When to use PYBOSSA? Case studies on crowdsourcing for Dutch

Presentation · December 2018
DOI: 10.13140/RG.2.2.329588.847

2 authors:

Peter Dekker
Instituut voor de Nederlandse Taal

Tanneke Schoonheim
Instituut voor de Nederlandse Taal

Some of the authors of this publication are also working on these related projects:

enetCollect: European NETwork for COmbining Language LEarning with Crowdsourcing Techniques (COST Action CA16105) View project

BSc thesis: Determining Dutch dialect phylogeny using Bayesian inference View project

All content following this page was uploaded by Peter Dekker on 11 December 2018.

The user has requested enhancement of the downloaded file.
When to use PYBOSSA?
Case studies on crowdsourcing for Dutch

Peter Dekker
Tanneke Schoonheim
Instituut voor de Nederlandse Taal
Introduction of INT

Instituut voor de Nederlandse Taal (Dutch Language Institute)

- Scholarly institute in the field of the Dutch language
- Central position in the Dutch-speaking world (the Netherlands, Flanders (Belgium), Surinam, Dutch Caribbean)
- Developer, keeper and distributor of corpora, lexica, dictionaries and grammars
- Making resources about the Dutch language accessible for researchers and the general public
Introduction of INT: Documenting language

Current projects at INT:

- Contemporary and historical dictionaries, corpora and lexica
- Grammar portal, spelling database, terminology lists
Introduction of INT: Language learning

Development and hosting of products for educational purposes, such as

- Bilingual dictionaries (New Greek, Portuguese, Estonian)
- Dutch Word Combinations
- Corpus of Elementary Dutch
Research question on experimenting with PYBOSSA

- How can we use crowdsourcing for documenting language and language learning?

- When and how can PYBOSSA be applied for these purposes?
Our experiments: Taalradar

- Blends (september 2018)
- Neologisms (november 2018-present)
- Language variation (november 2018-present)
Blends: Data

Blend
- Compound of two words, where parts of the words are lost
- Signifies a new meaning, related to the words it consists of

Examples:
- *glamping* (glamour + camping)
- *mup* (millenial + yup)

Can the crowd help in recognizing and analyzing blends?
Blends: Crowd

Where did we find the crowd?

- Newsletter Instituut voor de Nederlandse Taal (3873 subscribers, 519 clicks)
- Congress Internationale Vereniging voor Neerlandistiek
- Announcements on linguistics blog, Twitter, LinkedIn

Result: 549 participants, spread across two tasks
PYBOSSA installation

- Hosted version (crowdcrafting.org) vs hosting on own server
  - Own server: existing infrastructure at INT, more flexibility
- Clear installation guide on PYBOSSA website
- Complexity: PYBOSSA consists of multiple software packages
- Ansible script: recipe for reproducible installation
User details

Herkennen van blends 1: Contribute

Hieronder kunt u enkele persoonlijke gegevens invullen, die ons extra inzichten geven in ons onderzoek. Deze worden anoniem opgeslagen, gekoppeld aan uw antwoorden, zichtbaar voor bezoekers van dit platform. De velden die u niet wilt invullen, kunt u leeg laten.

Leeftijd:

Geslacht:

Zeg ik niet

Woonplaats:

Verzend!
Blends analysis

Analyseren van blends: Contribute

Uit welke twee volledige, bestaande woorden is het volgende woord opgebouwd?

adware

Tip: door op ENTER te drukken, beweegt u van het eerste naar het tweede woord, en van het tweede woord naar de volgende opdracht.
Results: Blends analysis

n = 326
### Results: Blends analysis for *preferendum*

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>referendum, prefereren</td>
<td>154</td>
</tr>
<tr>
<td>referendum, preferentie</td>
<td>60</td>
</tr>
<tr>
<td>referendum, pre</td>
<td>16</td>
</tr>
<tr>
<td>[Don’t know]</td>
<td>11</td>
</tr>
<tr>
<td>referendum, preferent</td>
<td>8</td>
</tr>
</tbody>
</table>

- Multiple word forms (noun, verb) for *prefer*
- Multiple interpretations (at least when word is presented without context):
  - referendum + to prefer
  - referendum + pre
**General experiences with PYBOSSA**

**Benefits**
- Freedom when developing tasks
- Share tasks with other researchers. Our repository: [https://github.com/INL/taalradar](https://github.com/INL/taalradar)
- Account system and loading/saving tasks handled by PYBOSSA
- Quick answers from developers via bug tracker

**Drawbacks**
- No ready-made translation for all languages
- No kiosk mode: multiple anonymous logins from same computer not allowed
- No built-in possibility to stop after number of tasks and show end screen
- User cannot go back to a previous task
PYBOSSA for linguistic crowdsourcing?

PYBOSSA designed for "pure" crowdsourcing

Problems when using PYBOSSA for linguistic crowdsourcing

- No built-in support for asking user details
  - Has to be presented as task
  - User details openly visible
- User identification by IP address: problem when using in class room
Alternatives: Survey tools

- Survey tools can be alternative
  - If you do not need the power and customizability of PYBOSSA
  - If mentioned drawbacks are problematic for you
- Google Forms: bad for privacy, data stored on external server
- Open source: installation and data on own server
  - TellForm
  - Lime Survey: limited free version
  - JD Esurvey
Our experiments: Taalradar

- Blends (september 2018)
- Neologisms (november 2018-present)
- Language variation (november 2018-present)

Nieuwe woorden: Bijdragen

(...) Syrische revolutie een ramp zou worden en een sektarische burgeroorlog zou genereren. Een afroderij voor elke revolutionair die er ook maar van droomt iets aan het status-quo te veranderen. (...)

Dit woord zal lang gebruikt blijven worden. ○ Ja ○ Nee
Dit woord zal door veel verschillende mensen (bijv. jong, oud) in veel verschillende situaties (bijv. gesprek, krant) gebruikt worden. ○ Ja ○ Nee
Crowdsourcing for documenting language: Automating lexicography

- Automate lexicographic workflow for neologisms

Neologism workflow at INT:
1. New words automatically collected from newspapers and websites
2. Collect judgments on durability of neologisms via crowdsourcing
3. Lexicographer selects neologisms, added to:
   a. neologism portal (all neologisms)
   b. ANW dictionary (rooted neologisms only)
Crowdsourcing for language learning

First experiments to get to know PYBOSSA, more focused on language documentation in general than on language learning in particular

Use crowdsourcing to make dictionary material accessible for language learning

- Add associated words to a given list of entries (horse: riding, saddle, pony, tail, chess, gymnastics, .... to make clusters of sense related associations, for instance for helping people with dysfasy/afasy
Conclusion

Is PYBOSSA useful for linguistic crowdsourcing?

- Is PYBOSSA useful for crowdsourcing for educational purposes?
- Is PYBOSSA useful for crowdsourcing for language learning?

Yes, powerful platform, with its own strengths and drawbacks. But you always have to be aware of what you ask from the crowd.

We are happy to share our tasks or collaborate!