Towards Finite-State Morphology of Kurdish

SINA AHMADI, Insight Centre for Data Analytics, National University of Ireland Galway, Ireland

HOSSEIN HASSANI, University of Kurdistan Hewlêr, Iraq

Morphological analysis is the study of the formation and structure of words. It plays a crucial role in various tasks in Natural Language Processing (NLP) and Computational Linguistics (CL) such as machine translation and text and speech generation. Kurdish is a less-resourced multi-dialect Indo-European language with highly inflectional morphology. In this paper, as the first attempt of its kind, the morphology of the Kurdish language (Sorani dialect) is described from a computational point of view. We extract morphological rules which are transformed into finite state transducers for generating and analyzing words. The result of this research assists in conducting studies on language generation for Kurdish and enhances the Information Retrieval (IR) capacity for the language while leveraging the Kurdish NLP and CL into a more advanced computational level.

1 INTRODUCTION

Morphology is the study of the structure of words. The morpheme is a phonemically defined segment of speech or set of segments of speech with a constant range of meaning. Morphological analysis is a central task in language processing that can take a word as input and detect the various morphological entities in the word and provide a morphological representation of it.

In the case of morphologically rich languages, multiple types and levels of information may be available at the word level. The lexical information for each word form may be augmented with information concerning the grammatical function of the word in the sentence, pronominal clitics, its grammatical relations to other words, inflectional affixes, and so on [Tsarfaty et al. 2013]. In Kurdish, many of these notions are expressed by inflectional affixes. For instance, the stem of a transitive verb can be preceded or succeeded by affixes to indicate the subject, the direct object, tense and grammatical mood. Figure 1 illustrates the morphological tree of the verb "*hełimnegirtbûnewe*" meaning "(I) had not retaken them".

Analysing morphologically rich languages is a challenging task regardless of the analysis technique implemented. Our main objective in this paper is to demonstrate the capacity of Finite-State Transducers (FSTs) as a method towards

Authors' addresses: Sina Ahmadi, sina.ahmadi@insight-centre.org, Insight Centre for Data Analytics, National University of Ireland Galway, Galway Business Park, Galway, Ireland, H91 AEX4; Hossein Hassani, University of Kurdistan Hewlêr, 30M Avenue, Erbil, Iraq, hosseinh@ukh.edu.krd.

```
                                        Verb
                        ┌────────────────┴────────────────────┐
                     Affix                                     │
                  ┌────┴────┐                    ┌─────────────┴──────────┐
                Affix     Affix                Verb                     Affix
                  │         │           ┌────────┴────────┐               │
                 heł       im         Verb              Affix            ewe
                                ┌───────┴──────┐          │
                              Verb          Auxiliary    (i)n
                        ┌───────┴──────┐       │
                      Affix          Root      bû
                        │             │
                       ne           girt
```
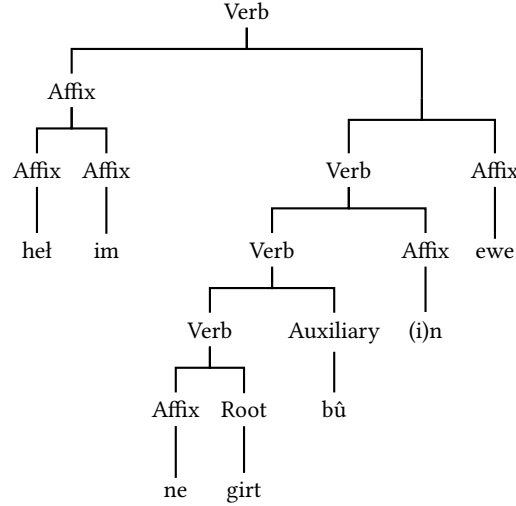
Fig. 1. Morphological tree of "hełimnegirtbûnewe"

computational morphology in the Kurdish language. We focus on Sorani Kurdish as one of the widely spoken and written dialects of Kurdish [Salavati and Ahmadi 2018]. Given the complexity of the Kurdish morphology, we have only covered the major inflectional morphemes in the current study, namely nouns, verbs, adjectives, and adverbs. This project is also an attempt to address three items of the identified gaps in the Kurdish BLARK (Basic Language Resource Kit) [Hassani 2018], namely morphological analysis, morphological synthesis, and text generation.

In our approach to the morphological study of Kurdish, we follow the trend which is followed in the Natural Language Processing (NLP) realm. This approach has been described in the related literature, for example, by Jacquemin and Tzoukermann [Jacquemin and Tzoukermann 1999]. That is, we are aware of the discussion and debates among the linguists upon the meaning of morphemes and lexemes and also the debate that is going on among the linguists in how to interpret these concepts [Beard and Volpe 2005; Nemo 2003]. However, we focus on the practical outcome which could help in the advancement of the resources and tools for Kurdish NLP.

Finite-State Transducers (FSTs) are well-known apparatus in various sectors of computational field including NLP and CL [Karttunen 2000], in a variety of tasks such as machine translation [Civera et al. 2005; Forcada et al. 2011] and computational morphology [Altantawy et al. 2011]. In comparison to the traditional computational morphology based on lexicon, FSTs are proven to be more effective instruments concerning morphologically rich languages [Altantawy et al. 2011]. One of the main obstacles that have hindered the progress in morphological analysis for Kurdish was lack of electronic lexicographic resources. Recently, Ahmadi et al. [Ahmadi et al. 2019] presented three electronic lexicographic resources for Kurdish covering the Hawrami, Sorani and Kurmanji dialects with enriched information such as part-of-speech. Therefore, as a preliminary study, we could construct a set of FST-based rules for four grammatical categories in Sorani Kurdish, namely nouns, adjectives, adverbs and verbs. For this purpose, We use Stuttgart Finite-State-Transducer (SFST) [Schmid 2005] as the experimental environment.

The rest of this paper is organized as follows: Section 2 reviews the literature and addresses the related work. Section 3 presents the Kurdish morphology. In section 4, we describe the implementation of our morphological analyser for

various morphological categories of Sorani Kurdish. And finally, the paper is concluded in Section 5 and a few idea for future research are proposed.

2 RELATED WORK

Computational morphology using FSTs has been a topic of research since the 1980s [Karttunen and Beesley 2001], and a variety of tools have been developed for this purpose [Beesley and Karttunen 2003]. The Finite-State Morphology (FSM) has covered both morphological analysis and generation aspects of computational morphology addressing various languages some of which have been well-equipped by NLP and CL tools and some not. While resourceful languages such English [Beesley and Karttunen 2003; Minnen et al. 2001] and German [Schmid 2005] are front-runners in FSM studies, we also observe similar scholarly attempts regarding other languages which might not be considered as widely-studied as English or German such as Uralic languages [Novák 2015], Arabic [Soudi et al. 2007] and Persian [Arabsorkhi and Shamsfard 2006; Megerdoomian et al. 2000]. This is also correct for less-studied languages such as Croatian [Mihajlović 2014]. These examples denote the existence of wider attention on the area of computational morphology in NLP and CL whether by using FSTs or by following other approaches.

Despite its crucial role in the language analysis and generation, computational morphology has not received noticeable attention in Kurdish NLP and CL except for a few efforts [Hassani 2018]. Walther and Sagot [Walther and Sagot 2010] and Walther et al. [Walther et al. 2010] present a methodology and preliminary experiments on constructing a morphological lexicon for Sorani and Kurmanji in which a lemma and a morphosyntactic tag are associated with each known form of the word. For this purpose, the Alexina framework [Sagot 2010] is used.

Computational morphology of Kurdish has been partially addressed in related NLP and CL tasks. Hosseini et al. [Hosseini et al. 2015] suggest a formulation for Sorani morphology used in creating a Sorani lexicon. Salavati et al. [Salavati and Ahmadi 2018] report challenges in Sorani Kurdish lemmatization and spelling error correction due to morphological complexity. [Gökırmak and Tyers 2017] carries out a morphological analysis for creating a dependency treebank for Kurmanji Kurdish.

The aforementioned situation indicates that the lack of Kurdish computational morphology is an area that not only requires but also deserves a sizeable effort by and collaboration among interested scholars.

3 KURDISH MORPHOLOGY

Kurdish is an Indo-European language which is spoken by approximately 30 million people in different countries [Ahmadi et al. 2019]. Haig and Matras [Haig and Matras 2002] provide a brief description of the structural properties of Kurdish. As a multi-dialect language, Kurdish dialects have different grammatical features and vocabulary sets [Hassani and Medjedovic 2016]. The differences in grammar and vocabulary vary among the dialects [Haig and Öpengin 2014; Jügel 2014]. In several cases, the differences are significant while in the others, they are trivial [Hassani and Medjedovic 2016]. Equally important, the language is written in different scripts with no standard orthography [Ahmadi 2019]. This enforces the computational morphology for Kurdish to be dialect-focused, at least in the first steps.

Kurdish is considered a morphologically rich language for which grammatical relations are indicated by changes in the word forms and modifying morpheme [Littell et al. 2016; Zivingi and Sulaiman 2019]. Unlike most Indo-European languages which are considered fusional languages, Kurdish is also characterized as a partially agglutative language [Khalid and Hamamorad 2015]. Agglutination refers to a linguistic process in which various word forms are created by stringing morphemes together.

Our focus in this paper is on four grammatical categories, namely verbs, nouns, adjectives, and adverbs. These categories are briefly described as follows.

3.1 Nouns

The absolute form of a noun in Kurdish is the form without any affixes which represents a generic meaning of the word [Thackston 2006]. This form is the lemma provided in the dictionary. The inflection of nouns is mostly carried out with suffixes to indicate number, definiteness, indefiniteness, demonstratives and gender. Unlike Kurmanji and Hawrami dialects, Sorani and Southern Kurdish dialects do not specify genders through morphological inflection. However, there are a few exceptions in some of the subdialects of the two latter dialects, such as the Mukryani subdialect of Sorani where nouns can have genders in specific cases. For instance, in the sentences "*deçime małe Zeynebî*" and "*le kin Ferhadê*", "*Zeyneb*" and "*Ferhad*" respectively as feminine and masculine proper names are inflected with the "î" and "ê" suffixes to represent the gender.

Noun form	Singular	Plural
absolute	*naw* *derga*	- -
indefinite	*naw**êk*** *derga**yek***	*naw**an*** *derga**yan***
definite	*naw**eke*** *derga**ke***	*naw**ekan*** *derga**kan***
demonstrative	*em naw**e*** *em derga**ye***	*em naw**ane*** *em derga**yane***

Table 1. Inflected forms of "*naw*" (name) and "*derga*" (door) in Sorani Kurdish

Table 1 describes the inflection of the two words "*naw*" (name) and "*derga*" (door) where the morphemes are highlighted in bold. As a result of the inflection, phonological changes, such as dropping a vowel or adding an auxiliary one, may happen when two vowels emerge consecutively. For instance, such a change can be observed in "*dergaye*" where "*y*" appears between the lemma "*derga*" and the demonstrative suffix "*e*".

3.2 Adjectives and adverbs

Adjectives in Sorani Kurdish are inflected according to the modified noun. Predicative adjectives in Kurdish are mostly inflected based on their syntactic role. In addition, the comparative and superlative degrees of an adjective are respectively made by suffixes "*tir*" and "*tirîn*". On the other hand, the attributive adjectives follow a richer morphological representation depending on the inflection of the noun. The main noun-adjective construction in Sorani Kurdish is carried out using the *Izafa* morpheme. The Izafa is an "*î*" or "*y*" (following a vowel) appearing between the noun and the adjective. However, there are other types of making attributive adjectives which are differing in the noun-adjective construction and the placement of the definiteness or number suffixes [Thackston 2006]. Table 2 provides an example of the attributive adjective "*ciwan*" (beautiful) for the lemma "*guł*" (flower) in various forms.

Although there are adverbs which are words in their absolute form, in most cases, adverbs are inflected forms of the absolute form of an adjective or noun in Kurdish. This process is particularly made by using the suffix "*ane*" or the prefix with "*be*". For instance, the two adverbs "***be**tûndî*" and "*tund**ane***" with the root "*tundî*" (noun, quickness) and "*tund*" (adjective, quick).

Noun form	Loose-Izafa		Close-Izafa	
	Singular	Plural	Singular	Plural
absolute	*gułî ciwan*	-	*gułe ciwan*	-
indefinite	*gułêki ciwan*	*gułanî ciwan*	*gułe ciwanêk*	*gułe ciwanan*
definite	*gułeke ciwan*	*gułekanî ciwan*	*gułe ciwaneke*	*gułe ciwanekan*
demonstrative	*em gułe ciwane*	*em gułe ciwanane*	*em gułe ciwane*	*em gułe ciwanane*

Table 2. Inflected forms of the adjective "*ciwan*" (beautiful) with the noun "*guł*" (flower) in Sorani Kurdish

3.3 Verbs

In Sorani Kurdish, verbs agree with their subject in number and person and in some cases, with their object as well. For instance, in the verb "*dîtimî*" meaning "(I) saw you (singular)", *dît*, the past stem of the verb "*dîtin*" (to see), is inflected based on the enclitic pronominal logical object *î* and the agent affix *im*. In addition, Kurdish is a split-ergative language where subject of an intransitive verb behaves like the object of a transitive verb and differently from the agent of a transitive verb [Comrie 1989]. Precisely, the ergative case is marked on agents and verbs of transitive verbs in past tenses [Bynon 1979].

Verbs have two stems in the past and present. Although the past stem can be extracted by removing the infinitive suffix "*in*", the present tense is irregularly derived from the infinitive form of the verb.

4 IMPLEMENTATION

After a thorough study of our intended four grammatical categories, i.e., noun, adjectives, adverbs, and verbs, we defined their morphological features in plain language. As the result, we could extract 171 inflection rules which enabled us to obtain a deeper insight into the morphological rules for implementing transducers.

Our implementation is realized in the SFST transducer specification language which is based on regular expressions with extended functionalities such as concatenation and variable definition [Schmid 2005]. The SFST compiler creates FSTs by concatenating and filtering morphemes and mapping the resulting analysis strings to the surface realizations according to phonological rules [Schmid et al. 2004]. SFST not only analyses the composing morphemes of the input word with feature decorations, but it also has a generation mode which makes the transducers reversible. This mode is particularly important for tasks relying on text generation. The following shows an example in the two modes for the word "*xwardim*" (I ate):

```
analyze> xwardim
xward<verb-transitive-past-stem><past-1s>
generate> xward<verb-transitive-past-stem><past-1s>
xwardim
```

The properties of base word forms are defined in a lexicon which is based on Ahmadi et al.'s work [Ahmadi et al. 2019]. In the case of nouns, adjectives and adverbs, we considered the absolute form of the word as the base form. However, this was not the case of words with derivational morphemes which was out of the scope of the current study. Regarding the verbs, two base forms are defined: the past stem and the present stem. In addition, the transitivity case of the verbs is specified as different morphological rules apply for transitive and intransitive verbs in past tenses. Figure 2 illustrates a finite-state transducer of the verb *xwardin* (to eat) in the past simple tense:

Description	Tag	PoS
Connected possessive pronoun (person 1-6)	<1s>	*im*
	<2s>	*it*
	<3s>	*î*
	<1p>	*man*
	<2p>	*tan*
	<3p>	*yan*
Copula (verbs for person 1-6)	<C1>	*im*
	<C2>	*î, ît*
	<C3>	*heye, hes, e*
	<C4>	*în, heyn*
	<C5>	*in, hen*
	<C6>	*in, hen*
Imperative marker (verb)	<IMP>	*bi, b*
Negative imperative (verb)	<NMP>	*me*
Negative marker (verb)	<NEG>	*ne, na*
Subjunctive marker (verb)	<SUB>	*bi*
Plural suffix (noun)	<PL>	*an, gel, ha, at*
Definite marker (noun)	<DEF>	*eke*
Indefinite marker (noun)	<IND>	*êk, yek*
Progressive marker (verb)	<CON>	*de, e*
Relative adjectives	<RA>	*î, y, în, yin, çî, nok, ko, û, île, yile, emenî*
Comparative marker (adjective)	<COMP>	*tir*
Superlative marker (adjective)	<SUP>	*tirîn*
Adverb marker (adverb)	<AD>	*ane, be, an*
Diminutive form (noun)	<DIM>	*çe, ke, ik, ko, oke, oł, ołe, ołik, ołke, ełe, elûke, yekołe, îlane, île, ûlke, ûle, le, łe*

Table 3. Some of the inflectional morphemes and clitics used in our finite-state analyser
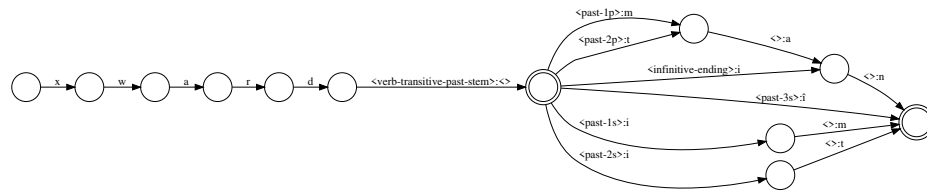
Fig. 2. A finite-state transducer analysing and generating simple past tense of the transitive verb *xwardin* (to eat) in Sorani Kurdish

In the case of the verb, we define a basic FST composing of the verb root (in present or past) suffixed with person and number affixes. This FST enables us to reduce redundancy in defining rules as different combination of prefixes such as *bi*, *ne* and *de* respectively as progressive, negative and imperative prefixes, would be feasible. Having said that, the ergative cases are not covered in this initial study. Table 3 provides some of the inflectional morphemes which are used as variable in our transducers.

Kurdish is written using different scripts in which Persian-Arabic and Latin are the two most widely used. The challenges in the processing of Persian-Arabic texts [Ahmadi 2019], particularly the missing "*i*" vowel, also known as *Bizroke* among Kurdish scholars, is the main reason that in the current work we focused on Latin-based script.

5 CONCLUSION AND FUTURE WORK

In this paper, we reported our progress in utilizing FST for the morphological analysis of Kurdish. We presented the four main categories of the Kurdish morphology, namely verbs, nouns, adjectives and adverbs. This research equips Kurdish NLP and CL with an advanced tool and resource which improves the quality of the Kurdish language processing tasks.

In this preliminary study, our developed tool can analyse and generate rather simple morphological forms. Therefore, as future work, we are aiming to cover more linguistic processes such as ergativity, and other grammatical categories in Sorani Kurdish. Moreover, analysing morphological derivation of Kurdish should be a priority, particuarly in the case of verbs which are mostly formed based on derivational morphemes. Another limitation of the current study is the absence of syntactical features which may modify morphological forms in a sentence.

In our implementation, we only used the Latin script of Kurdish as it is the least ambiguous one in comparison to the Persian-Arabic script [Ahmadi 2019]. We believe that the current study will pave the way for evaluating machine transliteration systems in future as well. In addition, we believe that the computation analysis of the other dialects of Kurdish will pave the way for further progress in the field.

Our tool and resource will be publicly available for non-commercial use under the CC BY-NC-SA 4.0 licence upon the acceptance of the paper.

REFERENCES

Sina Ahmadi. 2019. A Rule-based Kurdish Text Transliteration System. *Asian and Low-Resource Language Information Processing (TALLIP)* 18, 2 (2019), 18:1–18:8.

Sina Ahmadi, Hossein Hassani, and John P. McCrae. 2019. Towards Electronic Lexicography for the Kurdish Language. In *Proceedings of the eLex 2019 conference*. Brno: Lexical Computing CZ, s.r.o., Sintra, Portugal, 881–906.

Mohamed Altantawy, Nizar Habash, and Owen Rambow. 2011. Fast yet rich morphological analysis. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*. 116–124.

Mohsen Arabsorkhi and Mehrnoush Shamsfard. 2006. Unsupervised Discovery of Persian Morphemes. In *Demonstrations*. https://www.aclweb.org/anthology/E06-2023

Robert Beard and Mark Volpe. 2005. Lexeme-morpheme base morphology. In *Handbook of word-formation*. Springer, 189–205.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology* (1 ed.). Center for the Study of Language and Information.

Theodora Bynon. 1979. The ergative construction in Kurdish. *Bulletin of the School of Oriental and African Studies* 42, 2 (1979), 211–224.

Jorge Civera, Juan M Vilar, Elsa Cubel, Antonio L Lagarda, Sergio Barrachina, Francisco Casacuberta, and Enrique Vidal. 2005. A novel approach to computer-assisted translation based on finite-state transducers. In *International Workshop on Finite-State Methods and Natural Language Processing*. Springer, 32–42.

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation* 25, 2 (2011), 127–144.

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)*. 64–72.

Geoffrey Haig and Yaron Matras. 2002. Kurdish linguistics: a brief overview. (2002).

Geoffrey Haig and Ergin Öpengin. 2014. Introduction to Special Issue-Kurdish: A critical research overview. *Kurdish Studies* 2, 2 (2014), 99–122.

Hossein Hassani. 2018. BLARK for multi-dialect languages: towards the Kurdish BLARK. *Language Resources and Evaluation* 52, 2 (2018), 625–644.

Hossein Hassani and Dzejla Medjedovic. 2016. Automatic Kurdish dialects identification. *Computer Science & Information Technology* 6, 2 (2016), 61–78.

Hawre Hosseini, Hadi Veisi, and Mohammad Mohammadamini. 2015. KSLexicon: Kurdish-Sorani Generative Lexicon. [In Farsi].

Christian Jacquemin and Evelyne Tzoukermann. 1999. NLP for term variant extraction: synergy between morphology, lexicon, and syntax. In *Natural language information retrieval*. Springer, 25–74.

Thomas Jügel. 2014. On the linguistic history of Kurdish. *Kurdish Studies* 2, 2 (2014), 123–142.

Lauri Karttunen. 2000. Applications of finite-state transducers in natural language processing. In *International Conference on Implementation and Application of Automata*. Springer, 34–46.

Lauri Karttunen and Kenneth R Beesley. 2001. A short history of two-level morphology. *ESSLLI-2001 Special Event titled" Twenty Years of Finite-State Morphology* (2001).

Hewa Salam Khalid and Atta Mostafa Hamamorad. 2015. New Forms and Functions of Subordinate Clause in Kurdish. *International Journal of Kurdish Studies* 1, 2 (2015), 1–14.

Patrick Littell, Kartik Goyal, David R Mortensen, Alexa Little, Chris Dyer, and Lori Levin. 2016. Named entity recognition for linguistic rapid response in low-resource languages: Sorani Kurdish and Tajik. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 998–1006.

Karine Megerdoomian et al. 2000. *Persian computational morphology: A unification-based approach*. Computing Research Laboratory, New Mexico State University.

Milena Mihajlović. 2014. A computational morphology approach to Croatian noun inflection—the case of gender assignment. *Wiener Linguistische Gazette* (2014), 120–127. Issue Special Issue 78A.

Guido Minnen, John Carroll, and Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering* 7, 3 (2001), 207–223.

François Nemo. 2003. Morphemes and lexemes versus "Morphemes or Lexemes?". In *Mediterranean Morphology Meetings*, Vol. 4. 195–208.

Attila Novák. 2015. *A model of computational morphology and its application to Uralic languages*. Ph.D. Dissertation. Pázmány Péter Katolikus Egyetem.

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)*.

Shahin Salavati and Sina Ahmadi. 2018. Building a Lemmatizer and a Spell-checker for Sorani Kurdish. *arXiv preprint arXiv:1809.10763* (2018).

Helmut Schmid. 2005. A programming language for finite state transducers. In *FSMNLP*, Vol. 4002. 308–309.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection.. In *LREC*. Lisbon, 1–263.

Abdelhadi Soudi, Günter Neumann, and Antal Van den Bosch. 2007. Arabic computational morphology: knowledge-based and empirical methods. In *Arabic computational morphology*. Springer, 3–14.

Wheeler M. Thackston. 2006. *Sorani Kurdish–A Reference Grammar with Selected Readings*. Harvard University.

Reut Tsarfaty, Djamé Seddah, Sandra Kübler, and Joakim Nivre. 2013. Parsing morphologically rich languages: Introduction to the special issue. *Computational linguistics* 39, 1 (2013), 15–22.

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)*.

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*.

Safia Zivingi and Rafik Sulaiman. 2019. Comparative Morphological Analysis of Diminutive and Augmentativein English, Kurdish and Arabic. *International Journal of Communication Research* 9, 2 (2019), 132–135.