# UniMorph 2.0: Universal Morphology

**Christo Kirov**[1], **Ryan Cotterell**[1], **John Sylak-Glassman**[1], **Géraldine Walther**[2]
**Ekaterina Vylomova**[3], **Patrick Xia**[1], **Manaal Faruqui**[4], **Sebastian Mielke**[1], **Arya D. McCarthy**[1]
**Sandra Kübler**[5], **David Yarowsky**[1], **Jason Eisner**[1], **Mans Hulden**[6]
[1]Johns Hopkins University, [2]University of Zurich, [3]University of Melbourne
[4]Google, [5]Indiana University, [6]University of Colorado
Baltimore, Zurich, Melbourne, New York, Bloomington, Boulder
{ckirov1, ryan.cotterell, jcsg, paxia, sjmielke, arya, yarowsky, eisner}@jhu.edu, geraldine.walther@uzh.ch
evylomova@gmail.com, mfaruqui@google.com, skuebler@indiana.edu, mans.hulden@colorado.edu

**Abstract**

The Universal Morphology (UniMorph) project is a collaborative effort to improve how NLP handles complex morphology across the world's languages. The project releases annotated morphological data using a universal tagset, the UniMorph schema. Each inflected form is associated with a lemma, which typically carries its underlying lexical meaning, and a bundle of morphological features from our schema. Additional supporting data and tools are also released on a per-language basis when available. UniMorph is based at the Center for Language and Speech Processing (CLSP) at Johns Hopkins University in Baltimore, Maryland. This paper details advances made to the collection, annotation, and dissemination of project resources since the initial UniMorph release described at LREC 2016.

**Keywords:** morphology, multilingual, lexical resources

## 1. Introduction

Complex morphology is ubiquitous among the languages of the world. For example, roughly 80% of languages use morphology to mark verbal tense and 65% mark nominal case (Haspelmath et al., 2005). While overlooked in the past, explicit modeling of morphology has been shown to improve performance on a number of downstream HLT tasks, including including machine translation (MT) (Dyer et al., 2008), speech recognition (Creutz et al., 2007), parsing (Seeker and Çetinoğlu, 2015), keyword spotting (Narasimhan et al., 2014), and word embedding (Cotterell et al., 2016b). This has led to a surge of new interest and work in this area (Durrett and DeNero, 2013; Ahlberg et al., 2014; Nicolai et al., 2015; Faruqui et al., 2016).

The Universal Morphology (UniMorph) project, centered at the Center for Language and Speech Processing (CLSP) at Johns Hopkins University is a collaborative effort to improve how NLP systems handle complex morphology across the world's languages. The project releases annotated morphological data using a universal tagset, the UniMorph schema. Each inflected form is associated with a lemma, which typically carries its underlying lexical meaning, and a bundle of morphological features from our schema. Additional supporting data and tools are also released on a per-language basis when available.

Kirov et al. (2016) introduced version 1.0 of the UniMorph morphological database, created by extracting and normalizing the inflectional paradigms included in Wiktionary (`www.wiktionary.org`), a large, broadly multi-lingual crowd-sourced collection of lexical data. This paper describes UniMorph 2.0. It details improvements in Wiktionary extraction and annotation, as well as normalization of non-Wiktionary resources, leading to a much higher quality morphological database. The new dataset spans 52 languages representing a range of language families. As in UniMorph 1.0, we provide paradigms from highly-inflected open-class word categories — nouns, verbs, and adjectives. Many of the included languages are extremely low-resource, e.g., Quechua, Navajo, and Haida. This data was used as the basis for the CoNLL 2017 Shared Task on Morphological Learning (`http://sigmorphon.org/conll2017`) (Cotterell et al., 2017).

## 2. Wiktionary Extraction

In Kirov et al. (2016), we introduced version 1.0 of the UniMorph morphological database, based on a very large-scale parsing and normalization of Wiktionary. Wiktionary is a broadly multilingual resource with many crowd-sourced morphological paradigms in the form of custom HTML tables. Figure 1 illustrates the challenge associated with extracting this data. Wiktionary is designed for human, rather than machine readability, and authors have extensive freedom in formatting data. This leads to wildly differing table layouts across languages which need to be converted to a consistent tabular format.

The extraction process developed for UniMorph 1.0 relied heavily on statistical, visual, and positional heuristics (Sylak-Glassman et al., 2015b) to:

1. Determine which entries in an HTML table are inflected forms and which are grammatical descriptors.
2. Link each inflected form with its appropriate descriptors.
3. Convert each set of linked descriptors into a universal feature annotation schema, described in detail in Sylak-Glassman (2016).[1]

This led to a large dataset of 952,530 unique noun, verb, and adjective lemmas across 350 languages. Unfortunately,

[1] `unimorph.github.io/doc/unimorph-schema.pdf`

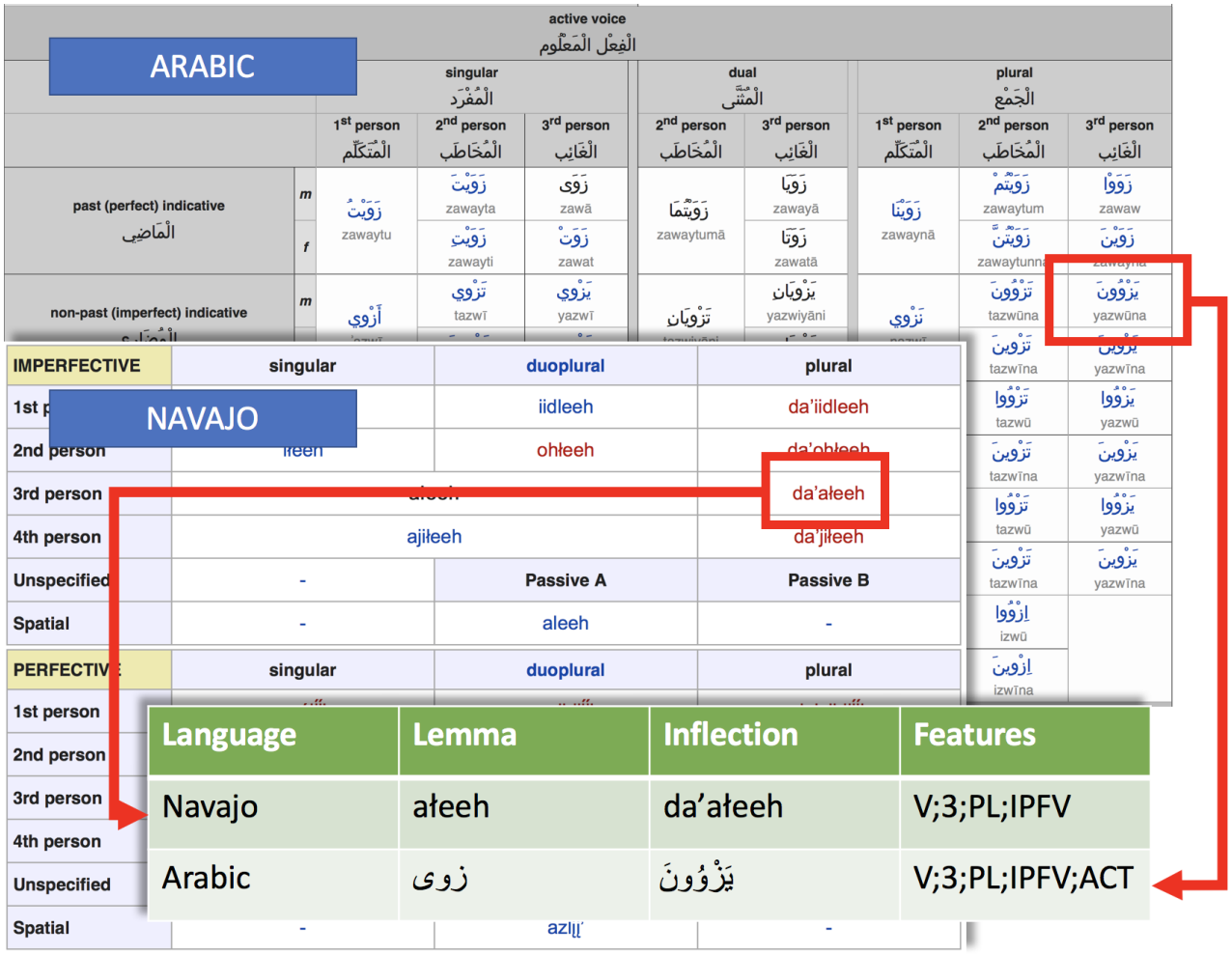


Figure 1: Paradigm extraction and normalization.

the UniMorph 1.0 dataset was very error-prone due to the inability of our heuristics to fully cover the degree of inconsistency found in Wiktionary. For many lemmas, inflected forms were systematically linked to incorrect feature vectors. To correct these errors, we noted that for each part-of-speech within a language in Wiktionary, authors use only a handful of distinct table layouts. Thus, it was sufficient for a human to verify and correct a single lemma parse from a particular layout, and apply those corrections to all similar lemma parses. A custom verification and correction process was created and applied to 8 languages (Arabic, Finnish, Georgian, German, Navajo, Russian, Spanish, and Turkish) in preparation for the SIGMORPHON 2016 Shared Task on Morphological Reinflection (Cotterell et al., 2016a).

For UniMorph 2.0, we noticed that the effort required to verify and correct an automatic parse of a Wiktionary table layout was greater than or equal to the effort required for a human to *directly* annotate a table with UniMorph features instead. Figure 2 illustrates this simplified process. Each language's HTML tables were parsed using Python's `pandas` library (`pandas.pydata.org`) and grouped according to their tabular structure and number of cells. Each group represents a different type of paradigm (e.g., regular verb).

For each group, a sample table was selected, and an annotator replaced each inflected form in the table with the appropriate UniMorph features. All annotation was compliant with the UniMorph Schema, which was designed to represent the full range of semantic distinctions that can be captured by inflectional morphology in any language (Sylak-Glassman et al., 2015a). The schema is similar in form and spirit to other tagset universalization efforts, such as the Universal Dependencies Project (Choi et al., 2015) and Interset (Zeman, 2008), but is designed specifically for typological completeness for inflectional morphology, including a focus on the morphology of especially low-resource languages. It includes over 200 base features distributed among 23 dimensions of meaning (i.e., morphological categories), including both common dimensions like tense and aspect as well as rarer dimensions like evidentiality and switch-reference. Despite the high coverage of the UniMorph tagset, for UniMorph 2.0, annotators were allowed to employ additional 'language specific' LGSPEC(1, 2, 3, etc.) features to mark any missing distinctions, or purely optional form variants that are not associated with a semantic difference. The Spanish imperfect subjunctive, for example, has two interchangeable forms (*-ra* and *-se*):

| | singular | plural |
|---|---|---|
| nominative | лéмма<br>lémma | лéммы<br>lémmy |
| genitive | лéммы<br>lémmy | лéмм<br>lémm |
| dative | лéмме<br>lémme | лéммам<br>lémmam |
| accusative | лéмму<br>lémmu | лéммы<br>lémmy |
| instrumental | лéммой, лéммою<br>lémmoj, lémmoju | лéммами<br>lémmami |
| prepositional | лéмме<br>lémme | лéммах<br>lémmax |

(a) Raw Wiktionary

| | singular | plural |
|---|---|---|
| nominative | лéмма | лéммы |
| genitive | лéммы | лéмм |
| dative | лéмме | лéммам |
| accusative | лéмму | лéммы |
| instrumental | лéммой | лéммами |
| prepositional | лéмме | лéммах |

(b) Unannotated Table

| | singular | plural |
|---|---|---|
| nominative | N;NOM;SG | N;NOM;PL |
| genitive | N;GEN;SG | N;GEN;PL |
| dative | N;DAT;SG | N;DAT;PL |
| accusative | N;ACC;SG | N;ACC;PL |
| instrumental | N;INS;SG | N;INS;PL |
| prepositional | N;ESS;SG | N;ESS;PL |

(c) Annotated Table

Figure 2: Annotation process

| *lemma* | form | features |
|---|---|---|
| *gravitar* | gravitaras | V;SBJV;PST;2;SG;LGSPEC1 |
| *gravitar* | gravitases | V;SBJV;PST;2;SG;LGSPEC2 |

As each example table is identical in structure to all members in the same layout group, annotating just one example allows mapping every inflected form in every table in the group to its corresponding morphological features. This minimizes the human annotation effort required per language, to the point that only 3 annotators were able to produce a complete initial dataset for 47 Wiktionary languages in a matter of days (data for these 47 languages, listed in Table 2, supplants the corresponding language data in the UniMorph 1.0 dataset).

Some of the extracted paradigms from Wiktionary were subject to additional post-processing. In particular, some Wiktionary tables contain multiple forms in the same cell. In the case of multiple forms, we separated them into their own entries. Looking at another Spanish example, we separate *tu* and *vos* forms corresponding to dialect differences in the choice of second person pronoun.

| | | |
|---|---|---|
| *gravitar* | gravitas(tú)<br>gravitás(vos) | V;IND;PRS;2;SG |
| *gravitar* | gravitas | V;SBJV;PST;2;SG;LGSPEC3 |
| *gravitar* | gravitás | V;SBJV;PST;2;SG;LGSPEC4 |

Finally, the content of all initial annotations was also verified as linguistically sensible by a second, larger set of adjudicators who were either native speakers of the language they reviewed or had significant expertise through research. The final dataset sizes are given by language in table 2.

## 3. Non-Wiktionary Data Sources

In additional to our large database of annotated inflected forms derived from Wiktionary, UniMorph 2.0 includes morphological data for several additional languages from non-Wiktionary sources. Data for Khaling, Kurmanji Kurdish, and Sorani Kurdish was derived from the Alexina project (Walther et al., 2013; Walther et al., 2010; Walther and Sagot, 2010).[2] Novel data for Haida, a severely endangered North American language isolate, was prepared by Jordan Lachler (University of Alberta). Basque language data was extracted from a manually designed finite-state morphological analyzer (Alegria et al., 2009). Data for all these additional languages was reformatted to match the Wiktionary-derived data using custom Python scripts. Any dataset-specific annotation was manually mapped to the UniMorph schema standard.

## 4. Supplementary Structured Data

As discussed in Kirov et al. (2016), we also mine additional structured data from Wiktionary. A number of Wiktionary pages contain lists of derived words under the HTML heading 'Related/Derived Terms' — 'sunflower' for example, appears on the list for the base lemma 'flower.' Furthermore, Wiktionary also contains tables of lemma translations. The English lemma 'flower' contains the translation entry 'Danish: blomstre.' As part of UniMorph 1.0, we collected an average of 3.42 derived terms per lemma across 76,038 lemmas, and an average of 3.54 translations per annotated lemma.

For UniMorph 2.0, we are releasing two additional resource types. First, only a subset of Wiktionary languages and lemmas contain embedded morphological tables. There are many more bare lemmas with no form of morphological annotation. We also scrape these lemmas, and provide a list of them along with their associated part of speech.

Second, for a number of languages in UniMorph, we provide multi-word English glosses for complex inflected wordforms. For example, the Spanish word *comprábamos* is mapped to the gloss '(we) were buying.' These glosses are generated for languages where adequately-sized lemma-to-lemma translation dictionaries are available, via the following general process:

1. Perform a generally language-independent conversion of UniMorph feature vectors to an English gloss template, e.g., V;1;PL;PST;IPFV → '(we) were VBG.' Here, VBG is a Penn Treebank tag which indicates that the template can be filled with the *-ing* form of an English verb.

2. Given an inflected lemma in the language with a particular feature vector and lemma translation, find the corresponding gloss template, e.g., *comprábamos*, *comprar*, V;1;PL;IPFV → 'buy: (we) were VBG'

3. Replace the English lemma placeholder in the template with the appropriately generated form of the English lemma, 'buy, (we) were VBG' → '(we) were buying'

[2] https://gforge.inria.fr/projects/alexina/

| Language | Inflections | Glosses |
|---|---|---|
| Amharic | 566553 | 1736981 |
| Farsi | 206711 | 582449 |
| Hausa | 55860 | 124492 |
| Hungarian | 2814006 | 9754197 |
| Oromo | 26690 | 246856 |
| Russian | 560067 | 2219960 |
| Somali | 451217 | 1144096 |
| Spanish | 153121 | 368636 |
| Ukrainian | 20288 | 41590 |
| Yoruba | 127833 | 356502 |
| **Total** | **4982569** | **16575759** |

Table 1: English glosses by language.

Generating complicated tenses of multi-word lemmata (e.g. "They will not have looked it up") and robustly generating appropriate English inflections for diverse and noisy translation dictionaries, are both a challenge and strength of this work.

Table 1 shows the a summary of the current resource sizes of selected languages, along with the number of distinct inflections covered, and the number of expanded phrasal glosses generated given multiple translations per lemma.

## 5. Community Features

Following the model of Universal Dependencies (UD),[3], UniMorph is intended to be a highly collaborative project. To that end, all data and tools associated with the project are released on a rolling basis with a permissive open source license. The main portal for the UniMorph project, which provides a high-level overview of project goals and activities, is `www.unimorph.org`. The hub for downloadable data and resources is `unimorph.github.io`. A full specification of the UniMorph annotation schema is available. For each language, the site indicates how many forms and paradigms have been extracted, the source of the data, and available parts of speech. The site is also designed to encourage community involvement. Each language is associated with a public issue tracker that allows users to discuss errors and issues in the available data and annotations. Interested users can also become part of the UniMorph mailing list.

Moving forward, we also intend to develop connections with other morphological resources. The Universal Dependencies project, for example, provides a token-level corpus complementary to the UniMorph type-level data. A preliminary survey of UD annotations shows that approximately 68% of UD features map directly to UniMorph schema equivalents. This set covers 97.04% of complete UD tags. Some UD features lie outside the current scope of UniMorph, which marks primarily morphosyntactic and morphosemantic distinctions. These include, for example, markers for abbreviated forms and foreign borrowings.

[3] `universaldependencies.org`

| Language | Family | Lemmata / Forms |
|---|---|---|
| Albanian | Indo-European | 589 / 33483 |
| Arabic | Semitic | 4134 / 140003 |
| Armenian | Indo-European | 7033 / 338461 |
| Basque | Isolate | 26 / 11889 |
| Bengali | Indo-Aryan | 136 / 4443 |
| Bulgarian | Slavic | 2468 / 55730 |
| Catalan | Romance | 1547 / 81576 |
| Czech | Slavic | 5125 / 134527 |
| Danish | Germanic | 3193 / 25503 |
| Dutch | Germanic | 4993 / 55467 |
| English | Germanic | 22765 / 115523 |
| Estonian | Uralic | 886 / 38215 |
| Faroese | Germanic | 3077 / 45474 |
| Finnish | Uralic | 57642 / 2490377 |
| French | Romance | 7535 / 367732 |
| Georgian | Kartvelian | 3782 / 74412 |
| German | Germanic | 15060 / 179339 |
| Haida | Isolate | 41 / 7040 |
| Hebrew | Semitic | 510 / 13818 |
| Hindi | Indo-Aryan | 258 / 54438 |
| Hungarian | Uralic | 13989 / 490394 |
| Icelandic | Germanic | 4775 / 76915 |
| Irish | Celtic | 7464 / 107298 |
| Italian | Romance | 10009 / 509574 |
| Khaling | Sino-Tibetan | 591 / 156097 |
| Kurmanji Kurdish | Iranian | 15083 / 216370 |
| Latin | Romance | 17214 / 509182 |
| Latvian | Baltic | 7548 / 136998 |
| Lithuanian | Baltic | 1458 / 34130 |
| Lower Sorbian | Germanic | 994 / 20121 |
| Macedonian | Slavic | 10313 / 168057 |
| Navajo | Athabaskan | 674 / 12354 |
| Northern Sami | Uralic | 2103 / 62677 |
| Norwegian Bokmål | Germanic | 5527 / 19238 |
| Norwegian Nynorsk | Germanic | 4689 / 15319 |
| Persian | Iranian | 273 / 37128 |
| Polish | Slavic | 10185 / 201024 |
| Portuguese | Romance | 4001 / 303996 |
| Quechua | Quechuan | 1006 / 180004 |
| Romanian | Romance | 4405 / 80266 |
| Russian | Slavic | 28068 / 473481 |
| Scottish Gaelic | Celtic | 73 / 781 |
| Serbo-Croatian | Slavic | 24419 / 840799 |
| Slovak | Slavic | 1046 / 14796 |
| Slovene | Slavic | 2535 / 60110 |
| Sorani Kurdish | Iranian | 274 / 22990 |
| Spanish | Romance | 5460 / 382955 |
| Swedish | Germanic | 10553 / 78411 |
| Turkish | Turkic | 3579 / 275460 |
| Ukrainian | Slavic | 1493 / 20904 |
| Urdu | Indo-Aryan | 182 / 12572 |
| Welsh | Celtic | 183 / 10641 |

Table 2: Total number of lemmata and forms available for each language in the morphological database.

## 6. Conclusion

As part of the UniMorph project, we are releasing the largest available database of high-quality morphological paradigms across a typologically-diverse set of languages. To create this dataset, we developed a type-based annotation procedure that enables extracting a large amount of data from Wiktionary with minimal effort from human annotators. The procedure successfully handles idiosyncratic variation in formatting across the languages in Wiktionary. UniMorph also prescribes a universal tagging schema and data formats that allow data to be incorporated from non-

Wiktionary data sources. The project welcomes community involvement, and all data and tools are released under a permissive open-source license at `unimorph.github.io`. UniMorph 2.0 data has already been used as the basis for the successful CoNLL 2017 Shared Task on Morphological Learning, the first shared task on morphology in the CoNLL community (Cotterell et al., 2017).

## 7. Bibliographical References


Ahlberg, M., Forsberg, M., and Hulden, M. (2014). Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 569–578, Gothenburg, Sweden, April. Association for Computational Linguistics.

Alegria, I., Etxeberria, I., Hulden, M., and Maritxalar, M. (2009). Porting Basque morphological grammars to *foma*, an open-source tool. In *International Workshop on Finite-State Methods and Natural Language Processing*, pages 105–113. Springer.

Choi, J., de Marneffe, M.-C., Dozat, T., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Nivre, J., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2015). Universal Dependencies. Accessible at: http://universaldependencies.github.io/docs/, January.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., and Hulden, M. (2016a). The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany, August. Association for Computational Linguistics.

Cotterell, R., Schütze, H., and Eisner, J. (2016b). Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1651–1660, Berlin, Germany, August. Association for Computational Linguistics.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2017). The CoNLL-SIGMORPHON 2017 shared task. In *CoNLL-SIGMORPHON 2017 Shared Task*.

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraclar, M., and Stolcke, A. (2007). Analysis of morph-based speech recognition and the modeling of out-of-vocabulary words across languages. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 380–387. Association for Computational Linguistics.

Durrett, G. and DeNero, J. (2013). Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195, Atlanta, Georgia, June. Association for Computational Linguistics.

Dyer, C., Muresan, S., and Resnik, P. (2008). Generalizing word lattice translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 1012–1020, Columbus, Ohio, June. Association for Computational Linguistics.

Faruqui, M., Tsvetkov, Y., Neubig, G., and Dyer, C. (2016). Morphological inflection generation using character sequence to sequence learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643, San Diego, California, June. Association for Computational Linguistics.

Haspelmath, M., Dryer, M., Gil, D., and Comrie, B. (2005). The world atlas of language structures (WALS).

Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very-large scale parsing and normalization of Wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3121–3126. European Language Resources Association (ELRA), May.

Narasimhan, K., Karakos, D., Schwartz, R., Tsakalidis, S., and Barzilay, R. (2014). Morphological segmentation for keyword spotting. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 880–885, Doha, Qatar, October. Association for Computational Linguistics.

Nicolai, G., Cherry, C., and Kondrak, G. (2015). Inflection generation as discriminative string transduction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 922–931, Denver, Colorado, May–June. Association for Computational Linguistics.

Seeker, W. and Çetinoğlu, O. (2015). A graph-based lattice dependency parser for joint morphological segmentation and syntactic analysis. *Transactions of the Association for Computational Linguistics*, 3:359–373.

Sylak-Glassman, J., Kirov, C., Post, M., Que, R., and Yarowsky, D. (2015a). A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In Cerstin Mahlow et al., editors, *Proceedings of the 4th Workshop on Systems and Frameworks for Computational Morphology (SFCM)*, Communications in Computer and Information Science, pages 72–93. Springer, Berlin, September.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., and Que, R. (2015b). A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 674–680, Beijing, July. Association for Computational Linguistics.

Walther, G. and Sagot, B. (2010). Developing a large-scale lexicon for a less-resourced language: General methodology and preliminary experiments on Sorani Kurdish.

In *Proceedings of the SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages (at LREC)*, Valetta, Malta. European Language Resources Association (ELRA).

Walther, G., Sagot, B., and Fort, K. (2010). Fast development of basic NLP tools: Towards a lexicon and a POS tagger for Kurmanji Kurdish. In *Proceedings of the 29th International Conference on Lexis and Grammar*, Belgrade.

Walther, G., Jacques, G., and Sagot, B. (2013). Uncovering the inner architecture of Khaling verbal morphology, September. Presentation at the 3rd Workshop on Sino-Tibetan Languages of Sichuan, Paris, September 2013.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *Proceedings of LREC 2008*, pages 213–218.