Wiktionary
The free dictionary

Wiktionary

Wiktionary:Frequency lists

Unless otherwise specified, the frequency lists linked from here count distinct orthographic words (not lemmas), including inflected and some capitalised forms. For example, the verb "to be" is represented by "is", "are", "were", and so on.

Suggestions for how to use these lists

Frequency lists have many applications in the realm of second language acquisition and beyond. One use for such lists in the context of the Wiktionary project is as an aid in identifying missing terms of high-frequency and thus, it is assumed, of high priority. Since English Wiktionary aims not just to be a mere database of lemmas, but a multi-directional, multi-lingual dictionary aimed at English speaking users, there are certain advantages to lists which include inflected forms as well. These forms reflect words as they are likely to be encountered and thus as they may be used in lookup.

Feel free to add definitions for words on these lists if you know the languages involved! Even better if you can include usage citations and references. If you are involved in another non-English language edition of Wiktionary, you might also consider implementing or expanding on this idea, if there is not already something similar in place. If you see a word in this list that is *clearly* out of place (wrong language, punctuation, superfluous capitalisation), you are welcome to remove it. While creating entries for words, please leave valid bluelinks in place as these pages may be copied for use with other language projects in the future.

However, this system is far from perfect due to the variable quality of the source data and the automated nature of processing. Thus a word's presence in any of these lists is merely an invitation for further investigation as to whether an entry is warranted. Please be mindful that there will be many words which

- are misspelt,
- are not commonly accepted words of the language for which they appear,
- are not words at all,
- or do not fulfil the Wiktionary Criteria for Inclusion.

Collocations may or may not warrant their own individual entries, and not necessarily in the exact form they appear here. As an aid to navigating this list, consider enabling the OrangeLinks.js gadget to reveal headword pages which exist (and so will still show a blue link) but which do not yet contain an entry for the relevant language. Please be mindful too that not all of the resources listed here are suitable for use directly in Wiktionary, mainly due to problems with licensing compatibilities.

Frequency lists

Adnyamathanha · Albanian · Alemannic German · Ancient Greek · Arabic · Belarusian · Bulgarian · Cantonese · Catalan · Cebuano · Czech · Danish · Dutch · Eastern Mari · English · Erzya · Esperanto · Estonian · Finnish · French · Galician · Georgian · German · Greenlandic · Hebrew · Hindi · Hungarian · Icelandic · Indonesian · Irish · Italian · Jamaican Creole · Japanese · Kabardian · Kapampangan · Kashubian · Khmer · Korean · Ladino · Latin · Latvian · Lithuanian · Low German · Lower Sorbian · Lü · Macedonian · Malay · Mandarin · Manx · Māori · Marshallese · Modern Greek · Nepali · North Frisian · Northern Sami · Norwegian · Odia · Ossetian · Palauan · Persian · Pitjantjatjara · Polish · Portuguese · Romanian · Russian · Sanskrit · Scots · Serbo-Croatian · Silesian · Slovak · Slovene · Spanish · Swahili · Swedish · Tagalog · Telugu · Thai · Turkish · Udmurt · Ukrainian · Upper Sorbian · Uyghur · Vietnamese · Welsh · West Frisian · Western Mari · Yiddish

See also

- Appendix:Swadesh lists
- Appendix:Vocabulary lists
- Appendix:Character frequency
- Wiktionary:List of languages
- Wiktionary:Multilingual statistics
- Category:Basic word lists by language
- Category:Basic word lists by family

External links

Resources covering many languages

- Word frequency lists from 10K up to 1M+ for 270+ languages, available for download as part of the Leipzig Corpora Collection (https://wortschatz.uni-leipzig.de/en/download/) (CC BY-4.0)
- 50K and larger word lists based on www.opensubtitles.org for 60+ Languages (https://github.com/hermitdave/FrequencyWords/) (CC BY-SA-4.0)
- Frequency lists (http://corpus.leeds.ac.uk/list.html) for English, Russian, Arabic, Chinese, French, German, Greek, Italian, Japanese, Portuguese and Spanish derived from corpora assembled by Leeds University's Centre for Translation Studies (CC BY-2.5)
- The wordfreq Python library (https://github.com/rspeer/wordfreq) contains large frequency lists for 40+ languages. (Data under various licence conditions (https://github.com/rspeer/wordfreq#license), some of which may be incompatible with Wiktionary.)
- Frequency lists for learners (http://corpus.leeds.ac.uk/serge/kelly/) of Arabic, Chinese, English, Greek, Italian, Norwegian, Polish, Russian and Swedish, available as part of the Kelly project. Swedish: (CC-BY-SA 3.0, LGPL 3.0); (all others: CC BY-ND-NC-SA 2.0, meaning they are incompatible with wiktionary)
- The SEAlang Library (http://sealang.net/) aims to collect lexical resources for the languages of South-East Asia. Resources are available for Balinese, Burmese, Indonesian, Javanese, Karen, Khmer, Lao, Malay, Maguindanao, Maranao, Mon, Shan, Thai, Vietnamese, among others. (Some resources are available under a generic CC license, however others are covered by copyright. You should check on an individual basis.)

- Wordlists in the CLARIN infrastructure (https://www.clarin.eu/resource-families/lexical-resources-wordlists) - just over half are monolingual lists in 10 languages (Dutch, Estonian, Finnish, German, Greek, Maltese, Ngbugu, Slovenian, Spanish, Swedish), while the other two dozen are in bilingual and multilingual combinations. (Some resources are available under a permissive or copyleft license, however others may be covered by copyright. You should check on an individual basis.)
- Gimenes, Manuel, and Boris New. "Worldlex: Twitter and blog word frequencies for 66 languages." *Behavior research methods*, 2015, pp. 1-10. PDF (https://drive.google.com/file/d/0B-sE9ac1ksCANWFVN3ZacHFWQ0k/view?resourcekey=0-ehEuEV6a8PFQiOzQq97cjw), data (http://www.lexique.org/?page_id=250).

Retrieved from "https://en.wiktionary.org/w/index.php?title=Wiktionary:Frequency_lists&oldid=83585274"