Revisiting and Amending Central Kurdish Data on UniMorph 4.0

Sina Ahmadi
Department of Computer Science
George Mason University, Fairfax, USA
ahmadi.sina@outlook.com

Aso Mahmudi
Faculty of Engineering and IT
University of Melbourne, Australia
aso.mehmudi@gmail.com

Abstract

UniMorph–the Universal Morphology project is a collaborative initiative to create and maintain morphological data and organize numerous related tasks for various language processing communities. The morphological data is provided by linguists for over 160 languages in the latest version of UniMorph 4.0. This paper sheds light on the Central Kurdish data on UniMorph 4.0 by analyzing the existing data, its fallacies, and systematic morphological errors. It also presents an approach to creating more reliable morphological data by considering various specific phenomena in Central Kurdish that have not been addressed previously, such as Izafe and several enclitics.

1 Introduction

Computational morphology, the study of word formation using computational methods, is one of the important tasks in natural language processing (NLP) and computational linguistics. This field has been one of the prevailing and longstanding tasks with many applications in syntactic parsing, lemmatization and machine translation (Roark and Sproat, 2007). There have been remarkable advances and paradigm shifts in approaches to analyze and generate morphology: starting from ad-hoc approaches in the earlier systems, then rule formalisms and finite-state models since the 1980s (Karttunen and Beesley, 2005) with the notable example of KIMMO two-level morphological analyzer (Karttunen et al., 1983), followed by statistical and classical machine learning since the 1990s as in (Goldsmith, 2001; Schone and Jurafsky, 2001), and more recently, approaches relying on neural network models since 2000s. Lastly, more robust techniques are proposed using monolingual data hallucination (Anastasopoulos and Neubig, 2019), transfer learning (Kann et al., 2017) and pretrained models (Hofmann et al., 2020).

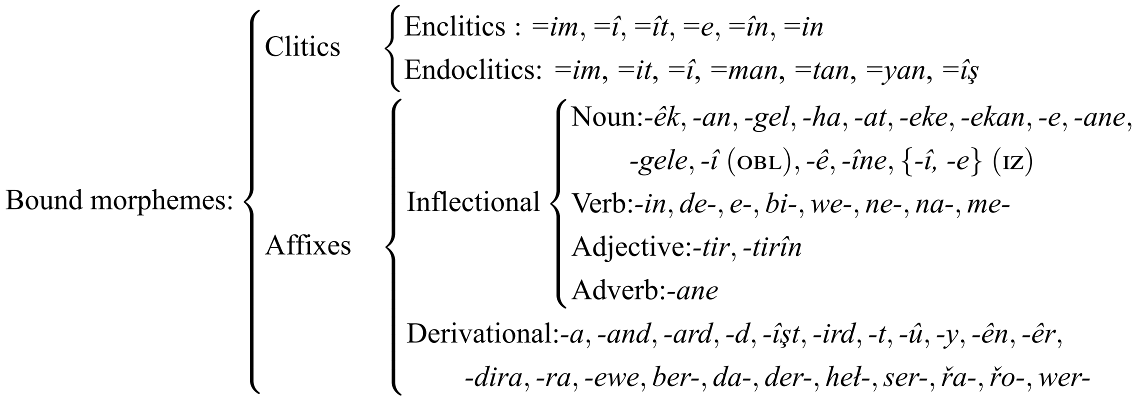
Unlike the progress in approaches, the dependence of systems on clean and reliable data, regardless of the size, for accurate morphological analysis and generation has not changed much. In order to bring together various linguistic communities to create datasets and incentivize further studies in the field, the UniMorph[1] (Batsuren et al., 2022) project has been a leading initiative in this vein. In UniMorph 4.0, the latest version of the project, there are 168 languages from various language families for which morphological data is provided according to the UniMorph schema (Sylak-Glassman, 2016). Additionally, the ACL Special Interest Group on Computational Morphology and Phonology (SIGMORPHON)[2] has played an important role to organize workshops and shared tasks using the UniMorph data. Some of the previous shared tasks focus on cross-linguality and context in morphology (McCarthy et al., 2019), unsupervised morphological paradigm clustering (Wiemerslage et al., 2021) and morphological inflection generation, segmentation, and interlinear glossing in this year's task.

One of the languages that is of interest in this paper and is also included in UniMorph is Central Kurdish, also known as Sorani (ckb). Central Kurdish, as a variant of the Indo-European language Kurdish, has a fusional morphology with several distinctive features due to its split-ergativity, erratic patterns in morphotactics and, several endoclitics used in verbal forms. These characteristics seem to be known to the UniMorph community, as described in Pimentel et al. (2021, p. 8). However, the current data available for Central Kurdish contains systematic errors and lacks coverage in morphological forms. The data is also provided in a script that is not used by Kurdish speakers, thus of no utility to downstream tasks in reality. Consequently, these result in poor performance of sys-

[1] https://UniMorph.github.io

[2] https://sigmorphon.github.io

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 38–48
July 14, 2023 ©2023 Association for Computational Linguistics



- Bound morphemes:
  - Clitics
    - Enclitics : *=im*, *=î*, *=ît*, *=e*, *=în*, *=in*
    - Endoclitics: *=im*, *=it*, *=î*, *=man*, *=tan*, *=yan*, *=îş*
  - Affixes
    - Inflectional
      - Noun:*-êk*, *-an*, *-gel*, *-ha*, *-at*, *-eke*, *-ekan*, *-e*, *-ane*, *-gele*, *-î* (OBL), *-ê*, *-îne*, {*-î, -e*} (IZ)
      - Verb:*-in*, *de-*, *e-*, *bi-*, *we-*, *ne-*, *na-*, *me-*
      - Adjective:*-tir*, *-tirîn*
      - Adverb:*-ane*
    - Derivational:*-a*, *-and*, *-ard*, *-d*, *-îşt*, *-ird*, *-t*, *-û*, *-y*, *-ên*, *-êr*, *-dira*, *-ra*, *-ewe*, *ber-*, *da-*, *der-*, *heł-*, *ser-*, *řa-*, *řo-*, *wer-*

Figure 1: A classification of Central Kurdish bound morphemes in the Latin-based script of Kurdish. Allomorphs and zero morphemes (∅) are not included.

tems that rely on the data in real scenarios.

Contributions This paper summarizes some of the salient features in Central Kurdish morphology. It also aims to discuss the main issues of Central Kurdish data on UniMorph 4.0. Moreover, the paper provides a new dataset of quality with considerable coverage and carries out experiments on the newly annotated data.

2 Central Kurdish Morphology

Kurdish is an Indo-European language spoken by over 25 million speakers in the Kurdish regions in Turkey, Iran, Iraq and Syria, and also by the Kurdish diaspora around the world (McCarus, 2007). Central Kurdish, also known as Sorani is the Kurdish variant that is mostly spoken by the Kurds within the Iranian and Iraqi regions of Kurdistan. Central Kurdish is a null-subject language and has a subject-object-verb (S-O-V) order and can be distinguished from other Indo-Iranian languages by its ergative-absolutive alignment which appears in past tenses of transitive verbs (Ahmadi and Masoud, 2020). In this section, we provide a brief description of Central Kurdish morphology by focusing on morphemes and morphological processes.

2.1 Bound Morphemes

Morphemes are classified into free and bound. While free morphemes are meaningful as they are, bound morphemes only carry meaning when affixed with other words. Bound morphemes are classified into two categories of affixes and clitics. Affixes and clitics are similar in the way that they cannot constitute a word and they lean on a prosodic host, i.e. a word for stress assignment. Clitics can appear with hosts of various syntactic categories while affixes only combine with syntactically-related stems (Haspelmath and Sims, 2013, p. 198). The clitics and affixes in Central Kurdish have been widely studied previously and have been shown to be challenging considering the general theory of clitics (W. Smith, 2014; Gharib and Pye, 2018). This problem is particularly observed with respect to the direct and oblique person markers which can appear in different positions within a word-form depending on the functionality. In this section, the clitics and affixes in Central Kurdish are described. Figure 1 provides the most frequent clitics and affixes in Central Kurdish.

2.1.1 Clitics

Clitics are categorized based on their position with respect to the host. A clitic is called proclitic and enclitic, if it appears before and after the host, respectively. There are two other forms of clitics which are non-peripherical and exist only among a few natural languages. If a clitic appears between the host and another affix, it is called a mesoclitic. A different type of non-peripheral clitic is endoclitic which appears within the host itself and is unique to a few languages around the world, such as Udi (W. Smith, 2014), Degema (Kari, 2002) and also Central Kurdish.

Central Kurdish has two types of endoclitics: pronominal makers, also introduced as mobile person markers by Walther (2012), and the emphasis endoclitic یش *=îş* which can be translated as ‘also’ or ‘too’ (Ahmadi et al., 2023). The pronominal endoclitics function as agent markers for transitive

0						girt						past stem of GIRTIN (to take, to get)
1						girt	im					I got.
2						girt	im	in				I got them.
3						girt	im	in	e			I got them to/with.
4						girt	im	in	e	ewe		I got them to/with again.
5						girt	îş	im	in	e	ewe	I got them also to/with again.
6			ne	îş	im	girt	in	e	ewe			I did not get them also to/with again.
7		ne	îş	im	de	girt	in	e	ewe			I was not getting them also to/with again.
8	da	îş	im	ne	de	girt	in	e	ewe			I was not taking down them also to/with again.

Table 1: The placement of the endoclitic *=îş* (in green boxes) and agent marker *=im* (in blue boxes) with respect to the base and each other in a verb form. Note that Central Kurdish is a null-subject language.

verbs in the past tenses or endoclitics as a patient marker for transitive verbs in the present tenses. This is due to the split ergativity feature of Central Kurdish where the agent and patient markers are specified differently. The following examples show the alignment in present and past tenses of کەوتن (KEWTIN, 'to fall') and گرتن (GIRTIN, 'to get'). The agent marker ن *-in* in intransitive present tenses (examples 1 & 3) serves as a patient marker in transitive past tenses (examples 2 & 4) due to ergativity while another morpheme یان *=yan* appears as the agent marker in the transitive verb (example 4).

(1) *dekew**in*** دەکەون
*de-kew-**in***
fall.PRS.PROG.INTR.3PL
'(they) are falling.'

(2) *degir**in*** دەگرن
*de-gir-**in***
get.PRS.PROG.TR.3PL
'(they) are getting.'

(3) *kewt**in*** کەوتن
*kewt-**in***
fall.PST.PROG.INTR.3PL
'(they) fell.'

(4) *girtyan**in*** گرتیانن
*girt=yan-**in***
get.PST.PROG.TR.ERG.3PL.3PL
'(they) got (them).'

Furthermore, the two endoclitic categories of Central Kurdish appear in an erratic pattern within a word form or a phrase. If a prefix appears before the stem of a transitive verb in the past tense, the agent marker postpends to the leftmost morpheme; in other cases, the agent marker appears after the verb stem following a varying morphotactic rule depending on the tense, mood, aspect and transitivity of the verb. Table 1 presents an example where the 1SG marker م (*=im*) and the emphasis endoclitic *=îş* appear after and before the host, i.e. گرت (*girt*), depending on the presence of other bound morphemes such as negation prefix نە (*ne-*) or the verbal particle دا (*da*). It is worth mentioning that this pattern may vary based on the Central Kurdish sub-dialects.

Moreover, the present form of the copula in Central Kurdish are also used as clitics with nouns and adjectives. Table 1 shows these as enclitics.

2.1.2 Affixes

In comparison to clitics, a higher number of bound morphemes in Central Kurdish belong to affixes. Affixes can be categorized into inflectional and derivational based on their ability to create new lexemes. The most frequent affixes in Central Kurdish appear as prefixes and suffixes. Some of the inflectional affixes of Central Kurdish belonging to open-class parts of speech, namely nouns, verbs, adjectives and adverbs, are shown in Table 2. In addition, Izafe particle *-î* and its allomorph *-e* which appear between a head and its dependents in a noun phrase are frequently used to create possessive constructions, as in ناوی من (*naw-î min* 'my name').

In addition to compound words, Central Kurdish relies on derivational morphemes to create new lexemes, particularly new verbal lexemes. To this end, verbal suffix ەوە (*-ewe*) and verbal particles such as دا (*da-*) and هەڵ (*hel-*) are used. It is worth noting that the passive form of verbs is derived from the verb stem by using را/رێ (*-ra/-rê*) or their allomorphs درا/درێ (*-dira/-dirê*) suffixes, unlike Kurmanji Kurdish which relies on periphrastic forms with HATIN (to come) (Ahmadi, 2021b).

In the following, we summarize some of the distinct features of affixes in Central Kurdish.

Nouns	Verbs	Adjectives	Adverbs
number (SG, PL)	**number** (SG, PL)	**number** (SG, PL)	**degree** (COMP, SUPL)
person (1, 2, 3)	**person** (1, 2, 3)	**degree** (COMP, SUPL)	
determiners (DEF, IND, DEM)	**mood** (IND, SBJV, IMP, COND)	**determiners** (DEF, IND, DEM)	
case (OBL, LOC, VOC)	**aspect** (PRF, IMP, PROG)		
gender (M, F)	**tense** (PST, PRS)		

Table 2: Inflectional features and values of Central Kurdish. It should be noted that the function of cases and genders vary among Sorani subdialects.

Discontinuous Morphemes A morpheme that gets interrupted by the insertion of another morphological unit is known as a discontinuous morpheme. Two categories of discontinuous morphemes exist in Central Kurdish: a) demonstratives "*em* ...*-e*" 'this.DEM' and '‘*ew* ...*-e*" 'that.DEM' and, b) circumpositions such as "*be* ...*-da*", "*le* ...*-da*" and "*bo* ...*-ewe*", respectively meaning 'through', 'in' and 'toward' where "..." refers to the position of another morphological unit between the two discontinuous morphemes. While the Latin-based orthography of Kurdish suggests writing such morphemes detached from the preceding word, they are usually concatenated in the Perso-Arabic-based script.

Postverbal Complement *-e* and Pronominal Adverb *-ê* In Central Kurdish, a verb that has the valency of a prepositional phrase with prepositions *be* 'to' or *bo* 'for' can take the postverbal complement *-e* to replace the preposition. In this case, it is compulsory for a noun phrase to come after the verb (Edmonds, 1961, p. 236). Furthermore, the pronominal adverb *-ê* can replace the antecedent prepositional object, and the postverbal complement *-e*, oblique pronoun, accusative nouns or locative adverb. This is particularly used with two verbs of DAN 'to give' and GEYIŞTIN 'to arrive'.

A more detailed description of Central Kurdish morphology, including adpositions and pronouns as free morphemes, is provided in (Ahmadi, 2021a) and (Naserzade et al., 2023).

3 Central Kurdish on UniMorph 4.0

In this section, we analyze the existing morphological data for Central Kurdish on UniMorph 4.0 and describe some of the current fallacies.

The UniMorph project provides a dataset for Central Kurdish that contains 24,316 word-forms.[3] This dataset was initially created within the Alexina Framework (Sagot, 2010) by Walther and Sagot (2010) and focuses on inflectional morphology by providing a set of forms of the paradigms of 252 lemmas with noun or verb part-of-speech tags. Overall, 33 morphological features based on UniMorph are used in the dataset, including LGSPEC1 and LGSPEC2 which are respectively used for Izafe morpheme *-î* and its allomorph *-e*. The number of features combined together is 226 features for all the word forms. In other terms, 0.98% of the word forms are assigned a unique combination of features. Analogous to the notion of 'leakage' in syntactic parsing (Krasner et al., 2022) that reveals the overlap of the train and test sets, such a repetitive usage of the features can cause an erroneously high performance of analysis models. As such, we believe that the dataset has very limited coverage of word forms and lacks diversity.

In the following, we categorize some of the major issues of the Central Kurdish data on UniMorph. Table 3 provides a few examples based on the dataset and categorizes their issues as well.

3.1 Unconventional Writing

Unlike Northern Kurdish which is mostly written in a Latin-based Kurdified script known as *Bedirxan*'s orthography, Central Kurdish is more conventionally written in a Perso-Arabic script. The Kurdish data on UniMorph is written in an unconventional Latin-based orthography that is not used in practice. Furthermore, the character <i> for phoneme /ɪ/ is not represented in the selected script, even though it is frequently used in many morphemes and undergoes various morphophonological alternations. This phoneme, also known as *Bizroke* (Ahmadi, 2019), is represented by <i> in the Latin-based script of Kurdish while is missing in the Perso-Arabic script. We transliterate the original forms in the dataset in the Latin-based script of Kurdish in Table 3.

[3] Available at `https://github.com/UniMorph/ckb`

Lemma	Feature	Form in UniMorph 4.0 (Incorrect)		Correct form	Issue
		Original	Transliterated		
aw 'WATER'	N;FOC	´awš	awş	awîş ئاویش	morphophonology
bûrîn 'FORGIVE'	V;PROG;IND;SG;3;PRS;PASS	debwwrrěět	debûrrêêt	debûrêt دەبوورێت	morphophonology
kirdin 'DO'	V;PROG;IND;SG;3;PRS	dekeě	dekeê	deka دەکا	morphophonology
bezandin 'DEFEAT (TR)'	V;PRF;SBJV;SG;1;NEG;PST	nembezandbwwayě	nembezandbuwayê	nembezandibuwaye نەمبەزاندبووایە	unknown morpheme *-yê*
bestin 'CLOSE (TR)'	V;PFV;SBJV;SG;1;PST	bbestmbayě	bibestimbayê	bimbestibaye بمبەستبایە	morphotactics
kirdin 'DO'	V;PROG;IND;PL;2;NEG;PRS;PASS	nakerěn	nakerên	nakirên ناکرێن	missing alternation
kokîn 'COUGH'	V;IMP;SG;NEG	mekok	mekok	mekoke مەکۆکە	missing morpheme *-e*

Table 3: Some of the categorical issues with the Central Kurdish data on UniMorph 4.0. The forms are transliterated into the conventional Latin-based script of Kurdish. The lemmata and the forms in the Perso-Arabic-based script of Kurdish are removed due to space limitations. The correct forms in both conventional scripts of Kurdish are reconstructed based on the features.

3.2 Morphotactics

As described in § 2, Central Kurdish has a complex morphotactics when it comes to verbs. This is also reflected in the inflection of verbal forms of the UniMorph dataset where some verbal word forms do not conform to the morphology of Central Kurdish and its dialects. This is particularly observed in transitive verbs in which the agent markers should appear before the verb stem and after the leftmost prefix in past tenses (see §2.1.1). However, this morphotactic rule is not systematically present in the verb forms. It is worth mentioning that this phenomenon is not the case in closely-related variants, i.e. Northern Kurdish and Southern Kurdish, or the closely-related language Persian. Therefore, we believe that the annotation was mistakenly and inaccurately carried out under the influence of such variants and languages.

3.3 Morphophonological Alternations

Many morphemes in Central Kurdish alter based on morphophonological rules. This is particularly the case of bound morphemes starting with a vowel, such as *-eke* as the singular definite marker and *-e* as a demonstrative suffix that respectively appear as *-ke*/*-yeke* and *-ye* depending on the preceding phoneme. In the UniMorph data, such alternations are not consistently taken into account. An eye-catching issue of this type is N;FOC which is associated with nouns that appear with the clitic *=îş*. The allomorph *=ş* of this clitic that appears after vowels seems to be universally used in the dataset regardless of the morphophonological rule. Therefore, word forms associated to this tag and other similar tags like N;LGSPEC2 are potentially wrong.

3.4 Incorrect Morphemes

A less severe problem of incorrect inflections is due to incorrect morphemes, particularly allomorphs. We believe that the unconventional script may have aggravated such issues. For instance, the singular imperative form of verbs, i.e. `v;imp;sg` are missing the suffix *-e* as in the incorrect form of *bbexš* (*bibexş*) instead of *bibexşe* (FORGIVE.IMP.2SG) and the morpheme *-yě* (*-yê*) is frequently and incorrectly used instead of the morpheme *-ye* to indicate the conditional mood of the verb. Nevertheless, such issues have been discussed, particularly concerning allomorphs, within the UniMorph community (Gorman et al., 2019).

Taking these issues into account, we estimate that 25% of the forms of Central Kurdish data on UniMorph 4.0 are incorrect.

4 Methodology

Given the fallacies of the Central Kurdish dataset on UniMorph 4.0, we believe that a new dataset is required for a thorough morphological analysis of this language. Although we correct the existing dataset on UniMorph 4.0, we also extend it with new lemmata and more complete paradigms. This measure was taken to ensure the quality of the forms based on a corpus and more importantly, in both conventional scripts of Kurdish, namely the

Perso-Arabic-based and the Latin-based scripts. In this section, we discuss our approach to creating a new dataset for Central Kurdish.

4.1 Modeling Central Kurdish on UniMorph

During the data preparation process, we noticed that the UniMorph schema described by Sylak-Glassman (2016) lacks several features that are commonly used in not only Central Kurdish but also, most Iranic languages, such as Izafe (Windfuhr, 2009). In the schema, the label LGSPEC with a consistent ID is considered for language-specific features. Using this, we also introduce a few features that are currently unsupported and map these new features to LGSPEC with an ID to be consistent with the current schema of UniMorph. Table 4 provides a list of such features.

Type	Function	Ours	UniMorph
Affix	Izafe	[IZAFE]	LGSPEC1
Affix	postverb adpositions	[E] [EE]	LGSPEC2
Affix	postverb adverbial /ewe/	[EWE1]	LGSPEC3
Affix	disc. adpositions	[DA],[RA], [EWE2]	LGSPEC4
Clitic	adverbial clitic	[ISH]	LGSPEC5
Clitic	demonstrative	[DEM]	LGSPEC6
Clitic	copula	[COP]	LGSPEC7
Clitic	pronominal markers (argument/possessive) on transitive past verbs	[PM]	LGSPEC8
Clitic	argument markers on noun/adjectives	[AM]	LGSPEC9

Table 4: Our proposed tags for the new Central Kurdish data in our dataset containing more customized tags and LGSPEC tags for the future versions of UniMorph

It is worth noting that in the current Central Kurdish data on UniMorph 4.0, LGSPEC1 and LGSPEC2 are respectively used for Izafa suffix <î/y> and its allomorph <e>. Similarly, the endoclitic *=îş* is specified as FOC. These are the only language-specific tags that are currently used in this dataset.

4.2 Finite-State Transducers

Relying on Naserzade et al. (2023)'s finite state transducers, we develop a morphological analyzer and generator that can handle all possible well-formed inflected forms of a given word in Central Kurdish. The analyzer takes a word and yields all possible morphological tags. Similarly, the generator takes as input a lemma and its part-of-speech tag, in addition to the past and present stems and transitivity for verbs, and inflects the lemma accordingly. The output words are formed according to Central Kurdish standard orthography and morphophonological rules. The number of forms with unique features is 3,032 for a general noun lemma, 9,096 for a gradable adjective, 3,180 for a transitive verb, and 636 for an intransitive verb. Figure 2 illustrates a transducer to generate noun forms.

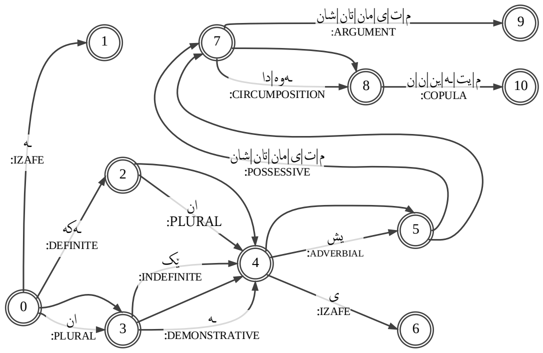


Figure 2: A finite-state transducer for generating nouns in Central Kurdish

4.3 Data Generation

Using the finite-state transducers, we generate two datasets containing a diverse set of word forms and part-of-speech tags as follows:

Gold-standard We first randomly extract 1,000 words from Veisi et al. (2020)'s corpus and then use the finite-state transducers to analyze them. Given that the transducers do not take word context into account, this step was followed by a manual verification to make sure that only the relevant analysis and tags are selected based on the context.

Silver-standard We also create another dataset that contains full paradigms for 10 nouns, 5 adjectives, 17 intransitive verbs (including three passive and two causative verbs) and 12 transitive verbs. As this dataset doesn't rely on context, we refer to it as silver-standard. These words are listed in Table A. To cover all morphophonological changes that occur in the inflectional forms, we select words having stems ending with a consonant, vowels <*a, e, ê, î, o, û*>, approximants <*y*> and <*w*/>, and diphthong <*wê*>. Note that vowels <*i*> and <*u*> do not occur in word-final positions.

During the generation processing, we set a few restrictions in our dataset. In the conjugation of transitive verbs, it is not possible to have both the subject and object pronouns in either the first or second person. This is due to the reflexive con-

Dataset	Noun		Adjective		Verb		Proper Noun		Other		Total	
UniMorph 4.0	1,729	141	0	0	22,291	112	0	0	0	0	24,020	253
Gold-standard	442	375	153	133	181	107	143	139	81	65	1,000	819
Silver-standard	30,320	10	45,447	5	35,116	25	0	0	0	0	110,883	40

Table 5: Number of inflected forms and unique lemmata (second column) by part-of-speech in Central Kurdish in the current dataset of UniMorph 4.0, our proposed datasets aggregated over all splits. The gold-standard presents a more diverse set of forms with part-of-speech tags for fewer lemmata while the silver-standard dataset presents full paradigms of more lemmata.

struction in Central Kurdish that does not commonly appear in the verb form. For example, **de-m=nas=im* '*I know me' and **de-tan-nas-ît* '*you know you' are ill-formed. For this purpose, the adverb *xo* 'self' is commonly used.

Table 5 summarizes the number of forms in our datasets in comparison to the current UniMorph 4.0 data. We present our datasets in both conventional scripts of Kurdish, the Arabic-based and Latin-based ones. The latter is more widely used for Northern Kurdish facilitating cross-dialectal comparisons. Moreover, we provide the corpus-based context of word forms and our customized tags in Table 4 in a separate dataset.

5 Analysis

5.1 Morphological Reinflection

To evaluate our datasets, we carry out an analysis on morphological reinflection introduced as the non-neural baseline for task 1 of SIGMORPHON 2018 that extracts lemma-to-form transformations heuristically (Cotterell et al., 2018). To do so, we first shuffled the datasets and created a 70–10–20 train–dev–test split. During the process, we made sure that identical samples were selected in the two scripts to make the comparison of performances valid. We then run the non-neural baseline using the train sets of the three datasets and evaluate the models on the three test sets. Table 6 presents the accuracy and average edit distance in the three datasets. Although it would have been interesting to compare the performance of the baseline system across test sets, e.g., training and testing on different datasets, such comparison could only be valid if the same set of tags has been used which is not the case in the current UniMorph data. Based on the results of the systems that participated in the SIGMORPHON 2021 Shared Task on morphological reflection (Pimentel et al., 2021), an accuracy of over 90% can be achieved.

Dataset (script)	**Accuracy**	**AED**
UniMorph 4.0	48.7%	0.97
Gold-standard (L)	63.5%	0.99
Gold-standard (A)	67.5%	0.88
Silver-standard (L)	61.2%	0.98
Silver-standard (A)	65.0%	0.75

Table 6: Experimental results of test sets on morphological re-inflection for the current UniMorph 4.0 in comparison to our datasets in terms of accuracy (higher is better) and average edit distance (lower is better). AED refers to average edit distance.

5.2 Error Analysis

In order to better understand the challenges of reinflection models, we manually checked the wrong outputs of the models trained and tested on our data to determine failure points. Since we have generated all possible inflectional forms of several lemmas and the data is shuffled before building the model, some complex forms do not occur in the train set. Therefore, the model failed to cover those forms. Another difficulty of the baseline model is in tackling the morphophonological changes. As we have covered stems with different final phoneme types, the majority of errors that have lower edit distance are in handling these changes. For example, the failure in alternating the indefinite suffix '*-êk*' to '*-yek*' after a vowel is a primary source of the errors.

In Kurdish, verbs have different past and present stems. For many verbs, the present stem is made by removing the final consonant or vowel of the past stem; for instance, the past and present stems of *girtin* 'to get' are *girt* and *gir*, respectively. However, numerous exceptions enforce computational studies to consider the present verb stems as irregular and look them up from a table, as in the present stems *lê* or *bêj* for *gutin* 'to say' and *xo* for *xwardin* 'to eat'. Analyzing the reinflectional errors showed that detecting such alternations is another major source of error.

Regarding the accuracy based on the scripts, the accuracy of the baseline on data written in the Latin-based (L) script is slightly lower than the Arabic-based script (A). This can be explained by the missing character *Bizroke* (*i*) in the Arabic-based script that plays an important role in Central Kurdish morphology (see §3.1) while the Latin-based character uses it.

5.3 Inflectional Synthesis Degree

As an additional analysis, we calculate the synthesis degree of inflected forms in Central Kurdish by averaging the number of morphemes per form in the gold and silver datasets. According to the degrees reported by Greenberg (1960), Central Kurdish has a relatively high synthetic degree of 2.22 comparable to Old English (2.12), Yakut (2.17), and Swahili (2.55). Among the selected part-of-speech tags, adjectives exhibit the highest level of synthesis as they can function with nominal affixes and clitics and also, a few other distinct ones such as *-tir* and *-tirîn* as comparative and superlative suffixes.

Although prefixing is not used in nouns and adjectives of Central Kurdish, verbs have a higher synthesis in prefixing, mainly due to the verbal prefixes related to negation such as *ne-*, *na-* and *me-* but also subjunctive *bi-* and progressive markers *e-* and *de-*. Moreover, transitive verbs show the highest ratios of synthesis in prefixing in comparison to intransitive verbs. This is due to the erratic patterns of pronominal endoclitics that may appear before or after the stem, while that's not the case in intransitive verbs (see §2.1.1).

It should be noted that these results are expected to be different in derivational morphology.

POS		Morpheme per form		
		pre-stem	post-stem	average
Noun		0	3.63	3.63
Adjective		0	4.30	4.30
Verb	INTR	1.05	2.32	1.68
	TR	1.65	2.46	2
Average		1.35	3.1	**2.22**

Table 7: Degree of synthesis in inflectional morphology of Central Kurdish based on our datasets

6 Conclusion

In this paper, we discuss some of the fallacies of the current data of Central Kurdish on UniMorph 4.0. We argue that the dataset is not only lacking coverage but also misrepresents Kurdish morphology by incorrect morphemes, unconventional writing and inaccurate morphotactics. Additionally, we propose a new dataset with a few additional labels for some of the features of Central Kurdish, such as Izafe and various clitics. Our dataset is generated using finite-state transducers with the human in the loop and are transliterated in the Latin-based script of Kurdish in addition to the Perso-Arabic-based ones. The transliteration of the word-forms facilitates comparative studies, particularly with Northern Kurdish which is mainly written in a Latin-based script. For each word-form, we also look it up in a corpus and provide the context in addition to the morphological features. Moreover, we create a baseline by training models in various setups and evaluating them on our dataset and the current Central Kurdish data on UniMorph 4.0. Finally, we suggest this dataset be added to the future version of UniMorph.

Limitations One of the limitations of our dataset is the lower number of word-forms belonging to a close-class part-of-speech as we chiefly focus on nouns, verbs (transitive and intransitive) and adjectives. On the other hand, we only include inflectional morphology without paradigms of word formation. Furthermore, we only address the morphology of the standard variety of Central Kurdish, i.e. that of Sulaymaniyah. We plan to extend our work to include other varieties of Central Kurdish along with derivational morphology. Given that Central Kurdish lacks a treebank, it will be compelling to bridge Central Kurdish morphology and syntax as well.

Another limitation of the current work is due to the UniMorph schema. Using the LGSPEC tag is not recommended for features that are found across languages but for those that are limited to specific languages (Sylak-Glassman, 2016, p.30). Given that some of the features of Central Kurdish, such as Izafe and pronominal copula, are also found in other closely-related languages, we believe that the current schema should be extended to use specific tags for such features or a better schema, akin to Guriel et al. (2022)'s hierarchical model, is needed for languages with rich morphology like Kurdish.

Acknowledgments

The authors are grateful for the constructive comments of the anonymous reviewers. Sina Ahmadi is generously supported by the National Science Foundation under DLI-DEL award BCS-2109578.

References

Sina Ahmadi. 2019. A rule-based Kurdish text transliteration system. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):1–8.

Sina Ahmadi. 2021a. A Formal Description of Sorani Kurdish Morphology. *arXiv preprint arXiv:2109.03942*.

Sina Ahmadi. 2021b. Hunspell for Sorani Kurdish Spell Checking and Morphological Analysis. *arXiv preprint arXiv:2109.06374*.

Sina Ahmadi, Antonios Anastasopoulos, and Géraldine Walther. 2023. A corpus-based study of endoclitic *=îş* in Kurdish. In *Book of abstracts of the the 56th Annual Meeting of the Societas Linguistica Europaea*, Athens, Greece. the 56th Annual Meeting of the Societas Linguistica Europaea.

Sina Ahmadi and Maraim Masoud. 2020. Towards Machine Translation for the Kurdish Language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 87–98.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, et al. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D McCarthy, Katharina Kann, Sabrina J Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2018. The conll–sigmorphon 2018 shared task: Universal morphological reinflection. *arXiv preprint arXiv:1810.07125*.

C. J. Edmonds. 1961. Kurdish Dialect Studies, Vol. Oxford University Press. *Journal of the Royal Asiatic Society*, 94(3-4).

Hiba Gharib and Clifton Pye. 2018. The clitic status of person markers in Sorani Kurdish. *University of Kansas Department of Linguistics*.

John Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2):153–198.

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 2019. Weird inflects but OK: making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 140–151. Association for Computational Linguistics.

Joseph H Greenberg. 1960. A quantitative approach to the morphological typology of language. *International journal of American linguistics*, 26(3):178–194.

David Guriel, Omer Goldman, and Reut Tsarfaty. 2022. Morphological reinflection with multiple arguments: An extended annotation schema and a Georgian case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Dublin, Ireland. Association for Computational Linguistics.

Martin Haspelmath and Andrea D Sims. 2013. *Understanding morphology*. Routledge.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2020. DagoBERT: Generating derivational morphology with a pretrained language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3848–3861, Online. Association for Computational Linguistics.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. One-shot neural cross-lingual transfer for paradigm completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1993–2003, Vancouver, Canada. Association for Computational Linguistics.

Ethelbert E Kari. 2002. On endoclitics: Some facts from Degema. *Journal of Asian and African Studies*, 63:37–53.

Lauri Karttunen and Kenneth R Beesley. 2005. Twenty-five years of finite-state morphology. *Inquiries Into Words, a Festschrift for Kimmo Koskenniemi on his 60th Birthday*, pages 71–83.

Lauri Karttunen et al. 1983. KIMMO: a general morphological processor. In *Texas Linguistic Forum*, volume 22, pages 163–186. Texas, USA.

Nathaniel Krasner, Miriam Wanner, and Antonios Anastasopoulos. 2022. Revisiting the effects of leakage on dependency parsing. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2925–2934, Dublin, Ireland. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, S. J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. *CoRR*, abs/1910.11493.

Ernst M McCarus. 2007. Kurdish morphology. *Morphologies of Asia and Africa*, 2:1021–1049.

Morteza Naserzade, Aso Mahmudi, Hadi Veisi, Hawre Hosseini, and Mohammad Mohammadamini. 2023. CKMorph: a comprehensive morphological analyzer for Central Kurdish. *International Journal of Digital Humanities*.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Brian Roark and Richard Sproat. 2007. *Computational approaches to morphology and syntax*, volume 4. OUP Oxford.

Benoît Sagot. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Patrick Schone and Dan Jurafsky. 2001. Knowledge-free induction of inflectional morphologies. In *Second meeting of the North American chapter of the association for computational linguistics*.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (UniMorph Schema). *Johns Hopkins University*.

Hadi Veisi, Mohammad MohammadAmini, and Hawre Hosseini. 2020. Toward Kurdish language processing: Experiments in collecting and processing the AsoSoft text corpus. *Digital Scholarship in the Humanities*, 35(1):176–193.

Peter W. Smith. 2014. Non-peripheral cliticization and second position in Udi and Sorani Kurdish. In *Paper under revision at Natural Language and Linguistic Theory*, https://user.uni-frankfurt.de/~psmith/docs/smith_non_peripheral_cliticization.pdf edition. (Date accessed: 12.05.2020).

Géraldine Walther. 2012. Fitting into morphological structure: accounting for Sorani Kurdish endoclitics. In *Mediterranean Morphology Meetings*, volume 8, pages 299–321. [Online; accessed 19-Mar-2019].

Géraldine Walther and Benoît Sagot. 2010. Developing a large-scale lexicon for a less-resourced language: General methodology and preliminary experiments on Sorani Kurdish. In *Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop)*.

Adam Wiemerslage, Arya D. McCarthy, Alexander Erdmann, Garrett Nicolai, Manex Agirrezabal, Miikka Silfverberg, Mans Hulden, and Katharina Kann. 2021. Findings of the SIGMORPHON 2021 shared task on unsupervised morphological paradigm clustering. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 72–81, Online. Association for Computational Linguistics.

Gernot Windfuhr. 2009. *The Iranian Languages*. Routledge London.

A Appendix

Noun	Adjective	Verb (present stem)	
		Intransitive	Transitive
dar / دار ‘tree’	*lar* / لار ‘crooked’	*kewtin* (*kew-*) / کەوتن ‘fall’	*girtin* (*gir-*) / گرتن ‘get’
pyaw / پیاو ‘man’	*zana* / زانا ‘shrewd’	*mirdin* (*mir-*) / مردن ‘die’	*birdin* (*bir-*) / بردن ‘take’
mey / مەی ‘wine’	*taze* / تازە ‘fresh’	*çûn* (*ç-*) / چوون ‘go’	*xwardin* (*xo-*) / خواردن ‘eat’
xesû / خەسوو ‘step-mother’	*namo* / نامۆ ‘weird’	*řoyştin* (řo-) / ڕۆیشتن ‘leave’	*biřîn* (*biř-*) / بڕین ‘cut’
masî / ماسی ‘fish’	*nwê* / نوێ ‘new’	*nûstin* (*nû-*) / نووستن ‘sleep’	*pêwan* (*pêw-*) / پێوان ‘measure’
ajawe / ئاژاوە ‘chaos’		*westan* (*west-*) / وەستان ‘stop’	*kirdin* (*ke-*) / کردن ‘do’
bira / برا ‘brother’		*pijmîn* (*pijm-*) / پژمین ‘sneeze’	*dan* (*de-*) / دان ‘give’
diro / درۆ ‘lie’		*tirsan* (*tirs-*) / ترسان ‘fear.CAUS’	*firoştin* (*firoş-*) / فرۆشتن ‘sell’
girê / گرێ ‘knot’		*birjan* (birjê-) / برژان ‘grill’	*gwastin* (*gwaz-*) / گواستن ‘carry’
gwê / گوێ ‘ear’		*biřan* (*biřê-*) / بڕان ‘cut’	*pařan* (*pařê-*) / پاڕان ‘beg’
		kizan (*kizê-*) / کڒان ‘singe’	
		leran (*lerê-*) / لەران ‘wobble’	
		geşan (*geşê-*) / گەشان ‘blow’	
		biran (*bir-*) / بران ‘carry.PASS’	
		pirsiran (*pirsir-*) / پرسران ‘ask.PASS’	
		niran (*nir-*) / نران ‘put.PASS’	
		bezîn (*bez-*) / بەزین ‘defeat.CAUS’	

Table A.1: Selected words for which full paradigms are generated and included in our dataset