# Crowdsourcing Language Resources for Dutch using PYBOSSA: Case Studies on Blends, Neologisms and Language Variation

**Peter Dekker, Tanneke Schoonheim**
Instituut voor de Nederlandse Taal (Dutch Language Institute)
{peter.dekker,tanneke.schoonheim}@ivdnt.org

**Abstract**

In this paper, we evaluate PYBOSSA, an open-source crowdsourcing framework, by performing case studies on blends, neologisms and language variation. We describe the procedural aspects of crowdsourcing, such as working with a crowdsourcing platform and reaching the desired audience. Furthermore, we analyze the results, and show that crowdsourcing can shed new light on how language is used by speakers.

**Keywords:** crowdsourcing, lexicography, neologisms, language variation

## 1. Introduction

Crowdsourcing (or: citizen science) has shown to be a quick and cost-efficient way to perform tasks by a large number of lay people, which normally have to be performed by a small number of experts (Holley, 2010; Causer et al., 2018). In this paper, we use the PYBOSSA (PB) framework[1] for crowdsourcing language resources for the Dutch language. We will describe our experiences with this framework, to accomplish the goals of language documentation and generation of language learning material. In addition to sharing our experiences, we will report on linguistic findings based on the experiments we performed on blends, neologisms and language variation.

For the Dutch language, crowdsourcing has been valuable in the past. We distinguish two types of approaches. On one hand, there are *fixed* tasks, where more or less one answer is correct. As fixed tasks, crowdsourcing has been applied to the transcription of letters from 17th and 18th-century Dutch sailors (Van der Wal et al., 2012) and historical Dutch Bible translations (Beelen and Van der Sijs, 2014).

On the other hand, there are *open tasks*, referred to in empirical sciences as *elicitation tasks*, where different answers by different users are welcomed, in order to capture variation. Examples of open tasks are *Palabras* (Burgos et al., 2015; Sanders et al., 2016), where lay native Dutch speakers were asked to transcribe vowels produced by L2 learners, and *Emigrant Dutch*[2], which tries to capture the language use of emigrant Dutch speakers. Of course, mixture forms between open and fixed tasks are possible.

The Dutch Language Institute strives to document the language as it is used, by compiling language resources (eg. dictionaries) based on corpora from different sources, such as newspapers and websites. Fixed-task crowdsourcing, such as transcription and correction, can help in this process. However, we see even greater possibilities for open-task crowdsourcing, asking speakers how they use and perceive the language, which we will explore in this paper.

Open-task crowdsourcing has been applied to lexicography for other languages, such as Slovene, where crowdsourcing was integrated in the thesaurus and collocation dictionary applications (Holdt et al., 2018; Kosem et al., 2018). On top of this goal of language documentation, we would like to use crowdsourcing to make language material available for language learners.

## 2. Method

As the basis for our experiments, we hosted an instance of PYBOSSA at our institute. We named our crowdsourcing platform *Taalradar* ('language radar'): this signifies both the 'radar' (overview) we would like to gain over the entire language through crowdsourcing, and the personal 'language radar' or linguistic intuition of contributors, which we would like to exploit. We ran two crowdsourcing rounds: in september 2018 and in november-december 2018.

### 2.1. Tasks

We designed four tasks, which are well-suited to reach our goals: documentation of the Dutch language and developing material for language learning. Since we would like to get a picture of the speakers of the language, we ask for user details (gender, age and city of residence) in all tasks. The tasks were created as Javascript/HTML files inside PB.

**Tasks 1 and 2: Blends analysis and recognition** Blends are compound words, formed "by fusing parts of at least two other source words of which either one is shortened in the fusion and/or where there is some form of phonemic or graphemic overlap of the source words" (Gries, 2004). An example of a blend in both English and Dutch is *brunch*, which consists of *breakfast* and *lunch*. For our experiments, we used blends collected for the Algemeen Nederlands Woordenboek (Dictionary of Contemporary Dutch; ANW) (Tiberius and Niestadt, 2010; Schoonheim and Tempelaars, 2010; Tiberius and Schoonheim, 2016).

We developed two tasks: *analysis* and *recognition* of blends. In the analysis task, contributors are presented with a blend, and asked of which source words this blend consists. No context of the blend is provided. 10 blends are

[1]Homepage: http://pybossa.com. DOI: https://doi.org/10.5281/zenodo.1485460

[2]http://www.meertens.knaw.nl/vertrokken-nederlands/

presented in total. Figure 1 shows the task as it is presented to the contributor.


Figure 1: Screenshot of the blends analysis task.

In the recognition task, contributors are presented with a citation from the ANW dictionary. 10 citations are presented. Contributors should recognize the blend in the citation. Every citation contains one blend, but we ask for "one or multiple blends" and present users with tree input fields to enter blends. We deliberately designed the task in this somewhat deceptive way, to see which other words are candidates for being perceived as blends.

**Task 3: Neologisms** In this task, contributors were asked to judge neologisms (new words) in a citation, on two criteria: *endurance of the concept* ("This word will be used for long time.") and *diversity of users and situations* ("This word will be used by different people [eg. young, old] in different situations [eg. conversation, newspaper]."). We selected these two criteria from the FUDGE test, a test to rate the sustainability of a neologism, which normally consists of 5 criteria (Metcalf, 2004). The neologisms and their citations were taken from newspaper material, which is used in the lexicographic workflow (see section 4.). From this corpus, sentences which contain a hitherto unknown word are extracted: these are possible neologisms, but can also be words that have been formed ad hoc. Lexicographers accept or reject a word as neologism. We presented 15 words in a citation to users: 5 which have been attested by lexicographers as neologisms, 5 which have been rejected as neologisms, and 5 unattested words.

**Task 4: Language variation** In this task, contributors are asked how they call a certain concept or how they would express a certain sentence. The goal is to chart dialectal variation, but also other kinds of language variation. We used a list of questions from Taalverhalen.be, a website which tries to chart language variation using questionnaires[3]. The list contains 16 questions: 9 questions on words for sweets, and 6 questions about the general vocabulary. An example of a question is: "How do you call VINEGAR?". On top of the user details we ask in other tasks (gender, age, city of residence), we also ask for province, mother tongue and educational level.

### 2.2. Audience

Our experiments were advertised via our institutional newsletter, which reaches 3891 subscribers with an interest in language. We assume this was the channel with the largest reach: in the first round, the newsletter article received 519 clicks, and in the second round, 65 clicks. In both rounds, we observed an increase in contributions after the release of the newsletter. Additionally, we attended two linguistics events, where we offered visitors the possibility to engage in our crowdsourcing experiments: the meeting of the international society of Dutch linguistics and Drongo festival, an event for the language sector in The Netherlands. Finally, we advertised our experiments via social media (Twitter, LinkedIn) and a Dutch linguistics blog.

[3] `http://taalverhalen.be`, maintained by Miet Ooms.

## 3. Results

The results section consists of two parts. We will first describe our experiences with PYBOSSA as a crowdsourcing platform. Then, we will report on the linguistic findings on the language phenomena we performed experiments on.

### 3.1. Experiences of crowdsourcing with PYBOSSA

Table 1 shows the number of contributors for each of the tasks. It can be observed that only a small number of visitors did not finish the whole task. This could be due to the small number of questions we offered per task. The tasks in the second round (november-december 2018) received less contributors than in the first round (september 2018), this could be due to a less prominent place of the announcement in our newsletter in the second round. In all experiments, more women than men participated. Also, participants with ages above 50 were well represented. More participants came from The Netherlands than from Flanders.

| Task | # started | # completed | period |
|---|---|---|---|
| Blends analysis | 326 | 305 | sept 2018 |
| Blends recognition | 223 | 209 | sept 2018 |
| Neologisms | 118 | 111 | nov-dec 2018 |
| Language variation | 114 | 108 | nov-dec 2018 |

Table 1: Number of contributors per task.

We will now discuss our experiences with the PYBOSSA platform. A strength of PB is the freedom it offers when designing a task: the whole interface can be written in HTML and Javascript and can be customized. This also makes it easy to share tasks with other researchers[4]. The account system and saving/loading of tasks is handled by PB, so this does not have to be implemented by the task developer. Responses of the PB authors on the bug tracker are quick and concise. It is clear that PB is mainly designed for fixed-task crowdsourcing, not focusing on variation and the details of the contributor. For open-task, linguistic purposes, some points require attention (at time of writing). Firstly, there is no built-in support for asking contributor details. We handled this by asking contributor details via a normal question. However, since all given answers are visible publicly in PB, this also applies to the details, which may not be ideal from a privacy perspective. Secondly, contributors cannot go back to a previous task and change their answers. Thirdly, multiple anonymous logins from the same computer are not allowed, making it harder to use PB on e.g. a trade fair. A workaround is possible, but not built in PB by default. Also, anonymous users are identified by IP address: this can cause problems when multiple anonymous contributors connect via a shared internet connection, such

[4] Our tasks can be downloaded from: `https://github.com/INL/taalradar`.

as in classroom use. Finally, there is no built-in possibility for a contributor to stop answering after a subset of the total number of questions available, and show an end screen.
All in all, PYBOSSA, is a convenient crowdsourcing tool, but has its limitations with regard to open-task crowdsourcing.

### 3.2. Linguistic findings

**Blends** For the blends analysis task, we compared the contributor answers to the attested analyses from the ANW. Contributors showed an average accuracy of 42%, with average accuracies per word ranging between 2-83%. Table 2 shows the given answers for the analysis of the blend *preferendum*. This shows that there is not always one correct analysis of a blend, when a related noun and verb can both be filled in as source word: while *prefereren* 'to prefer' + *referendum* is the attested analysis, *preferentie* 'preference' + *referendum* may also be an option. It is even more interesting to see that a number of contributors analyze this blend entirely differently than the attested analysis: they analyze the blend as *pre* 'before' + *referendum*.

| Answer | Frequency |
|---|---|
| **referendum, prefereren** | **154** |
| referendum, preferentie | 60 |
| pre, referendum | 16 |
| *do not know* | 11 |
| preferent, referendum | 8 |

Table 2: 5 most frequent answers given for analysis task for blend *preferendum*. Correct answer in bold.

For the blends recognition task, the contributor answers were compared to the ANW entry in which the citation occurs. Contributors had an average recognition accuracy of 87%, with average accuracies per word ranging between 54-97%. The accuracies are high: most blends are recognized correctly. Table 3 shows the given answers for the recognition of one specific blend: *twittie*. *twittie* 'twitter fight' is a blend of *twitter* and *fittie* 'fight' (slang). Most contributors correctly recognize this as blend. Many people however also perceive *fittie* (which does also occur in this citation) and *tweet* as blends, possibly because these words appear new or unknown.

| Answer | Frequency |
|---|---|
| **twittie** | **122** |
| twittie, fittie | 56 |
| fittie | 16 |
| tweet, twittie, fittie | 5 |
| *do not know* | 4 |

Table 3: 5 most frequent answers given for recognition task for blend *twittie*. Correct answer in bold.

**Neologisms** Table 4 shows the endurance and diversity judgments for the 15 words in the neologisms task. These results show that in general, neologisms rejected by lexicographers also receive lower crowd endurance scores. For diversity, this pattern is not as clear.

| Woord | Endurant | Diverse | Status |
|---|---|---|---|
| gendertransformatie | 91.2% | 70.2% | accepted |
| insectenafname | 83.5% | 55.7% | unattested |
| dreigingsmonitor | 79.6% | 54.0% | accepted |
| belevenisstad | 64.0% | 55.3% | accepted |
| vluchtelingenpraktijk | 62.8% | 46.9% | rejected |
| multimediamerk | 62.5% | 52.7% | unattested |
| zonnepriesteres | 52.2% | 17.4% | unattested |
| seniorenmodebranche | 47.4% | 28.9% | unattested |
| moeilijkheidsparadox | 45.2% | 21.7% | unattested |
| afradertje | 43.5% | 38.3% | rejected |
| tijdstrends | 38.3% | 27.8% | rejected |
| nachtnanny | 33.3% | 18.0% | accepted |
| lighttaks | 26.5% | 25.7% | accepted |
| korttheater | 20.4% | 15.0% | rejected |
| dieetopenbaring | 8.7% | 13.0% | rejected |

Table 4: Endurance and diversity judgments for the 15 words in the neologisms task, ordered by % endurance. Total number of contributors per word varies between 111 and 115. The rightmost column shows whether the word has been manually attested, and if so, has been accepted or rejected as neologism.

**Language variation** In the language variation task, we found that most people used the standard Dutch term to signify a word, only a minority of the given forms was a dialectal form. However, it is interesting to investigate the differences between Dutch and Flemish contributors. The number of contributors from The Netherlands (around 100 per question) is larger than the number of Flemish contributors (around 15 per question). Table 5 shows the relative frequencies of given answers for the concept TAKE A SEAT, split per language area. *ga lekker zitten* is very popular in The Netherlands, while *zet u* is only used in Flanders.

| Utterance | Flanders | The Netherlands |
|---|---|---|
| ga zitten | 31% | 38% |
| ga lekker zitten | 0% | 18% |
| neem plaats | 6% | 7% |
| zet u | 31% | 0% |
| pak een stoel | 6% | 4% |
| Total answers | 16 | 115 |

Table 5: Relative frequency of answers given for language variation task for concept TAKE A SEAT, per area. Top 5 results, sorted by overall absolute frequency.

These differences per area are observed for more questions. For example, a SWEET ON A STICK is referred to by many Flemish contributors as *lekstok*, whereas contributors from the Netherlands mainly use the form *lolly*. And WISHING A GOOD NIGHT is done by saying *slaap wel* in Flanders, while *welterusten* is used more in The Netherlands.

## 4. Future applications

**Integrating crowdsourcing into a lexicographic workflow** Our case study on neologisms shows the potential of crowdsourcing for lexicography. Crowdsourcing becomes

even more useful, if it becomes fully integrated into the lexicographic workflow. Currently, at the INT, newspaper material is fed in and sentences with unknown words are automatically extracted. Lexicographers then manually decide on inclusion in the dictionary. In an ideal workflow, the extracted sentences are automatically imported into a crowdsourcing application and shown to the public. Contributor judgments can help lexicographers in deciding on dictionary inclusion. A challenge will be to motivate a crowd to contribute over a long period of time. To maintain worflow stability, also in case of a temporary drop in crowd participation, crowd consultation will be an optional step in the workflow.

**Language learning** We have not yet performed crowdsourcing experiments for language learning, but we are looking into future directions which seem promising.
Crowdsourcing can be used to cluster word senses, which could help people with language or speech disabilities. Crowdsourcing has been used for word sense disambiguation before (Akkaya et al., 2010; Venhuizen et al., 2013), also specifically targeted at creating language learning material (Parent and Eskenazi, 2010). It would be worthwhile to apply this methodology to the ANW dictionary or the semantic lexixon DiaMaNT (Depuydt and De Does, 2018). Another idea could be to use crowdsourcing to select suitable learning sentences for collocations or proverbs from a corpus.

## 5. Conclusion

Our experiments have shown that crowdsourcing proves useful for documenting the Dutch language, and can be valuable for developing Dutch language learning material in the future. We used the PYBOSSA framework for our crowdsourcing experiments, which is very powerful, but also has its limitations when using it for linguistic purposes.

## 6. Acknowledgements

This work was supported by EU COST action CA160105 *enetCollect*, which is gratefully acknowledged. We thank Miet Ooms for supplying the questions for the language variation task. We thank our colleagues at the INT for valuable advices.

## 7. Bibliographical References

Akkaya, C., Conrad, A., Wiebe, J., and Mihalcea, R. (2010). Amazon mechanical turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 195–203. Association for Computational Linguistics.

Beelen, H. and Van der Sijs, N. (2014). Crowdsourcing de Bijbel. *Neerlandia / Nederlands van Nu*, (2-2014).

Burgos, P., Sanders, E., Cucchiarini, C., van Hout, R., and Strik, H. (2015). Auris populi: crowdsourced native transcriptions of Dutch vowels spoken by adult Spanish learners. *InterSpeech 2015. Dresden, Germany*, page 7.

Causer, T., Grint, K., Sichani, A.-M., and Terras, M. (2018). 'Making such bargain': Transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription.

Depuydt, K. and De Does, J. (2018). The Diachronic Semantic Lexicon of Dutch as Linked Open Data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May. European Language Resources Association (ELRA).

Gries, S. T. (2004). Shouldnt it be breakfunch? A quantitative analysis of blend structure in English. *Linguistics*, 42(3), January.

Holdt, Š. A., Čibej, J., Dobrovoljc, K., Gantar, P., Gorjanc, V., Klemenc, B., Kosem, I., Krek, S., Laskowski, C., and Robnik-Šikonja, M. (2018). Thesaurus of Modern Slovene: By the Community for the Community. *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 989–997, July.

Holley, R. (2010). Crowdsourcing: How and Why Should Libraries Do It? *D-Lib Magazine*, 16(3/4), March.

Kosem, I., Krek, S., Gantar, P., Holdt, Š. A., Čibej, J., and Laskowski, C. (2018). Collocations Dictionary of Modern Slovene. *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 989–997, July.

Metcalf, A. A. (2004). *Predicting new words: the secrets of their success*. Houghton Mifflin Harcourt.

Parent, G. and Eskenazi, M. (2010). Clustering dictionary definitions using Amazon Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 21–29. Association for Computational Linguistics.

Sanders, E., Burgos, P., Cucchiarini, C., and van Hout, R. (2016). Palabras: Crowdsourcing Transcriptions of L2 Speech. *International Conference on Language Resources and Evaluation (LREC) 2016. Portorož, Slovenia*, page 7.

Schoonheim, T. and Tempelaars, R. (2010). Dutch Lexicography in Progress, The Algemeen Nederlands Woordenboek (ANW). In *Proceedings of the XIV Euralex International Congress, Ljouwert, Fryske Akademy/Afuk, abstract*, page 179.

Tiberius, C. and Niestadt, J. (2010). The ANW: An online Dutch dictionary. *Proceedings of the XIV Euralex International Congress. Ljouwert, Fryske Akademy/Afuk*.

Tiberius, C. and Schoonheim, T. (2016). The Algemeen Nederlands Woordenboek (ANW) and its Lexicographical Process. *Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht "Internetlexikografie". Mannheim: Institut für Deutsche Sprache. (OPAL X/2014 ).*

Van der Wal, M. J., Rutten, G., and Simons, T. (2012). Letters as loot: Confiscated Letters filling major gaps in the History of Dutch. In Marina Dossena et al., editors, *Pragmatics & Beyond New Series*, volume 218, pages 139–162. John Benjamins Publishing Company, Amsterdam.

Venhuizen, N., Evang, K., Basile, V., and Bos, J. (2013). Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*.