Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1317–1325
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Wiktextract: Wiktionary as Machine-Readable Structured Data

Tatu Ylonen
University of Helsinki
Department of Digital Humanities / Language Technology
tatu@ylonen.org

Abstract

We present a machine-readable structured data version of Wiktionary. Unlike previous Wiktionary extractions, the new extractor, Wiktextract, fully interprets and expands templates and Lua modules in Wiktionary. This enables it to perform a more complete, robust, and maintainable extraction. The extracted data is multilingual and includes lemmas, inflected forms, translations, etymology, usage examples, pronunciations (including URLs of sound files), lexical and semantic relations, and various morphological, syntactic, semantic, topical, and dialectal annotations. We extract all data from the English Wiktionary. Comparing against previous extractions from language-specific dictionaries, we find that its coverage for non-English languages often matches or exceeds the coverage in the language-specific editions, with the added benefit that all glosses are in English. The data is freely available and regularly updated, enabling anyone to add more data and correct errors by editing Wiktionary. The extracted data is in JSON format and designed to be easy to use by researchers, downstream resources, and application developers.

Keywords: Wiktionary, extraction, dictionary, lexicon, and morphology, inflections, translation, etymology, dialects

1. Introduction

Wiktionary\(^1\) is undoubtedly the world’s largest freely available dictionary. While many partial extractors exist, no-one seems to have succeeded in fully converting Wiktionary into a computationally easy-to-use machine-readable format. This is largely because it is an unstructured wiki maintained by thousands of volunteers with varying technical skills. Tens of thousands of templates and programmed extension modules in the Lua programming language\(^2\) are used by the volunteers for formatting content. It is these templates and Lua modules, combined with the great variety of ways the information is encoded by the volunteers for hundreds of languages that have made general extraction so difficult. Lua modules are used to generate pronunciations, inflection tables, and even entire sets of glosses for many languages.

Over the last several years, we have developed new Python packages for computationally processing Wiktionary. The wikitextprocessor\(^3\) package implements parsing wikitext\(^4\) (the source format used for all Wiktionary and Wikipedia language variants), expanding templates, and executing Lua modules. The wiktextract\(^5\) tool implements converting the English Wiktionary into a machine-readable, structured format. The English Wiktionary has glosses in English, but it contains words and translations for hundreds of different languages.

We have furthermore implemented other tools that present the extracted data as a web site\(^6\) for testing, evaluation, post-processing, and for making the data available to others. The result of this work is a machine-readable version of Wiktionary, formatted as JSON\(^7\) to make it as easy as possible for others to use. We regularly update the extraction, so that any words added to Wiktionary or improvements made to existing words will soon appear in the extraction.

The extraction includes almost all the information from the English Wiktionary, including data for hundreds of languages (it is called the English Wiktionary because its glosses are in English; the vast majority of word entries are for languages other than English). Our extractor expands Wiktionary templates and executes its Lua modules. Word senses and glosses are extracted and partially parsed to identify word senses that are an inflection or an alternative form (e.g., abbreviation) of another word. Such senses are automatically linked to the related lemma. Pronunciations are extracted as IPA strings, sound files, and various language-specific formats. Word forms and inflections are extracted (including from inflection tables). Translations are extracted and parsed to identify the actual translation, gender/class, romanization, sense description, etc. Lexical and semantic relations between words are extracted and partially parsed to identify the actual linked word, romanizations, gender/class tags, sense descriptions, etc. Usage examples and etymological information are also collected in a structured form, and various special data items, such as Wikidata and Wikipedia links are also included.

Our contributions include the first known extractor capable of expanding Wiktionary templates and Lua modules, resulting in a more robust and maintainable extraction, and a lexical resource that contains nearly

---

1.https://wiktionary.org
2. https://www.lua.org
3. https://github.com/tatuylonen/wikitextprocessor
4. https://en.wikipedia.org/wiki/Help:Wikitext
5. https://github.com/tatuylonen/wiktextract
6. https://kaikki.org
7. https://www.ietf.org/rfc/rfc4627.txt
all of the information in the English Wiktionary in a format that is easy to use for both linguists and artificial intelligence researchers. Regular updates and the ability to add and fix things in Wiktionary provide additional flexibility and consistent tagging across languages makes it well suited for cross-lingual work.

2. Related Work

There have been a number of partial extractions from Wiktionary data previously. JWKTLE (Zesch et al., 2008) was one of the earliest extractors for Wiktionary data. UBY (Gurevych et al., 2012) integrated and aligned data from Wiktionary and several other lexical resources into a single unified database. Wik2dict (Ács, 2013) was a translation extraction tool that supports multiple languages. Wikpron (Lee et al., 2020) extracted pronunciation information for many languages.

Several researchers have developed language-specific extractors. GLAWI (Sajous and Hatbour, 2015) was an extractor for the French Wiktionary. ENGLAWI (Sajous et al., 2020) contains English words parsed from the English Wiktionary in XML format. It is limited to English words. knoWitary (Nastase and Straparava, 2015) is another extraction from the English Wiktionary. Zawilinski (Kurmas, 2010) extracted Polish words from the English Wiktionary. Pérez et al. (2011) presented an extractor for the Polish Wiktionary. Wikkit (Krizhanovsky and Lin, 2009) was an extractor for the Russian Wiktionary. Miletic (2017) demonstrated an extractor for the Serbo-Croatian Wiktionary. Various tools have been built to extract inflection tables from Wiktionary. Wikinflection (Metheniti and Neumann, 2018) initially attempted to process templates and Lua modules, but their attempt at processing Lua modules failed and they ended up just parsing HTML templates, ignoring inflection tables generated by Lua modules (they mention that inflection tables from 2927 templates out of 7068 could be parsed). We compare our extraction against Wikinfection in Section 5. IWNLP (Liebeck and Conrad, 2013) extracted paradigms from German inflection tables using a hand-coded implementation of a subset of Lua modules in C#. Kirov et al. (2016) parsed from HTML without capturing the arguments used to generate the paradigms (we discuss it in more detail in Section 5). Sennrich and Kunz (2014) used regular expressions on the German Wiktionary raw Wikitext. Krizhanovskaya and Krizhanovsky (2019) discussed reimplementing a Lua module in PHP for Veps inflection extraction. Dbnary (Sérasset and Tchechmedjiev, 2014) converted Wiktionary into an RDF linked data format, supporting extraction from multiple language editions of Wiktionary and disambiguating translations. We compare our extraction against Dbnary in Section 5. BabelNet (Navigli and Ponzetto, 2012) is an extensive resource that combines information from multiple sources. However, even though Wiktionary is sometimes mentioned as a source, BabelNet’s use of Wiktionary data currently seems limited. As far as we know, BabelNet does not include inflections, etymology, or dialectal tagging.

3. Wiktext Extractor

Wiktext is the first known extractor that can expand Wiktionary templates and execute Lua modules. There are 42,000 templates and 55,000 Lua modules totaling 3.4 million lines of Lua code in the English Wiktionary as of December 2021. New templates and Lua modules are constantly being defined, and Wiktext is usually able to handle them automatically. Many parts of Wiktionary pages are generated by Lua modules, and properly interpreting the Lua code is important for wide coverage, robustness, and maintainability. While some extractors have used the pre-generated Wiktionary HTML pages as the source, there is also useful information that is not easily accessible on the HTML pages, such as etymological relations and the templates and their arguments that are used to generate inflection paradigms.

Not having to worry about templates and Lua modules, we were able to focus on the information content and alternative ways of encoding it in Wiktionary. Simultaneous access to template arguments and the original source code helped solve tricky parsing issues. The detailed internal operation of Wiktext is beyond the scope of this paper. All the code and documentation is freely available in the Github repository at https://github.com/tatuylonen/wiktext.

In summary, Wiktext extracts the following kinds of data from the English Wiktionary:

- Words and word senses for hundreds of languages, with glosses in English. Certain glosses are further parsed to identify the lemmas for inflected forms, abbreviations, and other alternative forms. Usage examples are extracted and parsed into a citation, the actual example, and English translation where applicable.
- Translations (including those on separate translation pages). They are parsed to separate the actual translation from romanization, sense description, usage notes, etc. Translations are annotated with, e.g., grammatical gender, inflectional class, and/or dialect.
- Full inflection tables, whether the information is textually in the word head or in a conjugation/declension/inflection section. The templates that generate the inflected forms are also extracted.
- Etymological information in both human-readable text form and as a list of the templates used to gen-
erate it (providing easy machine-readable access to etymological relations).

- Pronunciations, hyphenations, and homonyms. This includes IPA\(^9\) pronunciations and sound files where available (about 942,000 sound files are currently included). Pronunciations generated programatically by Lua modules are also extracted (many languages use them extensively). Alternative pronunciations are generally annotated with the relevant dialect.

- Lexical and semantic relations (e.g., hypernyms, synonyms, derived terms). They are parsed to separate the linked word, romanizations, sense descriptions, English translations, etc. Relations are also extracted from separate thesaurus pages and merged with those provided in the word entry.

- Various topical and linguistic annotations, including grammatical gender or inflectional class, transliteration, countability, topic area, category links, etc. Inflected forms are annotated with machine-readable tags identifying the grammatical form.

3.1. Data Download Link

The full pre-extracted data sets are freely available for downloading at [https://kaikki.org/dictionary/rawdata.html](https://kaikki.org/dictionary/rawdata.html). We call this data the Witkextract raw data.

We also provide post-processed data on [https://kaikki.org](https://kaikki.org) and the site also contains a browsable and searchable version of the data. It is possible to download subsets of the data for individual languages, parts of speech, and various annotations. The extracted JSON for any word can be easily viewed.

We are adding more post-processing, including disambiguation, gloss parsing, and various kinds of augmentations from additional sources to the data on the web site, but the Witkextract raw data described in this paper will also remain available. Our own post-processing is just one of the downstream projects benefitting from the raw data.

The available data sets are typically updated every week based on the latest Wiktionary XML dump. New dumps seem to become available every 2-4 weeks. The extractor is also continuously improved, and bugs can be reported at [https://github.com/tatuylonen/witkextract/issues](https://github.com/tatuylonen/witkextract/issues).

4. Content and Structure of the Data

Our primary distribution format for the data is JSON (one JSON object per line), with usually one part of speech for a word in each object. This format is easy to parse in most programming languages. We use this format because it is easier to process on laptops with limited memory than if all data was in a JSON list (loading all of the data into Python simultaneously uses over 85 GB of memory).

The JSON object on each line describes either a redirect or a part of speech for a word (though it is possible for a word to have the same part of speech more than once, e.g., for different etymologies). The full documentation is in the Github repository.

Each word is represented by a JSON object (a key-value mapping, corresponding to a dictionary in Python). Each entry has the following fields: `word` is the word form, `lang` is the language name (e.g., English), `pos` is the part of speech (e.g., noun, name, verb, adj, adv), and `senses` is a list of word senses (JSON array). There are also several other possible fields, such as `translations` (for translations), `forms` (for inflected forms), `sounds` (for pronunciations), `categories` (for Wiktionary category links), hypernyms, synonyms, etc. Furthermore, `etymology_text` contains extracted human-readable etymology text, `etymology_templates` contains a list of templates describing potential etymological relations, and `inflection_templates` contains a list of inflection templates and their arguments used to generate inflection tables.

Each word sense is represented by a JSON object that typically has the following fields: `glosses` is expanded gloss strings in Unicode (with, e.g., chemical formulas and mathematical expressions converted to Unicode text) and templates expanded, `tags` contains morphological, syntactic, semantic, and regional annotation (e.g., archaic, declension-1, plural, participle, Australia), `alt_of` contains the base word for abbreviations and other alternative word forms, `form_of` contains the lemma for inflected forms, etc.

Translations are represented by a dictionaries with the following fields: `lang` is the language of the translation, `word` is the translated word, `alt` is an alternative script version of the translation (e.g., Hiragana for a Japanese Kanji word), `roman` is a romanization of the translation, `sense` is a string identifying the word sense the translation relates to, `english` contains the translation’s literal meaning or clarifies the sense in the translation’s language, and `tags` contains gender, inflection class, or dialectal annotation.

Pronunciations are dictionaries with the following fields: `ipa` for an IPA string, `mp3_url` for the URL of an MP3 sound file, and `tags` identifies dialectal or regional variant pronunciations (e.g., Received-Pronunciation, Australia).

The following truncated example illustrates this (see [https://kaikki.org/dictionary/All%20languages%20combined/meaning/ba/ba/baby.html](https://kaikki.org/dictionary/All%20languages%20combined/meaning/ba/ba/baby.html) for the complete data):

```json
{
  "categories": [
    "People", ...
  ],
```
5. Comparison with Other Wiktionary Extractions

Table 1 summarizes the extracted entries for various languages (as of December 2021). All words have English language glosses regardless of the language of the entry.

Table 2 summarizes the extracted translations. For many languages, additional word entries could be defined based on the translations. Many translations also include information about gender, inflection class, and semantic categories.

Table 3 compares Wiktextract raw data against two freely available extractions - Dbnary (Sérasset and Tchechmedjiev, 2014) and Wikinflection (Metheniti and Neumann, 2018). We took the latest available Wikinflection data from Github10. The Dbnary data are from the Dbnary dashboard11 and the inflected form counts computed from the dataset. We also discuss some other extractions in the text below.

Dbnary contains data for 22 languages, Wikinflection for 70 languages (some with minimal coverage). Wiktextract includes 70 languages with at least 10,000 lemmas, a total of 98 languages with at least 5,000 lemmas, and a total of 322 languages with at least 500 lemmas. These counts do not include translations, which could be used to synthesize additional entries for many languages.

Dbnary is a fairly extensive extraction. Notably, it extracts each language from its separate Wiktionary edition. Wiktextract, in contrast, extracts all languages from the English Wiktionary. Despite this, the coverage of Wiktextract appears comparable and sometimes better for lemmas.

Wiktextract captures all glosses in English, whereas Dbnary has glosses for each language in that language. We think that having all glosses in the same language makes downstream processing easier, as a single parser can then be used for extracting meaning from the glosses. Language-specific glosses might offer more nuances for humans, but we don’t think current lexical semantic representations are yet capable of capturing them and thus we think a uniform language for glosses is more useful for applications.

There are also other differences in gloss processing. Wiktextract formats chemical formulas into unicode, while Dbnary keeps them in ASCII: CH₃CONH₂ vs. CH₃CONH₂ (from “acetamine”). Similarly for mathematical formulas; e.g., glosses for “Gaussian function” and “Taylor series” are reasonably rendered into...
| Language   | Lemmas  | Non-lemma | Inflections | Nouns | Verbs | Adjs | Advs | Other |
|-----------|---------|-----------|-------------|-------|-------|-----|-----|-------|
| English   | 914 577 | 394 316   | 656 758     | 484 872 | 82 329 | 173 676 | 25 417 | 148 283 |
| Spanish   | 235 822 | 710 887   | 1348 507    | 65 047 | 134 964 | 22 595 | 424 000 | 8 976  |
| Italian   | 174 406 | 523 448   | 855 811     | 81 518 | 38 878 | 31 131 | 616 660 | 167 131 |
| Finnish   | 162 907 | 89 777    | 6618 259    | 114 633 | 16 310 | 16 979 | 7 184  | 7 801  |
| Chinese   | 202 766 | 120       | 560         | 84 521 | 32 601 | 10 846 | 297 2 | 71 826 |
| Japanese  | 160 420 | 1 660     | 771 804     | 72 582 | 16 413 | 2 990 | 203 4 | 66 401 |
| French    | 100 679 | 367 708   | 524 981     | 53 406 | 14 400 | 18 596 | 469 8 | 9 579  |
| Russian   | 92 253  | 375 116   | 1933 634    | 33 375 | 33 882 | 16 176 | 2 887 | 5 993  |
| German    | 89 662  | 446 734   | 2422 099    | 49 718 | 12 727 | 16 106 | 2 258 | 8 853  |
| Portuguese| 83 153  | 315 383   | 463 823     | 43 221 | 10 120 | 14 165 | 2 523 | 13 124 |
| Serbo-Croatian | 73 019  | 6 878    | 2272 886    | 35 669 | 15 668 | 11 309 | 5 352 | 5 021  |
| Dutch     | 71 798  | 74 954    | 341 447     | 37 634 | 14 173 | 7 307  | 1 500 | 11 184 |
| Polish    | 88 192  | 70 891    | 1257 515    | 31 440 | 17 789 | 8 533  | 2 413 | 28 017 |
| Romanian  | 85 093  | 17 870    | 842 108     | 37 549 | 7 640  | 12 913 | 914  | 26 077 |
| Latin     | 61 277  | 1368 199  | 2480 503    | 22 010 | 13 457 | 12 562 | 2 309 | 10 939 |
| Macedonian| 43 032  | 13 729    | 1196 987    | 18 304 | 10 497 | 8 784  | 3 432 | 2 015  |
| Middle English | 43 989 | 2 716    | 27 508     | 25 966 | 6 812  | 5 525  | 2 486 | 3 200  |
| Greek     | 44 128  | 52 272    | 368 070     | 24 106 | 6 396  | 8 647  | 1 063 | 3 916  |
| All others| 1730 640 | 1181 388  | 26960 519   | 859 443 | 247 466 | 176 857 | 42 148 | 404 726 |
| **Total** | 4457 813 | 6014 501  | 51363 779   | 2175 014 | 732 462 | 575 697 | 121996 | 852 644 |

Table 1:Wiktextract statistics on entries for various languages. *Lemmas* is the number of lemmas (entries that were not recognized as inflected forms of another word), *Non-lemma* is entries that are inflected forms of a lemma, *Inflections* is inflected forms extracted from word heads and inflection tables, and the other fields count the different parts of speech for the lemmas. The table is sorted by the sum of Nouns, Verbs, Adjs, and Advs.

| Language | Translations | Note | Sense | Target | Roman | Alt | Tags |
|----------|--------------|------|-------|--------|-------|-----|------|
| Finnish  | 140 686      | 1241 | 140 973 | 2945   | 9     | 47  | 3332 |
| German   | 133 821      | 772  | 133 756 | 1721   | 27    | 47  | 82 204 |
| Chinese  | 130 958      | 595  | 131 110 | 1383   | 79 037 | 1805 | 130 454 |
| Russian  | 122 410      | 1345 | 122 878 | 2961   | 121 424 | 1097 | 89 488 |
| French   | 102 900      | 530  | 102 759 | 777    | 29    | 29  | 60 712 |
| Spanish  | 98 838       | 312  | 98 553  | 691    | 7     | 14  | 57 745 |
| Portuguese| 86 192       | 197  | 86 102  | 303    | 9     | 14  | 53 022 |
| Italian  | 82 676       | 255  | 81 297  | 294    | 6     | 14  | 46 458 |
| Dutch    | 67 588       | 198  | 67 570  | 325    | 13    | 22  | 32 031 |
| Swedish  | 62 605       | 296  | 62 564  | 516    | 3     | 22  | 31 872 |
| Hungarian| 61 660       | 558  | 61 797  | 756    | 24    | 37  | 824 |
| Polish   | 61 552       | 218  | 61 556  | 449    | 7     | 11  | 44 910 |
| Japanese | 59 064       | 412  | 59 160  | 1118   | 56 053 | 29 390 | 878 |
| Greek    | 58 796       | 179  | 58 778  | 373    | 58 410 | 147  | 44 365 |
| Bulgarian| 54 443       | 38   | 54 447  | 258    | 54 081 | 118  | 29 593 |
| Norwegian| 52 093       | 117  | 52 088  | 347    | 1     | 8   | 43 618 |
| Serbo-Croatian | 50 438 | 74    | 50 449  | 174    | 319   | 28  | 45 575 |
| Czech    | 44 549       | 127  | 44 560  | 211    | 7     | 6   | 31 173 |
| Others   | 1 244 903   | 4932 | 1245 815 | 10 535 | 440 098 | 12 547 | 432 777 |
| **Total** | 2 716 172   | 12 396 | 2 716 212 | 26 137 | 809 564 | 45 403 | 1 261 031 |

Table 2: Wiktextract statistics on translations into various languages. *Translations* is number of translations extracted from English to that language; *Note* is number of usage notes additionally extracted. *Sense* counts translations and usage notes with source sense text and *Target* translations and notes with text limiting the non-English sense. *Roman* is the number of translations that have romanizations, *Alt* is the number of translations that have alternative forms (e.g., hiragana), and *Tags* is the number of translations with tags (e.g., gender, class, dialect).
Unicode in Wiktext extract, while they are totally garbled in Dbnary.

Many Wiktionary glosses begin with a parenthesized qualifier (e.g., a topic such as “(military)”). Wiktext extract decodes and canonicalizes such prefixes into topics and tags in separate fields, while Dbnary leaves them in the gloss. Wiktext extract also includes the original gloss – while the cleaned gloss may be easier for downstream processing, the original may be more suited for display to users. Wiktext extract also parses glosses to identify entries describing inflected forms and to link them with the related lemmas. Dbnary has no corresponding functionality; Wikinflection does not include glosses at all. Dbnary captures pronunciations but does not annotate them with dialectal tags and does not capture sound file names. Wiktext extract does not capture either. Wiktext extract captures both IPA and other types of pronunciations and includes URLs for downloading sound files for each word. WikiPron[12] includes pronunciations but not much of the other data. WikiPronunciationDict includes and canonicalizes pronunciations, using Wiktext extract data as the source for the English Wiktionary.

Dbnary captures some etymological relations but apparently not, e.g., cognates. It does not capture the human-readable etymology text. Wiktext extract captures both the human-readable text and the templates describing etymological relations.

Dbnary only captures inflection data for German, French, and English. However, at least the English inflections seem to have many systematic errors; for example, “house” (verb) has present participle “houses” and past “housed”, which are obviously incorrect (same for “browse”, “cope”, and many other similar verbs). It has “bosss” as the third-person present singular of the verb “boss” (and for “access”, “process”, etc.). It has a plural form “+” for the noun “fly” (and 3899 other words), past and past participle “flyed” for the verb “fly” (in the “to hit a ball” baseball sense; similarly for “dry”, “cry”); It gives “proded” and “proding” as the past and past participle of “prod” (similarly for “club”, “pit”). These errors are so common that they raise questions about the the English inflection data. We speculate that these errors could be due to incorrect or outdated manual reimplementation of the Lua code that generates the inflected forms. We did not look at the French or German inflection data.

Wikinflection includes inflections for many languages, but looking at the inflection data in Wikinflection, its Finnish inflections seem to be completely missing all plural noun forms, comparatives, superlatives, and forms with a possessive suffix. As seen in Table 3, it includes and canonicalizes inflections from all plural noun forms, comparatives, superlatives, and forms with a possessive suffix. As seen in Table 3, its coverage for several major languages is rather sporadic.

Neither Dbnary nor Wikinflection captures inflection template arguments.

Wiktext extract captures inflected forms from both word heads and inflection tables (declension and conjugation). It also captures mutation tables for various languages (e.g., Irish, Welsh). Table parsing includes expanding their templates and executing applicable Lua modules, and then parsing the tables and canonicalizing form descriptions into tags. It also imports annotations from footnotes (often used to indicate, e.g., archaic or polite forms). Wiktext extract also captures the template arguments and inflection classes used to generate the inflection tables, as this information is very helpful for inflection generation tools for some languages, such as Finnish; a Finnish verb can have up to 12,000 distinct forms [Karlsson, 1982] pp. 356-257), only a few dozen of which are included in Wiktionary’s

---

Table 3: Comparison between Dbnary, Wikinflection, and Wiktext extract raw data for a few languages. Dbnary does not distinguish lemmas and non-lemmas (i.e., inflected forms that have their own article entries); the Entries column includes both and corresponds to the sum of the Lemmas and Non-lemmas columns for Wiktext extract. Lemmas is base forms, Non-lemmas inflected forms with their own article entries, and Inflections is inflected forms (in Wiktext extract, extracted from word heads and inflection tables). Bold indicates the largest value (Entries compared against the sum of Lemmas and Non-Lemmas, both of latter bolded when bigger). Dbnary extracts each language from its language-specific Wiktionary edition, while Wikinflection and Wiktext extract all languages from the English Wiktionary.
Tagging in Wiktext is more extensive than the representation of inflected forms in Dbinary or Wikinflection. Over 2000 tags have been defined. Many thousands of ways of encoding tags in Wiktionary are canonicalized into the defined tags with reasonable uniformity across languages. Since Wiktext extracts inflections for a much wider set of languages than the other datasets, it needs more tags for encoding grammatical forms. Tags are assigned into tag categories, such as referent (definiteness, proximity, salience), degree (comparisons), person (including inclusive/exclusive “we”), number (including singular, plural, dual, trial, paucal, superplural, and collectivity/distributivity), object (for object concord tags), case, possession (alienable and inalienable possession, possessed concord tags), voice, tense, aspect, mood (with dozens of moods that do not exist morphologically in Indo-European languages), non-finite (for infinitives, participles, deverbal nouns, adjectives, and adverbs), polarity (for negation and conative), category (e.g., animateness, virility, countability), transitivity, register (e.g., degrees of formality and deference, vulgarity, colloquiality, slang), dialect (over a thousand regional and dialectal specifiers), and many others. Tags with spaces are currently used as a temporary extension mechanism for certain recognized but otherwise unimplemented constructs that may contain useful information (e.g., “of a bird” or “followed by for”). We hope that the tagging system will be useful for many kinds of cross-lingual research, language universals studies, and dialect studies.

We think that the JSON format used in Wiktext is easier and faster for most researchers to use than the RDF format in which Dbinary is distributed. Those working mostly with the semantic web might disagree. Wikidata\footnote{\url{https://wikidata.org}} is another dataset that includes data from several WikiMedia projects. It includes translations for words, but does not currently include inflection paradigms, or IPA pronunciations. It also does not include many of the word senses in Wiktionary. On the other hand, it includes significant additional information from several sources that is not included in Wiktionary.

Overall, it seems that extracting all languages from the English Wiktionary alone is a reasonable option and has the added benefit that all glosses are in the same language. However, there are translations, pronunciations, and inflections in the language-specific Wiktionary editions that are not yet available in the English Wiktionary. Some languages probably also have more words in the language-specific Wiktionaries, but surprisingly many languages already have equal or better coverage in the English Wiktionary. In fact, all languages that we evaluated, except French, had more entries [lemmas+non-lemmas] in the English Wiktionary than Dbinary had extracted from the corresponding

\textbf{Kirov et al. (2016)} presented another large-scale effort for parsing morphological paradigms from Wiktionary. They parsed from HTML tables from three editions of Wiktionary (English, French, and German). They mention having extracted data for 952,530 lemmas across 350 languages (cf. 4,457,813 lemmas from just the English Wiktionary in our extraction). They mention 11,006 French verbs in the English Wiktionary and 24,742 in the French Wiktionary; Table 1 shows we extracted 14,400 French verb lemmas from the English Wiktionary. This suggests that there are still many verbs in the French edition that have not yet been added to the English Wiktionary. For English they report 159,917 noun lemmas and 23,532 verb lemmas (respectively 484,872 and 82,329 in the Wiktext data). For Finnish they report 49,458 noun lemmas and 8,709 verb lemmas (respectively 114,633 and 16,310 in the Wiktext data).

Dbinary contains 2,628,675 translations from the English Wiktionary (Dec 2021). Wiktext raw data contains 2,716,172 translations extracted from the English Wiktionary. However, Dbinary includes additional translations extracted from other Wiktionary editions, and has a total of 8,296,362 translations.

Dbinary seems to extract translations only from translation tables that are directly on the word pages. However, the English Wiktionary has been moving translations of the most common words to separate translation pages (page titles ending in /translations). Wiktext merges translations from the translation pages into those found on word pages. Such translations are missing from Dbinary (e.g., for “woman”). Dbinary makes extra information in translations available in an unparsed usage field; Wiktext parses this information into tags and separate fields.

Wiktionary itself usually specifies translations for a part of speech rather than a word sense. It usually has a sense description associated with translations, but this description is often not identical to any gloss and often uses a different sense granularity. Dbinary heuristically disambiguates translations to the relevant word sense (in the source language, but apparently not in the target language). Wiktext itself does not do this disambiguation; however, we have implemented disambiguation of not just translations but also semantic relations and categories to word senses as a post-processing step; the post-processed data is available at \url{https://kaikki.org/dictionary/}. This approach lets anyone experiment with their own disambiguation. The Wiktext raw data makes all the data needed for disambiguation available for downstream tools.

Neither Dbinary nor Wikinflection extracts usage examples for words. Wiktext extracts them for all languages and splits them into the actual usage example, a reference, and an English translation (when present). The parsed usage examples are directly useful for training NLP applications.\footnote{\url{https://wikidata.org}}
language-specific Wiktionary. We expect the non-English coverage of the English Wiktionary to further improve going forward.

Parsing other Wiktionary editions with non-English glosses using the largely the same code is possible (template and Lua module mechanisms are the same). However, each edition uses a different set of templates and modules and somewhat different formatting. It would also be helpful to canonicalize tags, topics, etc. across editions. Thus implementing, testing, and maintaining each separate extractor is a significant effort.

6. Conclusion

Wiktionary is the most comprehensive free dictionary available, and Wiktextract converts it to rich structured machine-readable data with pronunciations, translations, inflection tables, lexical-semantic relations, etymology, and linguistic, semantic, and topical annotation. The English Wiktionary seems to have a better coverage of many languages than their respective language-specific Wiktionaries.

Expanding templates and Lua modules during extraction enables a more robust, maintainable, and complete extraction than prior approaches. The extracted data is freely available for download and is regularly updated, which lets users fix errors and add missing data by editing Wiktionary.

The resource will be useful for linguists, application developers, and for building hybrid artificial intelligence systems.

7. Bibliographical References

Ács, J. (2013). Building basic vocabulary across 40 languages. In Proceedings of the 6th Workshop on Building and Using Comparable Corpora, pages 52–58. Association for Computational Linguistics.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY - a large-scale unified lexical-semantic resource based on LMF. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 580–590. Association for Computational Linguistics.

Karlsson, F. (1982). Suomen kielen äänene- ja muotokirja. WSOY.

Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very large scale parsing and normalization of Wiktionary morphological paradigms. In 10th International Conference on Language Resources and Evaluation (LREC), pages 3121–3126. European Language Resources Association.

Krizhanovskaya, N. B. and Krizhanovsky, A. A. (2019). Semi-automatic methods for adding words to the dictionary of VepKar corpus based on inflectional rules extracted from Wiktionary. arXiv:2001.04719.

Krizhanovsky, A. and Lin, F. (2009). Related terms search based on WordNet / Wiktionary and its application in ontology matching. In Proceedings of the 11th Russian Conference on Digital Libraries (RCDL), Petrozavodsk, Russia.

Kurmas, Z. (2010). Zawilinski: A library for studying grammar in Wiktionary. In Proceedings of WikiSym. Association for Computing Machinery (ACM).

Lee, J. L., Ashby, L. F. E., Garza, M. E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A. D., and Gorman, K. (2020). Massively multilingual pronunciation modeling with WikiPron. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC), pages 4223–4228. Association for Computational Linguistics.

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 (Short Papers), pages 414–418. 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), pages 147–161.

Miletic, A. (2017). Building a morphosyntactic lexicon for Serbian using Wiktionary. In Journées d’étude toulousaines: Les interfaces en sciences du langage, pages 30–34. 6th edition, May.

Nastase, V. and Strapparava, C. (2015). knoWitiary: A machine readable incarnation of Wiktionary. International Journal of Computational Linguistics and Applications, 6(2), December.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence.

Pérez, L. A., Oliveira, H. G., and Gomes, P. (2011). Extracting lexical-semantic knowledge from the Portuguese Wiktionary. In Proceedings of the 15th Portuguese Conference on Artificial Intelligence (EPIA), pages 703–717. Springer.

Sajous, F. and Hathout, N. (2015). GLAWI: A free XML-encoded machine-readable dictionary built from the French wiktorynaire. In Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, pages 405–426.

Sajous, F., Calderone, B., and Hathout, N. (2020). ENGLAWI: From human- to machine-readable Wiktionary. In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC), pages 3016–3026. European Language Resources Association.

Sennrich, R. and Kunz, B. (2014). Znorge: A German morphological lexicon extracted from Wiktionary.
In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC), pages 1063–1067. European Language Resources Association.

Sérasset, G. and Tchechmedjiev, A. (2014). Dbnary: Wiktionary as linked data for 12 language editions with enhanced translation relations. In Proceedings of the 3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing, pages 67–71.

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).