Wikitext & Wiktionary studies

Inside-out parsing of Wikitext and extraction of a large multilingual machine-readable dictionary with translations, pronunciations, inflections, and more

Tatu Ylönen, Nov 11, 2020

Terminology

- **Wiktionary** – huge free multilingual dictionary (several language versions and the English version contains many languages)
- **Wikitext** – the format used in Wiktionary, Wikipedia, etc
- **MediaWiki** – the platform used to run these resources
- **WikiMedia** – the organization running the services
- **Template** – a macro facility in MediaWiki
- **Scribunto** – an extension for writing modules in **Lua** programming language

Background

- I was looking into Finnish inflection in 2018 and needed inflection classes
- Extracted inflection classes from Wiktionary and wrote a Finnish inflection generator: https://github.com/tatuylonen/wiktfinnish
- Last fall I turned the extractor into a general-purpose multi-language extractor: https://github.com/tatuylonen/wiktextract
- Maintenance headaches, so in September I decided to look into **expanding templates** and **Lua macros** and **rewriting the extractor**.
- This started as and still is a side-project, but may result in a useful resource.

Objective

- Create a free multilingual interlinked machine-readable dictionary
  - c. 10 million word senses
  - Hundreds of languages, dozens with decent coverage
  - Pronunciations, word senses, glosses, conjugations/declensions, translations, semantic linkages, semantic and topical annotation
  - English language glosses
  - Aim for high quality and accuracy
  - Automatically updates monthly
- Side product: general-purpose Python module for Wikitext processing

Related work

- Wikitext parsers: https://www.mediawiki.org/wiki/Alternative_parsers
- Wiktionary & combined extractions: BabelNet, UBY, Dbnary, ...
- Extractors for various languages: about two dozen papers...
- Translation extractors: there are a few
- Template/Lua processing: None that I know of (maybe parsoid)
- Sense alignment: several papers, done in BabelNet etc
- Current extractions generally lack conjugation/declension info, pronunciations, and don’t combine all types of data, or data formats/licensing is cumbersome
- Many current extractions and parsers have significant quality issues

What’s new in this

- As far as I know, this is the first Python package that can expand Scribunto Lua macros
- Full parsing and full/partial template expansion – previous Python parsers only seem to do partial parsing and no template expansion
- The scope of the extraction: conjugation/declension info, pronunciations, tags, translations, topical annotations, etc (various packages do some of it)
- Many issues can be fixed by just editing Wiktionary
- Hopefully, better quality of the extraction due to better methods

I don’t currently do

- Sense alignment of translations and glosses (several other resources do, algorithms are published)
- Alignment with WordNet, Wiktionary and other resources (cf. BabelNet)
- Alignment with image databases (cf. BabelNet)

Wikitext

- ===This is a subtitle===
- ‘‘‘**bold**‘‘‘ ‘‘*italic*‘‘ ‘‘‘‘‘***combined***‘‘‘ ‘‘*italic*‘‘
- [[page title|display title]]
- [external url]
- {{template|arg1|arg2|named=value}}
- {{{argname|default value}}}
- {{{#invoke:module|function name|args}}
- <gallery>…</gallery> … <nowiki />

In practice

- {{{{{templname|def}}}}|[[{{{{{#if:{{{1|}}}|{{{2}}}|{{{3}}}}}}}|{{{name|}}}}}]]}}
- Templates generate table starts / ends but not content; templates generate list items – not possible to parse their output before expansion
- Lua invocations generate wikitext
- Cannot parse Wikitext before expansion, don’t want to expand everything before parsing because templates contain useful infrmation (e.g., conjugation/declension parameters)

Parsing challenges

- Template syntax is **lexically ambiguous** left-to-right and right-to-left (cannot distinguish {{ {{{ }}} }}
- **Brackets and braces must be counted** to know what | splits
- Template/Lua processing is **first step**, and its output is then parsed as **second step**
- Whole dump file must be read before template/Lua expansion
- Lua code can and does read arbitrary other pages
- Templates and Lua macros generate mixed **HTML and Wikitext**
- Lua code relies on lazy argument expansion
- We want to do only **partial expansion** of templates

Parsing Wikitext template syntax inside-out

- The template syntax is unambiguous if parsed inside-out (Note: not related to PCFG parsing!)
- Basically, I repeatedly regexp-replace templates, arguments, and links by a unique magic character, saving the expansions
- Thus, the innermost construct is parsed first, then the next layer, etc.
- I think the parsing complexity is O(length * nesting depth), practically linear!
- *(I wonder what the class of ambiguous grammars with this property would be called… any suggestions?)*

Example

- {{{{{templname|def}}}}|[[{{{{{#if:{{{1|}}}|{{{2}}}|{{{3}}}}}}}|{{{name|}}}}}}]]}}
- {{<magic1>|[[{{{{{{#if:<magic2>|<magic3>|<magic4>}}}|<magic5>}}}}]]}}
- {{<magic1>|[[{{{<magic6>|<magic5>}}}]]}}
- {{<magic1>|[[<magic7>]]}}
- <magic9>

Parsing rest of it

- Once template syntax has been parsed, I parse subtitles, lists, etc. using a traditional tokenizer and a recursive-descent parser (with some magic for <nowiki /> etc)
- During parsing the magic characters are recursively replaced by their definition, which disambiguates them for the second phase parser
- Template expansion similarly recursively expands the magic characters

Invoking Lua modules

- I use **lupa** Python package to call Lua
- MediaWiki uses old version of Lua, which is incompatible with modern Lua in some ways; lupa cannot easily use old Lua
- I capture Lua loader function, load modules from cache, and use several regexp replaces and compatibility functions to make it run
- I wrote some sandbox code to safely run Lua, interface with Python side, template arguments, page fetching, etc., including compatibility modules for Scribunto libraries (mw.text, mw.title, etc)

Pages needing pre-expand

- Some templates produce Wikitext elements, such table start tags, table rows, or list items (some of these come from Lua code)
- Such templates need to be expanded before parsing tables, lists, etc.
- I use a separate template analyzing step that tries to identify such templates and triggers pre-expansion of such templates and those that cal them, before parsing
- I do not try to analyze which Lua code might produce such elements; Lua-based templates can be manually listed for pre-expansion if desired

Overall processing structure

- Scan the whole dump file, copying preprocessed pages to a cache file (i.e., extract templates, Lua modules)
- Reprocess pages of interest, pre-expanding, parsing, fully expanding e.g. glosses (this calls Lua modules as needed)
- This also allows fast development by saving the cache file from a previous run
- The framework handles parallelization to multiple cores; parallelism scales linearly

Performance

- Full English Wiktionary extraction currently around 7 hours (with 24 cores/48 hyperthreads)
- Produces about 6GB json file
- I also generate a website from the data, slicing & dicing the jsom file to smaller pieces by language, part-of-speech, linguistic tags, topics, etc. (all made available for download)
- The web site currently uses 135GB of disk space
- Extraction + web site updating takes about 12 hours

Project status

- **Still work in progress**
- github:tatuylonen/wikitextprocessor is the general Wikitext/dump parsing framework; it is now fairly stable but needs more Lua code
- github:tatuylonen/wiktextract is the new extractor, approaching completion; 2.0 now expected end of November and 2.1 in December/January
- Website for testing and downloading extracted data at dictionary.kaikki.org; this is still very much work in progress (translations, form-of links still mostly missing)

Metrics – for English Wiktionary only

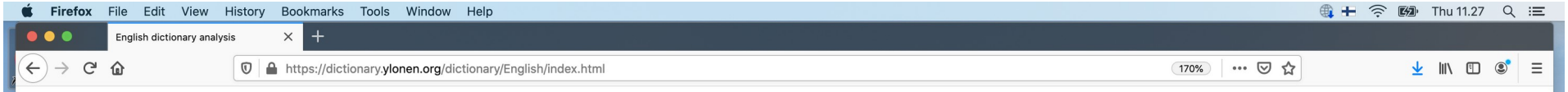
- 18 languages >100 000 senses
- 60 languages >10 000 senses
- Overall contains words for about 2000 languages
- Currently extracts 9.3 million senses total (some additional ones were still not extracted due to bugs as of yesterday, not sure how many)

List of languages extracted from English Wiktionary

XXX work in progress. Do not use this for anything important yet.

Available languages

- All languages combined (9278727 senses)
- Latin (1418947 senses)
- English (1195471 senses)
- Spanish (923075 senses)
- Italian (721193 senses)
- French (460449 senses)
- Russian (448320 senses)
- Portuguese (347592 senses)
- German (309347 senses)
- Latvian (252343 senses)
- Finnish (233857 senses)
- Arabic (176924 senses)
- Catalan (163415 senses)
- Chinese (136686 senses)
- Dutch (132653 senses)
- Japanese (131487 senses)



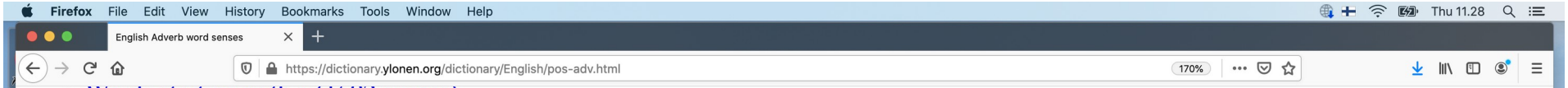
English dictionary analysis

Word lists

- All word forms (946734 distinct words)
- Words with at least 100 word senses (1 words)
- Words with at least 80 word senses (3 words)
- Words with at least 60 word senses (10 words)
- Words with at least 40 word senses (63 words)
- Words with at least 30 word senses (192 words)
- Words with at least 20 word senses (662 words)
- Words with at least 10 word senses (3463 words)
- Words with at least 5 word senses (14202 words)

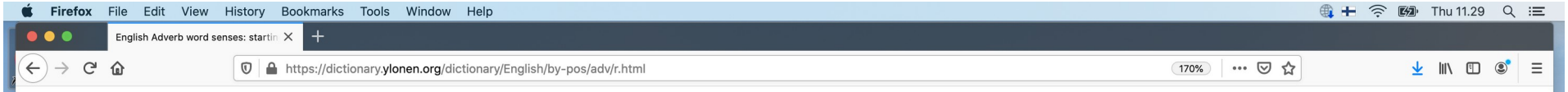
Word sense lists

- All word senses (1195471 senses)
- All non-inflected, non-alternative word senses (1195471 senses)
- Senses by grammatical tag (129 distinct tags)
- Senses by topic (17220 distinct topics)
- Senses with part-of-speech Abbreviation (839)
- Senses with part-of-speech Adjective (166550)
- Senses with part-of-speech Adverb (24138)
- Senses with part-of-speech Affix (4)
- Senses with part-of-speech Article (33)
- Senses with part-of-speech Character (1)



- Words starting with c (1592 senses)
- Words starting with d (1198 senses)
- Words starting with e (1130 senses)
- Words starting with f (989 senses)
- Words starting with g (524 senses)
- Words starting with h (935 senses)
- Words starting with i (1466 senses)
- Words starting with j (141 senses)
- Words starting with k (86 senses)
- Words starting with l (629 senses)
- Words starting with m (1100 senses)
- Words starting with n (909 senses)
- Words starting with o (807 senses)
- Words starting with p (1993 senses)
- Words starting with q (145 senses)
- Words starting with r (781 senses)
- Words starting with s (2500 senses)
- Words starting with t (1240 senses)
- Words starting with u (1584 senses)
- Words starting with v (301 senses)
- Words starting with w (574 senses)
- Words starting with x (15 senses)
- Words starting with y (94 senses)
- Words starting with z (44 senses)
- Words starting with à (8 senses)
- Words starting with æ (5 senses)
- Words starting with œ (2 senses)

Download JSON data for these senses (36.4MB)



English Adverb word senses: starting with r

781 senses starting with r

- rabbinically (Adverb) In a rabbinical way.
- rabidly (Adverb) In a rabid manner.
- racemosely (Adverb) In a racemose manner.
- racialistically (Adverb) In a racialistic manner.
- racially (Adverb) Relating to race.
- racily (Adverb) In a racy manner.
- racingly (Adverb) In a racing manner; at high speed.
- racistically (Adverb) racistly
- racistly (Adverb) in a racist manner
- rackingly (Adverb) So as to cause suffering.
- radiad (Adverb) Toward the radius.
- radially (Adverb) In a radial manner, outward from a center.
- radiantly (Adverb) In a manner that is radiant; glowingly.
- radiately (Adverb) In a radiate manner; with radiation or divergence from a centre.
- radiatingly (Adverb) So as to radiate, or spread outward in rays.
- radiationally (Adverb) In a radiational manner
- radiatively (Adverb) In a radiative manner
- radiatively (Adverb) With regard to radiation
- radicalistically (Adverb) In a radicalistic manner
- radically (Adverb) At the root.
- radically (Adverb) In a radical manner; fundamentally; very.
- radioactively (Adverb) Concerning radioactivity
- radioactively (Adverb) Using a radioactive substance
- radiobiologically (Adverb) By means of, or in terms of, radiobiology

"department" meaning — English dictionary

Noun

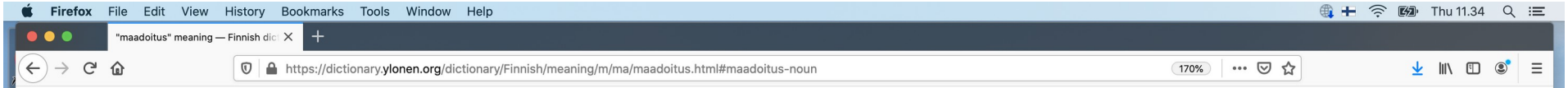
1. **Act of departing; departure.**
   Topic: obsolete
   Synonym: province, specialty(?), ministry(?)
   Derived forms: departmental(?), departmentally(?), Department of Redundancy Department(?), department store(?), fire department(?), interdepartmental(?), police department(?), state department(?), trouser department(?)

2. **A part, portion, or subdivision.**
   Synonym: province, specialty(?), ministry(?)
   Derived forms: departmental(?), departmentally(?), Department of Redundancy Department(?), department store(?), fire department(?), interdepartmental(?), police department(?), state department(?), trouser department(?)

3. **A distinct course of life, action, study, or the like.**
   Synonym: province, specialty(?), ministry(?)
   Derived forms: departmental(?), departmentally(?), Department of Redundancy Department(?), department store(?), fire department(?), interdepartmental(?), police department(?), state department(?), trouser department(?)

4. **A subdivision of an organization.**
   Synonym: province, specialty(?), ministry(?)
   Derived forms: departmental(?), departmentally(?), Department of Redundancy Department(?), department store(?), fire department(?), interdepartmental(?), police department(?), state department(?), trouser department(?)

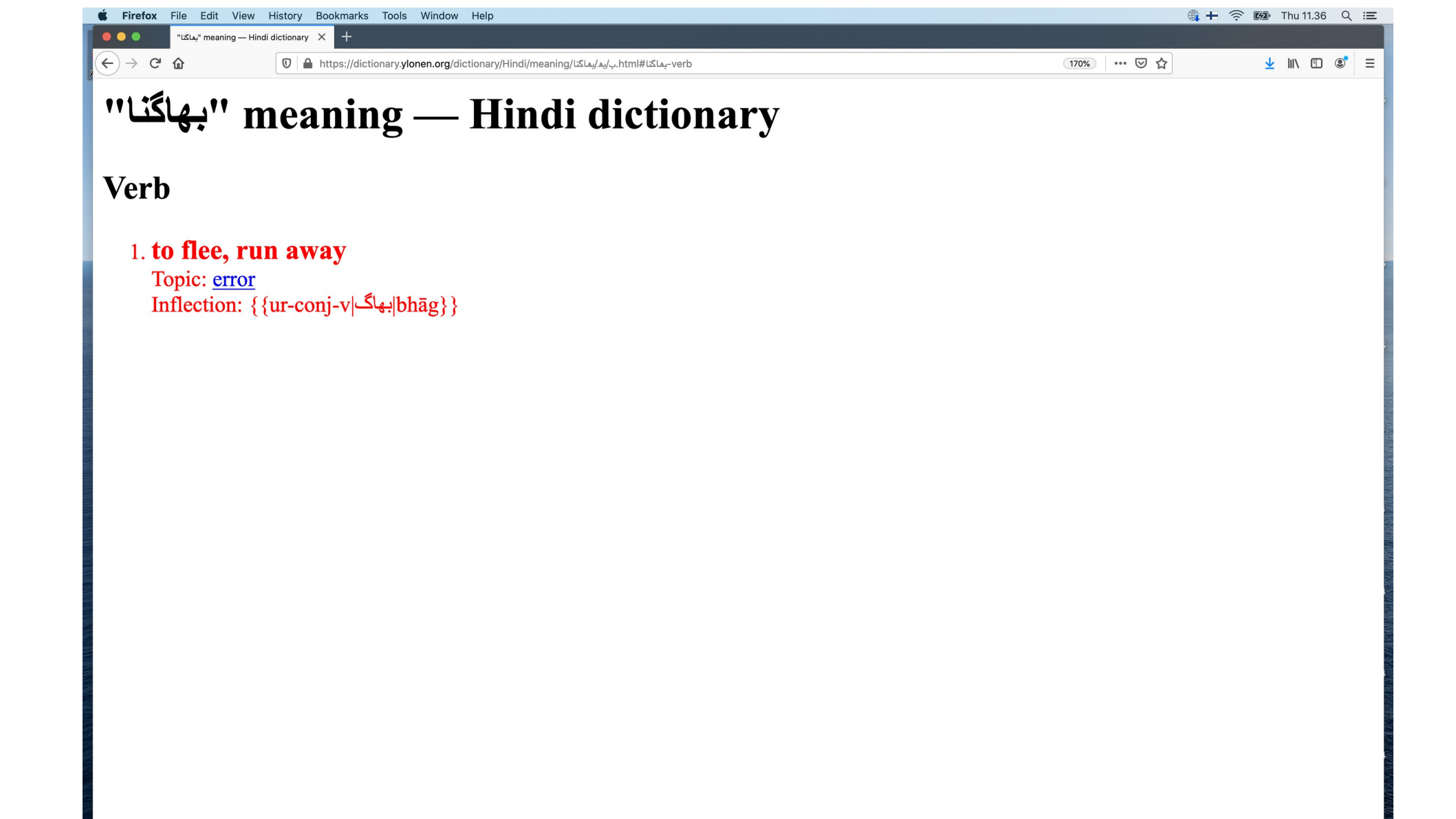
5. **A territorial division; a district; especially, in France, one of the districts into which the country is divided for governmental purposes, similar to a county in the UK and in the USA. France is composed of**



"maadoitus" meaning — Finnish dictionary

Noun

1. **earthing, grounding**
   Topic: electrical
   Topic: engineering
   Inflection: {{fi-decl-vastaus|maadoitu|a}}


Firefox File Edit View History Bookmarks Tools Window Help
Thu 11.36
"بھاگنا" meaning — Hindi dictionary
https://dictionary.ylonen.org/dictionary/Hindi/meaning/ب/بھ/بھاگنا.html#بھاگنا-verb
170%
"بھاگنا" meaning — Hindi dictionary
Verb
1. to flee, run away
Topic: error
Inflection: {{ur-conj-v|بھاگ|bhāg}}

"شائعة" meaning — Arabic dictionary

Noun

1. **rumor**
   Topic: feminine
   Inflection: {{ar-decl-noun|شَائِعَة|pl2=شَوَائع|pl=شَائِعَات}}

"仲裁" meaning — Chinese dictionary

Verb

1. **to arbitrate**
   Derived forms: 仲裁委員會, 仲裁人

All languages combined word senses marked with tag "childish"

Total 60 word senses

- 'splodey (Adjective) [English] Involving or reminiscent of an explosion or explosions.
- 'splodey (Adverb) [English] In a manner involving or reminiscent of an explosion.
- +QD+ (Adjective) [Portuguese] bestest, very best; superlative degree of D+
- begincement (Noun) [English] The beginning.
- bunny rabbit (Noun) [English] A bunny; a rabbit.
- bunny wunny (Noun) [English] A bunny; a rabbit.
- caca (Noun) [Portuguese] crap; excrement
- cachorrinho (Noun) [Portuguese] dog
- doggie (Noun) [English] Alternative spelling of doggy
- doggy (Noun) [English] A dog, especially a small one.
- drengelus (Noun) [Danish] boy cooties
- fishy wishy (Noun) [English] A fish.
- froggy (Noun) [English] A frog.
- gatinho (Noun) [Portuguese] cat
- ghostie (Noun) [English] ghost
- ghosty (Noun) [English] ghost (especially a non-frightening one)
- goosie (Noun) [English] goose
- hauveli (Noun) [Finnish] doggy, doggie
- horsy (Noun) [English] A child's term or name for a horse.
- jänöpupu (Noun) [Finnish] bunny rabbit
- kuža (Noun) [Slovene] doggy
- lambie (Noun) [English] A lamb.
- lickle (Adjective) [English] little
- lightning (Verb) [English] To produce lightning

Conclusion

- Comprehensive extraction from English Wiktionary
  - https://github.com/tatuylonen/wiktextract
- General framework MediaWiki dump file/Wikitext processing
  - https://github.com/tatuylonen/wiktitextprocessor
- I expect the pre-extracted dictionary resource to be ready for use in December
  - Web site (in progress): https://dictionary.kaikki.org
- For more info, email tatu.ylonen at helsinki.fi