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

Alice Millour, Karën Fort

June 20, 2019

Sorbonne Université

Outline

Working on non-standardized languages

Step 1: Existing experiments of CS for Alsatian

Step 2: Make use of non-standardized resources

Step 3: Integrating variation

Evaluations

Non-standardized languages

Non-standardized languages?

Non-standardized languages

Non-standardized languages?

Some examples:

- historical texts (standard was established later on)
- user generated content in French (standard exist but is not respected)
- **languages with a recent scriptural tradition**

Non-standardized languages

Non-standardized languages?

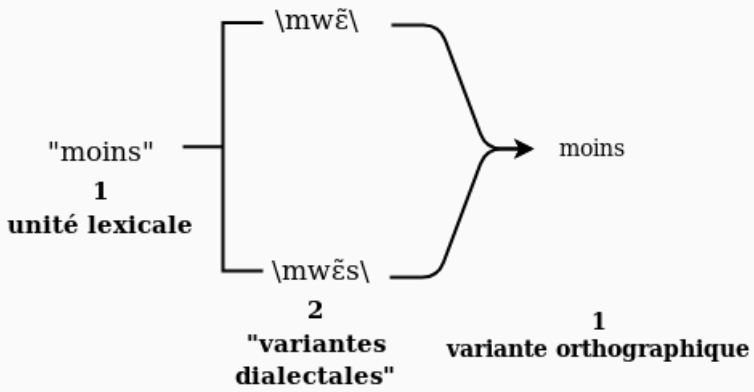
Some examples:

- historical texts (standard was established later on)
- user generated content in French (standard exist but is not respected)
- **languages with a recent scriptural tradition**

non-standardized languages present inter- and intra- speakers variation

The overlap of dialectal and spelling variations

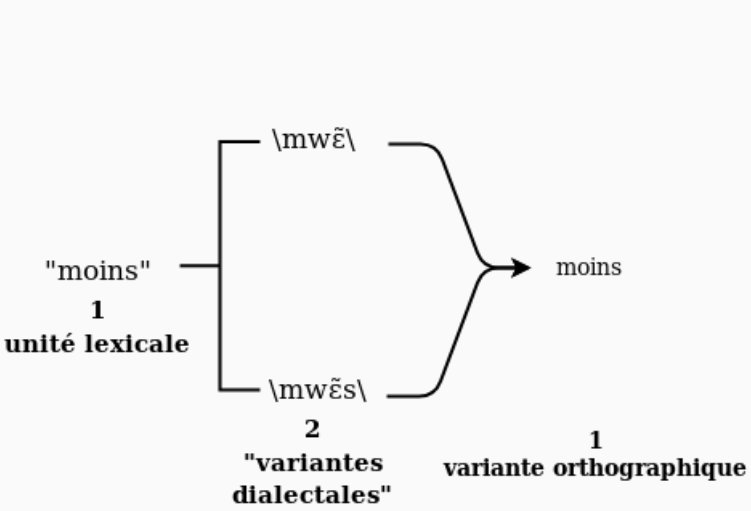
French lexical unit "*Moins*"

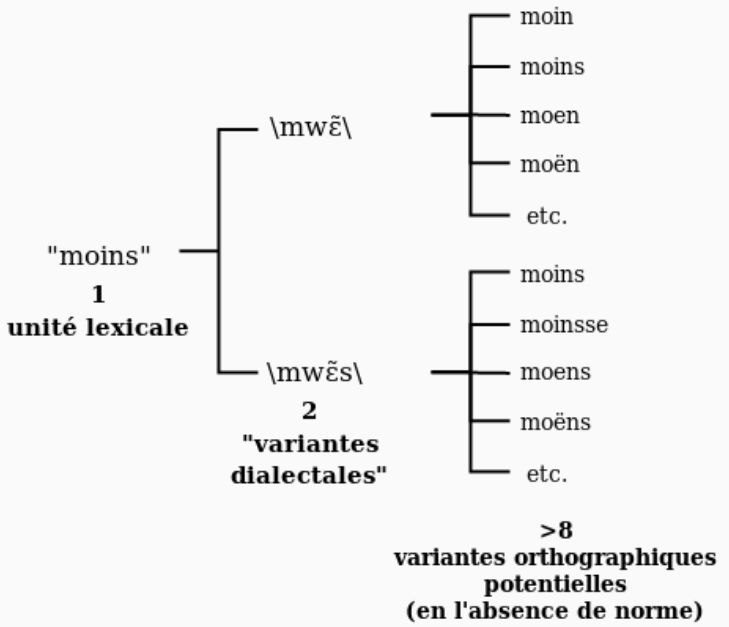


when a spelling standard exists

The overlap of dialectal and spelling variations

French lexical unit "*Moins*"



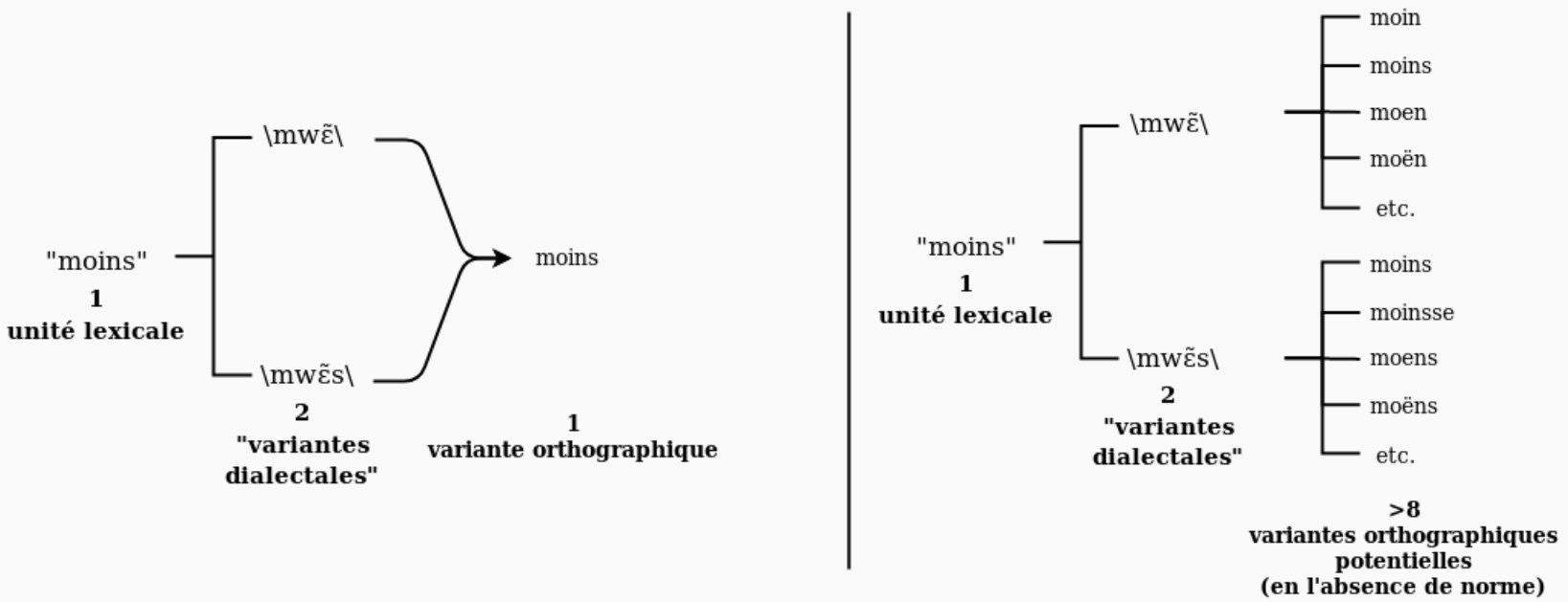


when a spelling standard exists

without a spelling standard

The overlap of dialectal and spelling variations

French lexical unit "*Moins*"



alternative written forms coexist

in terms of ML:

- **increase of OOV words proportion**
- **decrease of algorithms' performances**

Embracing variation

2 options :

1. build adecuate representative resources

Embracing variation

2 options :

1. build adecuate representative resources $\Rightarrow$ **unrealistic**

Embracing variation

2 options :

1. build adecuate representative resources ⇒ **unrealistic**
   Consequence: **we have to work with multi-variant linguistic resources**

Embracing variation

2 options :

1. build adecuate representative resources ⇒ **unrealistic**
   Consequence: **we have to work with multi-variant linguistic resources**
2. find another way to:
   - 2.1 acquire knowledge on the **mechanics** of the variation phenomena
   - 2.2 **integrate** the knowledge into the ML process

Embracing variation

2 options :

1. build adecuate representative resources ⇒ unrealistic
   Consequence: **we have to work with multi-variant linguistic resources**
2. find another way to:
   - 2.1 acquire knowledge on the **mechanics** of the variation phenomena
   - 2.2 **integrate** the knowledge into the ML process

How ?

Motivations for voluntary CS

make the most of the speakers' knowledge

common motivations:

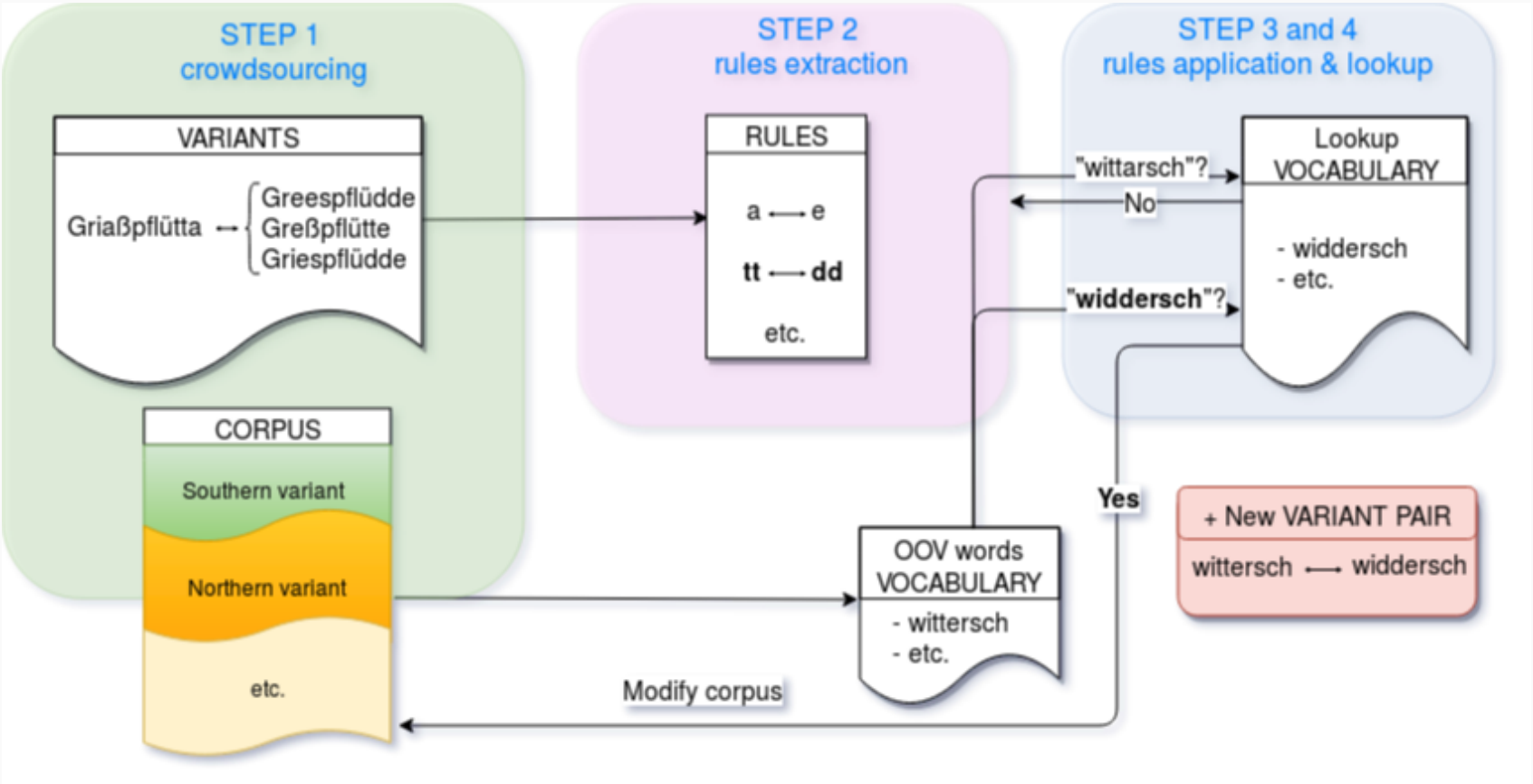
- lack of available resources

  (see `Prague Treebank`, 5 years, 600,000 $ [Böhmová et al., 2001])
  - raw and annotated linguistic resources
  - linguists
  - fundings
- accessibility to speakers

the case of non-standardized languages:

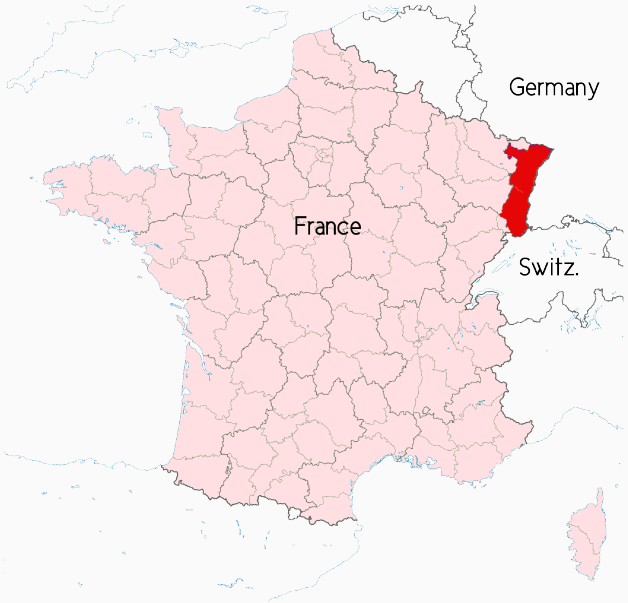
- **No expert can document all the existing variants for a given lexical item** ⇒ **necessity** to involve the languages' speakers

Overview of the process



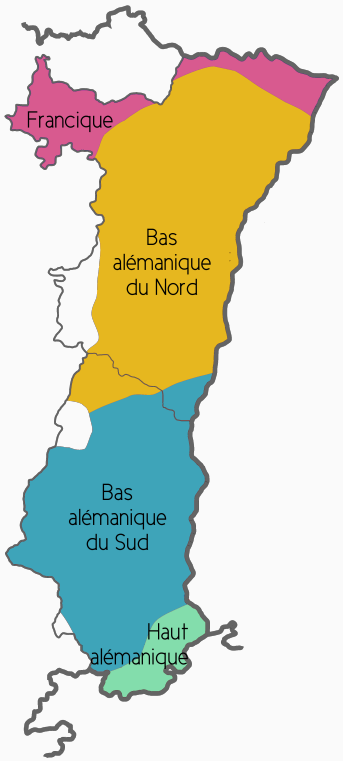
Outline

Elsässerditsch: a French “regional” language



- **Continuum** of Alemannic dialects
- **550,000 speakers** in 2004[Barre and Vanderschelden, 2004]
- **bilingual** population
- vulnerable (UNESCO)

The dialectal continuum and the spelling "standard"



7 to 8 identified variants

Mer müess mache dass d'Kerisch mittess im Dorf bliebt.

*Mer müess màche dàss d'Kìrisch mìtel im Dorf blibt.**

Mr müass màcha dàss d'Kîch mittess îm Dorf blibt.

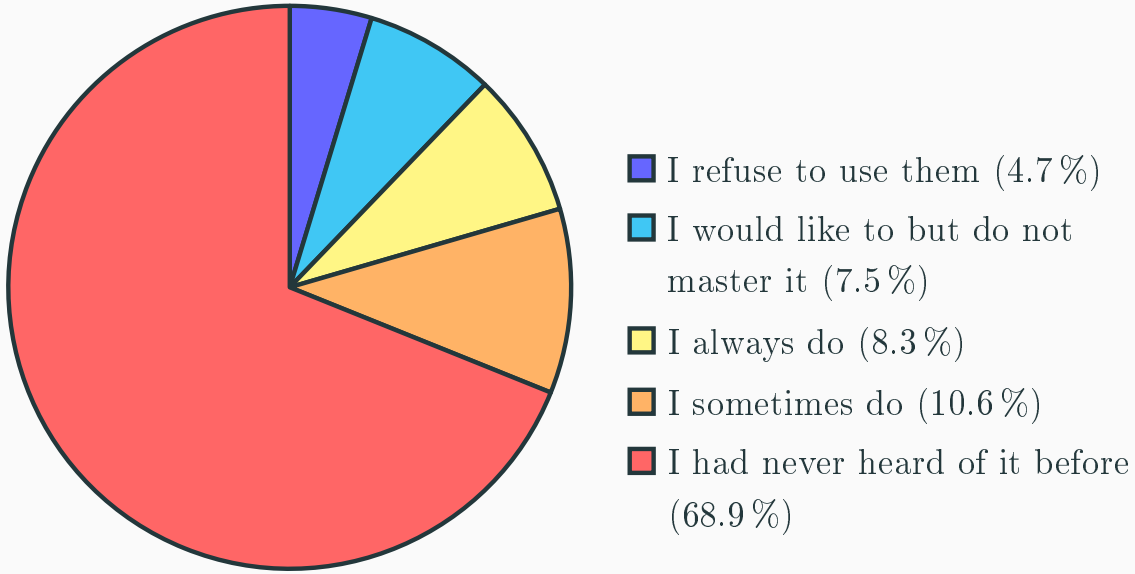
*M'r müess màcha dàss d'Kìch mìtel im Dorf blibt.**

*ORTHAL spelling system [Crévenat-Werner and Zeidler, 2008]

The dialectal continuum and the spelling "standard"

"When you write, do you follow the ORTHAL guidelines?"

[Millour and Fort, 2019]



- no consensual spelling standard
- no formal description of the variants
- high productivity of potentially out-of-vocabulary words

Kerisch ⇔ Kìrisch ⇔ Kîch ⇔ Kìch

First experiment: Bisame

One task: POS tag[a] open source existing raw corpus

`http://bisame.paris-sorbonne.fr` [Millour and Fort, 2018]

[a] See `http://universaldependencies.org/u/pos/all.html`, [Petrov et al., 2012].

First experiment: Bisame

Identified issues:

- lack of raw corpus
- underrepresentation of the variant(s) of the participants
- **sensitivity of the trained taggers to dialectal and spelling variation**

Second experiment: Recettes de Grammaire (Grammar's recipes)

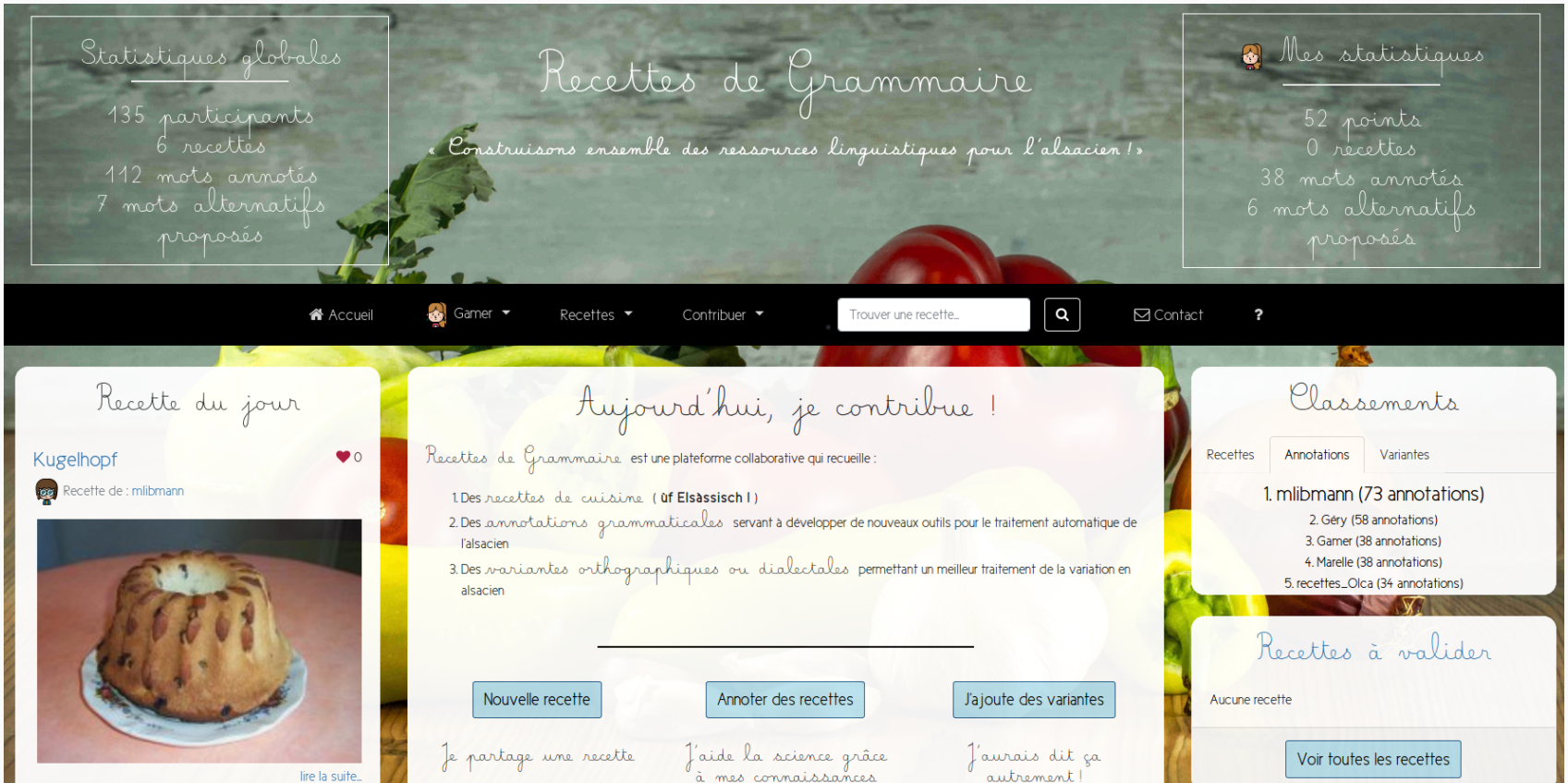
Three tasks:

- produce additionnal raw corpus (cooking recipes)
- annotate own writings
- **add dialectal and scriptural variants**

Second experiment: Recettes de Grammaire (Grammar's recipes)

Three tasks:

- produce additionnal raw corpus (cooking recipes)
- annotate own writings
- **propose dialectal and scriptural variants**



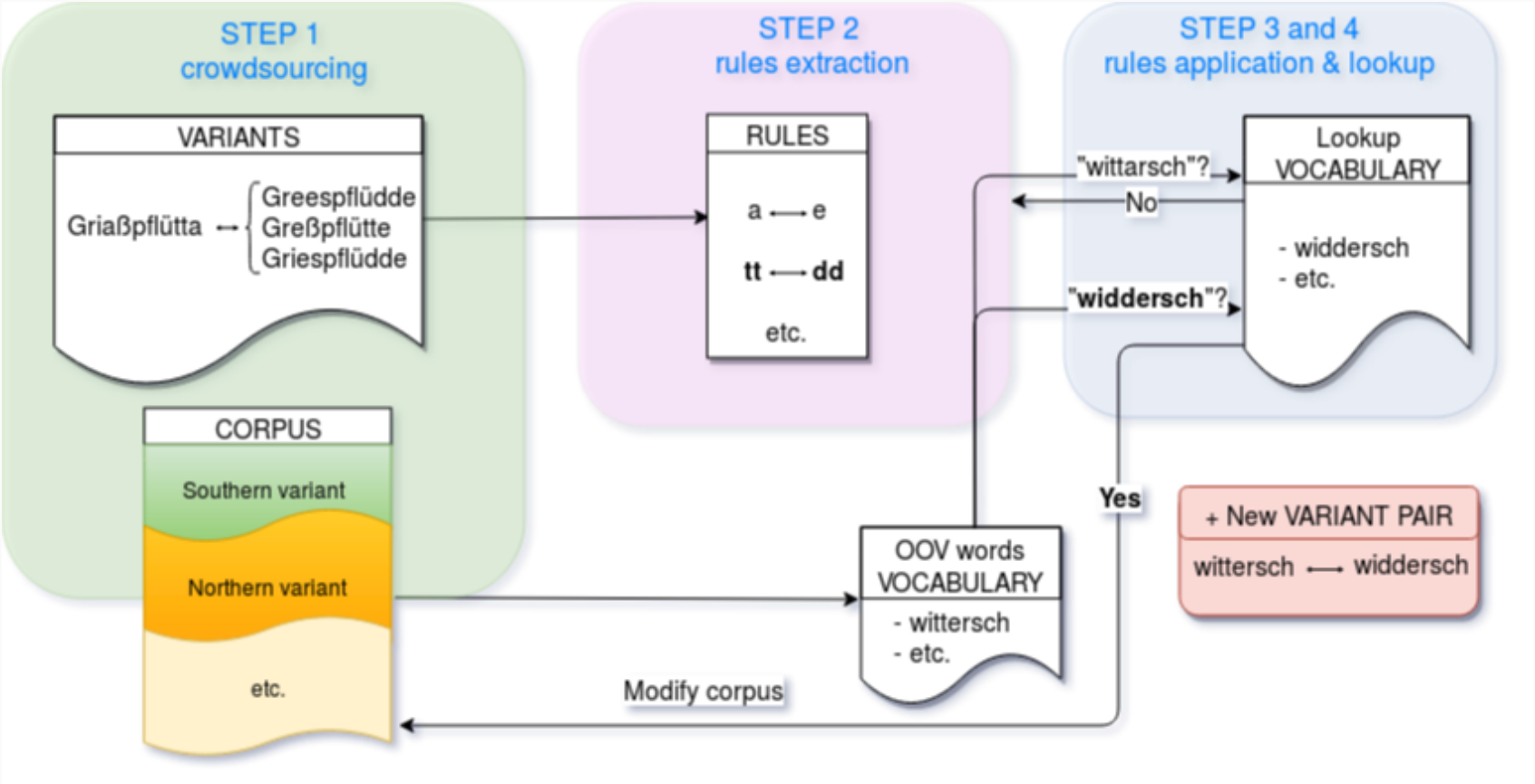
The resource produced: tuples of CS spelling variants

- 10 participants
- 145 words
- 367 variants (1 to 6 variants per word)

Example: {*bìtsi*, *bessel*, *béssel*} ("a bit of")

Outline

Step 2: Rules extraction



Alignment of crowdsourced variants

multi sequence alignment tool:

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

^	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".

Rule extraction

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

3 sets of rules extracted:

- force left and right contexts (L+R)
- force left context (L)
- force right context (R)

From (1) and (2), we extract 4 L+R rules:
LR ↔ LER ; RÌE ↔ RIE ; EWL ↔ EBL ; HE$ ↔ HA$
(+ 8 L rules and 8 R rules)

Results

from:

- 145 words
- 367 variants (1 to 6 variants per word)

we extract:

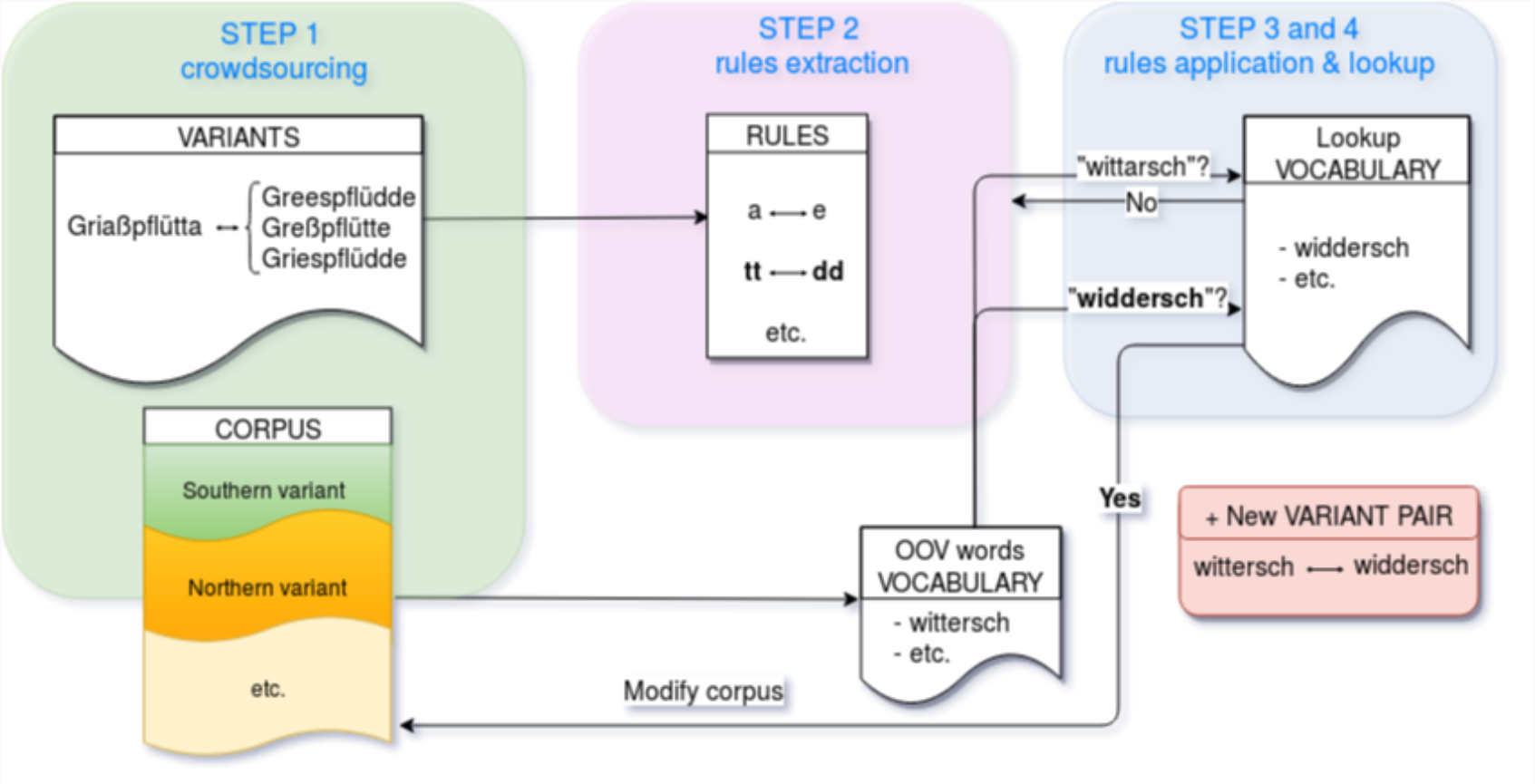
- 213 L+R rules
- 227 L rules
- 186 R rules

rules' frequencies vary

Outline

Overview of the process

Step 3: Rules application and lookup



Rules application

given:

- a vocabulary of known words V_{lookup}
- an unkown word $Word_{Unk}$ (size over 4 letters)

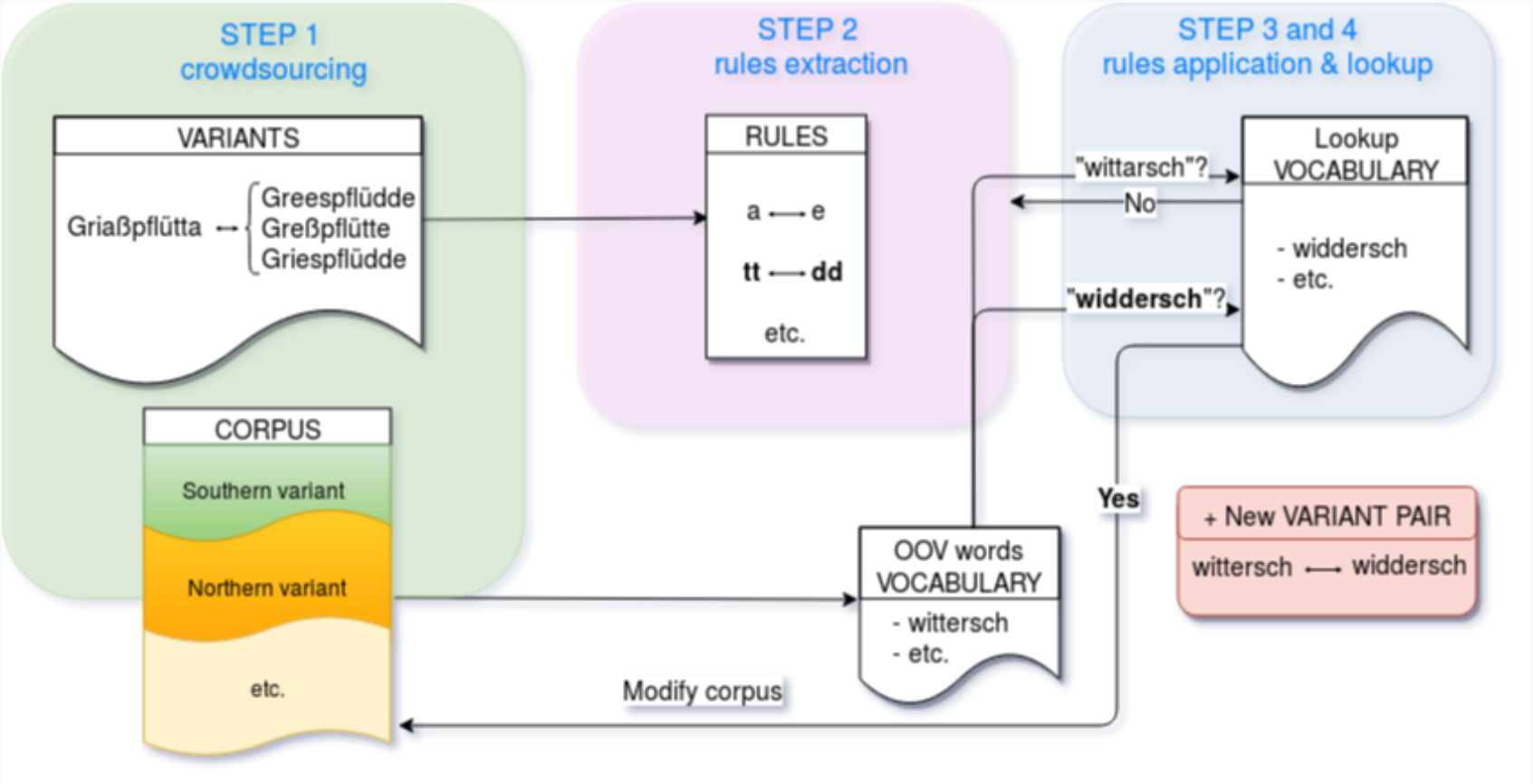
steps:

1. (optional) we filter $Word_{Unk}$ if it is a known proper
2. we select the rules that apply to $Word_{Unk}$: $\{R_{Word_{Unk}}\}$
3. we apply to $Word_{Unk}$ each *combination* of rules from $\{R_{Word_{Unk}}\}$

combination: given three rules A, B, C, the sequences of rules {A}, {B}, {C}, {A;B}, {A;C}, {B;C} and {A;B;C} are applied

a **brut force approach** that generates a list of potential variants for $Word_{Unk}$

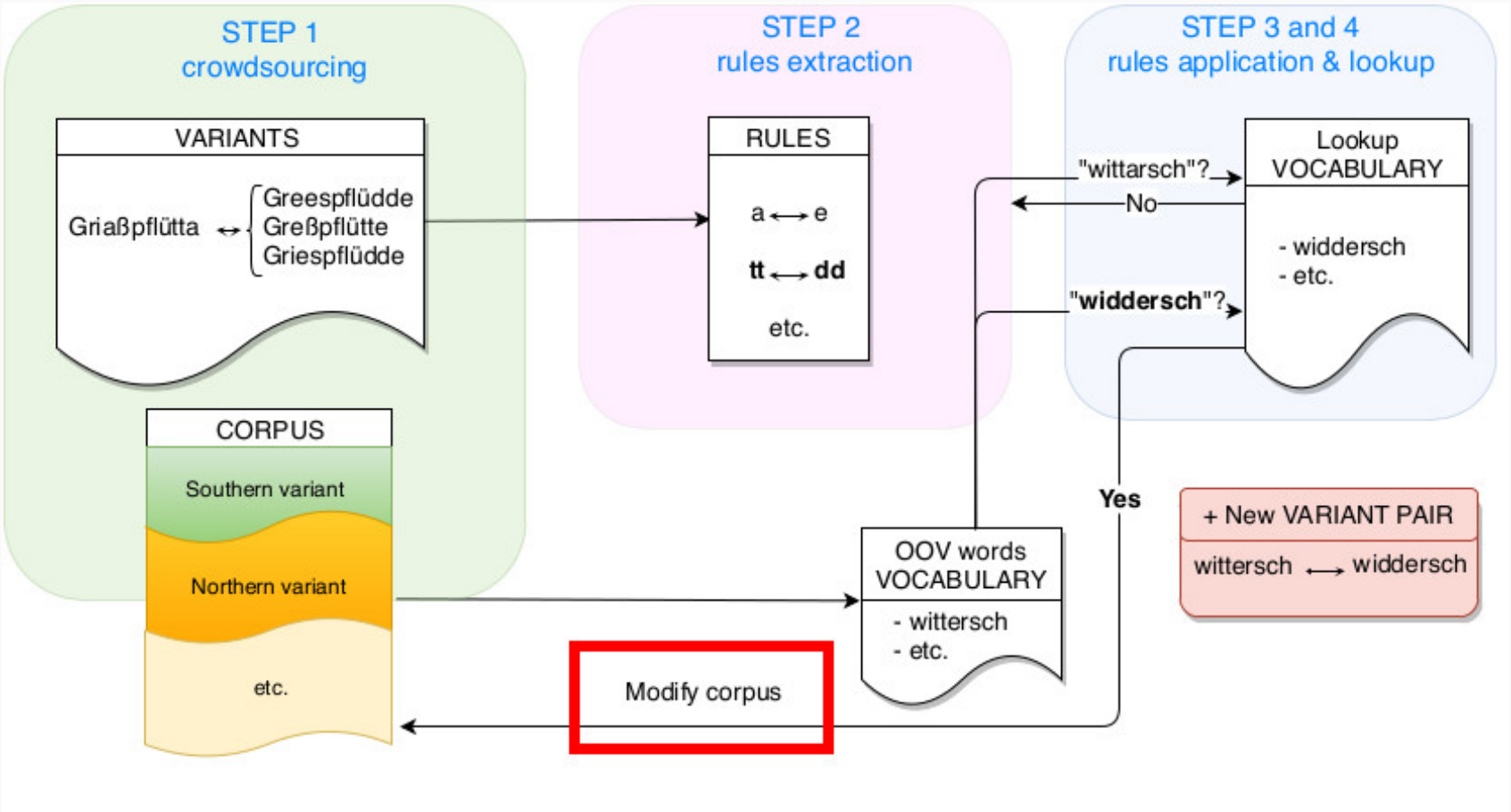
Lookup


STEP 1
crowdsourcing
VARIANTS
Griaßpflütta ⟷ Greespflüdde
Greßpflütte
Griespflüdde
CORPUS
Southern variant
Northern variant
etc.
STEP 2
rules extraction
RULES
a ⟷ e
tt ⟷ dd
etc.
STEP 3 and 4
rules application & lookup
Lookup VOCABULARY
- widdersch
- etc.
"wittarsch"?
No
"widdersch"?
Yes
+ New VARIANT PAIR
wittersch ⟷ widdersch
OOV words VOCABULARY
- wittersch
- etc.
Modify corpus

Outline

Objective

Match OOV words $Word_{Unk}$ with one of their known spelling variants



Objective

Match OOV words $Word_{Unk}$ with one of their known spelling variants

Available pos tagged corpora (total: **21,852 tokens**):

- `Crowdsourced Corpus` $C_{rowd}C$ [Millour and Fort, 2018]
- `Annotated Corpus for the Alsatian Dialects` $T_{rad}C$ [Bernhard et al., 2018]:

1. variants are generated for the $Word_{Unk}$ of the evaluation corpus (20%)
2. training corpus (80%) is used as the V_{lookup}

+ when a potential variant is discovered, $Word_{Unk}$ is replaced (corpus transposition)

Setup 1: Homogeneous setup

both training (∼17,500 tokens) and evaluation (∼4,350 tokens) corpora are **multi-variant**:

	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.

- 56 new variant pairs descovered on average
- +0.5% accuracy

Setup 2: Heterogeneous corpus

training and evaluation corpora are **mono-variant**:

- Northern variant: 4,880 tokens
- Southern variant: 7,690 tokens

	$N_{orth}C20$		$S_{outh}C20$	
$N_{orth}C80$			Before transp.	After transp.
Overall	0.853		**0.714**	**0.752**
OOV words	21%		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.

Setup 2: Heterogeneous corpus

training and evaluation corpora are **mono-variant**:

- Northern variant: 4,880 tokens
- Southern variant: 7,690 tokens

	$N_{orth}C20$		$S_{outh}C20$	
$N_{orth}C80$			Before transp.	After transp.
Overall	0.853		**0.714**	**0.752**
OOV words	21%		54%	52%
$S_{outh}C80$	Before transp.	After transp.		
Overall	**0.788**	**0.809**	0.864	
OOV words	51%	48%	29%	

Table 4: Accuracy of the model trained on mono-variant corpora, before and after the corpus transposition.

- higher impact on heterogeneous corpora (+ 1 to 4%)
- confirms the **necessity** of integrating knowledge about variants

Conclusions

the efficiency of the methodology depends on:

- the respective and relative sizes of the training and evaluation corpora
- the variation in variants existing between them

overall:

the performance of a tagging 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

Outline

Obtained resource

876 additional pairs of variants were discovered
during the experiments

60 were examined by an Alsatian teacher:

- 30 are actual variants
- 10 are erroneous matching we managed to correct (forcing case match and size over 4 letters)
- 13 are identical forms in different contexts (same POS), *e.g.*: *ìhm* (dative pronoun) / *ìrhem* (genitive pronoun), *kált* (feminine adjective) / *kálte* (masculine adjective), *wùrd* (future auxiliary) / *wärd* (conditionnal auxiliary)
- 7 are erroneous matching we were not yet able to correct *e.g. kräfti* (“strongly”, adverb) / *kräftiger* (“stronger”, adjective), *mine* (“mine”, determiner) / *meine* (“believe”, verb) etc.

Conclusions

the method:

- leads to **reduction of OOV proportion** hence improvement of POS tagging performances
- is **language independent** (currently adapted to Mauritian creole)
- feeds from being applied to unkown corpora
- is based on resources easy to produce by non expert speakers

Limitations

- the cost in time is high
- variation rules are hard to distinguish from morphological rules
- dialectal and spelling variations are uneasy to entangle

Thank you!

Vielmols merci!

Questions, comments?

References i


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


Bernhard, D., Erhart, P., Huck, D., and Steiblé, L. (2018).
Annotated corpus for the alsatian dialects.
Guide d'annotation, LiLPa, Université de Strasbourg.

References ii

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

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

References iii


Millour, A. and Fort, K. (2018).
Toward a Lightweight Solution for Less-resourced Languages: Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing.
In *11th International Conference on Language Resources and Evaluation (LREC'18)*, Miyazaki,.


Millour, A. and Fort, K. (2019).
À l'écoute des locuteurs : production participative de ressources langagières pour des langues non standardisées.

References iv

In *Revue TAL : numéro spécial sur les langues peu dotées (59-3)*. Association pour le Traitement Automatique des Langues.

Petrov, S., Das, D., and McDonald, R. (2012).
A universal part-of-speech tagset.
In *Actes de Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turquie. European Language Resources Association (ELRA).