Why do we Need Games? Analysis of the Participation on a Crowdsourcing Annotation Platform

Alice Millour
Université Paris-Sorbonne / STIH
alice.millour@paris-sorbonne.fr

Karën Fort
Université Paris-Sorbonne / STIH
karen.fort@paris-sorbonne.fr

1 Introduction

Annotated corpora are necessary both to develop and to evaluate natural language processing tools. However, building such corpora is notoriously expensive. For less-resourced languages, the (lack of) availability of language experts represents yet another obstacle to overcome. To address both issues, we developed a lightweight crowdsourcing platform, Bisame[1] (Millour et al., 2017), that aims at collecting part-of-speech (POS) annotations for less-resourced languages, testing it on a regional language from France: Alsatian (about 550,000 speakers). Crowdsourcing can take several forms, including microworking, citizen science and games with a purpose. We hypothesized that the speakers of a less-resourced language like Alsatian would be motivated to help developing linguistic resources for their own language, without the need to develop a full-fledged game.

2 Overview of the Platform

The tagset used for annotation is the Universal POS tagset, defined by Petrov et al. (2012). The annotation task is performed on a whole sentence. Participants must train on four reference sentences before producing annotations on the raw corpus. Depending on the results of the pre-annotation (performed by the `TreeTagger` (Schmid, 1997) for German, following the methodology of Bernhard and Ligozat (2013), and `MElt` (Denis and Sagot, 2010), which is regularly trained on the collected annotations, participants have to approve or reject a suggested annotation, or pick the correct tag in a shortlist of the most probable tags.[2] The only two gamified features that we introduced in this experiment are a leader board (since November 2016) and a progress bar (since January 2017), indicating the annotation state of the current corpus. The total numbers of participants and annotations are also displayed. As of end of February 2017, 42 participants produced 8,833 annotations in 59 days (between May 2016 and February 2017). Participation peaks were due to communication on `Facebook` or reminders by email.

3 Analysis of the Participation

Out of the 64 registered users who completed the training phase, only 42 actually produced annotations. Among them, 56% (10% of the annotations) spent one day on the platform, 33% (24% of the annotations) spent two or three days and 10% (66% of the annotations) spent more than three days. These observations show that, besides the complexity of attracting participants (motivation) we struggle to retain them on the platform (volition). They also tend to confirm a phenomenon that has already been described (Chamberlain et al., 2013): a minority of participants produces a lot. Nevertheless, while the total number of productive users has increased by 30% in the last nine months (from 31 to 42) the number of annotations per week has grown by more than 150%. Figure 1 presents the numbers of active users and produced annotations per week. Putting aside the peak observed in November, due to a unique user who produced more than 3,000 annotations in two days, annotations are nearly equally distributed between participants. Thus, we observe that in May 2016, 29 users produced 2,944 annotations, while in January 2017, half as many produced 7,302 annotations: the improvement in the interface design—easing the annotation task from choosing a tag within a list to approving or rejecting a suggested tag—and the features described earlier, are most probably responsible for the increase in the average number of annotations per user per week (from 87 to 316). This progress is encouraging

[1] See `http://bisame.herokuapp.com`.

[2] The probability of a tag is based on the confidence score associated to each annotation, which is equal to the confidence score of the participant (percentage of words correctly annotated on reference sentences).

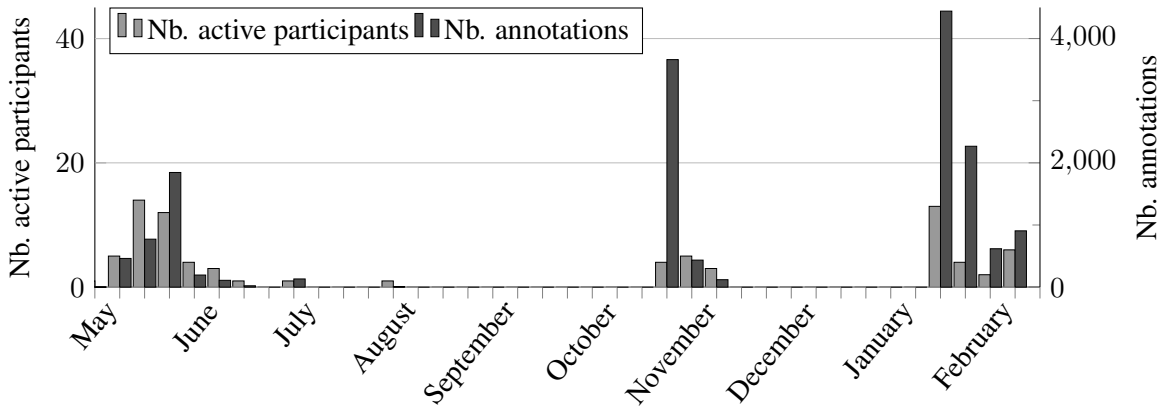

Figure 1: Number of productive participants and annotations per week (note the different scales).

and confirms the potential of gamification to retain participants. We noticed that participants tend to contribute to reach a numeric goal (leaving less than 100 unannotated words on the current corpus, for instance), and we observed that the leader board generated ephemeral competition between participants. Yet, the positive effects of the gamification features tend to vanish as the activity on the platform decreases, as the results of late February reveal.

4 Conclusion

We observed that helping develop language resources (therefore, natural language processing applications) for one's language is not enough of an incentive to produce the quantity of annotations we need to train a POS-tagger. Previous experiments, using crowdsourcing for natural disaster relief (Munro, 2013), showed that it is difficult to maintain the motivation of participants in voluntary crowdsourcing, even for potentially life-saving actions. However, we showed that gamification helps keeping the users participating. The quality of the collected annotations[3] and the progress made are promising. We therefore plan to tackle the issues discussed in section 3 by introducing new gamification features and allowing users to create their own text on the platform, following a suggestion from Liberman (2016).

References

Delphine Bernhard and Anne-Laure Ligozat. 2013. Es esch fàscht wie Ditsch, oder net? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand. In *Actes de TALARE 2013 : Traitement Automatique des Langues Régionales de France et d'Europe*, pages 209–220, Les Sables d'Olonne, France.

Jon Chamberlain, Karën Fort, Udo Kruschwitz, Mathieu Lafourcade, and Massimo Poesio. 2013. Using games to create language resources: Successes and limitations of the approach. In Iryna Gurevych and Jungi Kim, editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg.

Pascal Denis and Benoît Sagot. 2010. Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français. In *Traitement Automatique des Langues Naturelles : TALN 2010*, Montreal, Canada.

Mark Liberman. 2016. Oral histories: Linguistic documentation as social media. In *Actes de NIEUW: Novel Incentives and Engineering Unique Workflows*, Philadelphia, USA.

Alice Millour, Karën Fort, Delphine Bernhard, and Lucie Steiblé. 2017. Vers une solution légère de production de données pour le TAL : création d'un tagger de l'alsacien par crowdsourcing bénévole. In *TALN - Traitement Automatique des Langues Naturelles : TALN 2017*, Orléans, France.

Robert Munro. 2013. Crowdsourcing and the crisis-affected community: lessons learned and looking forward from mission 4636. *Journal of Information Retrieval*, 16(2).

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Actes de Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Helmut Schmid, 1997. *New Methods in Language Processing, Studies in Computational Linguistics*, chapter Probabilistic part-of-speech tagging using decision trees, pages 154–164.

[3]The annotations produced by the participants reach on average 93% in accuracy.