accessible as possible for field linguists, so it can run on standard computers without the need for GPUs. Our pipeline predicts common grammatical patterns robustly and, more generally, frequent combinations of morphemes and glosses. Although more advanced neural models do reach better results, our feature-based system still manages to be competitive and to provide interpretability. To foster further collaboration between field linguistics and NLP, we also provide some recommendations regarding documentation endeavours and release our pipeline code alongside.

1 Introduction

Language documentation aims to create and archive corpora alongside resources on a language usually classified as endangered. To do so, linguists carry out fieldwork and then process the collected data. Each annotation (e.g., transcribing the recordings, analysing the transcription) is mostly done manually; it is hence costly in terms of time and requires advanced linguistic knowledge. This is the 'transcription bottleneck' (Brinckmann, 2008), which underlines the gap between the amount of unannotated recordings and the fully annotated sentences. In this article, we focus on one of the central linguistic annotations, interlinear glosses, and aim to predict them automatically, to create a draft for the linguists to post-edit. It has been previously shown that such automation can actually help linguists both in terms of time and annotation quality (Baldridge and Palmer, 2009; Palmer et al., 2009).

1	Source	de	tirsī	kābrāy
2	Segmented	de	tirs=ī	kābrā–ī
3	Gloss	in	fear=EZ	fellow–OBL
4	Translation	out of the fear of the man		

Figure 1: Sentence annotated in the IGT format.

Figure 1 shows an example of an annotated sentence in the Interlinear Glossed Text format (IGT). The source sentence (1) is segmented into morphemes (2), the smallest meaningful units in the language. Each morpheme has a corresponding linguistic annotation, the gloss (3). We observe mainly two categories: grammatical glosses indicate the role of the morpheme (e.g., 'OBL' for oblique), while lexical glosses express its meaning (e.g., '*tirs*' for fear in English). Finally, the sentence is translated (4) in a meta-language used for the documentation (e.g., in English here).

Several languages and corpora have already been studied by the Natural Language Processing (NLP) community for the gloss generation task, for instance, during the SIGMORPHON Shared Task on interlinear glossing (Ginn et al., 2023). We focus, however, on the usability of an automatic glossing model in a real-life setting of an annotation workflow. This means that we take into account actual technical constraints that hinder the use of the most up-to-date NLP models.

To do so, we base our work on a corpus from one of the authors' fieldwork data (Asadpour, 2021) to enable linguistic analysis of the glossing. The studied language is Mukrī Kurdish, a variety of Central Kurdish, whose morphological complexity can be challenging. As a Kurdish language, it has a rich agglutinative system characterised by *ezafe* (linking) constructions, polypersonal agreement, and a variety of affixed, cliticised, and reduplicated

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 65–75

morphemes.

We present a simple pipeline using a feature-based model to label each source morpheme in Mukrī Kurdish with a gloss. Our work mostly focused on how to make such an NLP model more accessible for field linguists and closer to their workflow. Our model is indeed achieving performance around a few accuracy points behind state-of-the-art models, while it only requires stable Python dependencies with minimal computational resources (CPU of a standard computer). The pipeline can also output annotations in a format compatible with commonly used linguistic fieldwork tools.

Our contributions are as follows: (i) we release a feature-based minimal system for automatic glossing[1], (ii) we apply it to a manually annotated text from a real fieldwork corpus of one of the authors, and (iii) analyse the linguistic relevance of the predictions and learnt patterns.

Section 2 describes Mukrī Kurdish and the glossed corpus we studied. We explain our CRF pipeline methodology in Section 3. We present its performance and analyse the linguistic patterns learnt by the model in Section 4. We also point out a few recommendations for both field linguists and NLP practitioners in Section 5.

2 Language and fieldwork corpus

2.1 The language: Mukrī Kurdish

Mukrī Kurdish (also spelt Mukrīyānī) is primarily spoken in the northwestern region of Iran, specifically in middle and southern parts of West Azerbaijan and northern parts of Kurdistan provinces. The geographical area traditionally associated with Mukrī Kurdish is centred around the city of Mahābād (historically known as Sāblāx or Sāwjbłāx) and extends to surrounding cities, towns and villages, including Bokān, Pīrānšār, Sardašt, Šino and Naxada (Asadpour, 2021). This region, historically known as Mukrīyān, forms part of the larger Iranian Kurdistan area that borders Iraqi Kurdistan to the west.

Mukrī Kurdish belongs to the Central Kurdish (Sorānī) dialect group within the Indo-European language family. It is closely related to other Central Kurdish varieties spoken in both Iran and Iraq. However, it maintains distinctive features that set it apart from standard Sorānī as spoken in Silēmānīya or Hawlēr (Erbil) in Iraqi Kurdistan (Haig and Matras, 2002; Asadpour, 2021, 2022).

Among Central Kurdish varieties, Mukrī Kurdish has several distinctive characteristics. On the phonological aspect, certain vowel and consonant realisations differentiate it from standard Sorānī varieties, including retention of some archaic phonological features. On the lexical side, its unique vocabulary is influenced by its geographic position between different Kurdish dialect areas and contact with Jewish and Christian Neo-Aramaic, Armenian, and Azerbaijani Turkish communities (Asadpour, 2021).

Moreover, Mukrī Kurdish has a rich morphological structure with prefixes, suffixes, and enclitics. Correct morphological labelling requires an awareness of the surrounding context, such as in the example below:

Source	ne–	bird	–ī	=ewe
Gloss	NEG.PST–	take.PST	–2SG	=ASP
Translation	you did not take			

with a negation prefix *ne–*, a past verb stem *bird*, a person suffix *–ī*, and the aspectual enclitic *=ewe*. Verbal morphology, in particular, requires both left and right contexts for correct segmentation and interpretation. We note here that certain morphological markers are consistent and predictable both in form and position. For instance, verbs begin with mood/aspect prefixes (e.g., negation in the example), end with person suffixes (e.g., *–ī* for 2SG), and aspectual enclitics may also appear in the final position (e.g., *=ewe*).

2.2 Corpus preparation

We use the corpus collected through fieldwork by one of the authors (2004–in progress) in the Mukrī variety of Central Kurdish (Sorānī). The corpus includes narrative, conversational and procedural texts, ensuring diversity in genre and register. Annotation was done manually following the IGT format and Leipzig Glossing Rules (Lehmann, 2004; Bickel et al., 2008). Besides, the segmentation annotation tier marks morpheme boundaries with hyphens, while clitics are separated by equal signs (cf. tier 2 in Figure 1).

We split the corpus into training and test datasets (80:20) for our experiments. We also convert the sentences into the format used for the SIGMORPHON Shared Task (Ginn et al., 2023), with one sentence annotation tier per line. This notably ensures compatibility with tools devised for the Shared Task.

[1]The pipeline is released alongside a demonstration at: https://github.com/shuokabe/crf_glossing.

Table 1 displays the size of the fieldwork corpus of Mukrī Kurdish in terms of number of sentences (N_{sent}), number of words and morphemes for both tokens (N_{token}) and types (N_{type}).

		word		morpheme	
	N_{sent}	N_{token}	N_{type}	N_{token}	N_{type}
train	211	1,233	570	2,126	354
test	52	272	184	500	153

Table 1: Fieldwork corpus statistics for Mukrī Kurdish.

3 Gloss generation system

3.1 Gloss generation pipeline

We tackle the gloss generation task as a morpheme labelling task. We assume that the sentence has been previously segmented into morphemes. Interlinear glosses can hence be viewed as labels assigned to each morpheme.

Source	de	tirs=ī	kābrā–ī
Step I	IND	stem=EZ	stem–OBL
Step II	IND	UNK=EZ	fellow–OBL
True gloss	in	fear=EZ	fellow–OBL

Figure 2: Example output at each step from the model.

Our model can be decomposed into two steps, as presented in Figure 2. First, grammatical labels are predicted for each morpheme (step I), with lexical morphemes initially labelled as 'stem' placeholders. Then, these placeholder labels are replaced with actual lexical glosses using a simple dictionary built from frequent associations in the training data (step II).[2] When available, actual bilingual dictionaries or known morpheme-to-gloss mappings can be integrated to augment the lexical coverage in this step. For unknown lexical morphemes, the second step outputs the 'UNK' tag. Figure 3 summarises the pipeline.

As previously considered by (McMillan-Major, 2020; Barriga Martínez et al., 2021), our system is based on a Conditional Random Field (CRF) (Lafferty et al., 2001), which relies on local properties (or features) to predict a label. We use the default parameters in our experiments.

We use generic features to keep it adaptable to other languages, such as the current morpheme, its

[2]The dictionary contains one-to-one associations only, i.e., one source lemma can only have one possible lexical label.

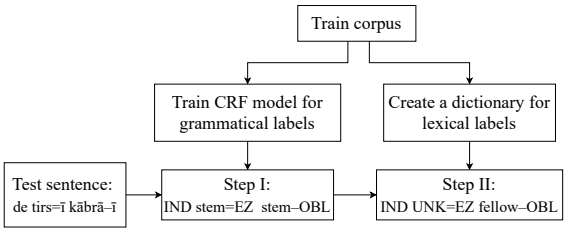

Figure 3: Glossing pipeline flowchart

immediate predecessors and successors, morpheme length, and boundary markers indicating whether the morpheme is separated by a hyphen (–) or an equal sign (=). An example list of features is presented in Appendix A.

3.2 Between simplicity and complexity

Technical requirements The main strength of our system is its simplicity, making it possible to run efficiently on CPUs rather than requiring GPUs. For instance, most participating submissions to the SIGMORPHON Shared Task (Ginn et al., 2023) used neural systems based on transformers (e.g., ByT5 (Xue et al., 2022)) or needed PyTorch to run (e.g., (Girrbach, 2023)'s winning system). In contrast, our approach pushes towards usability in real language documentation settings, where access to GPUs may be limited. This also means the model can run on common laptops within minutes, making it suitable for further integration into annotation workflows.

On the technical side, our CRF uses the sklearn-crfsuite library (Okazaki, 2007) in Python[3], which is widely used and stable. Besides this toolkit, our pipeline does not need any external packages.

Quality of the predictions However, this simplicity comes at a price. Compared to more advanced neural models, our pipeline shows lower overall performance, as shown in Section 4.1. It seems more adapted when the corpus is rather small, notably at the beginning of the annotation phase.

Furthermore, due to the pipeline approach, errors at step I impact the second step. In Figure 2, we see that the first morpheme is wrongly predicted with a grammatical tag (IND), although it should have been a 'stem' label for lexical glosses. Besides, even though our experiments show that lexical glosses can be relatively easily labelled with dictionaries in many cases due to the annotation

[3]https://sklearn-crfsuite.readthedocs.io/.

regularity in documentation corpora, this reliance means that unknown morphemes cannot be handled at all, as for the second morpheme ('tirs') which was never seen in the training corpus.

Interpretability and flexibility Another characteristic of our pipeline is that it allows for better interpretability, as developed in Section 4.4, given its feature-based nature. This can be helpful in understanding the patterns in the predictions and behaviour of the model, ensuring better transparency for the linguists compared to more black-box neural models. This follows previous analyses as in (Barriga Martínez et al., 2021; Okabe and Yvon, 2023a,b).

Besides, our current pipeline remains generic and only requires an annotated training dataset. When language-specific phenomena are known, the CRF can integrate them as additional features; when more annotations are made, the dictionary can be easily expanded for new words, and the CRF will be more robust.

In a nutshell, we chose to focus on a system with reduced technical complexity, trading performance for better accessibility, because we have real-life settings in mind. We recall that the purpose of automatic glossing is to *reduce* the proportion of manual workload by providing a first draft to start with for the linguist.

3.3 Workflow integration

More broadly than the glossing task, we strove to reduce the gap in the standard annotation workflow. For smoother integration, we created scripts to convert the predicted sentence annotations towards formats widely used in linguistic tools such as FieldWorks Language Explorer (FLEx) (Rogers, 2010), Toolbox[4], and ELAN (Wittenburg et al., 2006). This is to further reduce the friction of using yet another tool.

Below is how our feature-based pipeline can be put into practice in an existing framework for language documentation. Once the time-aligned audio recording is transcribed, with a consistent orthography, the sentences are segmented into words, but also into *morphemes*. The next step is to annotate a small batch of sentences with glosses; usually, the natural order of sentences is followed (e.g., each sentence of a recorded story), ensuring lexical consistency. Then comes the automatic glossing tool. Starting with as many training (i.e., fully annotated) sentences as possible, the model is applied to the rest of the corpus. The idea here is, naturally, to continue the annotation of sentences (possibly from the draft) and to compare the glosses. If specific linguistic phenomena are wrongly predicted systematically, dedicated features can be integrated into the CRF, or more sentences could be given. The latter solution also applies to lexical glosses since our approach depends on the coverage of the dictionary. Finally, the predictions are converted back to the format of the chosen annotation tool.

We note here that our approach does not solve the 'NLP gap' problem yet, as stated in (Gessler, 2022), since it runs separately and not concurrently from existing linguistic tools. It is, however, a step towards an actual integration in annotation software, where we reduced the technical constraints pertaining to the latest glossing models.

4 Experimental results

4.1 Comparison with the SIGMORPHON shared task on interlinear glossing

First, we compare our model with the most recent automatic glossing models to assess its quality in general. The SIGMORPHON Shared Task on interlinear glossing (Ginn et al., 2023) offered two tracks: the closed one only contained the source sentence with no segmentation information, while the open one notably had the morphological segmentation of the source sentence. The latter setting is closer to ours, where we have actual morphological boundaries of the source sentence.

The seven languages that were studied are diverse both geographically and linguistically (six language families). The released corpora are also of varying size, reflecting different stages of documentation (from 31 training sentences up to several thousand). We focus on six languages to test our model against the other submissions; we do not cover Arapaho, which had the largest corpus by far (40k sentences; 4 times the size of the second largest corpus).

We compare our model (`CRF+dict`) with a simple baseline, `dict`, which assigns the most frequent label seen in the training dataset. This replicates a dictionary-based functionality which is implemented in certain annotation tools (e.g., ELAN (Wittenburg et al., 2006) or FLEx (Rogers, 2010)).

Moreover, we present the results of the baseline model from the Shared Task (BASE_ST), based on a transformer architecture (Ginn, 2023), and the

[4] https://software.sil.org/toolbox/.

best performance reached during the Shared Task (mostly from (Girrbach, 2023); BEST_ST), usually a neural model. Additionally, we report the scores obtained by the state-of-the-art GlossLM model (Ginn et al., 2024), when it was fine-tuned on the corresponding language datasets (GlossLM$_{FT}$).

We evaluate the models according to the two main metrics used in the Shared Task: the accuracies computed at the word or morpheme levels.

	ddo	git	lez	ntu	nyb	usp
dict	65.3	28.1	81.2	81.5	64.4	72.8
CRF+dict	82.2	29.2	85.0	87.6	74.8	75.6
BASE_ST	75.7	16.4	34.5	41.1	84.3	76.6
BEST_ST	85.8	31.5	**85.4**	**89.3**	**88.0**	78.5
GlossLM$_{FT}$	**89.3**	**34.9**	71.3	81.5	87.7	**84.5**
Δ_{BEST}	-7.1	-5.7	-0.4	-1.7	-13.2	-8.9
dict	79.1	51.2	85.8	87.1	72.9	79.5
CRF+dict	89.2	51.8	**88.6**	91.7	82.0	82.0
BASE_ST	85.3	25.3	51.8	49.0	88.7	82.5
BEST_ST	92.0	**52.4**	87.6	**92.8**	**91.4**	84.5
GlossLM$_{FT}$	**92.8**	28.9	74.7	86.0	90.7	**86.4**
Δ_{BEST}	-3.6	-0.6	+1.0	-1.1	-9.4	-4.4

Table 2: Accuracy at the word (top rows) and morpheme (bottom) levels on the test set of the Shared Task. Best scores are in **bold**. Δ_{BEST} indicates the difference between our system and the best performance otherwise.

Table 2 presents the scores for both evaluation levels on the six corpora. First, we see that the glossing task can already be well achieved by a dictionary-based approach, as seen in the high accuracy reached by the dict baseline (except for Gitksan, git). This is due to the regularity in the annotations found in the dataset, especially for lexical morphemes and glosses. These units tend to be consistently annotated in the same way, which is one key assumption for our system.

Still, grammatical morphemes contain more variability with different glosses that could be attributed to the same unit; this is, hence, better captured by CRFs. We see a noticeable improvement of more than 16 points for Tsez (ddo), which is a morphologically rich language.

This approach is also consistently better than the Shared Task baseline (except for Uspanteko, usp). Moreover, despite its simplicity, it remains competitive with the best systems submitted to the shared task. Our model is only a few points behind the best models of the Shared Task, except in Nyangbo (nyb). Indeed, with an average accuracy of 72.4 for words and 80.9 for morphemes, the model would have been ranked fourth among eleven submissions. Compared to the current state-of-the-art GlossLM model, our system performs worse on their in-domain languages (i.e., which have more training data and, hence, were used for pre-training) but leads to notably better prediction on the languages with fewer data, such as Lezgi (lez) or Natugu (ntu).

More generally, we note that our CRF-based approach works better when the training data is smaller (git, lez, and ntu have below 800 training sentences), while more complex models naturally perform better with more sentences (Ginn et al., 2024). Hence, our system could help the early stages of documentation while alleviating some technical constraints.

4.2 Results on Mukrī Kurdish

Table 3 presents the accuracy scores on our Mukrī Kurdish corpus, computed using the same methodology as in the previous section. Given its size, we are in earlier documentation stages, where our system works relatively better (cf. Section 4.1).

	word	morpheme
dict	38.2	53.3
CRF+dict	50.7	64.1

Table 3: Accuracy on Mukrī Kurdish (top: word level, bottom: morpheme level).

We see for Mukrī Kurdish that the glossing performance with our model is still imperfect, actually in between the quality observed for Gitksan and Lezgi in Table 2. However, we notice a significant improvement over a pure dictionary-based approach, which is often used in linguistic annotation workflows.

Moreover, we additionally compare the usual precision, recall, and F-score separately for grammatical and lexical glosses. We notice that, as expected, the use of a CRF model improves the quality of grammatical label prediction (F-score of 48.3 to 66.3) due to their ambiguity. Our two-step pipeline also benefits the lexical tags thanks to a better separation of grammatical and lexical morphemes before replacement.

Among the 500 morphemes in the test set, 53 of them were tagged as UNK. This means that either the lexical morpheme was not seen in the training (in most cases), or the morpheme was wrongfully labelled with 'stem' instead of a grammatical tag.

Process Type	Error Rate (%)
Simple Affixation	9.60
Compounding	18.90
Cliticisation	14.20
Reduplication	37.10
Circumfixation	28.40
Infixation	33.80

Table 4: Error Rates by Morphological Process

Error rates (1 − accuracy) varied significantly across different morphological processes, as shown in Table 4. The lowest error rate was observed for simple affixation (9.60%), suggesting that the model effectively captures regular concatenative morphology even when trained with fairly few examples. For non-concatenative phenomena, however, performance deteriorated sharply, with reduplication showing the highest error rate at 37.10%, followed by infixation (33.80%) and circumfixation (28.40%). These results confirm that sequential models have particular difficulty with morphological operations involving copying, template relations, or internal alternation structures. They are harder to predict and thus require a more complex approach than a simple CRF modelling.

These findings are consistent with theoretical discussions in morphological typology (McCarthy, 1981; McCarthy and Prince, 1999), which distinguish between concatenative and non-concatenative morphology. Reduplication, in particular, involves correspondence constraints between base and copy elements, which are difficult to capture with the current surface-level statistical model alone.

4.3 Qualitative analysis

We discuss the linguistic peculiarities of Mukrī Kurdish and how they are reflected in the test data and predictions. Table 5 displays how ambiguous a given grammatical gloss is. We see that some highly systematic morphemes, such as *ne–*, *–eke*, and *=ewe*, appear consistently and should be easier to learn. In contrast, forms like *ī* have multiple roles (*ezafe*, 3SG, possessive, oblique), making them harder to disambiguate, and hence to predict.

As such, for canonical constructions, the pipeline showed strong performance, correctly identifying core and consistent morphemes such as *ezafe* markers, possessive suffixes, and common definite articles. For example, in the phrase *ser=ī yexdānē* (‘the door of the wardrobe’), the system accurately recognised ‘=ī’ as an *ezafe* or genitive marker linking the possessed noun (*ser*, ‘door [lit. head]’) to its possessor (*yexdānē*, ‘wardrobe’). This is consistent with the typical agglutinative structure found in many Iranian languages, where grammatical relations are expressed by postposed affixes (MacKenzie, 1961; Öpengin and Haig, 2014; Asadpour, 2022).

Error analysis We mainly noticed it struggles with under-represented (or absent) phenomena in the training corpus and ambiguous morphemes. For instance, discourse particles and switch-reference markers were poorly captured, especially in spoken narrative texts where such pragmatic features are prominent. The sentence in Figure 4 is a representative example.

S	[...] degeł	lē–de–de–ā	w
P	[...] with	at–IND–IND–3SG	PTCP
G	[...] with	PVB–IND–give.PRS–3SG	and

Figure 4: Example of wrong analysis. S: segmented source sentence, P: prediction from our system, G: gold glossed sentence.

In this case, the particle ‘*lēdedā*’ (‘lē–de–de–ā’) was wrongly analysed, and the conjunction ‘*w*’ was treated as a clitic rather than a full discourse element. As the gold standard shows, ‘*lēdedā*’ functions as a verb root combined with aspect markers around, while ‘*w*’ functions as a coordinating conjunction. This illustrates one of the limitations of a simple CRF-based model: longer dependencies are not well-captured. Since our features mainly look at the immediate neighbours of a morpheme, it still struggles with polymorphemic words, as in here.

A frequent error we saw concerned the treatment of agreement markers. For instance, the morpheme ‘ī’ was often assigned OBL1 instead of 3SG, which is likely due to overlapping surface forms.

Application to language documentation The results on the test data suggest that, despite the morphological complexity, many patterns in Mukrī Kurdish are consistent enough to be handled by automatic systems. Common and systematic affixes (especially for verbs and nouns) are good candidates for automatic glossing. However, ambiguous and pragmatic elements are likely to require manual review and correction. A tool that pre-annotates glosses based on these regularities can, however,

Label	Example	Description	Position	Consistency
IND (de–)	de–ke, de–łē, de–č–m	Indicative prefix	Verb-initial	High
IRR (bi–)	bi–hēn, bi–nūs, bi–ke	Irrealis prefix	Verb-initial	High
NEG (ne–)	ne–bird, ne–kew, ne–mā	Negation prefix	Verb-initial	High
PVB	heł–de–gir, lē–de–de, ber–de	Preverbal particles	Before verb	Medium
ASP (=ewe)	bird–ī=ewe, ke=ewe, dāte=ewe	Aspectual enclitic	Word-final	High
DEF (–eke)	kitēb–eke, čikołe–eke	Definite marker	Noun-final	High
PL (–ān)	kitēb–ān, žin–ān	Plural marker	After noun	High
OBL (–ī)	bird–ī, č–ī, goř–ī	Oblique case	Noun-final	Medium (ambig.)
EZ (=ī)	čend=ī, birā=ī	*Ezafe*	After noun	Medium (ambig.)
PRSNT	āhā, hā, hā	Presentative	Independent	Low
DISC	āhā, wiłāhī	Discourse markers	Variable	Low

Table 5: Consistency of 10 frequent grammatical labels in the test dataset.

notably reduce the burden on linguists by allowing them to correct rather than annotate from scratch.

This is in line with one of the author's feedback as a fieldworker. Using the CRF-trained model significantly reduced the time spent on routine glossing by pre-labelling frequent grammatical patterns and high-frequency morphemes with reasonable accuracy. This allowed him to focus more on irregular forms, novel constructions, and higher-level linguistic analysis.

4.4 Interpretation of the model

Since our system relies on a CRF, we can interpret the features and patterns that were learnt by the model. For instance, the left part of Table 6 displays the 10 most weighted local properties.

Feature		Transition	
source feature	gloss	gloss_1	$\rightarrow \text{gloss}_2$
morph: m	1SG	EZ	→ REFL
morph: ew	DEM	IND	→ –
morph: emin	1SG	INDF.PRO	→ INDF.PRO
morph: eto	2SG	PVB	→ –
morph: de	IND	=	→ 3SG
morph: t	2SG	VOC	→ RDP
morph: bi	IRR	IMP	→ DISC
morph: nā	NEG	–	→ OBL
morph: ēk	INDF	3SG	→ NEG3
morph: n	3PL	OBL1	→ POST1

Table 6: Left: top 10 features; right: top 10 label transitions learnt by the CRF.

We notice that key morphological patterns in Mukrī Kurdish were correctly identified. For instance, both the independent pronoun 'emin' and its bound form 'm' are associated with the first-person singular gloss. Other frequent and crucial grammatical morphemes are also learnt, such as the negation marker 'nā' or the indefinite suffix 'ēk'. Most of them are consistent annotations with little ambiguity and occur often. These associations are closely aligned with typological descriptions of Kurdish, where agglutination dominates, and each morpheme encodes a single grammatical meaning.

This supports usage-based theories of morphological acquisition (Bybee, 2010), which posit that speakers rely heavily on co-occurrence patterns to disambiguate morphological function. Our results also suggest that statistical models approximate native speakers' intuitions about morpheme function.

Similarly, the model learns label transitions; the most highly weighted ones are in the right part of Table 6. Some of the transitions highlight crucial morphosyntactic patterns. First, the transition from 'EZ' to 'REFL' captures a common construction in Mukrī Kurdish where reflexive pronouns often follow an *ezafe* marker. The model has also correctly identified some pronominal clitics appearing after a clitic boundary, as shown by the strong association between the clitic marker '=' and '3SG' (third-person singular). The transition from 'IND' (indicative) to '–' (morpheme boundary) reflects the morphological structure of Mukrī verbs, where the indicative prefix is typically followed by other verbal morphology (as in Figure 4). These patterns demonstrate that the model has actually captured central morphosyntactic regularities in Mukrī Kurdish, such as clitic placement or verbal morphology.

However, the model occasionally violated these constraints when exposed to less frequent patterns. This suggests that surface statistics, while informa-

tive, may not be sufficient to fully capture more complex morphosyntactic principles.

In short, while our pipeline performs reasonably well on regular morphological patterns represented in the training data, it struggles with rare constructions, phonologically conditioned allomorphy, and morphologically complex phenomena that require more global structural awareness. This is because the CRF relies on local statistical cues, which cannot handle rare, unseen or structurally divergent constructions. While the model does not explicitly learn abstract grammatical rules, it manages to infer recurrent associations between morphemes and their glosses based on distributional patterns present in the training data, which makes it effective for canonical morphology.

5 Recommendations for stakeholders

This article is the result of a collaboration between field linguistics and NLP; as such, we found a few recommendations for all parties involved or supporting language documentation, in line with (Flavelle and Lachler, 2023).

For field linguists, maintaining consistent segmentation and annotation conventions is essential for both humans and NLP models. On this point, following widely used conventions such as the Leipzig Glossing Rules (Lehmann, 2004; Bickel et al., 2008) can also help cross-lingual models, which might have seen the same grammatical glosses in other languages. In this regard, starting with a small but high-quality dataset is enough to start the first automatic gloss pipeline (e.g., the Gitksan corpus in the SIGMORPHON Shared Task had 31 sentences, and we have slightly more than 200 sentences).

For members of the language community, simplified interfaces and localised training materials can enable active participation in validation and annotation. Workshops to build consensus on terminology and validate results help to ensure cultural appropriateness and community ownership of digital resources.

For NLP researchers, the challenge is to improve the robustness of the model and to deal with more complex morphological phenomena while keeping in mind a real-life deployment of the glossing tool. Making the tools more user-friendly is also appreciated; specialised error analysis tools and visualisations would help to diagnose wrong predictions easily. Finally, a better evaluation protocol should be used to account for the error gravity; in the end, we aim at a system that helps rather than confuses the annotators.

6 Related Work

Interlinear gloss generation, in collaboration between linguistics and NLP for language documentation, has initially been explored with feature-based taggers. (Baldridge and Palmer, 2009) and (Palmer et al., 2009) both discuss the relevance and efficiency of active learning in such a context. They notably found that the benefit of better sampling techniques depends on the expertise of the annotators. (Samardžić et al., 2015) also applied a two-step pipeline with a tagger for grammatical glosses and a lexicon for the lexical glosses. Their experiments were, however, based on a much larger corpus for a better-documented language.

Moeller and Hulden (2018) show that CRFs are a reliable approach to predict *grammatical* glosses compared to a neural model for a corpus with 3,000 annotated words. Using the same methodology, Barriga Martínez et al. (2021) also find that CRFs outperform RNNs and biLSTMs on their corpus. Then, McMillan-Major (2020) proposes a pipeline combining two CRFs, one to predict from the source sentence and another one from the translation, an underexploited resource so far. All these methods are closely related to our methodology because CRFs are reliable in capturing local dependencies, especially in low-resource settings. However, due to the number of potential labels, lexical glosses cannot be predicted with CRFs alone.

This is one reason behind the consideration of neural models for glossing. Zhao et al. (2020) extend the methodology of (McMillan-Major, 2020) by considering both the source and translated sentences as inputs to a multi-source neural model (based on a transformer architecture; Vaswani et al., 2017).

Finally, a major milestone on the topic is the SIGMORPHON Shared Task on interlinear glossing (Ginn et al., 2023). Among the two possible tracks, the open one provided the morpheme-level segmentation of the source sentence. In this category, which is an easier task due to the additional information, the best performing model was (Girrbach, 2023), which trained a hard attention model. Two other submissions were also neural and based on transformers (Cross et al., 2023; He et al., 2023). Okabe and Yvon (2023b) have also compared their

feature-based systems against a simple CRF-based baseline; however, the former was not as accessible and convenient as our system, while the latter model was not released. The state-of-the-art for the task is currently achieved by the GlossLM model (Ginn et al., 2024), which also relies on the transformer architecture.

7 Conclusion

We have deployed an automatic glossing pipeline on a fieldwork corpus in Mukrī Kurdish, a Central Kurdish variety, to assess not only how it performs but also how usable such NLP tools are in practice. We have seen that our CRF-based system improved the prediction quality compared to the currently implemented full dictionary-based approach, which further reduces manual workload. It notably managed to learn the most frequent patterns while struggling with rarer phenomena and annotation, as expected. This is, however, not a major issue since any model output remains an annotation draft: they need to be corrected and controlled eventually. In our case, the system lowered the manual annotation effort noticeably, with a fairly robust reliability for repetitive annotations.

Even though our feature-based pipeline may not match the quality of state-of-the-art neural approaches (lagging by 3 points in accuracy on average in Table 2), it offers a more interpretable and adaptable alternative that is well-suited to early-stage documentation projects, such as for Mukrī Kurdish. We believe these characteristics outweigh the benefits of marginal gains obtained with more advanced models.

Finally, we are releasing the glossing pipeline under an open-source license to foster its use by both field linguists and NLP practitioners. We strove to provide a simple tool that can work with the usual infrastructure at hand in language documentation.

We stress again that this work, at the intersection of computational and documentary linguistics, aimed to bridge the gap between the vastly different technical environments of both fields. We also tried to lower the technical barrier by providing scripts to convert the annotations towards popular formats used in language documentation.

Our future work includes integrating the model into an actual annotation software so that it can be used even more easily. Moreover, we will also explore how performance can be improved by adapting known linguistic rules in the feature set, as some linguists already use rule-based processing to some extent.

Limitations

From the NLP perspective, the proposed model is not particularly novel, as similar models relying on CRFs were considered as a baseline for experiments. It does not reach a state-of-the-art performance either, given its simplicity. The model is, however, released not only to provide a fairly competitive yet simple baseline for future works in NLP but also to foster its use among field linguists. We believe, indeed, that the current pipeline can be integrated into actual annotation workflows, possibly after further simplifying user interaction with the model. Hence, our system choice is the result of a compromise between prediction quality and technical complexity.

From the linguistic side, some non-negligible errors remain in Mukrī Kurdish, which shows that the model cannot handle complex morphological patterns yet. For this article, we tried to release a model which could also be applied to other languages directly, i.e., without language-specific features. Thanks to the flexibility allowed by the features, the system can be better tailored to any language which will be studied.

Acknowledgments

We thank the anonymous reviewers for their comments. We are deeply grateful to all who have contributed their time and knowledge during Asadpour's fieldwork in the Mukrīyān region, which began in 2003 and has continued over the years. This research would not have been possible without their trust and generosity. Parts of the work related to the preparation and writing of this paper have received funding from the European Research Council (ERC) under grant agreement No. 101113091 – Data4ML, an ERC Proof of Concept Grant, supporting the contributions of Shu Okabe and Alexander Fraser. Asadpour's fieldwork and participation were conducted independently of this funding.

References

Hiwa Asadpour. 2021. Cross-dialectal diversity in Mukrī Kurdish I: Phonological and phonetic variation. *Journal of Linguistic Geography*, 9(1):1–12.

Hiwa Asadpour. 2022. *Typologizing word order variation in Northwestern Iran*. Ph.D. thesis, Goethe University Frankfurt, Frankfurt, Germany.

Jason Baldridge and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 296–305, Singapore. Association for Computational Linguistics.

Diego Barriga Martínez, Victor Mijangos, and Ximena Gutierrez-Vasques. 2021. Automatic interlinear glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43, Online. Association for Computational Linguistics.

Balthazar Bickel, Bernard. Comrie, and Martin Haspelmath. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Leipzig: Max Planck Institute for Evolutionary Anthropology, Department of Linguistics. https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

Caren Brinckmann. 2008. Transcription bottleneck of speech corpus exploitation.

Joan Bybee. 2010. *Language, Usage and Cognition*. Cambridge University Press.

Ziggy Cross, Michelle Yun, Ananya Apparaju, Jata MacCabe, Garrett Nicolai, and Miikka Silfverberg. 2023. Glossy bytes: Neural glossing using subword encoding. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 222–229, Toronto, Canada. Association for Computational Linguistics.

Darren Flavelle and Jordan Lachler. 2023. Strengthening relationships between indigenous communities, documentary linguists, and computational linguists in the era of NLP-assisted language revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 25–34, Dubrovnik, Croatia. Association for Computational Linguistics.

Luke Gessler. 2022. Closing the NLP gap: Documentary linguistics and NLP need a shared software infrastructure. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 119–126, Dublin, Ireland. Association for Computational Linguistics.

Michael Ginn. 2023. Sigmorphon 2023 shared task of interlinear glossing: Baseline model. *Preprint*, arXiv:2303.14234.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

Michael Ginn, Lindia Tjuatja, Taiqi He, Enora Rice, Graham Neubig, Alexis Palmer, and Lori Levin. 2024. GlossLM: A massively multilingual corpus and pretrained model for interlinear glossed text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12267–12286, Miami, Florida, USA. Association for Computational Linguistics.

Leander Girrbach. 2023. Tü-CL at SIGMORPHON 2023: Straight-through gradient estimation for hard attention. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 151–165, Toronto, Canada. Association for Computational Linguistics.

Geoffrey Haig and Yaron Matras. 2002. Kurdish linguistics: a brief overview. *STUF - Language Typology and Universals*, 55(1):3–14.

Taiqi He, Lindia Tjuatja, Nathaniel Robinson, Shinji Watanabe, David R. Mortensen, Graham Neubig, and Lori Levin. 2023. SigMoreFun submission to the SIGMORPHON shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 209–216, Toronto, Canada. Association for Computational Linguistics.

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

Christian Lehmann. 2004. Interlinear morphemic glossing. In *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung.*, volume 17 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, pages 1834–1857. Berlin & New York: W. de Gruyter.

David Neil MacKenzie. 1961. *Kurdish Dialect Studies*. Oxford University Press, London.

John J. McCarthy. 1981. A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry*, 12(3):373–418.

John J. McCarthy and Alan S. Prince. 1999. *Faithfulness and identity in Prosodic Morphology*, page 218–309. Cambridge University Press.

Angelina McMillan-Major. 2020. Automating gloss generation in interlinear glossed text. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 355–366, New York, New York. Association for Computational Linguistics.

Sarah Moeller and Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*,

pages 84–93, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shu Okabe and François Yvon. 2023a. LISN @ SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 202–208, Toronto, Canada. Association for Computational Linguistics.

Shu Okabe and François Yvon. 2023b. Towards multilingual interlinear morphological glossing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5958–5971, Singapore. Association for Computational Linguistics.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs).

Ergin Öpengin and Geoffrey Haig. 2014. Regional variation in kurmanji: A preliminary classification of dialects. *Kurdish Studies*, 2(2):143–176.

Alexis Palmer, Taesun Moon, and Jason Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 36–44, Boulder, Colorado. Association for Computational Linguistics.

Chris Rogers. 2010. Review of Fieldworks Language Explorer (FLEx) 3.0. In *Language Documentation & Conservation 4*, pages 78–84.

Tanja Samardžić, Robert Schikowski, and Sabine Stoll. 2015. Automatic interlinear glossing as two-level sequence classification. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 68–72, Beijing, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Xingyuan Zhao, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig, and Lori Levin. 2020. Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5397–5408, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A CRF features

Table 7 presents the features for the following sentence ('*out of the fear of the man*') at the fourth position[5] (first '**ī**'):

1	2	3	**4**	5	6	7
de	tirs	=	**ī**	kābrā	–	ī.

In general, we check the same local properties (source entity itself and its length) for the current (0), previous (-1), and next (+1) positions. Depending on the presence of a morpheme boundary, we also check the 'actual' previous morpheme (-2) to account for morpheme dependencies inside polymorphemic words.

position	feature	example for 'ī'
0	morpheme	ī
0	length	1
0	morpheme boundary?	False
-1	morpheme	=
-1	length	1
-1	morpheme boundary?	True
-2	morpheme	tirs
-2	length	4
+1	morpheme	kābrā
+1	length	5

Table 7: List of the computed features for a given entity. Position indicates the relative position compared to the entity (0: current position, -1: previous position, and +1: next position).

[5]We count source entities: both actual morphemes and morpheme boundaries.