Unsupervised Data Augmentation for Less-Resourced Languages with no Standardized Spelling

Alice Millour, Karën Fort

Unsupervised Data Augmentation for Less-Resourced Languages with no Standardized Spelling

Alice Millour and **Karën Fort**
Sorbonne Université / STIH - EA 4509
28, rue Serpente, 75006 Paris, France
alice.millour@etu.sorbonne-universite.fr,
karen.fort@sorbonne-universite.fr

Abstract

Non-standardized languages are a challenge to the construction of representative linguistic resources and to the development of efficient natural language processing tools: when spelling is not determined by a consensual norm, a multiplicity of alternative written forms can be encountered for a given word, inducing a large proportion of out-of-vocabulary words.

To embrace this diversity, we propose a methodology based on crowdsourcing alternative spellings from which variation rules are automatically extracted. The rules are further used to match out-of-vocabulary words with one of their spelling variants. This virtuous process enables the unsupervised augmentation of multi-variant lexicons without requiring manual rule definition by experts. We apply this multilingual methodology on Alsatian, a French regional language and provide (i) an intrinsic evaluation of the correctness of the obtained variants pairs, (ii) an extrinsic evaluation on a downstream task: part-of-speech tagging.

We show that in a low-resource scenario, collecting spelling variants for only 145 words can lead to (i) the generation of 876 additional variant pairs, (ii) a diminution of out-of-vocabulary words improving the tagging performance by 1 to 4%.

1 Natural Language Processing and Non-Standardized Languages

Non-standardized languages present a great productivity of spelling variants for a given word. The absence of standardized spelling points up the geographical and demographic variations that might exist and are otherwise smoothed down. This variability results in the coexistence of alternative written forms, hence in a large proportion of out-of-vocabulary words in the context of supervised machine learning.

In what follows, we first present our approach to generate spelling variant pairs based on an initial set of crowdsourced spelling variant pairs. This method is language independent and relies on resources that do not require expert knowledge, hence can easily be crowdsourced.

Second, we exemplify the use of such a method to reduce the proportion of unknown words that undermines supervised algorithms in the context of non-standardized languages.

1.1 Working with Multi-Variant Linguistic Resources

The question of variation in non-standardized languages naturally arises starting when one begins the process of corpus building (or collection). When dialectal and spelling variants overlap, inter- and intra-dialectal variations can be hard, not to say impossible, to untangle. In the following, we will design as "spelling variant" any variant due to either dialectal variation, spelling convention variation, or an accumulation of both.

Although one might chose to work on corpora produced in a controlled environment, in which the spelling conventions and writers are carefully chosen, this setup is unlikely to produce satisfying results on real-life data.

Producing linguistic resources, be it lexica, raw or annotated corpora, represents a cost that cannot be afforded for languages missing resources in the broad sense, including funding and experts. Crowdsourcing has proven to be a viable option to produce quality resources at a reduced cost (Chamberlain et al., 2013). Applying crowdsourcing to less-resourced non-standardized lan-

Proceedings of Recent Advances in Natural Language Processing, pages 776–784,
Varna, Bulgaria, Sep 2–4 2019.

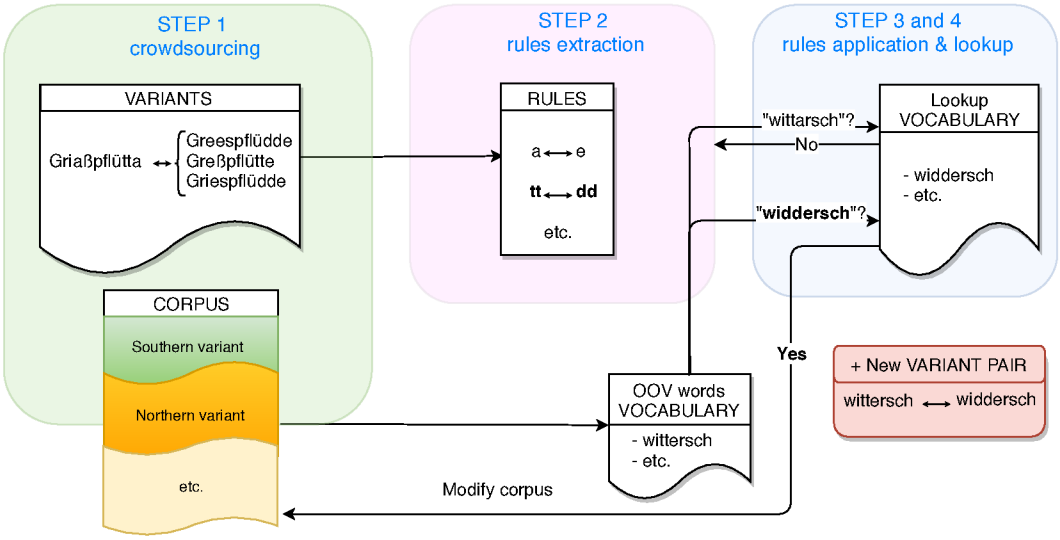

Figure 1: Data augmentation process.

guages presents additional difficulties such as accessibility to the speakers, or representativity of contents (Millour and Fort, 2018b).

Yet, when a community of speakers can be found on-line, it seems necessary to empower them to produce raw corpora and to document variability. In fact, the speakers appear to be, collectively, the only experts of the mechanisms at stake.

To meet this goal, we developed a crowdsourcing platform that collects two types of resources: (1) raw texts and (2) spelling variants on these texts. These resources are used to seed the unsupervised augmentation of the multi-variant lexicon following a process that we detail in Figure 1.

1.2 Process Overview

Given an existing linguistic resource (corpus, lexicon, or both) R_{Lookup} and a set of out-of-vocabulary words Voc_{OOV}, the process consists of four steps:

1. crowdsourcing spelling variant pairs,
2. automatic rules extraction,
3. application of the rules on elements of Voc_{OOV},
4. lookup of the resulting transformed spelling in R_{Lookup}.

These steps are detailed in sections 2 and 3 and illustrated with their application on Alsatian.

In the context of OOV words reduction in a given corpus, step 4 is followed by a transposition of those for which a variant has been identified in R_{Lookup}. Especially, in the context of supervised machine learning, one cannot expect to find all existing variants in a training corpus. By replacing an OOV word by one of its already known spelling variants, we make the most of the annotations we have at our disposal (see Section 4).

1.3 The Case of Alsatian

Alsatian is a French regional language counting 550,000 speakers in 2004 (Barre and Vanderschelden, 2004). This continuum of Alemannic dialects is an example of language in which the dialectal variants are not erased in the written form by any spelling system.

Initiatives such as the Orthal guidelines (Crévenat-Werner and Zeidler, 2008) have been developed to unify the Alsatian spelling while being respectful of its variations. Yet, these keep the variability (*Kìrisch* and *Kìch* are the Orthal version of *Kerisch* and *Kîch*, Northern and Southern possible versions for the word "church"), and are still unknown by a majority of Alsatian Internet users as shown by a recent survey (Millour, 2019).

For this reason, the dialectal variations (6 to 8 variants emerge from the continuum) combine

with the variety of spelling habits, which might depend, for instance, on the linguistic backgrounds of the speakers.

Also, since there exist an active community of on-line speakers, Alsatian is a good candidate for crowdsourcing experiments.

2 Crowdsourcing Spelling Variants

We developed a slightly gamified crowdsourcing platform, `Recettes de Grammaire`[1] which allows us to collect (i) raw corpora in the shape of cooking recipes, (ii) part-of-speech annotations on the recipes, and (iii) alternative spellings. The platform is language independent and its source code is freely available on GitHub[2] under the CeCILL v2.1 license.[3]

We do not differentiate variants due to a variation in dialects, in spelling or in an accumulation of these two factors during collection.

The addition of a new spelling variant can be performed: (i) by adding a variant to any word that is present on the platform by clicking on a word cloud on the main page (see Figure 2), (ii) by dynamically editing the written contents on the website thanks to a feature called "Personally, I would have said it like that!", illustrated on Figure 3.

These features enable the participants to modify the content they read and further annotate in a manner that suits their writing habits. In fact, feedback we received on previous experiments led on crowdsourcing part-of-speech annotations for Alsatian (Millour and Fort, 2018a) highlighted the fact that some participants felt unrepresented by the texts on the platform, and that annotating dialectal or spelling variants they are not familiar with was an obstacle hard to overcome.

The interface allows the participants to provide an alternative spelling for either a single word or a sequence of words. Although the latter facilitates the task for the participants, it sometimes leads to alternative spellings which number of words did not match the original version, hence could not be immediately aligned. In such cases and when possible, the alternative spellings were manually aligned with the original version.

So far, the collected resource contains 367 variants provided by 10 participants for 145 words (with two to six variants per word), e.g. {*bìtsi*, *bessel*, *béssel*}, "a bit of".

The only information we possess about these participants is the languages they speak and their place of origin, when they fill it in in their profile. Based on the information provided by 8 of them, we can assume that 3 to 4 dialectal areas are covered by the towns of origin of the participants. No assumption can be made regarding their proficiency in Alsatian.

The size of this resource does not allow us to perform direct lookup for any out-of-vocabulary word me might encounter. However, we can use the aligned variants to identify substitution patterns, and extract sets of rules we apply to any OOV word as described in the following section.

3 Unsupervised Data Augmentation

3.1 Rules Extraction

In the manner of (Prokić et al., 2009), we used ALPHAMALIG[4], a multi sequence alignment tool, to perform the alignment of our variants necessary to the extraction of substitution patterns. The tool requires to be provided with an alphabet of symbols, weighted with the match, mismatch, insertion and deletion scores of given characters. Since we have no *a priori* knowledge of these scores, the only assumption we made is that vowels are more likely to match vowels than consonants and vice versa. Insertion and deletion are given the same scores for all characters. An example of the alignment obtained for four crowdsourced variants is given in table 1.

^	G	A	L	-	R	**Ì**	**E**	**W**	L	**E**	K	**Ü**	**E**	C	H	**E**	$	(1)
^	G	A	L	**E**	R	**I**	**E**	**B**	L	**E**	K	**Ü**	**E**	C	H	**A**	$	(2)
^	G	A	L	**E**	R	-	**E**	**W**	L	**E**	K	**Ù**	-	C	H	**E**	$	(3)
^	G	A	L	-	R	**Ì**	**A**	**W**	L	**A**	K	**Ü**	**A**	C	H	**A**	$	(4)

Table 1: Alignment of four variants of the Alsatian (compound) word for "carrot cake".

From the produced alignments we can identify substitution patterns of different degrees of rigidity, depending on the size of the context. We extract three sets of rules which either force the matching of the left (L), right (R) or both contexts (L+R).

The ^ and $ characters, respectively representing the beginning and the end of a word, are in-

[1] "Grammar's Recipes", see `https://bisame.paris-sorbonne.fr/recettes`.

[2] See `https://github.com/alicemillour/Bisame/tree/recipes`.

[3] See `http://www.cecill.info/index`.

[4] Source code: `http://alggen.lsi.upc.es/recerca/align/alphamalig/intro-alphamalig.html`.

Figure 2: Spelling addition using the wordcloud. The word is shown in its context, with the proposed part-of-speech (if available).

Figure 3: Spelling addition (1) and visualization (2) (highlighted words present at least one additional variant).

terpreted as elements of context. The rules are extracted from each pair of the combination of the aligned variants. From the aligned variants (1) and (2) showcased in Figure 1, four L+R rules are extracted: LR ↔ LER ; RÌE ↔ RIE ; EWL ↔ EBL ; HE$ ↔ HA$. The eight left-and-right-context-only corresponding rules are deduced from the L+R rules.

Since the result we seek is not to normalize the spelling, each rule can be used in both directions which are considered equally frequent.

From the 367 variant pairs collected for Alsatian, we extracted 213 unique rules using the left and right contexts, 227 rules using the left context only, 186 rules using the right context only.

3.2 Variant Identification and Filtering

Given a vocabulary of known words V_{lookup}, the identification of potential variants of an OOV word includes (i) optional preliminary filtering, (ii) application of rules, and (iii) lookup:

1. preliminary filtering (optional): if the unknown word is identified as a known proper noun in the lexicon, it is ignored.

2. application of the rules: for each set of rules, L+R, L, R, used in this order, the subset of rules applying to the original OOV word $R_{original\_word}$ is identified, and ordered by rule frequency. From this subset, we apply on the OOV word each possible *combination* of rules, meaning that if three rules A, B, C apply, the sequences of rules {A},{B},{C},{A;B},{A;C},{B;C} and {A;B;C} are applied.

3. lookup: the sequence of rules apply until the produced form is matched with a word present in V_{lookup}.

Although this "brute-force" method generates a great quantity of noise, the filtering operated by V_{lookup} leads to the matching of OOV word with existing variant candidates.

Since part of the dialectal and spelling variation mechanisms may be similar to some of the language morphological rules (such as gender, number, conjugation or declension), the generated variant pairs should be manually checked in context.

This phenomenon is illustrated by the analysis of the pairs generated for Alsatian in Section 5.

4 Evaluation on a Downstream Task

To illustrate the benefits of the identification of variant pairs, we evaluate its impact on a downstream task: part-of-speech (POS) tagging.

Previous experiments on Alsatian from (Millour and Fort, 2018b) have shown that using multiple variants for training can lead to a drop of accuracy on the sections of the evaluation corpus which do not match the variants represented in the training corpus (-1.4% accuracy when a corpus of Strasbourg specific variant is added in the training of a Southern variant only).

In this context, we use our methodology to match OOV words from the evaluation corpus with their potential spelling variant appearing in the training corpus.

It is important to understand that this process is independent from the tagger, and occurs after it has been trained. The extraction of pairs is performed at the time of annotation on a previously **unseen** corpus.

4.1 Language Resources and Tools Used for Evaluation

Experiments in POS tagging Alsatian include our previous work (Millour and Fort, 2018b), which uses `MElt` (Denis and Sagot, 2012), a freely available sequence labeller achieving at best 84% accuracy when the variants in the training and the evaluation corpus are carefully controlled. Experiments using word embeddings have been also been carried on Alsatian by (Magistry et al., 2018), using a raw corpus of 200 000 tokens and reaching 91% accuracy.

In the following experiments, we chose to train `MElt`, which enables us to take advantage of available lexicons existing for Alsatian. The differential in performance is more interesting to us than the performance *per se*, which is why we chose not to focus on testing our methodology on other taggers.

Two POS-tagged corpora are available for Alsatian. Both are made of texts produced in an uncontrolled environment (such as `Wikipedia`[5]) and contain multiple variants of the language:

- The `Crowdsourced Corpus` (Millour and Fort, 2018b), $C_{rowd}C$, annotated by benevolent participants on a dedicated crowdsourcing platform `Bisame`[6] with the universal POS tagset (Petrov et al., 2012) extended with two categories: APPART (preposition-determiner contraction), and FM (foreign words). The corpus contains 9,282 tokens (439 sentences), and is available under CC BY-NC-SA license. The accuracy of the annotations provided by the benevolent participants has been evaluated to 93% (Millour and Fort, 2018b).

- The `Annotated Corpus for the Alsatian Dialects` (Bernhard et al., 2018a), $T_{rad}C$, annotated with the tagset described above, extended with the categories EPE (epenthesis) and MOD (modal verb) (Bernhard et al., 2018b). The corpus contains 12,570 tokens (533 sentences) and is available under CC BY-SA license. It was annotated manually by expert linguists.

We manually corrected $T_{rad}C$ to match the tagset used in $C_{rowd}C$.

The corpus resulting from the concatenation of the two corpora, $C_{oncat}C$, was used for the following experiments. We performed a cross validation on 4 subdivisions (80% used for training, $C_{oncat}C80$, 20% for the evaluation, $C_{oncat}C20$).

We also have at our disposal two lexica:

- a multi-variant lexicon $M_{ultiVar}L$ of 54,355 entries annotated with their POS, containing grammatical words (Bernhard and Ligozat, 2013), verbs from (Steiblé and Bernhard, 2016), and various entries from (i) the Office for Alsatian Language and Culture (OLCA)

[5] See `https://als.wikipedia.org`

[6] See `https://bisame.paris-sorbonne.fr`.

bilingual lexicons, (ii) the dictionary compiled by the Culture and Heritage of Alsace Association (ACPA), and (iii) a multilingual French-German-Alsatian dictionary (Adolf, 2006).

- the `Lexicon of Place Names in the Alsatian Dialects` which contains 1,346 entries (Bernhard, 2018), used during training only.

4.2 Application of the Methodology

Since the identification of potential variant pairs depends on the initial conditions of the experiment, *i.e.* the corpus, and optionally, the lexica used to train the model beforehand, we present two experiments in which these parameters vary.

For each experiment, we extract from the training corpus the vocabulary VT_lookup and from the external lexicon, the vocabulary VL_lookup. We use the set of rules presented in Section 3.1.

We prioritize the lookup in VT_lookup to further ease the evaluation of the generated pairs relying on the context.

If the length of the OOV word is less or equal to four characters (^ and $ excluded), only the L+R rules are applied: it has been observed in preliminary tests that shorter words were more likely to lead to erroneous matching such as *das* (determiner) */dass* (subordinating conjunction) or *dien* (auxiliary) */dene* (determiner). Additionally, we force the variant candidates to have the same letter case as the OOV word.

After variant pairs have been generated, the OOV words are replaced by their variant candidate, and the pre-trained model is applied on the transposed evaluation corpus. After the corpus has been tagged, the transposed words are replaced by they original form.

4.3 Experiment 1: Uncontrolled Setup

By "uncontrolled", we mean that training and evaluation corpora are both extracted from a shuffled corpus that contains multiple variants.

Our first model is trained with $C_{oncat}C80$ (17,136 words) and evaluated on $C_{oncat}C20$ (4,374 words) before and after its transposition. After the application of the three sets of rules, using both the vocabularies extracted from $C_{oncat}C80$ and $M_{ultiVar}L$ for the lookup, 56 variant pairs were discovered and the same number of words were transposed.

	Before transp.	After transp.
Overall	0.859	0.864
OOV words	24%	22%

Table 2: Accuracy of the model trained on multi-variant corpora, before and after the corpus transposition.

The proportion of OOV words was diminished by around 2% resulting in an improvement of the tagging performance of 0.5 points (see table 2). This minimal impact is expected since the performance on "known words" is around 10 points higher than on OOV words in this setup. In fact, considering the sizes of our corpora, lowering the number of OOV words of 100 is expected to improve the overall results of 0.2 points.

4.4 Experiment 2: Controlled Setup

By "controlled", we mean that training and evaluation each contain a specific variant of Alsatian selected in a multi-variant corpus. In the following, we compare homogeneous and heterogeneous setups, in which the training and evaluation corpora either contain the same or distinct variants of Alsatian.

To highlight the effect of our methodology in an heterogeneous context, met when no corpus of each possible variant is available, we manually split $C_{oncat}C$ in two sub-corpora $N_{orth}C$ (4,880 words) and $S_{outh}C$ (7,690 words) based on the frequencies of the -e and -a noun endings, which are specific of the Northern and Southern variants respectively.

The results of these experiments are presented in table 3.

Unsurprisingly, the best results are obtained when training and evaluation corpora are of the same variant. Yet, we can observe that in this setup, the effect of transposition to identified variants has a higher impact on the proportion of OOV words and the tagging performances.

The efficiency of the methodology largely depends on: (i) the respective and relative sizes of the training and evaluation corpora, (ii) the variation in variants existing between them.

This experiment shows that the performance of a tool trained on a given corpus can be improved by modifying the corpus it is applied on to match the vocabulary it was trained with.

	$N_{orth}C20$		$S_{outh}C20$	
$N_{orth}C80$			Before transp.	After transp.
Overall	0.853		0.714	**0.752**
OOV words	40%		54%	52%
$S_{outh}C80$	Before transp.	After transp.		
Overall	0.788	**0.809**	0.864	
OOV words	51%	48%	29%	

Table 3: Accuracy of the model trained on mono-variant corpora, before and after the corpus transposition.

5 Obtained Results

The newly created resource contains 876 pairs of variants, from which 400 were identified in the training corpus, and 476 in the lookup lexicon. The size of the created resource depends on the size of the lookup corpora and lexicon, and on the number of rules. The application of the method to any unpreviously seen text may increase the number of variant pairs.

A subset of 60 pairs of these automatically generated variant pairs were submitted to an Alsatian teacher, familiar with both the dialectal and spelling variants. The pairs were presented in the context of their sentence. The expertise of the teacher was used to measure the precision of the pairs, not the recall.

Among the 60 pairs:

- 30 were actual dialectal or spelling variants.
- 13 were pairs of different forms of identical words, *e.g.*: *ìhm* (dative pronoun) / *ìrhem* (genitive pronoun), *kált* (feminine adjective) / *kálte* (masculine adjective), *wùrd* (future auxiliary) / *wärd* (conditionnal auxiliary) etc.
- 10 were caused by erroneous matching we managed to correct by making the adjustments described in section 4.2, i.e. (i) forcing the case of potential variants to match the case of the original OOV word, (ii) limiting the application of rules considering only left or right context to words which size is over four characters.
- 7 were caused by erroneous matching we were not yet able to correct *e.g.* *kräfti* (“strongly”, adverb) / *kräftiger* (“stronger”, adjective), *mine* (“mine”, determiner) / *meine* (“believe”, verb) etc.

These results show that the generated variant pairs should be hand-checked, a task that can itself be crowdsourced, provided that we have access to the context of appearance of both elements.

By construction, the newly generated pairs will not provide additional substitution rules. Yet, they provide information on the frequency of the substitution patterns.

Additionally, the erroneous variant pairs manually filtered out can be used as counter examples of variants, and further used to train variant classifiers (see, for instance, (Barteld, 2017)).

6 Related Work

Dealing with non-standardized, less-resourced languages, takes us to the limits of NLP: first, we have no standard to rely on and not enough expert linguists to help us, and second, very few language resources are available for us to work with, even raw corpora. These two constraints are rarely met in the literature and, to our knowledge, the solution we propose has never been used before.

However, it closely relates to other experiments that involve at least a standard spelling and sometimes an expert linguist supervision. One such example is VARD 2, a tool that allows to manually and automatically standardize Early Modern English (Baron and Rayson, 2008, 2009). Another one concerns the Basque language (Etxeberria Uztarroz et al., 2014) and proposes a solution to map the variations of the language to the standard form using an existing morphological analyzer and a parallel corpus. Obviously, a lot of more or less recent publications concern the design and use of morphophonological rules, in particular in various flavors of FSTs, but most of them require the intervention of a highly-skilled linguist.

Among such publications, the work by Kimmo Koskenniemi on modeling regular correspondences between Finnish and Estonian is particularly inspiring (Koskenniemi, 2013), but inapplicable in our case. The same goes for the type

for work described in (Theron and Cloete, 1997), in which the rules are automatically extracted but with a known (morphological) goal.

The closest work to ours is that described in (Barteld, 2017), as it focuses on detecting spelling variants in Middle Low German unrelated to a standard. Yet, the described method requires the training of a classifier to filter the generated pairs. This classifier is based on a resource that contains 1,834 pairs of spelling variants, a resource that is unavailable for most non-standardized languages.

Regarding Alsatian more specifically, Bernhard (2014) aligns spelling variants relying on a multi-variant bilingual French-Alsatian lexicon annotated with part-of-speech and a phonetization of Alsatian. This high dependency on existing resources make this method challenging to adapt to other languages for which the only available experts are the very speakers of the language.

7 Conclusion

We have presented a method to automatically generate pairs of spelling variants based on a small subset of crowdsourced pairs.

The method does not require manual rules definition by experts and is language independent. The resources needed to perform variant pair detection can be easily produced by the speakers, who hold the knowledge of the of the variation mechanisms. The crowdsourcing of variants, unlike that of POS tags, requires no prior training.

In fact, even the expertise necessary for the validation of the variant pairs is about to be transferred to the participants of the crowdsourcing platform.

The originality of this methodology is that once the rules have been extracted, the process feeds from previously unseen texts. This is particularly useful in a less-resourced scenario where a raw corpus is being collected from various sources.

The code of both the gamified crowdsourcing platform and the variants generation is freely available on GitHub[7]. The created multi-variant lexicon is also available under a CC license.

We plan to extend this work to other non-standardized languages. We have started working on adapting the platform to Mauritian, a French-based Creole, the morphology of which is very different from that of Alsatian.

[7]See `https://github.com/alicemillour/Bisame/tree/recipes`.

Acknowledgments

We wish to thank the participants of `Bisame` and `Recettes de Grammaire` for their motivation and comments, as well as J-N. S. Kempf for his time and expertise on the Alsatian variants. We also thank B. Sagot who inspired us with the title of this paper.

References

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

Alistair Baron and Paul Rayson. 2008. Vard 2: A tool for dealing with spelling variation in historical corpora. In Aston University, editor, *Proceedings of the Postgraduate Conference in Corpus Linguistics,*. Birmingham, UK.

Alistair Baron and Paul Rayson. 2009. Automatic standardisation of texts containing spelling variation: How much training data do you need? In University of Liverpool, editor, *Proceedings of the Corpus Linguistics Conference*. Liverpool, UK.

Corinne Barre and Mélanie Vanderschelden. 2004. *L'enquête "étude de l'histoire familiale" de 1999 - Résultats détaillés*. INSEE, Paris.

Fabian Barteld. 2017. Detecting spelling variants in non-standard texts. In *Proceedings of Student Research Workshop (EACL 2017)*. Valencia, Spain.

Delphine Bernhard. 2014. Adding dialectal lexicalisations to linked open data resources: the example of alsatian. In *Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014)*. Reykjavik, Iceland, pages 23–29. https://hal.archives-ouvertes.fr/hal-00966820.

Delphine Bernhard. 2018. Lexicon of place names in the alsatian dialects. https://doi.org/10.5281/zenodo.1404873.

Delphine Bernhard, Pascale Erhart, Dominique Huck, and Lucie Steiblé. 2018a. Annotated corpus for the alsatian dialects. Guide d'annotation, LiLPa, Université de Strasbourg.

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 *Proceedings of TALARE (Traitement Automatique des Langues Régionales de France et d'Europe) (TALN'13)*. Les Sables d'Olonne, France, pages 209–220.

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steible, Pascale Erhart,

Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynés, Sophie Rosset, Jean Sibille, and Thomas Lavergne. 2018b. Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard. In *Proceedings of 11th edition of the Language Resources and Evaluation Conference (LREC 2018)*. Miyazaki, Japan. https://hal.archives-ouvertes.fr/hal-01704806.

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*, Springer Berlin Heidelberg, Theory and Applications of Natural Language Processing, pages 3–44. https://doi.org/10.1007/978-3-642-35085-6_1.

Danielle Crévenat-Werner and Edgar Zeidler. 2008. *Orthographe alsacienne - Bien écrire l'alsacien de Wissembourg à Ferrette*. Jérôme Do Bentzinger.

Pascal Denis and Benoît Sagot. 2012. Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Lang. Resour. Eval.* 46(4):721–736. https://doi.org/10.1007/s10579-012-9193-0.

Izaskun Etxeberria Uztarroz, Iñaki Alegría Loinaz, Mans Hulden, and Larraitz Uria Garin. 2014. Learning to map variation-standard forms in basque using a limited parallel corpus and the standard morphology. *Procesamiento del Lenguaje Natural* 52:13–20.

Kimmo Koskenniemi. 2013. Finite-state relations between two historically closely related languages. In Northern European Association for Language Technology, editor, *Proceedings of the workshop on computational historical linguistics at NODALIDA 2013*. Oslo , Norway, volume 18, pages 43–53.

Pierre Magistry, Anne-Laure Ligozat, and Sophie Rosset. 2018. Étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux. In *Proceedings of Conférence sur le Traitement Automatique des Langues Naturelles*. Rennes, France. https://hal.archives-ouvertes.fr/hal-01793092.

Alice Millour. 2019. Getting to Know the Speakers: a Survey of a Non-Standardized Language Digital Use. In *Proceedings of 9th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, Poland. https://hal.archives-ouvertes.fr/hal-02137280.

Alice Millour and Karën Fort. 2018a. À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées. In *Revue TAL : numéro spécial sur les langues peu dotées (59-3)*, Association pour le Traitement Automatique des Langues.

Alice Millour and Karën Fort. 2018b. Toward a Lightweight Solution for Less-resourced Languages: Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing. In *Proceedings of 11th International Conference on Language Resources and Evaluation (LREC'18)*. Miyazaki, Japan. https://hal.archives-ouvertes.fr/hal-01790615.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey.

Jelena Prokić, Martijn Wieling, and John Nerbonne. 2009. Multiple sequence alignments in linguistics. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 18–25. http://dl.acm.org/citation.cfm?id=1642049.1642052.

Lucie Steiblé and Delphine Bernhard. 2016. Towards an Open Lexicon of Inflected Word Forms for Alsatian: Generation of Verbal Inflection. In *JEP-TALN-RECITAL 2016*. Paris, France, volume 2 of *Proceedings of la conférence conjointe JEP-TALN-RECITAL 2016, volume 2 : TALN*, pages 547–554. https://hal.archives-ouvertes.fr/hal-01338411.

Pieter Theron and Ian Cloete. 1997. Automatic acquisition of two-level morphological rules. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, ANLC '97, pages 103–110. https://doi.org/10.3115/974557.974573.