



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Open-Source Morphologies and Crowd-Sourcing Lexicography

at FSCONS 2013

Tommi A Pirinen

<tommi.pirinen@helsinki.fi> /

<flammie@gentoo.org>

November 13, 2013





Outline

Part 1: Crowd-Sourcing and Lexical Data Concepts and Experiences

- Introduction: Concepts

- Crowd-sourcing: uses and issues

Part 2: Productising Research Results

- Introduction

- Some examples

- Requirements for a Software Product



Myself and relevant projects

- **Academically: Tommi A Pirinen** `http://www.helsinki.fi/%7etapirine/`, see also **Open science / reproducible research** at `http://github.com/flammie/purplemonkeydishwasher`
- **in FLOSS e.g., Flammie**
`http://dev.gentoo.org/%7eflammie/`
- **Open source morphology for Finnish**
`http://code.google.com/p/omorfi/`, #omorfi on Freenode
- **hfst-ospell** `http://hfst.sf.net/`, #hfst
- **apertium machine translation, simple4all text-to-speech, localisation etc. ...**



Outline

Part 1: Crowd-Sourcing and Lexical Data Concepts and Experiences

Introduction: Concepts

Crowd-sourcing: uses and issues

Part 2: Productising Research Results

Introduction

Some examples

Requirements for a Software Product



Morphology

- originally in linguistics: inflect words
- in a broad sense: classifying words, inflectional suffixes, etc.
- E.g., *hundarnas* = hund + ar + na + s = dog, common gender, needs ar as plural suffix, possessive form (the dogs')
- derivation and compounding can create infinitely many infinitely long words which must be predicted e.g. paternal grandfathers in Finnish (isä, isänisä, isänisänisä, isän... isä)
- *Finite-State Morphology* I work with, is capable of much more complex language systems
- to reach a system dealing with this we need data about words, leading to...



Lexicography

- “Dictionary writing”, in this context more like data harvesting
- Collect all words
- How do they inflect (i.e., which are the valid forms of the word)
- How do they operate with other words in sentence (syntax)
- What do words mean, how do you translate them (semantics)
- Everything else

Example of trad. dictionary

[Oxford English Dictionary, 3rd ed., s.v. set]

Set /set/ *verb*

set¹ ► *verb* (**sets**, **setting**; past and past participle **set**)
1 [with obj. and usu. with adverbial] put, lay, or stand (something) in a specified place or position: Delaney set the mug of tea down | Catherine set a chair by the bed.
■ (**be set**) be situated or fixed in a specified place or position: the village was set among olive groves on a hill. ■ represent (a story, play, film, or scene) as happening at a specified time or in a specified place: a private-eye novel set in Berlin. ■ mount a precious stone in (something, typically a piece of jewellery): a bracelet set with emeralds. ■ mount (a precious stone) in something. ■ Printing arrange (type) as required. ■ Printing arrange the type for (a piece of text): article headings will be set in Times fourteen point. ■ prepare (a table) for a meal by placing cutlery, crockery, etc. on it in their proper places. ■ (**set something to**) provide (music) so that a writer can be produced in a musical form: a form of music. ■ Bell-ringing mo



One example of Digital Dictionary

[our Finnish omorfi database s.v. *asettaa* (set)]

```
asettaa ['V_VIEROITTAA'] VERB 53 C
False False None False False None
aset asett0aa^backC None weaken back
False False False None False False
False False asettaa
```




Crowd-sourcing

- Getting lots of people to work on same project
- Wikipedia is the best success story here
- Ideal for lexicography: no special skills needed, all native speakers know words of their language
- There are projects for dictionary building as well: Wiktionary, Omegawiki, . . . (not as huge success stories, yet)



Outline

Part 1: Crowd-Sourcing and Lexical Data Concepts and Experiences

Introduction: Concepts

Crowd-sourcing: uses and issues

Part 2: Productising Research Results

Introduction

Some examples

Requirements for a Software Product



Uses of Crowd-Sourcing in Morphology Development

- New words come and go all the time: *crowd-sourcing*, *facebooking*, . . . , and we need all of them ASAP
- Collecting new features and information bits for words never ends
- semantics: is it (can it be) human, sentient, edible, female, location, corporation, mass nouns
- popularity: common word, rare, obscure
- style and usage: dialects, curse words, academic, computer, medicine
- . . . approx. each new application for language model needs new data



Issues in Crowd-Sourcing Lexicographies

1. Using data (long) after it has been built by harvesting, scraping, etc. requires lots of work
2. Inputting well-structured data in system not designed for it is cumbersome and error prone
3. That is, Wiktionary is really just an attempt of using something designed for writing encyclopedic prose in structured dictionaries
4. Wiktionaries are never stable, trying to use data from outside the system requires tracking changes in conventions
5. Newer systems attempted to bridge the gaps have not been successful either (Omegawiki, ...)

Example of Wiktionary Page

The image shows a browser window displaying the Finnish Wiktionary page for the word "asettaa". The browser's address bar shows the URL "fi.wiktionary.org/wiki/asettaa". The page title is "asettaa – Wikisanakirja". The browser's navigation bar includes icons for back, forward, refresh, and home, along with a search bar and a star icon. Below the navigation bar, there are several tabs: "News", "Uusi välilehti", "Mikro Værksted...", "Popular", and "Dept. of Linguis".

The main content area is divided into two columns. The left column contains a sidebar with the following items:

- Ohje
- Lahjoitukset
- Työkalut
 - Tänne viittaavat sivut
 - Linkitettyjen sivujen muutokset
 - Toimintosivut
 - Tulostettava versio
 - Ikilinkki
 - Sivun tiedot
 - Viitetiedot

The right column contains the main content, which is the entry for the word "asettaa". The entry is titled "Suomi" and "Verbi". The word "asettaa" is followed by the information "(53-C) (taivutus)". The entry lists three conjugations:

- laittaa, panna, sijoittaa paikalleen
Hän asetti maljakon pöydälle.
- määrätä, määrittää
Hän asettaa liian suuria vaatimuksia itselleen.
- seisauttaa, taituttaa, rauhoittaa



And its Source...

```
===Verbi===
```

```
{{fi-verbi|as|ettaa|muistaa|C}}
```

```
# [[laittaa]], [[panna]], [[sijoittaa]] pa
```

```
#:''Hän asetti maljakon pöydälle.''
```

```
# [[määrätä]], [[määrittää]]
```

```
...
```

```
====Käännökset====
```

```
{{kohta|1|laittaa, panna, sijoittaa paikal
```

```
*englanti: [[put]], [[place]], [[set]], mo
```

```
*hollanti: [[aanbrengen]]
```



Scraping the Data From Wiktionary

1. find section for Finnish words
2. find each definition
3. find and translate something like

```
fi-verbi|as|ettaa|muistaa|C into  
asettaa V_MUISTAA VERB 53 C...
```

- e.g., when I last wrote the script for scraping this data, `fi-verbi|as|ettaa|muistaa|C` was `fi-verb|53|C`



Example 2: Omegawiki

- database approach for storing data in well structured form
- getting data would be easier and more consistent
- still quite cumbersome to edit
- lacks some central pieces of information for Finnish and most other langs than English, e.g., inflection classification

Example of Omegawiki Page

The screenshot shows a web browser window with the URL www.omegawiki.org/Expression:set. The page title is "Kieli: englanti" (Language: English) and the main heading is "Substantiivi" (Noun). The page lists five definitions of "set" in English, each preceded by a right-pointing triangle (▶). The definitions are: "A matching collection of things of the same kind.", "A collection of various objects for a particular purpose.", "An object made up several parts.", "(set theory) A well-defined collection of mathematical objects (called elements or members) that can be distinguished from other objects.", and "An association or group of people, usually meeting socially." Below the definitions is the section "Verbi" (Verb), which contains one definition: "To set or place an object in a different place than it original was." Underneath the verb definition is a section for "Lexical annotations" with a sub-section "Ominaisuus Arvo" (Property Value) showing "sanaluokka" (word class) as "verbi" (verb). At the bottom, there is a partially visible section for "Määritelmä" (Definition).

www.omegawiki.org/Expression:set

News Uusi välilehti Mikro Værksted... Popular Dept. of Linguist... Academic Phra... HolidayPirates -...

Etusivu
Visual Dictionary
Satunnainen sivu
Tuoreet muutokset
OmegaWiki blog

Contributing

Ohje
Kahvihuone
Development
Donate to OmegaWiki

Työkalut

Tänne viittaavat sivut
Linkitettyjen sivujen muutokset
Toimintosivut
Sivun tiedot

Kieli: englanti

Substantiivi

- ▶ **set**: A matching collection of things of the same kind.
- ▶ **set**: A collection of various objects for a particular purpose.
- ▶ **set**: An object made up several parts.
- ▶ **set**: (set theory) A well-defined collection of mathematical objects (called elements or members) that can be distinguished from other objects.
- ▶ **set**: An association or group of people, usually meeting socially.

Verbi

- ▼ **set**: To set or place an object in a different place than it original was.

▼ Lexical annotations

Ominaisuus Arvo
sanaluokka verbi

▼ Määritelmä



Quality Issues in Crowd-Sourcing

- people know lots of their native languages but not always enough
- some contributors are language learners
- vandalism
- Two ways currently used to cope with this: python scripts, regexes etc. to check some sanity
- Automatic tests with the final software and free texts: do new additions work somewhat like old words, etc.
- In the end it all falls down to expert reviews again



Conclusions (questions): How to Proceed?

- How to combine popularity of Wiktionary with forms and structure of Omegawiki?
- Improve user interfaces?
- Better access to wiki data?
- Feedback from databases to Wiktionary?
- Answers? Questions?



HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Open-Source Morphologies and Crowd-Sourcing Lexicography

at FSCONS 2013

Tommi A Pirinen

tommi.pirinen@helsinki.fi /

flammie@gentoo.org

November 13, 2013





Myself and relevant projects (again)

- Academically: Tommi A Pirinen <http://www.helsinki.fi/%7etapirine/>, see also Open science / reproducible research at <http://github.com/flammie/purplemonkeydishwasher>
- in FLOSS e.g., Flammie <http://dev.gentoo.org/%7eflammie/>
- Open source morphology for Finnish <http://code.google.com/p/omorfi/>, #omorfi on Freenode
- hfst-ospell <http://hfst.sf.net/>, #hfst
- apertium machine translation, simple4all text-to-speech, localisation etc. ...



Outline

Part 1: Crowd-Sourcing and Lexical Data Concepts and Experiences

Introduction: Concepts

Crowd-sourcing: uses and issues

Part 2: Productising Research Results

Introduction

Some examples

Requirements for a Software Product



Case of Creating Spell-Checkers for Less-Resourced Languages

- Research work: Moving from open source morphologies to efficient finite-state spell-checkers
- Including conversion from existing formats to something equivalent of finite-state automata (e.g., from hunspell and its predecessors)
- At the moment: Software exists, is usable in enchant, libreoffice, etc., but not available in distros



Research Programming in Slashdot

Larry Page and Sergey Brin are Lousy Coders:

`<http://slashdot.org/story/13/11/01/
1324209/`

`larry-page-and-sergey-brin-are-lousy-coder`

“Google engineering boss Craig Silverstein recalls in the book. ‘I had to deal with their legacy code from the Stanford days and it had a lot of problems. They’re research coders: more interested in writing code that works than code that’s maintainable.’ ”



Current Research Methodology

1. Research problem (issues in current spell-checking)
2. Idea for solution (scribbled notes and formulas)
3. Proof-of-concept implementation (hacky code)
4. Experimentation (one-off measurements)
5. Publication
6. ... Research projects, funding etc. end here, all results get abandoned

Results, data, code, is all published in open science terms.



Suggested Continuation

1. ... Publication
2. Software Development (from hacky code to real library)
3. Integration to Real World Software
4. Distribution
5. Maintenance
6. Profit



Outline

Part 1: Crowd-Sourcing and Lexical Data Concepts and Experiences

Introduction: Concepts

Crowd-sourcing: uses and issues

Part 2: Productising Research Results

Introduction

Some examples

Requirements for a Software Product



Example from Early Part of my Project

hunspell2fst, would be rather important in business of replacing hunspell:

1. Transforming existing hunspell dictionaries into more efficient finite-state spell-checkers
2. Few obscure formulas:
3. Then some flex and yacc code and scripts to transform hunspell data files in around 10 commands
4. Measured some improvement over hunspell on most of the languages
5. Published in 2010 in an IEEE journal
6. The collection of scripts used is probably unusable now



Compare to: End of my Thesis Project

- Full working spell-checking, faster than hunspell, more efficient in most cases (but likely less stable)
- Integration to common open source software: LibreOffice, Mozilla, enchant (GTK+) (via software library voikko)
- Standard installation but turned off by default, requires manual work and not in current distributions, but packages exist



Outline

Part 1: Crowd-Sourcing and Lexical Data Concepts and Experiences

Introduction: Concepts

Crowd-sourcing: uses and issues

Part 2: Productising Research Results

Introduction

Some examples

Requirements for a Software Product



Product Requirements Unmet by Typical Scientific Software

- Stability: no error checking, no crash guarding, ... since software is only used in protected env. by experts without malicious intents
- Licencing: Academic licence restrictions are strictly against GNU definition of Freedom; smaller discrepancies, e.g., Debian legal does not allow even GPLv2 and Apache2 on same software
- Standards: GNU standards, not only for licence but installation procedures, packaging
- User interfaces: GNU, Gnome, KDE, ... integration
- Documentation: Academic paper is not code documentation or so forth



Other Issues

- Software maintenance in Linux distributions requires committed people to work on it (e.g., I only have access to gentoo's web since lack of activity etc.)
- Getting access to Linux distribution systems requires social engineering
- for some distros and products external repositories, overlays, ppa's, help, but they are not feasible for the most important target group of spell-checkers



Social Issues with Linguistics vs. FLOSS Hackers

- Niche products (limited use scientific software, small languages' support) may be frowned upon by software engineers. E.g.:
- “Well, that’s a valid enhancement request, of course, but something must to be done to prevent filling the text language dropdown with **such rubbish languages**, making it hard to use.” —a maintainer comment to bug report asking for Kumyk support in LibreOffice
- Similar attitude is common for any non-English related language support request



Windows Support? And other systems; Android, Mac OS X?

- Windows support usually requires commercial contracts, non-free implementations, NDAs
- For spell-checking, Windows 8 (as far as I've heard), Android 4, Mac OS X are gradually opening the access to spell-checking components that can be used to replace or extend system libraries
- In general, software product maintenance could be passed over from scientist to hobbyists and commercial workers,



Conclusion: Questions? Answers?

- Open science and FLOSS is not enough for all (any?) academic projects to become products (in FLOSS envs even)
- Scientists are scarce resource for software development, maintenance, distribution...
-



Even More Links and References

- <http://github.com/flammie/purplemonkeydishwasher/2013fscons/>
- <http://wordpress.let.vupr.nl/reproducingnlpresearch/>
- <https://sourceforge.net/p/hfst/code/HEAD/tree/trunk/conversion-scripts>
- <https://sourceforge.net/p/hfst/code/HEAD/tree/trunk/hfst-ospell>