Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms

Christo Kirov, John Sylak-Glassman, Roger Que, David Yarowsky
Center for Language and Speech Processing, Johns Hopkins University
Baltimore, MD 21218
ckirov@gmail.com, jcsg@jhu.edu, query@jhu.edu, yarowsky@jhu.edu

Abstract

Wiktionary is a large-scale resource for cross-lingual lexical information with great potential utility for machine translation (MT) and many other NLP tasks, especially automatic morphological analysis and generation. However, it is designed primarily for human viewing rather than machine readability, and presents numerous challenges for generalized parsing and extraction due to a lack of standardized formatting and grammatical descriptor definitions. This paper describes a large-scale effort to automatically extract and standardize the data in Wiktionary and make it available for use by the NLP research community. The methodological innovations include a multidimensional table parsing algorithm, a cross-lexeme, token-frequency-based method of separating inflectional form data from grammatical descriptors, the normalization of grammatical descriptors to a unified annotation scheme that accounts for cross-linguistic diversity, and a verification and correction process that exploits within-language, cross-lexeme table format consistency to minimize human effort. The effort described here resulted in the extraction of a uniquely large normalized resource of nearly 1,000,000 inflectional paradigms across 350 languages. Evaluation shows that even though the data is extracted using a language-independent approach, it is comparable in quantity and quality to data extracted using hand-tuned, language-specific approaches.

Keywords: Wiktionary, Morphology, UniMorph, Multilingual Resources

1. Introduction

Wiktionary[1] is one of the largest sources of lexemes (or lemmas) with morphological paradigms across a wide range of languages, and exhibits substantial ongoing growth. This makes it an ideal source of data for broadly cross-lingual machine learning in NLP, with potential applications in machine translation, information extraction, and many other tasks. However, the human-focused, crowd-sourced nature of Wiktionary presents a number of challenges to gathering and using this data.

Wiktionary is composed of a number of editions, with each edition written for readers of a particular language (e.g., the French edition is written for French readers). Each *edition* covers lemmas in many *languages* (i.e., both the French and English editions of Wiktionary contain Swahili lemmas). There is no fixed set of standards for how lexical information should be presented. This leads to many formatting idiosyncrasies across editions, languages, and lemmas.

Our aim is to find and exploit any regularities in how authors choose to represent lemmas and morphological paradigms to extract and standardize the information in Wiktionary into a form suitable for machine processing. Our current focus for this effort is capturing inflectional morphology, although we also extract data relevant to derivational morphology and word-to-word translation. To achieve standardized output, we develop an original, robust parsing algorithm that operates directly on the surface HTML of Wiktionary, and requires minimal edition-specific or language-specific tuning. The output of our generalized approach produces data of similar quality to previous fine-tuned language-specific extractors, although in much greater quantities. Our code and data will be made available under an open-source license as part of the UniMorph project.[2] Data will be provided in a tabular format that includes columns identifying the language, base lemma, inflected form, and grammatical features.

2. Extracting Inflectional Paradigms from Wiktionary

Wiktionary contains a large amount of inflectional morphology data, typically in the form of paradigm tables. Unfortunately, Wiktionary pages are written for human consumption, and there are no fixed standards for how morphological data should be presented or coded. Paradigm table layouts are inconsistent across Wiktionary editions, languages, and, in some cases, even lemmas. The grammatical descriptors used to describe morphological forms are also not consistently defined or applied, and are left to the discretion of the Wiktionary authors. We applied a novel multidimensional parsing approach to overcome these inconsistencies (Sylak-Glassman et al., 2015b). This ultimately resulted in a series of tuples containing an inflected form, its lemma (i.e., citation form), and a vector of language-independent, standardized morphological features. All work in this paper is based on a snapshot of Wiktionary from June 20, 2015, provided as an OpenZIM (.zim) archive by the Wikimedia Foundation.[3]

2.1. Extraction from HTML Tables

Figure 1 shows pieces of a typical Wiktionary inflection table (the French verb *ouvrir*, 'to open') across three editions of Wiktionary. Here, we focus on the English edition. The cells are divided into inflected forms, and grammatical descriptors that indicate the morphological features of the form. An initial challenge for our parser was to

[1] http://www.wiktionary.org

[2] Temporarily located at http://ckirov.github.io/UniMorph/

[3] https://dumps.wikimedia.org/

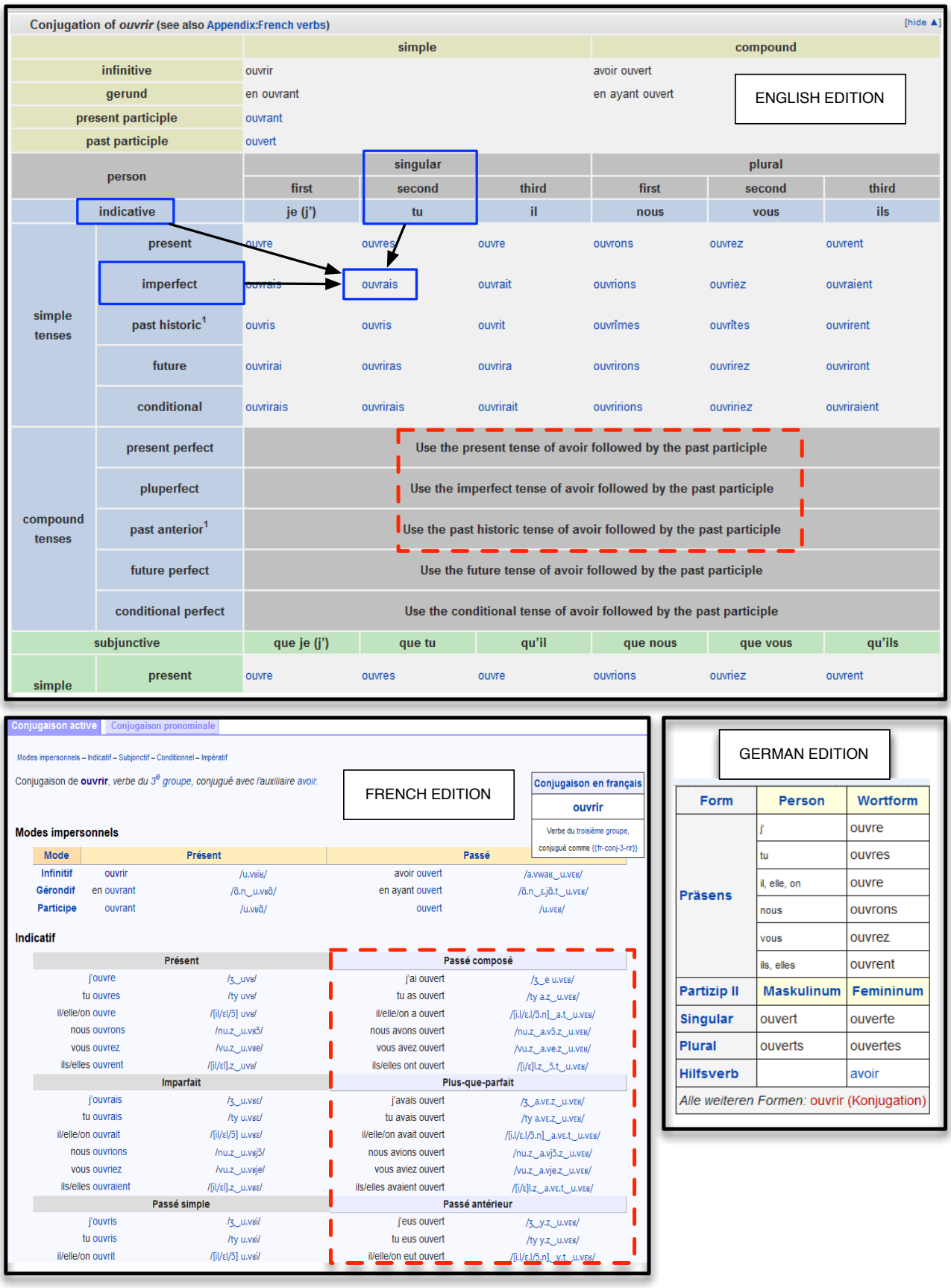

Figure 1: Comparison of the representation of the French verb *ouvrir* ('open') in the English, French, and German editions of Wiktionary.

distinguish which cells are forms, and which are descriptors. We noted that although Wiktionary's inflection tables have many different layouts, grammatical descriptors such as **singular** tended to appear on many pages, while inflected forms such as *ouvrons* appeared much less frequently, usually only once. We exploited this tendency by counting the number of pages in which each distinct cell text in a Wiktionary edition appeared, and then manually determined a cutoff point for each language above which any cell with matching text was considered to be a grammatical descriptor.

Next, we matched each inflected form cell with the headers immediately to its north, west, and northwest. This resulted in a list ordered by distance from the content cell. Headers that spanned multiple rows or columns (e.g., **singular** and **indicative** in Figure 1) were assigned to all content cells in those rows or colums. Figure 1 shows an example where *ouvrais* (determined to be a form of the verb *ouvrir* based on HTML headers external to the table) is associated with **tu**, **second**, **singular**, **imperfect**, and **indicative**.

2.2. Mapping Inflected Forms to Universal Features

The grammatical descriptors discovered by the frequency-based preprocessing step were manually assigned sets of features from the UniMorph Schema (previously the Universal Morphological Feature Schema), which was designed to represent the meanings 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 to represent inflectional morphology in any language, especially low-resource languages. It includes over 212 features distributed among 23 dimensions of meaning (i.e., morphological categories), which include both common dimensions like tense and aspect as well as rarer dimensions like evidentiality and switch-reference.

All inflected forms found by our scrape of Wiktionary were assigned complete UniMorph Schema vectors by looking up each of their Wiktionary descriptors using the manual mapping described above. Conflicts within a dimension of meaning were resolved using a positional metric that privileged descriptors nearer to the inflected form in its originating table.

Ultimately, the process of extraction and mapping yielded 952,530 unique noun, verb, and adjective lemmas across 350 languages (of which 130 had more than 100 lemmas), with each inflected form of the lemma described by a vector of features from the UniMorph Schema. Table 5 shows the counts of noun, verb, and adjective lemmas, along with their average inflectional complexity, for the set of languages in Wiktionary with over 1,000 lemmas.

3. Verifying and Correcting Extracted Paradigms

Although assessing the linguistic accuracy of the content of Wiktionary is beyond the scope of this work, we would like to ensure that our automated extraction method accurately retrieves the information that Wiktionary authors have entered.

Just as grammatical descriptors are likely to be re-used in multiple inflection tables, authors are likely to re-use table layouts. Each table or parenthetical list has a 'signature' consisting of the set of row, column, and corner descriptors that our parser has extracted. These signatures represent unique table layouts, and there are only a small number of such signatures in any given language (123.35 on average — many fewer than the total number of lemmas). Since our extraction method is deterministic, all lemmas with the same signature will be parsed in the same way. Thus, it is sufficient for a human to verify and correct a single lemma with a particular signature to determine that all lemmas with that same signature are correctly extracted. Furthermore, verification and correction consists of writing rules in a consistent syntax that can be automatically and directly applied to the parsed data, producing corrected output without needing to modify the parser code. To the extent that Wiktionary authors re-use table layouts for future lexical entries, the rules can also be re-used as Wiktionary itself is updated over time. The rule syntax enables editors to remove erroneous table types or table entries, and to add or remove grammatical descriptors associated with full tables or specific inflected forms.

The amount of human effort required to complete the verification and correction process varies significantly by language, from a few minutes to several hours. The effort required depends on the number, size, and initial accuracy of the table layouts that are present in a language. Many languages contain table layouts that do not correspond to morphological paradigms and can be safely removed from the dataset without further editing. These include, for example, language-specific versions of the periodic table of elements and tables containing cardinal numbers and their ordinal equivalents. The human effort required also increases when the language is in a script that is unfamiliar to the editor.

A verified and corrected subset of the entire Wiktionary database is currently available as part of the SIGMORPHON 2016 Shared Task on Morphological Reinflection.[4] This includes data for the 8 languages shown in Table 1.

Language	# Lemmas
Arabic	5,383
Finnish	81,845
Georgian	9,142
German	34,371
Navajo	605
Russian	22,422
Spanish	37,380
Turkish	4,356

Table 1: Wiktionary subset for SIGMORPHON 2016.

4. Extraction of Derivational Paradigms and Lemma Translations

Although our focus so far has been on extracting inflectional paradigms, we also mined additional information

[4] Accessible at: http://ryancotterell.github.io/sigmorphon2016/

from Wiktionary pages that will be useful for downstream NLP tasks. A small number of pages in the English edition of Wiktionary contain lists of words under the HTML headings 'Related/Derived Terms.' For example, 'sunflower' appears in the list of derived terms for the base lemma 'flower.' For each lemma in each language, we record any associated terms. At present, an average of 3.42 derived terms are extracted from each of the 76,038 lemmas (occurring across all languages) that contain associated terms. Table 5 indicates the number of lemmas in a language with associated derivations, as well as the mean and standard deviation of the number of derivations for each lemma. At a later time, this data could be used to develop automatic analysis and generation of derivational morphology.

For certain lemmas, Wiktionary also contains tables of translated terms. These all follow the same format, listing pairs composed of a language and the lemma translated into that language. For example, the translation table associated with the English lemma 'flower' contains the entry 'Danish: blomstre.' For any lemma page that contains a translation table, we record all listed translations. In our current extraction, lemmas that contain these tables have an average of 3.54 translations each. Table 5 shows the number of lemmas in a language with associated translations, as well as the mean and standard deviation of the number of translations for each lemma.

5. Comparison to Previous Approaches to Wiktionary Extraction

Other authors have previously extracted small portions of Wiktionary to generate task-specific training data. Liebeck and Conrad (2015) developed a limited extractor specific to German nouns and verbs, which they used to train a lemmatization system. On a slightly larger scale, Durrett and DeNero (2013) created finely-tuned extractors designed specifically for German, Spanish, and Finnish verbal and nominal inflectional paradigms in order to generate training sets for their automatic inflection learner.

Table 2 shows that our own system, UniMorph, which operates in a language-agnostic manner, was able to extract a number of paradigms comparable to these language-specific extractors. Overall, we were able to generate suitable[5] training data for 123 language-POS pairs across 88 languages, compared to Durrett and DeNero's 5 pairs across 3 languages. The average number of inflected forms collected per paradigm differs across systems, as Liebeck and Conrad (2015) only considered forms which fit certain Wiktionary templates and Durrett and DeNero (2013) extracted only paradigms for which they could obtain a fixed set of forms (their software requires all training paradigms to be equal in size). In contrast, our system extracts all paradigms, regardless of completeness.[6]

Furthermore, the quality of our data was similar to that of the finely-tuned data extracted by Durrett and DeNero (2013). We were able to reformat our data to serve as input to their inflection learner, keeping only paradigms with a fixed set of forms. Table 3 compares performance of the two datasets.

Lang/POS	D&D	L&C	UM
German N	2764	52092	21746
Forms/Lemma	8	7.94	5.10
German V	2027	6033	3606
Forms/Lemma	27	8.38	103.00
Spanish V	4055	N/A	5982
Forms/Lemma	57	N/A	66.39
Finnish N/Adj	40589	N/A	56693
Forms/Lemma	28	N/A	32.49
Finnish V	7249	N/A	8709
Forms/Lemma	53	N/A	128.46

Table 2: Comparison of extracted data for Durrett and DeNero (2013), Liebeck and Conrad (2015), and UniMorph.

	Paradigm Match		Indiv. Form Match	
Lang/POS	D&D	UM	D&D	UM
German V	85.0%	78.5%	96.2%	93.0%
Spanish V	95%	96.5%	99.7%	99.1%
Finnish V	87.5%	61.0%	96.4%	94.1%

Table 3: Performance of Durrett and DeNero's (2013) paradigm completion software trained on their finely-tuned data and the data extracted by the general UniMorph methods presented here.

6. Extending to International Editions

While much of the information in different editions of Wiktionary overlaps due to each edition describing many of the same languages, editions may contain complementary data in the form of different lemmas or more complete paradigms. Figure 1 highlights differences in the representation of French verbs across the English, French, and German editions of Wiktionary. For example, while the English edition lists certain verb forms using only human-readable instructions (e.g., "Use the present tense of avoir followed by the past participle"), the French edition lists the forms explicitly. Furthermore, both the English and French editions offer much better form coverage than the German edition. Merging all available editions therefore has the potential to fill gaps in coverage. Table 4 shows the increased coverage of French, German, and Spanish verbs possible when the English edition is merged with the French or German editions.

However, using international editions poses unique challenges. In particular, international editions use language names and grammatical terms that differ from those in the English edition. As a first pass at this problem, we attempt to translate foreign grammatical descriptors into English using Google Translate. We then look up the translated terms in our mapping files as normal. Unfortunately, due to the limitations of Google Translate, this method permits coverage of only a small portion of international edition data.

[5] In this context, suitable means 200 or more fully inflected lemmas.

[6] Note that paradigms may be incomplete not only because of partial entry by Wiktionary authors, but also due to being systematically defective. For example, Latin deponent verbs have only morphologically passive forms.

Edition →	English	French	German
French V	11,006	24,742 (19,282)	1,249 (102)
Forms/Verb	35	102	9
German V	3,606	446 (169)	5,470 (2,993)
Forms/Verb	103	92	535
Spanish V	5,982	N/A	55 (5)
Forms/Verb	66	N/A	117

Table 4: Lemma coverage provided by foreign editions. Additional lemmas not in the English edition are in parentheses. At the time this data was extracted, the French edition did not contain Spanish verb forms.

A more complex solution with the potential for greater coverage is to use a consensus-based mapping of the grammatical terms. For any foreign feature descriptor, we can find all the inflected forms in the foreign edition that have that descriptor and also exist in the English edition. We can then assign the most common shared feature label among the English forms to the foreign descriptor. This work is left for future releases of the Wiktionary database.

7. Conclusion

We have presented a generalized method of extracting lexical and morphological information from Wiktionary. The method works on surface HTML, and successfully handles the extensive variations in edition, language, lemma, and author-specific idiosyncrasies inherent in the Wiktionary format. The method produced high-quality inflectional paradigm data for 952,530 lemmas across 350 languages, contributing a uniquely large, cross-lingual normalized resource of high-quality morphological information that can support downstream NLP tasks across a very wide range of the world's languages.

8. Bibliographical References

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.

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. Association for Computational Linguistics, Atlanta.

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 (ACL-IJCNLP)*, pages 414–418, Beijing, July. Association for Computational Linguistics.

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.

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

Lang	N Lemmas	N Forms/Lemma	V Lem.	V For./Lem.	ADJ Lem.	ADJ For./Lem.	# Der.	M/SD	# Tr.	M/SD
English	159917	1.04	23532	3.11	42552	2.02	21027	4.21/11.22	243913	6.77/20.32
Italian	43142	1.18	17246	21.60	20270	2.37	4624	3.01/14.45	78162	1.62/1.25
Finnish	49458	32.41	8709	128.46	7235	33.05	5280	3.53/5.14	52424	1.46/1.16
French	24508	1.10	11006	34.55	9653	2.13	3589	2.77/5.65	41761	1.57/1.15
Serbo-Croatian	25231	12.51	9230	66.00	8714	93.73	1176	2.01/0.20	46639	1.77/1.32
Japanese	29471	1.84	7708	35.62	882	25.87	1188	4.91/8.77	33820	1.88/1.78
Spanish	19529	1.12	5982	66.39	6853	2.39	1873	2.55/1.67	31480	1.64/1.28
German	21746	5.11	3606	103.00	5469	99.22	2856	3.27/3.74	32303	1.53/1.15
Portuguese	19073	1.16	4384	78.96	6170	3.33	358	1.96/0.93	14438	1.70/1.19
Dutch	19954	1.79	3531	27.03	3163	13.44	2943	3.02/3.63	18342	1.70/1.29
Esperanto	14191	2.90	1799	59.66	8878	3.03	1109	2.89/3.10	11632	1.25/0.60
Latin	9128	12.62	4586	112.76	7358	33.33	1978	2.67/2.72	22848	2.39/2.20
Russian	13679	12.44	2646	27.68	2382	25.93	617	3.25/3.28	22560	2.26/2.01
Swedish	8799	7.49	2250	15.57	1862	10.88	514	3.14/3.47	14698	1.67/1.28
Hungarian	8542	47.14	1730	88.50	1575	7.07	4199	3.77/3.24	15845	1.56/1.15
Greek	7377	8.19	1551	10.43	1721	30.70	459	2.39/1.73	0	0.00/0.00
Polish	7128	13.47	939	95.33	1815	35.96	1372	2.43/1.10	10900	1.42/0.92
Georgian	6316	13.46	67	75.25	2663	7.82	131	2.07/0.56	12482	1.67/1.40
Irish	7405	11.54	622	73.07	265	3.30	849	2.00/0.36	10540	1.89/2.05
Catalan	4790	1.14	1403	63.90	1949	2.65	408	2.05/0.45	8637	1.39/0.98
Latvian	4093	12.88	510	49.57	3013	27.96	292	1.99/0.14	7682	1.73/1.29
Armenian	5096	33.62	838	209.56	1298	36.70	242	2.07/0.30	9066	2.08/2.04
Icelandic	4988	13.35	1219	73.64	756	80.80	1693	3.08/3.82	8452	1.85/1.44
Romanian	3830	4.25	1233	42.83	874	17.57	922	2.28/1.36	8003	1.97/1.68
Norwegian Bokmål	4036	3.36	555	4.97	811	2.46	118	3.40/7.13	2145	1.47/0.98
Korean	4009	1.09	867	64.69	324	48.97	748	4.59/5.09	7734	1.90/3.92
Scottish Gaelic	3912	1.91	621	4.56	160	6.24	0	0.00/0.00	0	0.00/0.00
Norwegian Nynorsk	3587	2.96	401	8.29	674	2.99	118	3.40/7.13	2145	1.47/0.98
Ido	2522	1.00	2110	17.56	6	1.00	371	2.15/1.00	2598	1.29/0.67
Old French	2818	3.58	1489	61.91	329	7.53	0	0.00/0.00	0	0.00/0.00
Old Armenian	2338	15.78	1460	46.57	652	15.65	0	0.00/0.00	0	0.00/0.00
Danish	3259	7.96	585	5.37	606	2.46	292	2.36/1.83	5852	2.05/1.95
Macedonian	639	9.60	1408	40.85	2382	11.38	26	2.00/0.73	5697	1.56/0.93
Turkish	3729	56.42	497	114.69	104	80.90	652	2.60/1.18	6192	1.40/1.01
Czech	1380	14.19	1277	42.20	1269	54.61	2911	2.82/1.91	19576	1.31/0.77
Jèrriais	2965	1.04	116	1.03	549	2.32	0	0.00/0.00	0	0.00/0.00
Asturian	2470	1.09	373	62.12	678	3.77	34	1.85/0.60	2630	1.35/0.91
Arabic	2022	25.90	954	149.49	449	47.78	245	2.24/0.99	4688	3.33/3.82
Ancient Greek	2197	16.02	563	321.08	610	41.56	0	0.00/0.00	0	0.00/0.00
Luxembourgish	2043	1.02	922	15.02	355	13.25	53	2.51/1.13	4148	1.55/1.11
Faroese	2143	18.07	651	35.59	418	30.49	223	3.11/3.92	3393	1.77/1.35
Slovene	2205	17.58	177	7.22	663	47.70	99	2.03/0.22	0	0.00/0.00
Bulgarian	1646	6.54	715	71.78	582	26.03	69	3.17/2.59	3003	3.10/3.21
Galician	2426	1.14	434	90.03	2	1.00	95	1.79/0.90	3572	1.35/0.81
Albanian	1560	6.76	797	39.48	345	2.39	314	2.17/0.72	4051	1.72/1.14
Persian	1800	2.76	474	116.46	316	3.73	407	2.01/0.23	6088	1.95/1.86
Middle French	1395	1.09	516	58.79	387	2.12	0	0.00/0.00	0	0.00/0.00
Manx	1616	2.72	414	2.36	234	2.72	751	2.03/0.24	5611	2.00/2.22
Old English	1022	6.43	712	20.13	306	58.04	0	0.00/0.00	0	0.00/0.00
Hebrew	1512	2.26	76	35.57	399	2.85	228	2.16/0.87	5187	1.79/1.37
Venetian	1147	1.18	373	32.81	291	2.55	43	2.49/1.21	2144	1.72/1.05
Ukrainian	1625	14.42	64	22.81	41	25.51	11	1.91/0.51	961	1.94/1.63
Hindi	1526	2.10	116	211.97	43	12.00	101	2.16/0.97	3125	2.45/2.22
Estonian	1343	20.45	192	121.62	135	12.97	82	2.10/0.64	3154	1.31/0.77
Welsh	1213	2.24	212	57.54	180	4.43	112	2.28/0.85	2058	1.64/1.23
Adyghe	1511	9.85	5	10.00	78	9.74	1	4.00/0.00	2703	1.41/0.84
Volapük	1449	12.63	0	0	107	12.00	0	0.00/0.00	0	0.00/0.00
Classical Syriac	1453	24.40	0	0	33	12.52	0	0.00/0.00	0	0.00/0.00
Lithuanian	879	13.66	284	41.68	191	86.09	210	2.00/0.07	2204	1.35/0.88
Norman	1092	1.03	44	1.07	158	1.68	8	2.00/0.00	0	0.00/0.00
Malay	966	1.33	143	1.44	135	1.03	45	2.98/3.99	0	0.00/0.00
Crimean Tatar	1154	6.02	59	4.98	21	6.00	0	0.00/0.00	0	0.00/0.00
Romansch	1012	1.03	11	39.91	177	2.98	65	2.00/0.00	0	0.00/0.00
Afrikaans	821	1.31	169	3.11	202	1.97	42	2.10/1.78	1214	1.42/1.06
Urdu	1074	1.72	54	215.31	18	11.72	50	2.02/0.14	1881	2.94/3.07
Slovak	945	11.40	31	28.42	158	48.53	228	2.32/0.96	2170	1.28/0.67
Sanskrit	893	29.82	7	23.57	196	72.47	20	2.00/0.55	1674	5.47/6.47
Ladin	495	1.00	286	17.93	228	2.99	7	1.57/0.73	1234	1.48/0.89

Table 5: Wiktionary data quantity by language, sorted by number of noun lemmas. Abbreviations in table header: ADJ = adjective, # Der. = number of derived terms, For. = forms, Lem. = lemma(s), M/SD = mean / standard deviation, N = noun, # Tr. = number of translated terms, V = verb.