

enetCollect WG3&5 meeting Leiden, 24-25 October 2018

Recognizing blends: First experiments with PYBOSSA

Peter Dekker
Tanneke Schoonheim
Instituut voor de Nederlandse Taal



cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

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
- Developer, keeper and distributor of corpora, lexica, dictionaries and grammars
- Provider of necessary building blocks of the study of Dutch



Introduction of INT

Current staff of INT:

- 17 (computational and corpus) linguists, lexicographers, terminologists, 4 linguistic assistants and trainees
- 5 software engineers, 1 system administrator
- 5 administration and communication





Introduction of INT

Current projects at INT:

- Contemporary and historical dictionaries and dictionary portals
- Contemporary and historical corpora and lexica
- Grammar portal, spelling database, terminology lists
- Infrastructure, tools and data for linguistic research (CLARIN)



Introduction of INT

Relatively new projects at INT:

Development and hosting of products for educational purposes, such as

- Bilingual dictionaries (New Greek, Portuguese, Estonian)
- Dutch Word Combinations
- Corpus Eenvoudig Nederlands (Corpus of Elementary Dutch)

Can crowdsourcing help us developing these and other educational products?



Research objective

Crowdsourcing:

- Task solved by public: answer unknown
- User details not important

Traditional (socio)linguistic research/survey:

- Answer of task known beforehand, in many cases
- User details important

Our research objective: combination

- PYBOSSA can be suitable, more solutions exist



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



Experiments with 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* (millennial + yup)

Blends in English: Gries, S. T. (2004). Shouldn't it be breakfunch? A quantitative analysis of blend structure in English. *Linguistics*, 639-668.



Experiments with blends: Data

Blends are part of two related projects at INT:

- Algemeen Nederlands Woordenboek (ANW; Dictionary of Contemporary Dutch)
- Neologism portal (upcoming)

Neologism Workflow at INT:

1. Data from newspapers and websites
2. Processed automatically, new words put aside
3. Lexicographer selects neologisms, creates entries in
 - a. neologism portal (all neologisms)
 - b. ANW dictionary (rooted neologisms only)



Experiments with blends: Crowd

Where did we find the crowd?

- Newsletter Instituut voor de Nederlandse Taal
- Congress Internationale Vereniging voor Neerlandistiek



Experiments with blends: Jobs

Can the crowd help in recognizing and analyzing blends?

Two jobs created in PYBOSSA:

- Blend recognition
 - Recognize blends in a text
- Blend analysis
 - Analyze the words a blend consists of

10 tasks per job



User interface design

Freedom in UI design: design using HTML and Javascript

PYBOSSA only loads and saves tasks from database



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:

Woonplaats:

 Verzend!

Blends analysis

/instituut voor de
Nederlandse taal/

Community

Projects

Create

About

admin ▾

Analyseren van blends: Contribute

Blend 1 van 10

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

adware

Woord 1:

☐ Weet ik niet

Woord 2:

☐ Weet ik niet

Volgende

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

Blends recognition

/instituut voor de
Nederlandse taal/

Community

Projects

Create

About

admin ▾

Herkennen van blends 1: Contribute

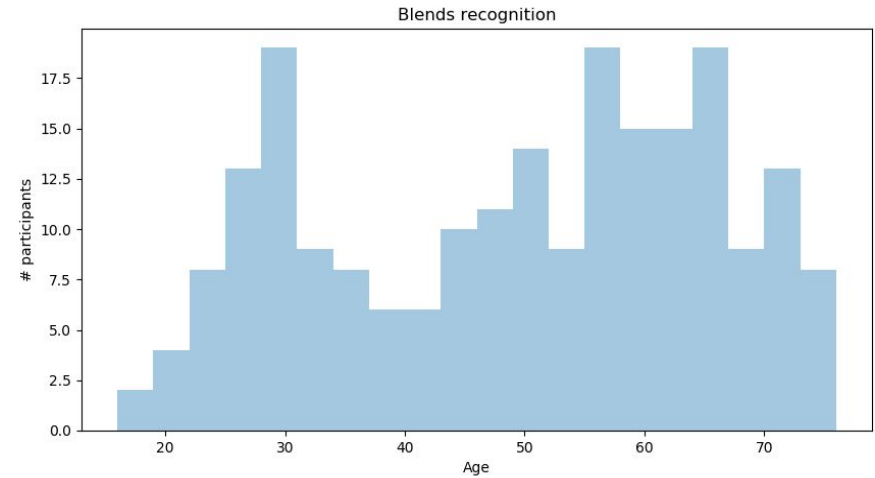
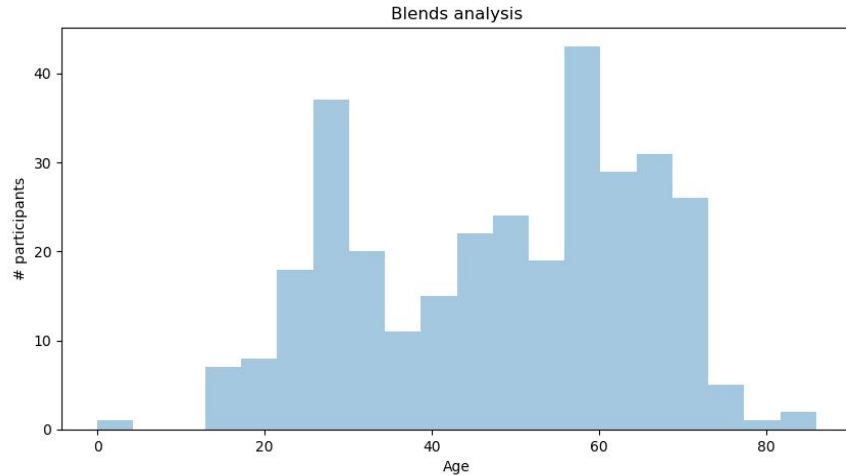
Tekst 1 van 10

Welk woord of welke woorden (minimaal 1) in deze tekst herkent u als blends?

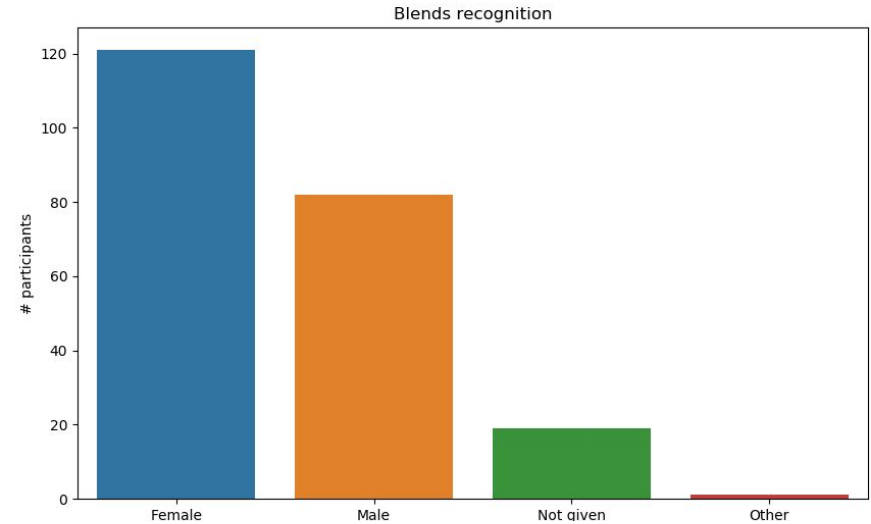
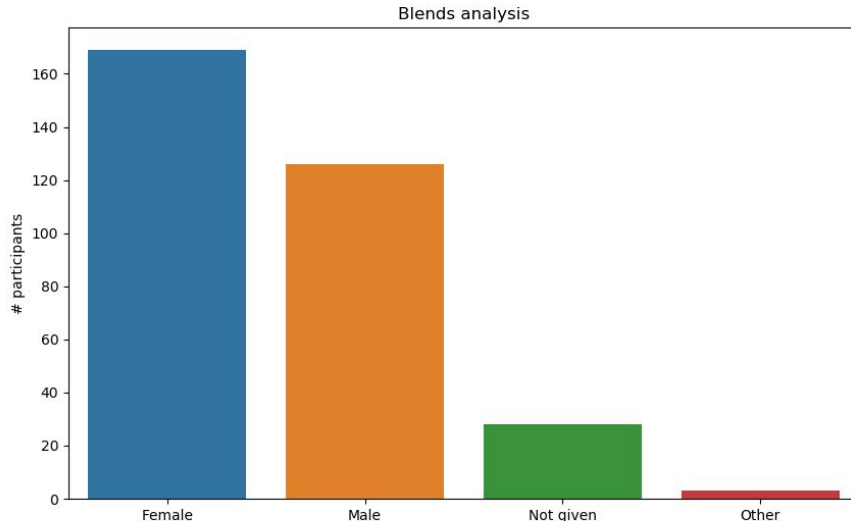
De manier waarop Farrah haar moeder toen behandelde, staat Minaj totaal niet aan. En wat doe je dan? Dan begin je een fittie op Twitter. Een Twittie. De Amerikaanse zangeres nam het voortouw en plaatste een tweet over het gedrag van Farrah, wat wellicht niet heel eerlijk is aangezien Farrah in 2009 nog niet volwassen was.

Volgende

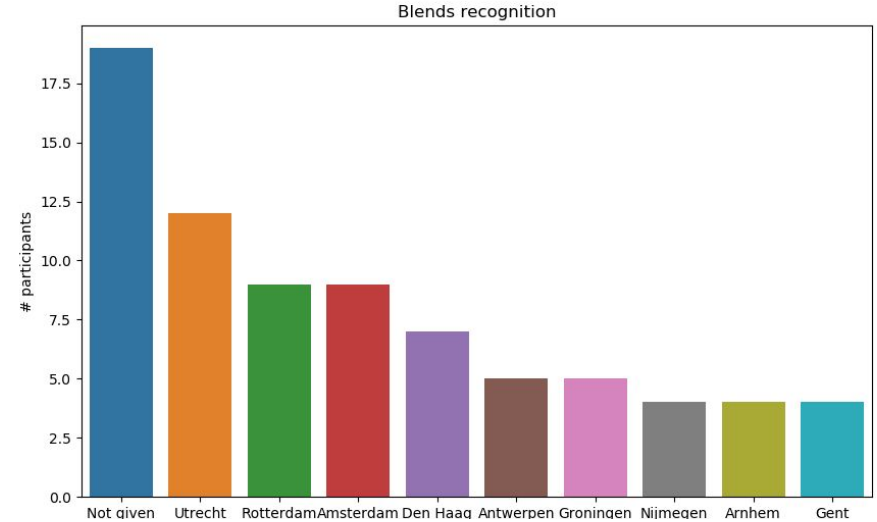
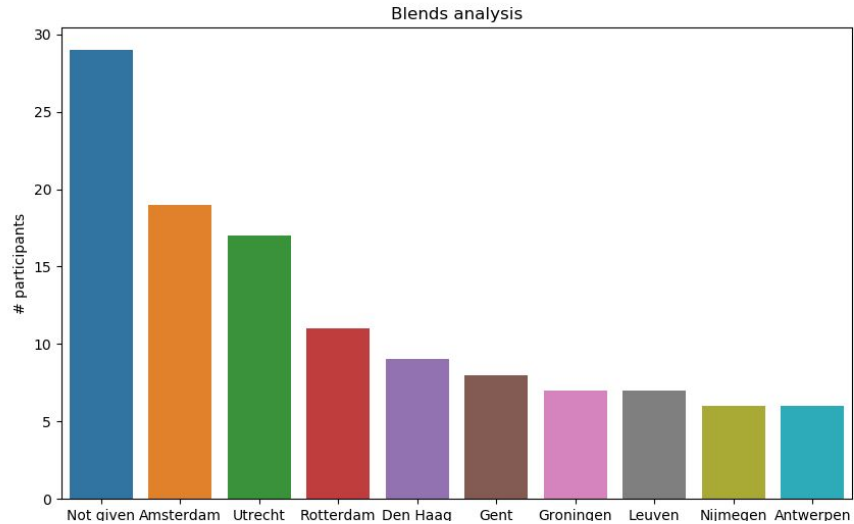
Results: Age



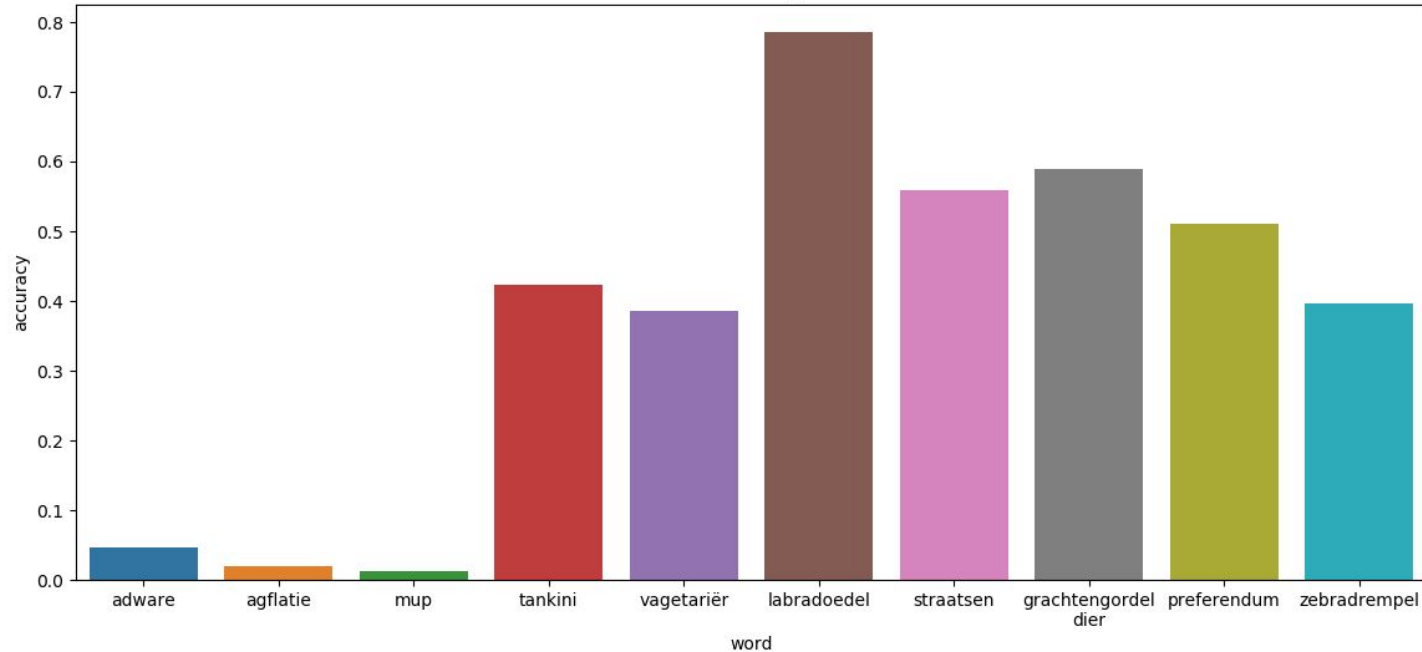
Results: Gender



Results: Location



Results: Blends analysis



n = 326

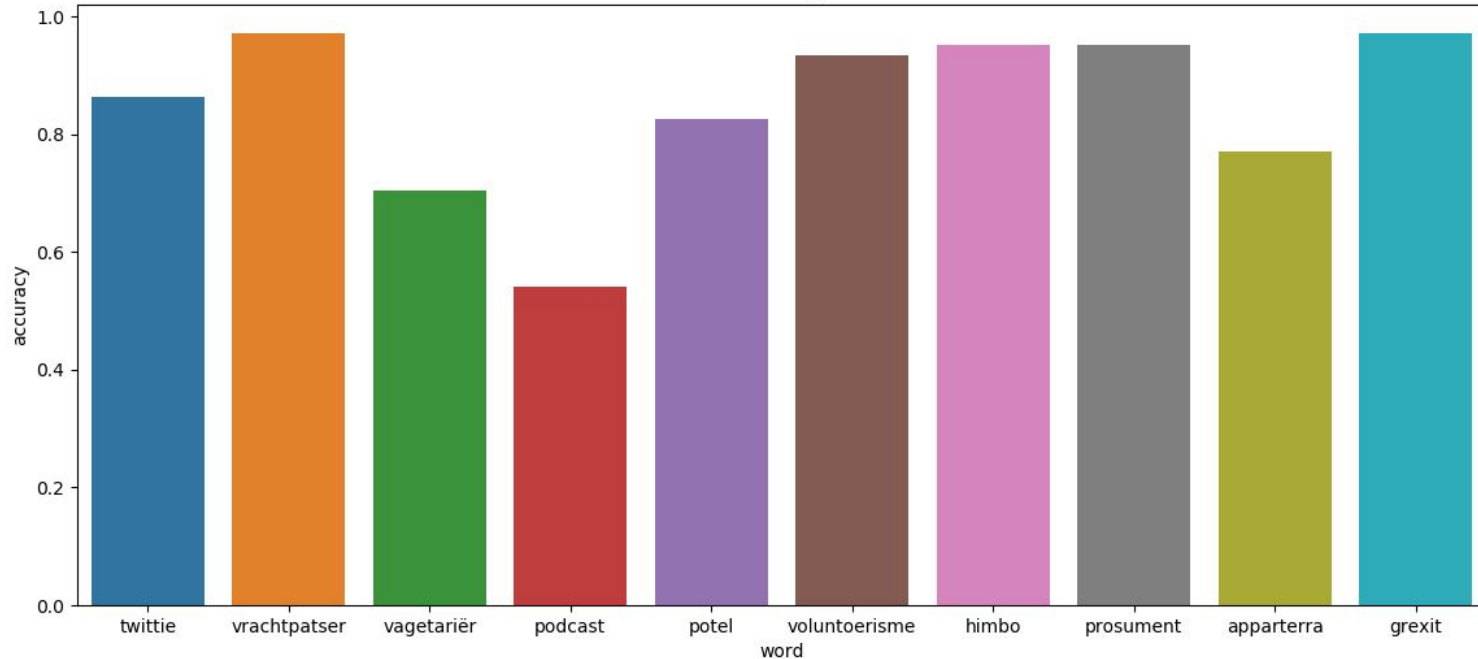


Results: Blends analysis for *preferendum*

<i>Analysis</i>	<i>Frequency</i>
referendum, prefereren	154
referendum, preferentie	60
referendum, pre	16
[Don't know]	11
referendum, preferent	8

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

Results: Blends recognition



n = 223



Results: Blends recognition for *twittie*

<i>Recognized blends</i>	<i>Frequency</i>
twittie	122
twittie, fittie	56
fittie	16
twittie, tweet, fittie	5
[Do not know]	4

- twittie: twitter + fittie 'fight' (slang)
- fittie itself also occurred in text:
misinterpreted as blend
- More input fields (3) than real blends per task (1):
stimulates giving more blends

User feedback

- English language of PYBOSSA, while tasks are about Dutch
- Too many buttons
- Task not always clear
- Make welcome page attractive

/instituut voor
de Nederlandse
taal/

Taalradar

Taalradar is een initiatief van het [Instituut voor de Nederlandse Taal](#) om kennis te verzamelen over het Nederlands door gebruik te maken van het taalgevoel van sprekers van de taal (crowdsourcing). Op dit moment draaien we een experiment om ervaring op te doen, waar u aan mee kunt doen!

Het experiment loopt tot en met 30 september 2018.

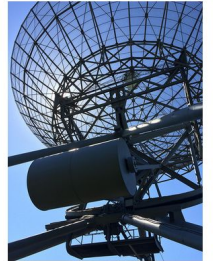
Meedoen

[Maak eerst een account aan](#), en kies vervolgens een taak hieronder. Als u een taak aanklikt, krijgt u verdere uitleg.

- [Analyseren van blends](#): in deze taak moet u bepalen uit welke twee woorden een samenstelling bestaat
- [Herkennen van blends in tekst](#): in deze taak moet u blends, een bepaald type samenstelling, herkennen in een tekst

Contact

Heeft u vragen, opmerkingen of suggesties over Taalradar? Mail dan naar servicedesk@ivdnt.org.





Experiences with PYBOSSA

Benefits

- Freedom when developing tasks
- Share tasks with other researchers
- Everything else (account system and loading/saving tasks) handled by PYBOSSA
- Quick answers from developers via bug tracker

Drawbacks

- No ready-made translation for all languages
- PYBOSSA not designed for asking user details
- When uploading large number of tasks, there is no clear end of job (you have to code that yourself)
- User cannot easily go back to a previous task
- User identification by IP address does not always work

Possible alternative for some purposes: Google Forms



Future experiments

- Neologisms and dialects
- User detail prediction as reward

Interesting issue:

- Is PyBossa better suited for these tasks than for instance Google forms?

Conclusion

Is crowdsourcing useful for the analysis of blends?

Yes, because it gives an insight in how blends are interpreted by non-linguists.

Is PYBOSSA useful for this kind of crowdsourcing?

Yes, powerful platform, with its own strengths and drawbacks.

