Toward a Lightweight Solution for Less-resourced Languages: Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing

Alice Millour, Karën Fort
STIH - EA 4509, Université Paris-Sorbonne, France
{alice.millour, karen.fort}@paris-sorbonne.fr

Abstract

We present here the results of an experiment aiming at crowdsourcing part-of-speech annotations for a less-resourced French regional language, Alsatian. We used for this purpose a specifically-developed slightly gamified platform, `Bisame`. It allowed us to gather annotations on a variety of corpora covering some of the language dialectal variations. The quality of the annotations, which reach an averaged F-measure of 93%, enabled us to train a first tagger for Alsatian that is nearly 84% accurate. The platform as well as the produced annotations and tagger are freely available. The platform can easily be adapted to other languages, thus providing a solution to (some of) the less-resourced languages issue.

Keywords: crowdsourcing, less-resourced languages, POS tagging, Alsatian

1. Introduction

Despite the progress made in unsupervised learning, manually annotated corpora are still necessary both to develop and to evaluate natural language processing (NLP) tools. However, building such corpora is notoriously expensive (see, for example (Böhmová et al., 2001)). For less-resourced languages, the (lack of) availability of language experts represents yet another obstacle to overcome.

We hypothesized that contributing to the creation of NLP tools for their language would be an incentive for non-expert speakers, especially for languages known for their speakers' activism, to participate in such a project. We therefore developed a lightweight crowdsourcing platform, `Bisame`[1], which enables participants to collaboratively produce part-of-speech (POS) annotations. We tested it on a French regional language with the appropriate specificities, i.e. activism and easy access to the Internet: Alsatian.

We present here the related work performed on Alsatian, POS tagging less-resourced languages and crowdsourcing linguistic annotations. We then describe the methodology we used and the results we obtained. Finally, we discuss the limits of this work.

2. Related Work

2.1. Alsatian

Alsatian is a generic term for the continuum of Germanic dialects spoken in Alsace and part of Moselle, two diglossic regions where French and Alsatian dialects coexist.[2] Beyond this generic glotonym lies a set of dialectal subsystems (Malherbe, 1983). In fact, low Alemannic, the main variant of Alsatian, can itself be divided into two subsets: Northern (NV) and Southern (SV) low Alemannic. Besides, Strasbourg presents a slightly different variant of Northern low Alemannic (STRV), tinted with Franconian, as evidenced by the use of *Dienschtdàj* for "Tuesday", instead of the Alemannic *Zischdig*. Thus, the 7 (phonetically different) forms *Baum*, *Bààm*, *Bœm* (SV); *Bauim*, *Bàuim*, *Bäum* (NV); *Baam* (STRV) can be found for the word "tree". Additionally to the dialectal variation and although the Orthal spelling guidelines (Crévenat-Werner and Zeidler, 2008) have existed since 2006, no orthographic convention is recognized yet as the legitimate standard among the speakers. Consequently, elided forms such as *m'r* for *mir* ("we"), or *d'* for *die* ("the") coexist depending on the spelling habits of the writer and independently of the variant.

This two level variation scheme results in a great heterogeneity of the existing written forms for a given word.

So far, works on POS tagging Alsatian corpora are very exploratory: only Bernhard and Ligozat (2013) have proposed a method based on the transposition of grammatical words into German, enhancing the performances of the German tagger on the transposed corpora. This method reaches up to 85% accuracy but has a low potential for improvement.

2.2. POS Tagging Less-resourced Languages

A variety of methods have been developed to overcome the data scarcity issue regarding the POS tagging task. They may involve additional resources such as bilingual corpora used with recurrent neural networks (see for instance (Zennaki et al., 2016)), annotation projection (see for instance (Agić et al., 2016)), or weakly supervised methods (see for instance (Li et al., 2012), showing the benefits of using the `Wiktionary` as an external resource). No such resources are available for Alsatian[3], we therefore decided to rely only on freely available raw corpora.

2.3. Crowdsourcing Linguistic Annotations

Crowdsourcing consists in issuing an open call encouraging people (a "crowd") to participate in producing data (encyclopaedia entries, a drawing, a vote etc.), today mainly through the Internet. Many taxonomies of crowdsourcing have been proposed (see (Geiger et al., 2011) for an

[1] See: `http://bisame.paris-sorbonne.fr`.

[2] In spite of the decrease of the family linguistic transmission, a study registered 550,000 Alsatian speakers (Barre and Vanderschelden, 2004).

[3] The Alsatian `Wiktionary` has disappeared, being merged with the Alemannic `Wikipedia`. The Alsatian section of this `Wikipedia` contains around 50,000 words once very similar articles have been excluded.

overview). We suggest to consider it along two axes (Fort, 2016): i) the remuneration (or not) of the activity and ii) the awareness of the participants that they are producing data (as this can be hidden underneath a playful interface, for instance). This typology allows to distinguish between transparent, voluntary crowdsourcing (for instance, Wikipedia), microworking platforms which propose rather transparent tasks and a micro-remuneration (such as Amazon Mechanical Turk) and games with a purpose, that more or less hide the task being performed (like, for example, in Phrase Detectives[4] (Poesio et al., 2013) or ZombiLingo[5] (Guillaume et al., 2016)).

Amazon Mechanical Turk has been used by many researchers, directly or through CrowdFlower, including to have POS tags annotation produced (Hovy et al., 2014). In addition to the ethical issues it raises (Fort et al., 2011), this kind of platform is not adapted to the languages we target, as very few (if not none) microworkers are fluent speakers of these languages. Besides, microworking platforms do not allow to train annotators, only to test them.

Games with a purpose have proven efficient at getting good quality linguistic data at lower cost than traditional methods (Chamberlain et al., 2013). Yet, developing a full-fledged game is a long-term endeavor which requires a range of skills (Web development, gaming mechanisms knowledge, user experience design, advertising) and has to be made profitable in the long run.[6] We thus chose to develop a lightweight platform with very few playful elements so far, namely a basic point system, a scoreboard, and a progress bar indicating the evolution of the annotation state of the corpus. We are not aware of any other voluntary crowdsourcing application for POS tagging annotation.[7] However, related tasks have been successfully achieved by volunteers, such as the annotation of suicide notes (Pestian et al., 2012) or text message translation in an humanitarian emergency context (Munro, 2013).

Finally, there exist some generic platforms for citizen science, such as Crowd4U[8] or Zooniverse[9], but none presents a linguistic application yet.[10]

3. Methodology

3.1. Tagset

For the sake of adaptability, we chose to work with the universal POS tagset introduced by (Petrov et al., 2012) (see Appendix I), which synthesizes the tagsets of 22 languages and can easily be adapted to the needs of each language.[11] In fact, the only modification we initially made was to have the X category ("Others", a catch-all category hard to interpret) to match only the cases of code-switching which cannot be analyzed as loan words.

We further became aware that the contractions that we do not automatically split (such as *am*: *an+dem* ("at the")) generated confusion and frustration among the participants. We thus introduced the ADP+DET category, and accordingly corrected the existing annotations to abide by this new tagset.

3.2. Corpora and Lexicons

The training corpus (T), annotated via the platform, consists of 333 sentences, adding up to 6,878 tokens. The low quantity of raw corpus as well as the urge to produce a freely available annotated corpus have forced us to follow a pragmatic approach and to gather an "opportunistic" corpus (McEnery and Hardie, 2011). By definition, this generates a bias in term of content: 80% of the corpus is made of articles from the Alemannic Wikipedia, 20% being a novel kindly provided by a participant-author.

Our evaluation corpus (E) is made of 4 texts adding up to 1,468 tokens (102 sentences), manually annotated by expert linguists from LiLPa, in Strasbourg. As shown in Table 1, both corpora are made of at least two variants.

Name	Nb. Sentences (Nb. tokens)	Content
T_{SV}	267 (5,110)	Wikipedia articles
T_{STRV}	66 (1,768)	Novel
E_{SV}	47 (875)	Wikipedia articles
$E_{\text{NV},1}$	26 (362)	Theater piece
$E_{\text{NV},2}$	29 (231)	Recipes

Table 1: Description of the training and evaluation corpora.

We have also integrated two lexicons to the tagger training process: i) a lexicon of grammatical words (determiners, pronouns, prepositions, conjunctions, particles) and frequent verbs and adverbs, summing up to 322 entries, which has been compiled by Bernhard and Ligozat (2013), ii) a lexicon with more than 40,000 entries from the Office for Alsatian Language and Culture (OLCA) bilingual lexicons , a bilingual dictionary compiled by the Culture and Heritage of Alsace Association (ACPA) and a multilingual French-German-Alsatian dictionary (Adolf, 2006).

The integration of these various sources increases the coverage of the dialectal and scriptural variants. We can for instance find the following entries for the word "elbow": *Elleböje* (OLCA), *Ellaboja* (OLCA), *Elleboje* (ACPA), *Ällabooga* (ACPA).

3.3. Preprocessing of the Corpora

The gathered texts were tokenized using a specific Python script, which we completed when cases of wrong segmentation –due to unknown spelling habits– were brought to our attention by the participants. For instance *'r* can either be considered as a separated token when placed after a verb (e.g. *hät'r*, "he has") or as part of a token containing an elided vowel (e.g. *d'r*, "the").

Both the training and evaluation corpora have been pre-annotated with two taggers: i) the Stanford POS

[4] See: http://anawiki.essex.ac.uk/phrasedetectives/.

[5] See: http://www.zombilingo.org.

[6] For more details, see (Lafourcade et al., 2015).

[7] Such a platform might exist though, as some did not get any dedicated scientific publication, like LanguageQuiz: http://quiz.ucomp.eu/.

[8] See: https://crowd4u.org.

[9] See: https://www.zooniverse.org.

[10] A specific platform is under development as we write this paper: https://lingoboingo.org/.

[11] See: http://universaldependencies.org/u/pos/all.html.

`Tagger` (Toutanova et al., 2003) applied to the texts after a transposition of grammatical words in German, following the methodology defined in (Bernhard and Ligozat, 2013). and ii) `MElt` (Denis and Sagot, 2012), that we regularly trained on the annotated corpus. These pre-annotations were used to provide suggestions to the participants: when the taggers disagree, the two categories they produce are proposed to the participants, while when the agree, the consensual tag can be directly validated: this fastens the annotation process on frequent and weakly ambiguous words and has led to an increase of annotated sequences during one annotation session.[12]

3.4. Annotating with `Bisame`

To be granted access and actually produce annotations, participants must go through a four sentences training phase during which they must annotate correctly every token. The production phase also consists in annotating a sequence of 4 sentences, from which one is taken from the evaluation corpus and is used to give the participant a confidence score at the end of each sequence. The confidence score given to a participant P having produced $NbAnn_{Ref}$ annotations on sentences coming from C_{Ref} is the ratio of correct categorizations:

$$Score_P = \frac{NbAnn_{Ref,Correct}}{NbAnn_{Ref}}$$

We set $ScoreAnn_{T,P,C_i}$, the confidence score for an annotation produced by P with the category C_i, to $Score_P$ at the time of the annotation. Each token being annotated by different annotators with potentially concurrent tags, we further decide on a unique category C_T: a confidence score for each category C_i is calculated averaging the scores of the Ann_{T,P_j,C_i} produced:

$$Score_{T,C_i} = \frac{\sum_j ScoreAnn_{T,P_j,C_i}}{\sum_{i,j} ScoreAnn_{T,P_j,C_i}}$$

and $C_T = \arg\max_i(Score_{T,C_i})$. For instance, in the sentence *Dr Mentelin hàt sina Stroßburger Drukaréi grinda.* ("Mentelin has founded his Strasbourg printing house"), the token $T =$ *Stroßburger* has been annotated with three different tags $\{C_{i=1..3}\}$ by five participants, $\{P_{j=1..5}\}$. Table 2 illustrates this case, that results in choosing $C_{Stroßburger}$ =ADJ, which is the correct tag.

C_i	$ScoreAnn_{T,P_j,C_i}$	$Score_{T,C_i}$
PROPN	0.935	0.24
ADJ	0.875	**0.69**
	0.846	
	0.938	
NOUN	0.25	0.07

Table 2: Choosing the most probable tag for a given token.

4. Results

4.1. Annotated Corpus

So far, 202 people have created an account, 72 have completed the training phase, and 46 have collaboratively produced 18,917 annotations. The platform was released in May 2016 but the annotations have mainly been produced during short periods of time adding up to 73 days. Our experience confirms a well-known phenomenon described in (Chamberlain et al., 2013): a minority of participants contributes a lot (in our case the 10 most active participants produced almost 90% of the annotations).

Among the annotations, 8,244 were added on the evaluation corpus, thus enabling us to evaluate the quality of the collected annotations. The accuracy and weighted average F-score for these annotations reach 93%. This quality is above that obtained by Hovy et al. (2014) (80% accuracy) when crowdsourcing POS tags on Twitter data *via* `CrowdFlower`. This shows the benefits of being able to train the participants when designing a crowdsourcing task of this kind. Accordingly with Guillaume et al. (2016), we observe that the quality of the annotation raises with participation: with twice as many annotations, the average F-score has raised by more than 40%. The remaining 10,673 annotations were used to annotate a raw corpus of 6,878 tokens (Millour and Fort, 2017).

4.2. Tagger Performance

	E_{SV}	$E_{\text{NV},1}$	$E_{\text{NV},2}$	E_{SV}+$E_{\text{NV},1}$+$E_{\text{NV},2}$
T_{SV}	**83.7**	78.7	71.3	77.9
Unk. Tokens	40%	65%	62%	52%
$T_{\text{SV}} + T_{\text{STRV}}$	82.3	**82.8**	**71.8**	**79.1**
Unk. Tokens	40%	37%	61%	47%

Table 3: Accuracy of the trained taggers.

Table 3 shows the accuracy of the two `MElt` taggers we trained, according to the training corpus we used. The first corpus contains only the Southern variant, while the second experiment includes the corpus written in the Strasbourg variant, which is closer to Northern low Alemannic. The addition of the lexicons described in Section 3.2. to the training of `MElt` led to an increase of nearly 31% of the accuracy on unknown words, reaching 70% on average, and to improve the overall performances by 6%. Unsurprisingly, the best performance (83.7%) is reached when the training (T_{SV}) and evaluation (E_{SV}) corpora are made of the same variant of the language. We observe that adding the Strasbourg variant to the training process positively impacts the performance on the Northern variant corpora while leading to a drop in accuracy on the Southern variant corpus. This emphasizes the need to integrate the variation to the training process, and not to treat Alsatian dialects as a whole.

Overall, an analysis of the F-scores per tag shows that the lowest performances (lower than 0.5) concern the less represented tags (PART, SCONJ, SYM, ADP+DET) as well as the X category which represents 2% of the evaluation corpus and shows a F-score of 0.3. While the tag proportions are roughly the same between the training and the test corpus (both taken as a whole), as only the PROPN and X are slightly overrepresented, the distribution of tags is not balanced among the evaluation corpora.[13] In particular, the category PROPN is proportionally 2.2 times more present in E_{SV} than in the training corpus and is mistaken for NOUN

[12]The two taggers agree in 50% of the cases, and this consensual tag is corrected by the participants in 12% of the cases.

[13]The full tag distributions can be found in Appendix II.

in almost 60% of the cases. Another frequent error is the confusion between AUX and VERB (in 30% of the cases) and between VERB and ADJ (in 25% of the cases).
Our performance on Alsatian is lower than what has been obtained with similar amounts of resources for other languages: for instance, with 100 sentences (less than 3,000 tokens), Fort and Sagot (2010) trained a tagger (MElt) reaching 86.6% accuracy on section 23 of the Penn Treebank. On a less canonical language, Vergez-Couret et al. (2014) showed that the Talismane parser (Urieli, 2013) can be trained for Occitan (another French regional language) to reach 89% accuracy with 2,500 tokens, using a lexicon of 225,000 entries. We therefore think that beyond dialectal variation, orthographic inconsistencies might be the source of the lower performances obtained on Alsatian.

5. Discussion

5.1. Corpus Size and Evaluation

The small size of our evaluation corpus has led us to assess the quality of the crowdsourced annotations on a very small set of cases. The corpora also revealed unbalanced in terms of tag proportion. We are now looking into ways of building a minimal reference corpus specifically designed to address the difficulties inherent to POS tagging. In order to collect more freely available corpora we are also planning to ask participants to contribute to raw corpus building within the platform, following a suggestion from Liberman (2016).
Another consequence of the lack of reference corpus is that we could not evaluate the tagger we trained on a corpus which had not already been used to train the participants to the task, or to evaluate them. Consequently, a bias exist in our evaluation of the tagger. Moreover, Fort and Sagot (2010) have shown that pre-annotation can have a negative impact on the quality of annotations, especially on less-trained participants. This bias has not been evaluated so far on our platform. Finally, we identified recurring errors in the participants annotations due either the task complexity or to unclear guidelines. The most frequent errors concern a confusion between ADJ and ADV categories, and AUX which is mistaken for VERB in 75% of the cases. We also noticed a confusion between code-switching (annotated with X) and loanwords (annotated with their part of speech). We thus plan to work on adapting the guidelines we provide to tackle these difficulties.
Improving our user evaluation method is also necessary, as a manual inspection of the annotations has revealed that a participant with a high confidence score (0.95, the average being 0.82) had produced some bad quality annotations (some were recurring errors, others were due to the bad automatic translations provided by his browser), without significant impact on his confidence score.

5.2. Motivating the Participants

With respect to our crowdsourcing experience, the hypothesis we initially made regarding the motivation of the speakers could not be completely validated: official structures and local media have turned out to be ineffective to advertise the platform, and recruitment using social networks has revealed time consuming. The platform is not a game, and motivating the participants to contribute and to come back to contribute again has been challenging. So far, only 37% of the participants came back on the platform at least twice. Nonetheless, we have observed that adding a scoreboard and a progress bar have had a positive effect on participation and quantity of data produced during a session (Millour and Fort, 2017). We are therefore considering the development of a couple of simple gamification features to make the annotation task less tedious. In fact, according to the feedbacks we gathered from the participants, we believe that diversifying the available tasks and enhancing the community feeling within the platform to develop the *social* incentive (Poesio et al., 2013) would help heading towards a more autonomous platform.

5.3. Dealing with Variations

This experiment showed that, at least from a NLP point of view, Alsatian should not be considered as a unified language, as we observed that a multi-dialect training corpus can lead to a drop in performance on some portions of the evaluation corpus. Moreover, the use of external lexicons has proven to be efficient to enhance performance. Our future work will therefore involve focusing on crowdsourcing a multi-variant tag dictionary to complement our token-supervised (labeled sentences) with a type-supervised (tag dictionaries) approach, as defined and suggested by Garrette and Baldridge (2013). We also intend to address orthographic variation in the manner of Samardzic et al. (2015) on Swiss German, normalizing to a single representation before training the tagger.

6. Conclusion

Thanks to the voluntary crowdsourcing platform we developed, we collected 18,917 annotations, thereby building the first open source POS annotated corpus for Alsatian. We used this corpus to develop the first tagger specific to Alsatian. The quality of the annotations gathered (93%) as well as the tagger performance (reaching 83.7% in accuracy) show that our method is valid. Nonetheless, some improvements with regard to both language specific considerations and methodological points should be provided to tackle the obstacles and biases we identified and discussed in Section 5. The platform source code is freely available on GitHub[14] under the CeCILL v2.1 license.[15] It can be adapted to any languages (an instance already exists for Guadeloupean Creole) for which i) a minimal reference, ii) an open source raw corpus, iii) adapted annotation guidelines and, if available, iv) a baseline tagger exist. Both the corpora and the tagger model are freely available under the CC BY-NC-SA license.[16]

Acknowledgements

We wish to thank all the participants of Bisame for their motivation and valuable comments, as well as D. Bernhard and L. Steiblé (LiLPa) for their advice and for providing us with the tokenizer and the annotated reference corpus.

[14] See: https://github.com/alicemillour/Bisame.
[15] See: http://www.cecill.info/.
[16] See: https://bisame.paris-sorbonne.fr/corpora.

Appendix

Appendix I

Open classes	ADJ	ADV	INTJ	NOUN	PROPN	VERB		
Closed classes	ADP	AUX	CONJ	DET	NUM	PART	PRON	SCONJ
Others	SYM	X	PUNCT					

Table 4: The universal POS tagset.[17]

Appendix II

	T_{SV} (`Wikipedia` articles)	T_{STRV} (Novel)	E_{SV} (`Wikipedia` articles)	$E_{\text{NV},1}$ (Theater piece)	$E_{\text{NV},2}$ (Recipes)
ADJ	6%	3%	3%	3%	6%
ADP	12%	7%	10%	7%	9%
ADP+DET	3%	2%	4%	3%	2%
ADV	6%	5%	5%	7%	6%
AUX	4%	6%	6%	2%	0%
CCONJ	4%	4%	3%	6%	7%
DET	11%	8%	13%	8%	12%
INTJ	0%	0%	0%	1%	0%
NOUN	17%	10%	14%	9%	21%
NUM	3%	0%	4%	0%	3%
PART	1%	1%	0%	1%	0%
PRON	6%	6%	3%	12%	2%
PROPN	3%	1%	8%	4%	0%
PUNCT	13%	34%	10%	19%	13%
SCONJ	1%	1%	0%	2%	1%
SYM	1%	0%	0%	0%	0%
VERB	10%	9%	8%	11%	18%
X	1%	0%	9%	4%	0%

Table 5: Tag distribution in the training and evaluation corpora.

7. References

7.1. Bibliographical References

Adolf, P. (2006). *Dictionnaire comparatif multilingue: français-allemand-alsacien-anglais*. Midgard, Strasbourg, France.

Agić, Ž., Johannsen, A., Plank, B., Martínez, H. A., Schluter, N., and Søgaard, A. (2016). Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Barre, C. and Vanderschelden, M. (2004). *L'enquête "étude de l'histoire familiale" de 1999 - Résultats détaillés*. INSEE, Paris.

Bernhard, D. and Ligozat, A.-L. (2013). Hassle-free POS-Tagging for the Alsatian Dialects. In Marcos Zampieri et al., editors, *Non-Standard Data Sources in Corpus Based-Research*, ZSM Studien, pages 85–92. Shaker. Volume 5.

Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2001). The Prague dependency treebank: Three-level annotation scenario. In Anne Abeillé, editor, *Treebanks: Building and Using Syntactically Annotated Corpora*. Kluwer Academic Publishers.

Chamberlain, J., Fort, K., Kruschwitz, U., Lafourcade, M., and Poesio, M. (2013). Using games to create language resources: Successes and limitations of the approach. In Iryna Gurevych et al., editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 3–44. Springer Berlin Heidelberg.

Crévenat-Werner, D. and Zeidler, E. (2008). *Orthographe alsacienne - Bien écrire l'alsacien de Wissembourg à Ferrette*. Jérôme Do Bentzinger.

Denis, P. and Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Lang. Resour. Eval.*, 46(4):721–736, December.

Fort, K. and Sagot, B. (2010). Influence of pre-annotation on POS-tagged corpus development. In *Proc. of the ACL Linguistic Annotation Workshop (LAW)*, pages 56–63, Uppsala, Sweden, July.

Fort, K., Adda, G., and Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Fort, K. (2016). *Collaborative Annotation for Reliable Natural Language Processing*. Focus series. ISTE Wiley.

Garrette, D. and Baldridge, J. (2013). Learning a part-of-speech tagger from two hours of annotation. In *Proc. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-13)*, pages 138–147, Atlanta, GA, USA, June.

Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., and Schader, M. (2011). Managing the crowd: Towards a taxonomy of crowdsourcing processes. In *Proc. of the Americas Conference on Information Systems (AMCIS)*, Detroit, MI, USA, August.

Guillaume, B., Fort, K., and Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proc. of the 26th International Conference on Computational Linguistics (COLING): Technical Papers*, pages 3041–3052, Osaka, Japan, December.

Hovy, D., Plank, B., and Søgaard, A. (2014). Experiments with crowdsourced re-annotation of a pos tagging data set. In *Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 2: Short Papers)*, pages 377–382, Baltimore, MD, USA, June.

Lafourcade, M., Joubert, A., and Le Brun, N. (2015). *Games with a Purpose (GWAPS)*. Wiley-ISTE, July.

Li, S., Graça, J. a. V., and Taskar, B. (2012). Wiki-ly supervised part-of-speech tagging. In *Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CoNLL)*, pages 1389–1398, Jeju Island, Korea, July.

Liberman, M. (2016). Oral histories: Linguistic documentation as social media. In *Novel Incentives and Engineering Unique Workflows (NIEUW)*, pages 38–39, Philadelphia, PA, USA, October.

Malherbe, M. (1983). *Les langages de l'humanité (une encyclopédie des 3000 langues parlées dans le monde)*. Collection Bouquins. Laffont.

McEnery, T. and Hardie, A. (2011). *Corpus Linguistics:*

[17] We reproduce here the classification proposed by Petrov et al. (2012) without validating all their choices: some "closed classes", such as prepositions (ADP), being highly productive.

Method, Theory and Practice. Cambridge Textbooks in Linguistics. Cambridge University Press.

Millour, A. and Fort, K. (2017). Why do we Need Games? Analysis of the Participation on a Crowdsourcing Annotation Platform. In *Proc. of Games4NLP*, Valencia, Spain, April.

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

Pestian, J. P., Matykiewicz, P., and Linn-Gust, M. (2012). What's in a note: Construction of a suicide note corpus. *Biomedical Informatics Insights*, 5:1–6.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proc. of Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May.

Poesio, M., Chamberlain, J., Kruschwitz, U., Robaldo, L., and Ducceschi, L. (2013). Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):3:1–3:44.

Samardzic, T., Scherrer, Y., and Glaser, E. (2015). Normalising orthographic and dialectal variants for the automatic processing of swiss german. In *Proc. of the 7th Language and Technology Conference*, Poznań, Poland, November.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 173–180, Edmonton, Canada, May.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Toulouse le Mirail-Toulouse II University.

Vergez-Couret, M., Urieli, A., and Foix, F. (2014). Postagging different varieties of occitan with single-dialect resources. In *Proc. of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 21–29, Dublin, Ireland, August.

Zennaki, O., Semmar, N., and Besacier, L. (2016). Inducing Multilingual Text Analysis Tools Using Bidirectional Recurrent Neural Networks. In *Proc. of International Conference on Computational Linguistics (COLING)*, Osaka, Japan, December.

7.2. Language Resource References

Millour, Alice and Fort, Karën. (2017). *Crowdsourced POS Tagged Corpus of Alsatian*.