Ouverture

Come creare un libero dizionario per il controllo ortografico in lingua Sarda

www.speling.org

Jacob Sparre Andersen <sparre@crs4.it>

Please interrupt me *when* I'm speaking too fast or otherwise am difficult to understand.
Questions are welcome at *any time* during the talk.

Overview

- **The principles behind the system.**
- How to work as a proof-reader on a "speling.org" based dictionary.
- How to manage the technical and editorial work on a "speling.org" based dictionary.
- The to-do list of the "speling.org" developers.

The principles behind the system

`speling.org`

- ... è una sistema per la creazione di dizionari elettronici.
- ... è bassata su una idea di cooperazione stilo Open Source.
- ... accepts that individual proof-readers can make mistakes[a].

[a] The most active Danish proof-readers have an estimated error-rate close to one out of thousand.

The principles behind the system

- Since we assume that the individual creators of the dictionary entries are imperfect, we don’t work with dictionary entries which are edited directly if somebody locate a mistake. – Wikipedia style.
- Instead we work with a system of voting, where the object the creators work directly on are votes for or against having some information assigned to an entry in the dictionary.

The principles behind the system

Benefits:

- All proof-reading results are stored – both the positive and the negative ones.
- There is less back-and-forth editing of the entries.
- The quality of the dictionary will gradually improve and mistakes by individual contributors has less effect.

The principles behind the system

Drawbacks:

- Since proof-readers have to act, even when the entries are correct, it takes more work to create a dictionary.

The principles behind the system

Creating a spell-checking word-list from the proof-reading results (votes):

- The system counts up how many positive and negative votes each string has received.
- A threshold number of votes is selected, such that strings with that many more positive than negative votes are considered correctly spelled words.

The principles behind the system

Gradually improving quality

For a spell-checking word list to be useful, it has to fullfill two requirements:

- It should contain all (as many as possible) of the words the writer uses.
- It should contain no (as few as possible) misspelled words.

The principles behind the system

Gradually improving quality

As each string in the dictionary is proof-read more and more times, we can increase the threshold number of votes without shrinking the resulting word-list, but increasing our certainty that the strings **in** the list are correctly spelled words.

Overview

- The principles behind the system.
- **How to work as a proof-reader on a "speling.org" based dictionary.**
- How to manage the technical and editorial work on a "speling.org" based dictionary.
- The to-do list of the "speling.org" developers.

How to work as a proof-reader ...

Proof-reading is primarily done by e-mail:

- You subscribe to a daily collection of words to be proof-read,
- you proof-read the words, and
- you send them back to the `speling.org` server,
- which adds up the votes and updates the word-list.

How to work as a proof-reader ...

Example proof-reading e-mail:

```
Reply-To: proof-reading@sc.speling.org
Subject: [SC] Words for proof-reading

# Proof-reading key: 892253cde850f90b5ba45

WORD: filuferru
STATUS: ?
EDITOR: Jacob Sparre Andersen

WORD: snaps
STATUS: ?
EDITOR: Jacob Sparre Andersen
```

How to work as a proof-reader ...

We reply and change the ? to + for correct words and to - for incorrectly spelled words:

```
To: proof-reading@sc.speling.org
Subject: Re: [SC] Words for proof-reading

> # Proof-reading key: 892253cde850f90b5ba
>
> WORD: filuferru
> STATUS: +
> EDITOR: Jacob Sparre Andersen
>
> WORD: snaps
> STATUS: -
> EDITOR: Jacob Sparre Andersen
```

How to work as a proof-reader ...

Our reply will then count as one extra vote for the string `filuferru` and one extra vote against `snaps`.

- If this means that `filuferru` now has enough positive votes, the word will be included in the next version of the word-list.
- If this means that `snaps` (simile a filuferru, ma danese) doesn't have enough positive votes, the word will be removed in the next version of the word-list. – If it should ever have appeared in the first place.

How to work as a proof-reader ...

Unless the word-list is produced with a threshold of one more positive than negative vote, a single proof-reader can not on his/her own add a word to the dictionary.

This is of course a bit annoying, but it is an effect of taking into account that the proof-readers can make mistakes.

How to work as a proof-reader ...

The proof-reading messages can contain many more fields than the three I showed in the example[a], but since the handling of information beyond raw word-lists for spell-checking is likely to change soon, I will not cover the full extent of the format here.

[a] ANTONYM, AUTHORITY, CATEGORY, CLASS, COMMENT, COMPOSITE-WORD, CONJUGATION, CONJUGATION-RULE, CORRECTION, DATE, DESCRIPTION, EXAMPLE, HYPHENATION, ROOT, SOURCE, SOURCE-YEAR, SYNONYM, TRANSLATION-DE-WORD, TRANSLATION-EN-WORD, TRANSLATION-FO-WORD, TRANSLATION-FR-WORD, TRANSLATION-IT-WORD, TRANSLATION-NO-WORD and TRANSLATION-SV-WORD.

How to work as a proof-reader ...

Alternate proof-reading methods:

- There is a prototype Gtk+ based proof-reading tool (which I presented in my last GULCh talk).
- Automated harvesting of user additions to the Aspell and Ispell dictionaries (`http://www.speling.org/#dictionary_feedback`).
- Palm Pilot interface.

Overview

- The principles behind the system.
- How to work as a proof-reader on a "speling.org" based dictionary.
- **How to manage the technical and editorial work on a "speling.org" based dictionary.**
- The to-do list of the "speling.org" developers.

How to manage the technical ...

The first step in managing the technical and editorial work on a `speling.org` based dictionary is of course to install the software.

- Download the source code from `http://www.speling.org/`.
- Unpack the source code and move to the created directory.
- Configure:
  `PREFIX=/opt/speling.org ./configure`

How to manage the technical ...

(install continued)

- Make: `make`
- Install: `sudo make install`
- Fix `PATH`:
  `export PATH=/opt/speling.org/bin:$PATH`

How to manage the technical ...

Once the system is installed, a dictionary should be configured, `make_new_dictionary sc`[a], and populated:

```
cat corpus/* | tr ’ ’ ’\n’ | sort -u \
 | words_to_ds \
 > /var/speling.org/sc/incoming.ds/corpus
```

[a] "sc" is the ISO 639-1 two-letter language code for Sardinian.

How to manage the technical ...

Dictionary data are by default stored under `/var/speling.org/<language code>/`.

New editor proof-reading reports (in .ds format) should be put in `/var/speling.org/<language code>/incoming.ds/`.

The program `update_dictionaries` reads the new proof-reading reports and generates updated word-lists.

How to manage the technical ...

The program `send_words_to_proof-reading` is used to send proof-reading e-mails out to the subscribing proof-readers.

I use Procmail to intercept, filter and archive the proof-reading reports as they arrive. The file `dot.procmailrc` is an example of how this can be done.

How to manage the technical ...

Getting words from the World Wide Web
Crúbadán.

Once you have set up a system for receiving proof-reading messages by e-mail, you might want to get in touch with Kevin Patrick Scannell who runs Crúbadán (http://borel.slu.edu/crubadan/). It is possible that he has data for your language in Crúbadán, so he can set the system up to send you messages with possible words for your dictionary, when it finds them on the net.

How to manage the ... editorial ...

Using authoritative sources

As an editor of a `speling.org` dictionary, you have the option of using the `AUTHORITY` field in the proof-reading format to cite authoritative sources (commonly recognized dictionaries, experts, etc.) of the information you report to the system:

```
WORD: husholdning
STATUS: +
AUTHORITY: Retskrivningsordbogen, 3. udgav
EDITOR: Jacob Sparre Andersen
```

Overview

- The principles behind the system.
- How to work as a proof-reader on a "speling.org" based dictionary.
- How to manage the technical and editorial work on a "speling.org" based dictionary.
- **The to-do list of the "speling.org" developers.**

To-do list for "speling.org"

The current version of `speling.org` is fine for creating a plain word-list for spell-checking, but it is insufficient when it comes to creating a proper dictionary with grammatical information, synonyms, explanations of words, etc.

The problem is that the current format only is expressive enough to **add** this extra information, not to correct it, if it is wrong.

To-do list for "speling.org"

- Make a Debian package with `speling.org`.
- Define a more expressive format for adding and correcting extra information.
- Write a tool for converting from the old to the new format.
- Reimplement the system with the new source format.
- Write web and and graphical client-side tools for the proof-readers.

Credits

- The `speling.org` system was developed in cooperation with Henrik Christian Grove and Peter Makholm.
- The `speling.org` logo was designed by Hans Schou.

- Dansk Sprognævn (The Danish Language Council) provoked me to start the project.

Links

- Source code for the `speling.org` system:
  `http://www.speling.org/`
- A running `speling.org` system:
  `http://da.speling.org/`
- Crúbadán - a source for minority language corpora: `http://borel.slu.edu/crubadan/`
- Wikipedia - a different, open way of creating dictionaries:
  `http://it.wikipedia.org/wiki/Wikipedia`