Wikinflection Corpus: A (Better) Multilingual, Morpheme-Annotated Inflectional Corpus

Eleni Metheniti* †, Günter Neumann ‡
*CLLE-CNRS, †IRIT-CNRS , ‡DFKI
* †Toulouse, France ‡Saarbrücken, Germany
eleni.metheniti@{*univ-tlse2.fr, †irit.fr}, ‡neumann@dfki.de

Abstract

Multilingual, inflectional corpora are a scarce resource in the NLP community, especially corpora with annotated morpheme boundaries. We are evaluating a generated, multilingual inflectional corpus with morpheme boundaries, generated from the English Wiktionary (Metheniti and Neumann, 2018), against the largest, multilingual, high-quality inflectional corpus of the UniMorph project (Kirov et al., 2018). We confirm that the generated Wikinflection corpus is not of such quality as UniMorph, but we were able to extract a significant amount of words from the intersection of the two corpora. Our Wikinflection corpus benefits from the morpheme segmentations of Wiktionary/Wikinflection and from the manually-evaluated morphological feature tags of the UniMorph project, and has 216K lemmas and 5.4M word forms, in a total of 68 languages.

Keywords: inflection, morphological inflection, inflectional corpus, inflection annotation

1. Introduction

Inflection is the linguistic process in which a word acquires morphological features, which allow it to create syntactic dependencies with its context or express an additional nuance without changing the word's core meaning, e.g. number, time. To create inflection on a word-surface level, different languages make different choices on what transformations will occur to the stem of the word, what morphemes will be added (at the end, start, middle, and how many of them). Knowledge on the way a language creates inflections is crucial, in order to be able to identify and create different forms of the base word, the *lemma*, (re-inflection) and be able to reverse the process (lemmatization).

In this paper, we are describing our method of creating a multilingual inflectional corpus, using two available inflectional corpora; the corpus from the UniMorph project (Kirov et al., 2018), and the corpus we generated from the paper and code of the Wikinflection project (Metheniti and Neumann, 2018). The UniMorph corpus is the largest to date inflectional corpus, including 108 languages and ∼10.7 word forms, and has been extensively evaluated and enriched with other resources. However, it only includes entire word forms, without information on a subword level; meanwhile, Wikinflection generates a corpus of 6.4M words in ∼140 languages, and offers the morpheme segmentations of word forms as they exist in the English Wiktionary templates. We test the robustness and quality of the generated Wikinflection corpus by generating an iteration of the Wikinflection corpus and running our old evaluation script. In addition, we also evaluate Wikinflection with the UniMorph corpus. Results show that Wikinflection is of lower quality than UniMorph, due to the lack of manual evaluation, with the most serious problem being the inability to capture all grammatical tags for some word forms.

From our findings, we were also able to create a new corpus, the Wikinflection corpus, from the generated corpus of Wikinflection and the UniMorph corpus, by evaluating Wikinflection with UniMorph. Our corpus has 216K lemmas and 5.4M word forms, in 68 languages, with the morpheme segmentations for every word form from Wikinflection and evaluated on word forms with UniMorph and using the morphological feature tags of UniMorph, converted to the Universal Dependencies tag set (Nivre et al., 2018). Our corpus is released on GitHub[1].

2. Previous Work

There is a limited amount of corpora with inflectional information made available to the NLP community, because manual segmentation and evaluation are difficult, costly and time-consuming. Automatic segmentation is favoured, with semi-supervised or unsupervised methods (Cotterell et al., 2015; Ruokolainen et al., 2016; Cotterell et al., 2016). Only a few corpora have a significant number of entries that are annotated for morphological inflections and with segmented words; *CELEX* for English, Russian, German and Dutch (Baayen et al., 1996), the *Tübingen Treebank of Written German* for German (Telljohann et al., 2004), *Korpus 2000* for Danish (Asmussen, 2001), *Corpus Of Serbian Language (CSL)* for Serbian (Kostič, 2001), *Stockholm Umeå Corpus* for Swedish (Ejerhed et al., 2006), the dataset of the *Morpho Challenge* for English, Finnish and Turkish (Kurimo et al., 2010), *Italian Content Words v3* for Italian (Grella, 2018).

A widely-used source for gathering data and creating corpora is Wiktionary, the open-access, crowd-sourced multilingual dictionary of the Wikimedia foundation. The Wiktionary, in its many source languages, offers various linguistic information on the target words; while some is easily accessible from the XML dump files, such as translations (Acs et al., 2013) and lexical-semantic information (Zesch et al., 2008), other information is not explicitly present, but is generated. The conjugation and declension of lemmas is, unfortunately, not readily available via

[1] github.com/lenakmeth/Wikinflection-Corpus

the parsing tools and the resources made available by the Wikimedia foundation, but is created with the use of *inflectional templates* and *inflectional modules*. Every time the online HTML page of a lemma is accessed, the inflectional templates (human-readable templates of inflectional classes, created by the Wiktionary community) and modules (machine-readable code in `Lua`) linked to this lemma are used to generate the inflectional tables in the page. Therefore, the inflectional paradigm is not explicitly embedded in the static XML file that generates the online page[2].

However, there have been some attempts to mine Wiktionary for its inflectional information. *IWNLP* (Liebeck and Conrad, 2015) is a parser for the German Wiktionary, which is able to access the lemma's inflectional template and recreate its inflectional paradigm. Their method involved re-implementing the `Lua` modules for inflectional templates of Wiktionary into `C#`, and then using these inflectional templates alongside lemmas to generate inflectional paradigms with inflectional segmentations (as given by the templates). While they achieved very high quality and accuracy, their method requires great effort and is only used on a fraction of the German words in Wiktionary.

The *Universal Morphology* (UniMorph) project is a long-standing project which has released the largest multilingual inflectional corpus to date (UniMorph 2.0), generated by the English Wiktionary (Kirov et al., 2018). Their approach to extracting the relevant information from the Wiktionary is different than that of Liebeck and Conrad (2015); they used the static HTML pages of Wiktionary instead of the offline resources, which allowed them to capture word forms that are explicitly written and not generated by a template or word forms which are dictionary entries but not lemmas. These word forms, however, are not segmented to their morphemes, as such information is only available through templates in Wiktionary. They annotated their inflectional paradigms with their own UniMorph schema (Sylak-Glassman et al., 2015), which aims to capture all morphological features of human languages in one unified notation. Unlike the 1.0 version of UniMorph (Kirov et al., 2016), the UniMorph 2.0 corpus is generated by the inflectional tables of the lemma, the tables were grouped based on similarity, and human annotators evaluated, annotated and corrected the pairs of word forms and generated morphological feature tags. This ensured that the UniMorph 2.0 corpus has gold-standard word forms and high-quality annotations for all the 400K inflectional paradigms and 10.7M word forms present.

Another approach to mine the English Wiktionary for inflectional information is *Wikinflection* (Metheniti and Neumann, 2018). Our approach made use of the static HTML inflectional tables of the Wiktionary, to recreate inflectional templates which could then be associated (with the Wiktionary's dynamic links) and used with lemmas in the Wiktionary XML dump file, to generate inflectional paradigms with inflectional morpheme boundaries, where the segments added by the template are considered to be the word's inflectional morphemes. We also offered a script to randomly evaluate the inflectional templates, by randomly selecting one corresponding lemma and inflectional paradigm for each template, and checking the lemma's HTML page on Wiktionary. Before evaluation, our generated corpus has 225.453 inflectional paradigms and 1.708 inflectional templates, generating 8.426.480 inflected words, in a total of 199 languages. After performing some random evaluations, we reported different numbers for each evaluation; random evaluation 3 returned 210.172 inflectional paradigms and 1.521 templates, and 6.024.077 word forms for 138 languages. We have thoroughly documented the shortcomings of generating inflections without extensive and human evaluation, mostly due to the conflicting styles and templates used across different target languages in the English Wiktionary; our approach to massively gather information from the diversely structured tables of the Wiktionary led sometimes to partial loss of morphological tags. In addition, we were also critical of our own method of evaluation, because it is prone to errors and returns different results in every evaluation run.

3. Re-evaluating Wikinflection

In our previous paper, we (Metheniti and Neumann, 2018) presented our method of generating inflections by reverse-engineering the process in which the Wiktionary server generates inflections on command, every time the page of a dictionary entry is loaded. From the Wiktionary XML dump file, we find words that are lemmas and have inflectional information, we gather the dynamic links that connect a lemma to its corresponding inflectional template, and then we look up the HTML pages for these inflectional templates. With the parsed templates and the information provided in the dynamic link (stem, stem allomorphs, phonetic additions), we expect to be able to exactly recreate the inflectional paradigms as presented in the HTML page of a lemma, and maintain the morpheme boundaries among the stem/stem allomorph and the inflectional morphemes.

A method of evaluation, however, is necessary; for each of the 1.708 unique inflectional templates, we randomly choose a lemma associated with that template, we look up the online HTML page for the lemma in Wiktionary, and we remove any generated word forms which were not found in the HTML page. This method of evaluation was selected because there are not large enough corpora for all the 199 languages for which Wikinflection has generated paradigms, and even in high-resource languages, some inflected types are very rare. However, this method of evaluation does not guarantee gold-standard quality; as the authors document, three different executions of our evaluation script produced corpora of different sizes; for example, when the template for Latin second declension was evaluated with the noun *campus* "campus", all word forms were deemed correct and thus no corrections were made in the template, but when it was evaluated with the proper noun *Herostratus*, the word forms associated with the plural number were not found because the proper noun does not exist in plural.

We decided to run the Wikinflection code, generate the corpus via the Metheniti and Neumann (2018) script and

[2]We have previously explained in detail the process of generating inflections in Wiktionary in Metheniti and Neumann (2018).

Corpus type	*No. languages*	*Inflectional templates*	*No. Lemmas*	*No. Words*
Non-evaluated	149	1.810	274.798	9.320.503
Wiktionary Evaluation	140	1.614	254.712	6.447.613
UniMorph Evaluation	68	977	216.624	5.410.746

Table 1: The size of the generated Wikinflection corpus (number of languages present, number of (correct and updated) inflectional templates, number of lemmas and number of inflected word forms. The first row refers to the generated corpus from Wikinflection, without evaluation, the second row refers to the Wikinflection corpus after the one random evaluation we performed using Wiktionary and our Wikinflection script, and the third row refers to the evaluation we performed using the UniMorph corpus.

then, at first, evaluate with the evaluation script we previously used and provided in the Wikinflection repository [3]. We used the latest English Wiktionary XML dump file (November 21, 2019) alongside the `Python3` code. The results of the generated corpus, before evaluation, are shown in the first row of Table 1; we assume that our numbers are higher than the ones reported for the generated corpus in Metheniti and Neumann (2018) (mentioned in Section 2.) because the Wiktionary is constantly adding new dictionary entries and improving the existing templates. We then ran the evaluation script as provided by the authors, and we report the results in the second row of Table 1. The numbers are, foreseeably, lower than in the non-evaluated corpus, and are close to the evaluation runs that the authors have reported in our previous work. However, we still cannot guarantee that these evaluated, generated paradigms are of high quality; the evaluation log showed us which random lemma was picked for each inflectional template and how many word forms were corrected from each template. We manually checked a few of the corrected templates, by accessing the HTML pages of the inflectional templates and the chosen lemmas, and we noticed the problems with the evaluation. For example, the inflectional template `es-conj-ír`[4] for verbs ending in *-ír* in Spanish; even after the evaluation, there were four duplicate forms of the infinitive form in the template. On the other hand, the inflectional template `fi-decl-käsi-kulkija`[5] for Finnish nouns was erroneously found entirely incorrect, because the randomly chosen word for this template, *Uusi-Kaledonia*[6] "New Caledonia", is a proper noun without plural, and different rules in Wiktionary applied outside of inflectional templates implement these transformations in the lemma's page.

Since, in order to perform an evaluation on all the Wikinflection languages, we need a source of inflected forms as big and multilingual as in the Wiktionary, we considered that we could use the UniMorph corpus for evaluation, as it is also created with the use of the English Wiktionary, is larger than the Wikinflection corpus, and is manually evaluated. We downloaded the current versions of the repositories for each available language, and in Table 2 we are presenting the current size of each language's corpus. We ran the evaluation in the same method followed as in Metheniti and Neumann (2018); we are randomly selecting one corresponding lemma for every non-evaluated template, we are generating its inflectional paradigm, and then we are checking the existence of each word form in the UniMorph corpus. We then check, for each word, whether its morphological tags from the generated template and the morphological tags of UniMorph are a (partial) match[7], and if they are, we overwrite them with the UniMorph features in the corresponding position of the template. In order to perform this, we have to convert the UniMorph schema to the Universal Dependencies schema (Nivre et al., 2018), because Wikinflection is built on Universal Dependencies. We used a conversion list provided by McCarthy et al. (2018) in their Github repository [8].

We are presenting the results of our evaluation with UniMorph in Table 1 as well. First of all, we notice a drop in the remaining languages, because UniMorph has 108 languages compared to the 149 in Wikinflection, and the reason that we have even fewer is that UniMorph was enriched with sources other than Wiktionary (e.g. corpora in Italian, Grella (2018)), or some languages in Wiktionary do not have any inflectional templates with significant inflectional information (e.g. Ancient Greek). The number of inflectional templates has been halved, but we can deduce that the remaining templates are of good quality since they have been evaluated with UniMorph, and we also ensured to get rid of the duplicate word forms in the templates. We notice that the Wikinflection corpus lacks some high-resource languages present in UniMorph, such as French, or has a dramatically smaller number of lemmas and word forms, such as in Arabic and Portuguese. This is probably due to the lack of or the bad structure of the inflectional templates for these languages in Wiktionary, which Wikinflection was not able to parse. However, we notice that for some languages (e.g. Ingrian, Veps) the evaluated Wikinflection has more word forms than UniMorph; this is because we evaluate the templates, and generate the word forms. It is possible that the UniMorph project has not fully queried all the HTML pages for dictionary entries for a language, but since we are using the XML file, we have access to all dictionary entries and can create inflections for whichever lemmas are linked with inflectional templates.

In order to check the improvement in quality that the evaluation with UniMorph brought, we decided to pick one of the randomly chosen words from the previous Wiktionary evaluation, so that we can compare all the evaluated and non evaluated outputs with a human evaluation. We se-

[3] `github.com/lenakmeth/Wikinflection`

[4] `en.wiktionary.org/wiki/Template:es-conj-%C3%ADr`

[5] `en.wiktionary.org/wiki/Template:fi-decl-k%C3%A4si-kulkija`

[6] `en.wiktionary.org/wiki/Uusi-Kaledonia`

[7] We are not looking for a one-to-one match of the UniMorph and the WIkinflection tags, as long as the word form matches. Wikinflection's tags are problematic, and it would be impossible to get a perfect match.

[8] `github.com/unimorph/ud-compatibility`

lected[9] the Spanish verb *sofreír* "to fry lightly", which is linked to the inflectional template `es-conj-ír` which was problematic in the Wiktionary evaluation, and we examined the paradigms from the corpora and the evaluations. We are presenting four different paradigms of this lemma; the non-evaluated lemma from Wikinflection (Table 4), the paradigm from the UniMorph corpus (Table 5), the Wiktionary-evaluated, generated paradigm from Wikinflection (Table 6) and the UniMorph-evaluated generated paradigm (Table 7). First, we notice the many duplicate forms in the Wikiflection paradigms, a problem which also exists in the Wiktionary-evaluated corpus, but is dealt with with our UniMorph evaluation. Second, we notice that in Wikinflection, both in the non-evaluated and the evaluated versions, there are word forms with incomplete tags; by copying the UniMorph tags, we are ensuring the complete morphological annotation of each word form. Overall, our generated and evaluated paradigm is not complete, but the word forms present are unique, correct and fully annotated both for morpheme boundaries and morphological features.

4. A new corpus: Wikinflection + UniMorph

Following this evaluation process, we believe that it would be beneficial to release the result of evaluating Wikinflection with UniMorph as an open-access resource. This corpus includes the segmentations of inflectional morphemes for every word form, as found in the Wiktionary templates and in Wikinflection, is evaluated with the use of the manually-evaluated UniMorph, and has UniMorph's manually-annotated morphological tags. We opted for the use of the Universal Dependencies 2 schema, as opposed to the original UniMorph, since we have already converted the tags to the UD tags on the evaluation stage, and also because currently the UD schema is more commonly used in NLP applications (e.g. dependency parsers). The full list of languages and lemmas and word forms per language can be found in Table 3. We are also going to make use of the ISO 639-3 language codes, as UniMorph does, to ensure uniform language tags; both the Wiktionary and Wikinflection use different versions of the ISO 639 protocol for languages (e.g. the tag *fi* for Finnish is from ISO 639-1, and ISO 639-3 uses *fin*). An example of the format of our corpus can be seen in Table 7, for the lemma *sofreír*.

As happened during the the evaluation of Wikinflection with the Wiktionary, we run in the risk of false negatives, when evaluating an inflectional template with a random lemma that possibly does not have all forms of a template. Therefore, we ran the evaluation twice, and the corpus and the numbers we report in Table 3 are the results of two evaluation runs with UniMorph and their combination.

5. Discussion

Our work started as an effort to assess the quality of the Wikinflection corpus, the largest inflectional corpus to date (openly available) to include inflectional morpheme boundaries. We discovered that the corpus, because it is automatically generated, has several weaknesses, especially when compared to UniMorph, a carefully-evaluated inflectional corpus from a long-standing research project. However, we were able to evaluate Wikinflection with the use of UniMorph, and discovered that a significant portion of Wikinflection is on par with the UniMorph standards. From our evaluation results, we used the intersection of the two corpora to create a new one, with the strengths of each: the morpheme boundaries of Wikinflection and the gold-standard morphological tags of UniMorph.

We are aware of the weaknesses of our work; our inflectional paradigms are not complete, because of the method the inflectional information is parsed by and included in Wikinflection. We also aim to perform a more in-depth analysis of the corpora, on a manual level, to ensure that there are no false positives/negatives cause by the inflectional templates of Wiktionary and Wikinflection. However, we are confident that the quality of the Wikinflection corpus, after the evaluation with UniMorph, is on par with the UniMorph standards, for the present word forms and morphological feature tags.

Finally, we would like to address the fact that the inflectional morphemes, in Wiktionary and subsequently in Wikinflection and our corpus as well, are in most cases composites and not broken down to the smallest possible units; for example, *sofreiremos* "we will fry lightly" (see Tables 4-7) is decomposed to *sofre-iremos*, but should have been analyzed as:

$$\textit{sofre-ir}_{[+future]}\textit{-emos}_{[+1st.pers.plural]}$$

This is a weakness of the way inflectional templates in Wiktionary are crafted, and if we would like to deal with this issue (which could prove serious for agglutinative languages such as Finnish), we would need language-specific knowledge and many resources, to ensure gold-standard quality. We aim to address this in a following edition of our corpus, but even with this problem, we see merit in releasing our work to be freely used by the NLP community, especially since our corpus currently includes a meaningful number of word forms for many low-resource languages (Ingrian, Pashto, Classic Syriac, Veps).

6. Acknowledgements

This work was partially funded by the European Union's Horizon 2020 grant agreements No. 731724 (iREAD) and No. 777107 (Precise4Q), and by the BMBF project DeepLee ((01IW17001). We would like to sincerely thank Arya McCarthy for his help during the last months of our project, and we would also like to thank Nabil Hathout and Tim Van de Cruys for their support.

7. Bibliographical References

Acs, J., Pajkossy, K., and Kornai, A. (2013). Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, August. Association for Computational Linguistics.

Asmussen, J. (2001). Korpus 2000. *Korpuslingvistik (NyS30)*.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1996). Celex2. *Linguistic Data Consortium, Philadelphia*.

[9] Out of the randomly evaluated words and templates, we chose a language and grammar which we are familiar with, in order to also be able to evaluate the word forms personally.

Cotterell, R., Müller, T., Fraser, A., and Schütze, H. (2015). Labeled morphological segmentation with semi-Markov models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 164–174, Beijing, China, July. Association for Computational Linguistics.

Cotterell, R., Kumar, A., and Schütze, H. (2016). Morphological segmentation inside-out. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2330, Austin, Texas, November. Association for Computational Linguistics.

Ejerhed, E., Källgren, G., and Brodda, B. (2006). Stockholm-Umeå corpus version 2.0. *Stockholm University, Dep. of Linguistics and Umeå University, Dep. of Linguistics*.

Grella, M. (2018). Italian content words v3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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'16)*, pages 3121–3126, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S., McCarthy, A., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2018). UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Kostič, D. (2001). Kvantitativni opis strukture srpskog jezika – Korpus srpskog jezika. [Quantitative description of Serbian language structure – Corpus of Serbian language.]. *Belgrade, Serbia: Institute for Experimental Phonetics and Speech Pathology, Belgrade and Laboratory for Experimental Psychology, University of Belgrade*.

Kurimo, M., Virpioja, S., Turunen, V. T., et al. (2010). Proceedings of the Morpho Challenge 2010 Workshop. In *Morpho Challenge Workshop; 2010; Espoo*. Aalto University School of Science and Technology.

Liebeck, M. and Conrad, S. (2015). IWNLP: Inverse Wiktionary for natural language processing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 414–418, Beijing, China, July. Association for Computational Linguistics.

McCarthy, A. D., Silfverberg, M., Cotterell, R., Hulden, M., and Yarowsky, D. (2018). Marrying Universal Dependencies and Universal Morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101, Brussels, Belgium, November. Association for Computational Linguistics.

Metheniti, E. and Neumann, G. (2018). Wikinflection: Massive Semi-Supervised Generation of Multilingual Inflectional Corpus from Wiktionary. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018), December 13–14, 2018, Oslo University, Norway*, number 155, pages 147–161. Linköping University Electronic Press.

Nivre, J., Abrams, M., Agić, Ž., et al. (2018). Universal Dependencies 2.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ruokolainen, T., Kohonen, O., Sirts, K., Grönroos, S.-A., Kurimo, M., and Virpioja, S. (2016). A comparative study of minimally supervised morphological segmentation. *Computational Linguistics*, 42(1):91–120, March.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., and Que, R. (2015). 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 (Volume 2: Short Papers)*, pages 674–680, Beijing, China, July. Association for Computational Linguistics.

Telljohann, H., Hinrichs, E., and Kübler, S. (2004). The Tüba-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Language	*ISO 639-3*	*Words*	*Lemmas*
Adyghe	**ady**	20475	1666
Old English	**ang**	42425	1867
Arabic	**ara**	140003	4134
Mapudungun	**arn**	783	26
Asturian	**ast**	29797	436
Azerbaijani	**aze**	8004	340
Bashkir	**bak**	12168	1084
Belarusian	**bel**	16113	1027
Bengali	**ben**	4443	136
Tibetan	**bod**	353	65
Breton	**bre**	2294	44
Bulgarian	**bul**	55730	2468
Catalan	**cat**	81576	1547
Czech	**ces**	134527	5125
Old Church Slavonic	**chu**	4148	152
Central Kurdish	**ckb**	22990	274
Cornish	**cor**	469	9
Crimean Tatar	**crh**	7514	1230
Kashubian	**csb**	509	37
Welsh	**cym**	10641	183
Danish	**dan**	25503	3193
German	**deu**	179339	15060
Lower Sorbian	**dsb**	20121	994
Modern Greek	**ell**	199763	11906
English	**eng**	115523	22765
Estonian	**est**	38215	886
Basque	**eus**	11889	26
Faroese	**fao**	45474	3077
Persian	**fas**	37128	273
Finnish	**fin**	2490377	57642
French	**fra**	367732	7535
Middle French	**frm**	36970	603
Old French	**fro**	123374	1700
North Frisian	**frr**	3204	51
West Frisian	**fry**	1429	85
Friulian	**fur**	8071	168
Galician	**gal**	36801	486
Scottish Gaelic	**gla**	781	73
Irish	**gle**	107298	7464
Manx	**glv**	14	1
Middle High German	**gmh**	708	29
Middle Low German	**gml**	1513	52
Gothic	**got**	0*	0*
Ancient Greek	**grc**	41593	2431
Haida	**hai**	7040	41
Serbo Croatian	**hbs**	840799	24419
Hebrew	**heb**	13818	510
Hindi	**hin**	54438	258
Hungarian	**hun**	490394	13989
Armenian	**hye**	338461	7033
Icelandic	**isl**	76915	4775
Italian	**ita**	509574	10009
Ingrian	**izh**	1099	50
Greenlandic	**kal**	368	23
Kannada	**kan**	6402	159
Georgian	**kat**	74412	3782
Kazakh	**kaz**	357	26
Kabardian	**kbd**	3092	250
Khakas	**kjh**	1200	75
Khaling	**klr**	156097	591
Northern Kurdish	**kmr**	216370	15083
Karelian	**krl**	682	20
Latin	**lat**	509182	17214
Latvian	**lav**	136998	7548
Lithuanian	**lit**	34130	1458
Livonian	**liv**	3987	203
Ladin	**lld**	180	7656
Ludian	**lud**	400*	124*
Macedonian	**mkd**	168057	10313
Maltese	**mlt**	3584	112
Murrinhpatha	**mwf**	1110	29
Neapolitan	**nap**	1808	40
Navajo	**nav**	12354	674
Low German	**nds**	0*	0*
Dutch	**nld**	55467	4993
Norwegian Nynorsk	**nno**	15319	4689
Norwegian Bokmål	**nob**	19238	5527
Occitan	**oci**	8316	174
Livonian languages	**olo**	3187*	203*
Old Saxon	**osx**	22287	863
Polish	**pol**	201024	10185
Portuguese	**por**	303996	4001
Pashto	**pus**	6945	395
Quechua	**que**	180004	1006
Romanian	**ron**	80266	4405
Russian	**rus**	473481	28068
Sanskrit	**san**	33847	917
Old Irish	**sga**	1089	49
Slovenian	**slv**	60110	2535
Northern Sami	**sme**	62677	2103
Spanish	**spa**	382955	5460
Albanian	**sqi**	33483	589
Swahili	**swc**	10092	100
Swedish	**swe**	78411	10553
Classical Syriac	**syc**	3652	160
Tatar	**tat**	7832	1283
Telugu	**tel**	1548	127
Tajik	**tgk**	77	75
Turkmen	**tuk**	810	68
Turkish	**tur**	275460	3579
Ukrainian	**ukr**	20904	1493
Urdu	**urd**	12572	182
Uzbek	**uzb**	1260	15
Venetian	**vec**	18227	368
Veps	**vep**	33196*	868*
Votic	**vot**	1430	55
Classical Armenian	**xcl**	97181	4300
Norman	**xno**	280	5
Yiddish	**yid**	7986	803
Zulu	**zul**	49119	566
TOTAL		**10688113**	**385573**

Table 2: The languages of the UniMorph 2.0 corpus and their current statistics. With an asterisk (*) are marked the languages for which there are available resources on the Github repository of the project (`github.com/unimorph`) but no statistics were mentioned in the project's website (`unimorph.github.io`) as of November 21, 2019. We collected the relevant repositories and added the counts to this table. Please note that the UniMorph project claims to have corpora for 110 languages, but only 108 are currently available (accessed November 21, 2019).

Language	*ISO 639-3*	*Words*	*Lemmas*
Adyghe	**ady**	56	4
Old English	**ang**	56833	4332
Arabic	**ara**	36	6
Asturian	**ast**	30141	373
Belarusian	**bel**	378	35
Tibetan	**bod**	24	8
Bulgarian	**bul**	344	108
Catalan, Valencian	**cat**	2042	28
Czech	**ces**	32	3
Church Slavic	**chu**	1252	124
Welsh	**cym**	25423	401
Danish	**dan**	12040	5999
German	**deu**	47388	5234
Lower Sorbian	**dsb**	18844	914
Modern Greek	**ell**	4	1
English	**eng**	74	25
Estonian	**est**	85200	1878
Basque	**eus**	78	2
Faroese	**fao**	41816	3446
Persian	**fas**	10668	94
Finnish	**fin**	3046391	94609
Middle French	**frm**	632	34
Old French	**fro**	15	3
Irish	**gle**	48055	10118
Middle High German	**gmh**	84	9
Middle Low German	**gml**	495	40
Gothic	**got**	13447	1052
Ancient Greek	**grc**	366	72
Hindi	**hin**	188	96
Hungarian	**hun**	467958	24099
Armenian	**hye**	3715	317
Icelandic	**isl**	488	88
Ingrian	**izh**	19712	64
Kalaallisut, Greenlandic	**kal**	544	34
Kannada	**kan**	9	3
Georgian	**kat**	73833	3978
Kazakh	**kaz**	1281	183
Kabardian	**kbd**	56	4
Khakas	**kjh**	24	6
Latin	**lat**	187389	13451
Latvian	**lav**	88841	8061
Lithuanian	**lit**	29932	1497
Livonian languages	**liv**	412	289
Macedonian	**mkd**	39805	4881
Maltese	**mlt**	64	36
Neapolitan	**nap**	108	36
Low German	**nds**	1478	238
Dutch	**nld**	19816	4710
Occitan	**oci**	808	404
Old Saxon	**osx**	15167	838
Polish	**pol**	142805	4878
Portuguese	**por**	1059	353
Pashto	**pus**	3227	469
Romanian	**ron**	7879	1507
Old Irish	**sga**	752	88
Slovenian	**slv**	113	15
Northern Sami	**sme**	45693	3460
Spanish, Castilian	**spa**	712070	7277
Albanian	**sqi**	7307	207
Swedish	**swe**	14891	1761
Classic Syriac	**syc**	58676	1692
Turkmen	**tuk**	1092	91
Turkish	**tur**	2823	883
Ukrainian	**ukr**	42	3
Urdu	**urd**	722	223
Veps	**vep**	12300	820
Votic	**vot**	528	66
Classical Armenian	**xcl**	4981	566
TOTAL		**5410746**	**216624**

Table 3: The languages and sizes of our corpus, created by the intersection of the Wikinflection and UniMorph corpora.

Word	Template	Features	POS	Prefix	Suffix	Infix	Stem
sofreír	es-conj-ír	VerbForm=Inf	VERB	-	ír	-	sofre
sofreír	es-conj-ír	VerbForm=Inf	VERB	-	ír	-	sofre
sofreír	es-conj-ír	VerbForm=Inf	VERB	-	ír	-	sofre
sofreír	es-conj-ír	VerbForm=Inf	VERB	-	ír	-	sofre
sofreír	es-conj-ír	VerbForm=Inf	VERB	-	ír	-	sofre
sofreiendo	es-conj-ír	VerbForm=Ger	VERB	-	iendo	-	sofre
sofreiendo	es-conj-ír	VerbForm=Ger	VERB	-	iendo	-	sofre
sofreiendo	es-conj-ír	VerbForm=Ger	VERB	-	iendo	-	sofre
sofreiendo	es-conj-ír	VerbForm=Ger	VERB	-	iendo	-	sofre
sofreiendo	es-conj-ír	VerbForm=Ger	VERB	-	iendo	-	sofre
sofreído	es-conj-ír	VerbForm=Ger;Number=Sing	VERB	-	ído	-	sofre
sofreído	es-conj-ír	Number=Sing	VERB	-	ído	-	sofre
sofreída	es-conj-ír	Number=Sing	VERB	-	ída	-	sofre
sofreída	es-conj-ír	Number=Sing	VERB	-	ída	-	sofre
sofreídos	es-conj-ír	VerbForm=Ger;Number=Plur	VERB	-	ídos	-	sofre
sofreídos	es-conj-ír	Number=Plur	VERB	-	ídos	-	sofre
sofreídas	es-conj-ír	Number=Plur	VERB	-	ídas	-	sofre
sofreídas	es-conj-ír	Number=Plur	VERB	-	ídas	-	sofre
sofreo	es-conj-ír	VerbForm=Ger;Number=Sing; Person=3;Mood=Ind;Tense=Pres	VERB	-	o	-	sofre
sofrees	es-conj-ír	Number=Sing;Person=1;Mood=Ind;Tense=Pres	VERB	-	es	-	sofre
sofreís	es-conj-ír	Number=Sing;Person=1;Mood=Ind;Tense=Pres	VERB	-	ís	-	sofre
sofree	es-conj-ír	Number=Sing;Person=2;Mood=Ind;Tense=Pres	VERB	-	e	-	sofre
sofreímos	es-conj-ír	Number=Plur;Person=3;Mood=Ind;Tense=Pres	VERB	-	ímos	-	sofre
sofreís	es-conj-ír	Number=Plur;Mood=Ind;Tense=Pres	VERB	-	ís	-	sofre
sofreen	es-conj-ír	Number=Plur;Mood=Ind;Tense=Pres	VERB	-	en	-	sofre
sofreía	es-conj-ír	VerbForm=Ger;Number=Sing; Person=3;Mood=Ind;Aspect=Imp;Tense=Imp	VERB	-	ía	-	sofre
sofreías	es-conj-ír	Number=Sing;Person=1;Mood=Ind; Aspect=Imp;Tense=Imp	VERB	-	ías	-	sofre
sofreía	es-conj-ír	Number=Sing;Person=2;Mood=Ind; Aspect=Imp;Tense=Imp	VERB	-	ía	-	sofre
sofreíamos	es-conj-ír	Number=Plur;Person=3;Mood=Ind; Aspect=Imp;Tense=Imp	VERB	-	íamos	-	sofre
sofreíais	es-conj-ír	Number=Plur;Mood=Ind;Aspect=Imp;Tense=Imp	VERB	-	íais	-	sofre
sofreían	es-conj-ír	Number=Plur;Mood=Ind;Aspect=Imp;Tense=Imp	VERB	-	ían	-	sofre
sofreí	es-conj-ír	VerbForm=Ger;Number=Sing;Person=3; Mood=Ind;Tense=Past	VERB	-	í	-	sofre
sofreíste	es-conj-ír	Number=Sing;Person=1;Mood=Ind;Tense=Past	VERB	-	íste	-	sofre
sofreió	es-conj-ír	Number=Sing;Person=2;Mood=Ind;Tense=Past	VERB	-	ió	-	sofre
sofreímos	es-conj-ír	Number=Plur;Person=3;Mood=Ind;Tense=Past	VERB	-	ímos	-	sofre
sofreísteis	es-conj-ír	Number=Plur;Mood=Ind;Tense=Past	VERB	-	ísteis	-	sofre
sofreieron	es-conj-ír	Number=Plur;Mood=Ind;Tense=Past	VERB	-	ieron	-	sofre
sofreíré	es-conj-ír	VerbForm=Ger;Number=Sing;Person=3; Mood=Ind;Tense=Fut	VERB	-	iré	-	sofre
sofreírás	es-conj-ír	Number=Sing;Person=1;Mood=Ind;Tense=Fut	VERB	-	irás	-	sofre
sofreírá	es-conj-ír	Number=Sing;Person=2;Mood=Ind;Tense=Fut	VERB	-	irá	-	sofre
sofreíremos	es-conj-ír	Number=Plur;Person=3;Mood=Ind;Tense=Fut	VERB	-	iremos	-	sofre
sofreíréis	es-conj-ír	Number=Plur;Mood=Ind;Tense=Fut	VERB	-	iréis	-	sofre
sofreirán	es-conj-ír	Number=Plur;Mood=Ind;Tense=Fut	VERB	-	irán	-	sofre
sofreiría	es-conj-ír	VerbForm=Ger;Number=Sing;Person=3;Mood=Cnd	VERB	-	iría	-	sofre
sofreirías	es-conj-ír	Number=Sing;Person=1;Mood=Cnd	VERB	-	irías	-	sofre
sofreiría	es-conj-ír	Number=Sing;Person=2;Mood=Cnd	VERB	-	iría	-	sofre
sofreiríamos	es-conj-ír	Number=Plur;Person=3;Mood=Cnd	VERB	-	iríamos	-	sofre
sofreiríais	es-conj-ír	Number=Plur;Mood=Cnd	VERB	-	iríais	-	sofre
sofreirían	es-conj-ír	Number=Plur;Mood=Cnd	VERB	-	irían	-	sofre
sofrea	es-conj-ír	VerbForm=Ger;Number=Sing;Person=3; Mood=Sub;Tense=Pres	VERB	-	a	-	sofre
sofreas	es-conj-ír	Number=Sing;Person=1;Mood=Sub;Tense=Pres	VERB	-	as	-	sofre
sofreásvos	es-conj-ír	Number=Sing;Person=1;Mood=Sub;Tense=Pres	VERB	-	ásvos2	-	sofr
sofrea	es-conj-ír	Number=Sing;Person=2;Mood=Sub;Tense=Pres	VERB	-	a	-	sofre
sofreamos	es-conj-ír	Number=Plur;Person=3;Mood=Sub;Tense=Pres	VERB	-	amos	-	sofre
sofreáis	es-conj-ír	Number=Plur;Mood=Sub;Tense=Pres	VERB	-	áis	-	sofre
sofrean	es-conj-ír	Number=Plur;Mood=Sub;Tense=Pres	VERB	-	an	-	sofre
sofreiera	es-conj-ír	VerbForm=Ger;Number=Sing;Person=3; Mood=Sub;Aspect=Imp;Tense=Imp	VERB	-	iera	-	sofre
sofreieras	es-conj-ír	Number=Sing;Person=1;Mood=Sub; Aspect=Imp;Tense=Imp	VERB	-	ieras	-	sofre
sofreiera	es-conj-ír	Number=Sing;Person=2;Mood=Sub; Aspect=Imp;Tense=Imp	VERB	-	iera	-	sofre
sofreiéramos	es-conj-ír	Number=Plur,Person=3;Mood=Sub; Aspect=Imp;Tense=Imp	VERB	-	iéramos	-	sofre
sofreierais	es-conj-ír	Number=Plur;Mood=Sub;Aspect=Imp;Tense=Imp	VERB	-	ierais	-	sofre
sofreieran	es-conj-ír	Number=Plur;Mood=Sub;Aspect=Imp;Tense=Imp	VERB	-	ieran	-	sofre
sofreiese	es-conj-ír	VerbForm=Ger;Number=Sing;Person=3; Mood=Sub;Aspect=Imp;Tense=Imp	VERB	-	iese	-	sofre
sofreieses	es-conj-ír	Number=Sing;Person=1;Mood=Sub; Aspect=Imp;Tense=Imp	VERB	-	ieses	-	sofre
sofreiese	es-conj-ír	Number=Sing;Person=2;Mood=Sub; Aspect=Imp;Tense=Imp	VERB	-	iese	-	sofre
sofreiésemos	es-conj-ír	Number=Plur,Person=3;Mood=Sub; Aspect=Imp;Tense=Imp	VERB	-	iésemos	-	sofre
sofreieseis	es-conj-ír	Number=Plur;Mood=Sub;Aspect=Imp;Tense=Imp	VERB	-	ieseis	-	sofre
sofreiesen	es-conj-ír	Number=Plur;Mood=Sub;Aspect=Imp;Tense=Imp	VERB	-	iesen	-	sofre
sofreiere	es-conj-ír	VerbForm=Ger;Number=Sing;Person=3;Mood=Sub	VERB	-	iere	-	sofre
sofreieres	es-conj-ír	Number=Sing;Person=1;Mood=Sub	VERB	-	ieres	-	sofre
sofreiere	es-conj-ír	Number=Sing;Person=2;Mood=Sub	VERB	-	iere	-	sofre
sofreiéremos	es-conj-ír	Number=Plur;Person=3;Mood=Sub	VERB	-	iéremos	-	sofre
sofreiereis	es-conj-ír	Number=Plur;Mood=Sub	VERB	-	iereis	-	sofre
sofreieren	es-conj-ír	Number=Plur;Mood=Sub	VERB	-	ieren	-	sofre
sofree	es-conj-ír	Number=Sing;Person=1;Mood=Imp;Polarity=Pos	VERB	-	e	-	sofre
sofreí	es-conj-ír	Number=Sing;Person=1;Mood=Imp;Polarity=Pos	VERB	-	í	-	sofre
sofrea	es-conj-ír	Number=Sing;Person=2;Mood=Imp;Polarity=Pos	VERB	-	a	-	sofre
sofreamos	es-conj-ír	Number=Plur;Person=3;Mood=Imp;Polarity=Pos	VERB	-	amos	-	sofre
sofreíd	es-conj-ír	Number=Plur;Mood=Imp;Polarity=Pos	VERB	-	íd	-	sofre
sofrean	es-conj-ír	Number=Plur;Mood=Imp;Polarity=Pos	VERB	-	an	-	sofre
no sofreas	es-conj-ír	Number=Sing;Person=1;Mood=Imp;Polarity=Neg	VERB	no	as	-	sofre
no sofrea	es-conj-ír	Number=Sing;Person=2;Mood=Imp;Polarity=Neg	VERB	no	a	-	sofre
no sofreamos	es-conj-ír	Number=Plur;Person=3;Mood=Imp;Polarity=Neg	VERB	no	amos	-	sofre
no sofreáis	es-conj-ír	Number=Plur;Mood=Imp;Polarity=Neg	VERB	no	áis	-	sofre
no sofrean	es-conj-ír	Number=Plur;Mood=Imp;Polarity=Neg	VERB	no	an	-	sofre

Table 4: The lemma *sofreír* in non-evaluated Wikinflection.

Lemma	Word	Features
sofreír	no sofriáis	V;NEG;IMP;2;PL
sofreír	no sofriamos	V;NEG;IMP;1;PL
sofreír	no sofrían	V;NEG;IMP;3;PL
sofreír	no sofrías	V;NEG;IMP;2;SG
sofreír	no sofría	V;NEG;IMP;3;SG
sofreír	sofreíais	V;IND;PST;2;PL;IPFV
sofreír	sofreíamos	V;IND;PST;1;PL;IPFV
sofreír	sofreían	V;IND;PST;3;PL;IPFV
sofreír	sofreías	V;IND;PST;2;SG;IPFV
sofreír	sofreía	V;IND;PST;1;SG;IPFV
sofreír	sofreía	V;IND;PST;3;SG;IPFV
sofreír	sofreídas	V.PTCP;PST;FEM;PL
sofreír	sofreída	V.PTCP;PST;FEM;SG
sofreír	sofreídos	V.PTCP;PST;MASC;PL
sofreír	sofreído	V.PTCP;PST;MASC;SG
sofreír	sofreíd	V;POS;IMP;2;PL
sofreír	sofreímos	V;IND;PRS;1;PL
sofreír	sofreímos	V;IND;PST;1;PL;PFV
sofreír	sofreirán	V;IND;FUT;3;PL
sofreír	sofreirás	V;IND;FUT;2;SG
sofreír	sofreirá	V;IND;FUT;3;SG
sofreír	sofreiréis	V;IND;FUT;2;PL
sofreír	sofreíremos	V;IND;FUT;1;PL
sofreír	sofreiré	V;IND;FUT;1;SG
sofreír	sofreiríais	V;COND;2;PL
sofreír	sofreiríamos	V;COND;1;PL
sofreír	sofreirían	V;COND;3;PL
sofreír	sofreirías	V;COND;2;SG
sofreír	sofreiría	V;COND;1;SG
sofreír	sofreiría	V;COND;3;SG
sofreír	sofreír	V;NFIN
sofreír	sofreísteis	V;IND;PST;2;PL;PFV
sofreír	sofreíste	V;IND;PST;2;SG;PFV
sofreír	sofreís	V;IND;PRS;2;PL
sofreír	sofreí	V;IND;PST;1;SG;PFV
sofreír	sofriáis	V;SBJV;PRS;2;PL
sofreír	sofriamos	V;POS;IMP;1;PL
sofreír	sofriamos	V;SBJV;PRS;1;PL
sofreír	sofrían	V;POS;IMP;3;PL
sofreír	sofrían	V;SBJV;PRS;3;PL
sofreír	sofrías	V;SBJV;PRS;2;SG
sofreír	sofría	V;POS;IMP;3;SG
sofreír	sofría	V;SBJV;PRS;1;SG
sofreír	sofría	V;SBJV;PRS;3;SG
sofreír	sofriendo	V.CVB;PRS
sofreír	sofrían	V;IND;PRS;3;PL
sofreír	sofrierais	V;SBJV;PST;2;PL;LGSPEC1
sofreír	sofriéramos	V;SBJV;PST;1;PL;LGSPEC1
sofreír	sofrieran	V;SBJV;PST;3;PL;LGSPEC1
sofreír	sofrieras	V;SBJV;PST;2;SG;LGSPEC1
sofreír	sofriera	V;SBJV;PST;1;SG;LGSPEC1
sofreír	sofriera	V;SBJV;PST;3;SG;LGSPEC1
sofreír	sofriereis	V;SBJV;FUT;2;PL
sofreír	sofriéremos	V;SBJV;FUT;1;PL
sofreír	sofrieren	V;SBJV;FUT;3;PL
sofreír	sofrieres	V;SBJV;FUT;2;SG
sofreír	sofriere	V;SBJV;FUT;1;SG
sofreír	sofriere	V;SBJV;FUT;3;SG
sofreír	sofrieron	V;IND;PST;3;PL;PFV
sofreír	sofrieseis	V;SBJV;PST;2;PL
sofreír	sofriésemos	V;SBJV;PST;1;PL
sofreír	sofriesen	V;SBJV;PST;3;PL
sofreír	sofrieses	V;SBJV;PST;2;SG
sofreír	sofriese	V;SBJV;PST;1;SG
sofreír	sofriese	V;SBJV;PST;3;SG
sofreír	sofríes	V;IND;PRS;2;SG
sofreír	sofríe	V;IND;PRS;3;SG
sofreír	sofríe	V;POS;IMP;2;SG
sofreír	sofrío	V;IND;PRS;1;SG
sofreír	sofrió	V;IND;PST;3;SG;PFV

Table 5: The lemma *sofreír* in UniMorph.

Word	Template	Features	POS	Prefix	Suffix	Infix	Stem
sofreír	es-conj-ír	VerbForm=Inf	VERB	-	ír	-	sofre
sofreír	es-conj-ír	VerbForm=Inf	VERB	-	ír	-	sofre
sofreír	es-conj-ír	VerbForm=Inf	VERB	-	ír	-	sofre
sofreír	es-conj-ír	VerbForm=Inf	VERB	-	ír	-	sofre
sofreír	es-conj-ír	VerbForm=Inf	VERB	-	ír	-	sofre
sofreído	es-conj-ír	VerbForm=Ger;Number=Sing	VERB	-	ído	-	sofre
sofreído	es-conj-ír	Number=Sing	VERB	-	ído	-	sofre
sofreís	es-conj-ír	Number=Sing;Person=1;Mood=Ind;Tense=Pres	VERB	-	ís	-	sofre
sofreímos	es-conj-ír	Number=Plur;Person=3;Mood=Ind;Tense=Pres	VERB	-	ímos	-	sofre
sofreís	es-conj-ír	Number=Plur;Mood=Ind;Tense=Pres	VERB	-	ís	-	sofre
sofreía	es-conj-ír	VerbForm=Ger;Number=Sing;Person=3;Mood=Ind;Aspect=Imp;Tense=Imp	VERB	-	ía	-	sofre
sofreías	es-conj-ír	Number=Sing;Person=1;Mood=Ind;Aspect=Imp;Tense=Imp	VERB	-	ías	-	sofre
sofreía	es-conj-ír	Number=Sing;Person=2;Mood=Ind;Aspect=Imp;Tense=Imp	VERB	-	ía	-	sofre
sofreíamos	es-conj-ír	Number=Plur;Person=3;Mood=Ind;Aspect=Imp;Tense=Imp	VERB	-	íamos	-	sofre
sofreíais	es-conj-ír	Number=Plur;Mood=Ind;Aspect=Imp;Tense=Imp	VERB	-	íais	-	sofre
sofreían	es-conj-ír	Number=Plur;Mood=Ind;Aspect=Imp;Tense=Imp	VERB	-	ían	-	sofre
sofreí	es-conj-ír	VerbForm=Ger;Number=Sing;Person=3;Mood=Ind;Tense=Past	VERB	-	í	-	sofre
sofreíste	es-conj-ír	Number=Sing;Person=1;Mood=Ind;Tense=Past	VERB	-	íste	-	sofre
sofreímos	es-conj-ír	Number=Plur;Person=3;Mood=Ind;Tense=Past	VERB	-	ímos	-	sofre
sofreísteis	es-conj-ír	Number=Plur;Mood=Ind;Tense=Past	VERB	-	ísteis	-	sofre
sofreíré	es-conj-ír	VerbForm=Ger;Number=Sing;Person=3;Mood=Ind;Tense=Fut	VERB	-	iré	-	sofre
sofreírás	es-conj-ír	Number=Sing;Person=1;Mood=Ind;Tense=Fut	VERB	-	irás	-	sofre
sofreírá	es-conj-ír	Number=Sing;Person=2;Mood=Ind;Tense=Fut	VERB	-	irá	-	sofre
sofreíremos	es-conj-ír	Number=Plur;Person=3;Mood=Ind;Tense=Fut	VERB	-	iremos	-	sofre
sofreíréis	es-conj-ír	Number=Plur;Mood=Ind;Tense=Fut	VERB	-	iréis	-	sofre
sofreírán	es-conj-ír	Number=Plur;Mood=Ind;Tense=Fut	VERB	-	irán	-	sofre
sofreíría	es-conj-ír	VerbForm=Ger;Number=Sing;Person=3;Mood=Cnd	VERB	-	iría	-	sofre
sofreírías	es-conj-ír	Number=Sing;Person=1;Mood=Cnd	VERB	-	irías	-	sofre
sofreíría	es-conj-ír	Number=Sing;Person=2;Mood=Cnd	VERB	-	iría	-	sofre
sofreíríamos	es-conj-ír	Number=Plur;Person=3;Mood=Cnd	VERB	-	iríamos	-	sofre
sofreíríais	es-conj-ír	Number=Plur;Mood=Cnd	VERB	-	iríais	-	sofre
sofreírían	es-conj-ír	Number=Plur;Mood=Cnd	VERB	-	irían	-	sofre
sofreí	es-conj-ír	Number=Sing;Person=1;Mood=Imp;Polarity=Pos	VERB	-	í	-	sofre
sofreíd	es-conj-ír	Number=Plur;Mood=Imp;Polarity=Pos	VERB	-	íd	-	sofre

Table 6: The lemma *sofreír* in Wiktionary-evaluated Wikinflection.

Word	Template	Features	POS	Prefix	Suffix	Infix	Stem
sofreír	es-conj-ír	VerbForm=Inf	VERB	-	ír	-	sofre
sofreído	es-conj-ír	Tense=Past;Gender=Masc;Number=Sing	VERB	-	ído	-	sofre
sofreída	es-conj-ír	Tense=Past;Gender=Fem;Number=Sing	VERB	-	ída	-	sofre
sofreídos	es-conj-ír	Tense=Past;Gender=Masc;Number=Plur	VERB	-	ídos	-	sofre
sofreídas	es-conj-ír	Tense=Past;Gender=Fem;Number=Plur	VERB	-	ídas	-	sofre
sofreís	es-conj-ír	Mood=Ind;Tense=Pres;Person=2;Number=Plur	VERB	-	ís	-	sofre
sofreímos	es-conj-ír	Mood=Ind;Tense=Past;Person=1;Number=Plur;Aspect=Perf	VERB	-	ímos	-	sofre
sofreía	es-conj-ír	Mood=Ind;Tense=Past;Person=3;Number=Sing;Aspect=Imp	VERB	-	ía	-	sofre
sofreías	es-conj-ír	Mood=Ind;Tense=Past;Person=2;Number=Sing;Aspect=Imp	VERB	-	ías	-	sofre
sofreíamos	es-conj-ír	Mood=Ind;Tense=Past;Person=1;Number=Plur;Aspect=Imp	VERB	-	íamos	-	sofre
sofreíais	es-conj-ír	Mood=Ind;Tense=Past;Person=2;Number=Plur;Aspect=Imp	VERB	-	íais	-	sofre
sofreían	es-conj-ír	Mood=Ind;Tense=Past;Person=3;Number=Plur;Aspect=Imp	VERB	-	ían	-	sofre
sofreí	es-conj-ír	Mood=Ind;Tense=Past;Person=1;Number=Sing;Aspect=Perf	VERB	-	í	-	sofre
sofreíste	es-conj-ír	Mood=Ind;Tense=Past;Person=2;Number=Sing;Aspect=Perf	VERB	-	íste	-	sofre
sofreísteis	es-conj-ír	Mood=Ind;Tense=Past;Person=2;Number=Plur;Aspect=Perf	VERB	-	ísteis	-	sofre
sofreiré	es-conj-ír	Mood=Ind;Tense=Fut;Person=1;Number=Sing	VERB	-	iré	-	sofre
sofreirás	es-conj-ír	Mood=Ind;Tense=Fut;Person=2;Number=Sing	VERB	-	irás	-	sofre
sofreirá	es-conj-ír	Mood=Ind;Tense=Fut;Person=3;Number=Sing	VERB	-	irá	-	sofre
sofreiremos	es-conj-ír	Mood=Ind;Tense=Fut;Person=1;Number=Plur	VERB	-	iremos	-	sofre
sofreiréis	es-conj-ír	Mood=Ind;Tense=Fut;Person=2;Number=Plur	VERB	-	iréis	-	sofre
sofreirán	es-conj-ír	Mood=Ind;Tense=Fut;Person=3;Number=Plur	VERB	-	irán	-	sofre
sofreiría	es-conj-ír	Mood=Cnd;Person=3;Number=Sing	VERB	-	iría	-	sofre
sofreirías	es-conj-ír	Mood=Cnd;Person=2;Number=Sing	VERB	-	irías	-	sofre
sofreiríamos	es-conj-ír	Mood=Cnd;Person=1;Number=Plur	VERB	-	iríamos	-	sofre
sofreiríais	es-conj-ír	Mood=Cnd;Person=2;Number=Plur	VERB	-	iríais	-	sofre
sofreirían	es-conj-ír	Mood=Cnd;Person=3;Number=Plur	VERB	-	irían	-	sofre
sofreíd	es-conj-ír	Polarity=Pos;Mood=Jus;Person=2;Number=Plur	VERB	-	íd	-	sofre

Table 7: The lemma *sofreír* in UniMorph-evaluated Wikinflection.