Adding Dialectal Lexicalisations to Linked Open Data Resources

The Example of Alsatian

Delphine Bernhard

May 26 2014

CCURL 2014

Workshop

UNIVERSITÉ DE STRASBOURG

lilpa
linguistique, langues, parole

Outline

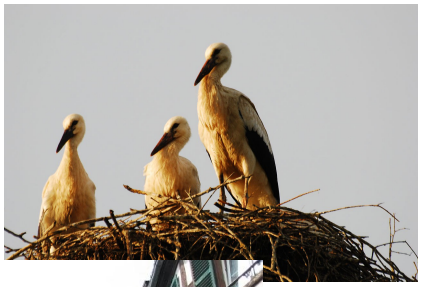
The Alsace Region

	Iceland	Alsace
Area	103,001 km2	8,280 km2
Population	325,671	1,815,488

Source : Wikipedia

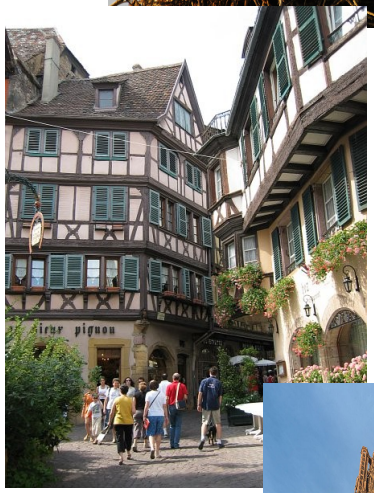
Some clichés about Alsace

Storik



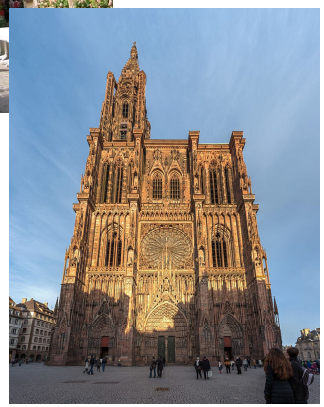
Flàmmeküeche

Köjelhopf



Fàchwarkhüss

Minschter



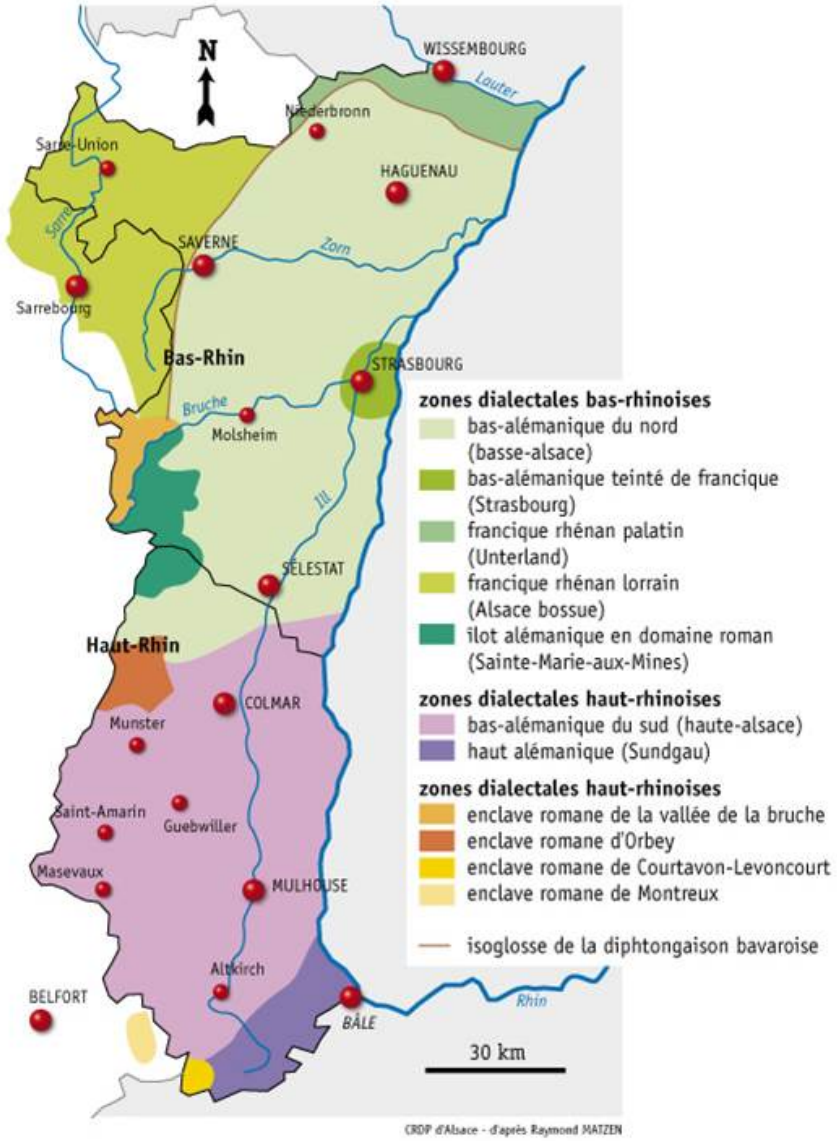
Brattschtall

Sürkrüt

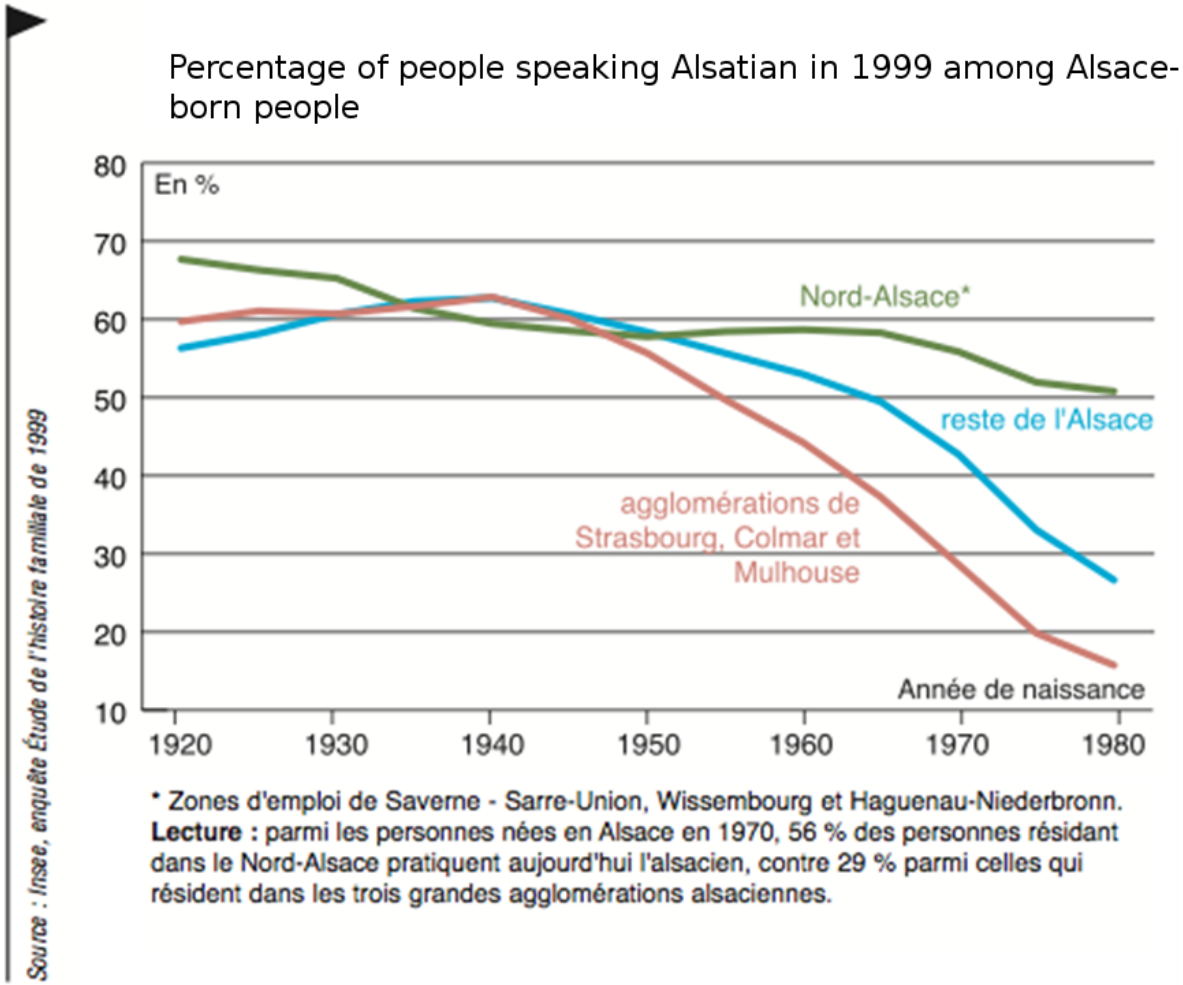
Source : Wikipedia

Linguistic Situation in Alsace

- Official language : French
- Germanic (Alemannic and Franconian) dialects spoken since the fifth century AD
- Decreasing influence of German

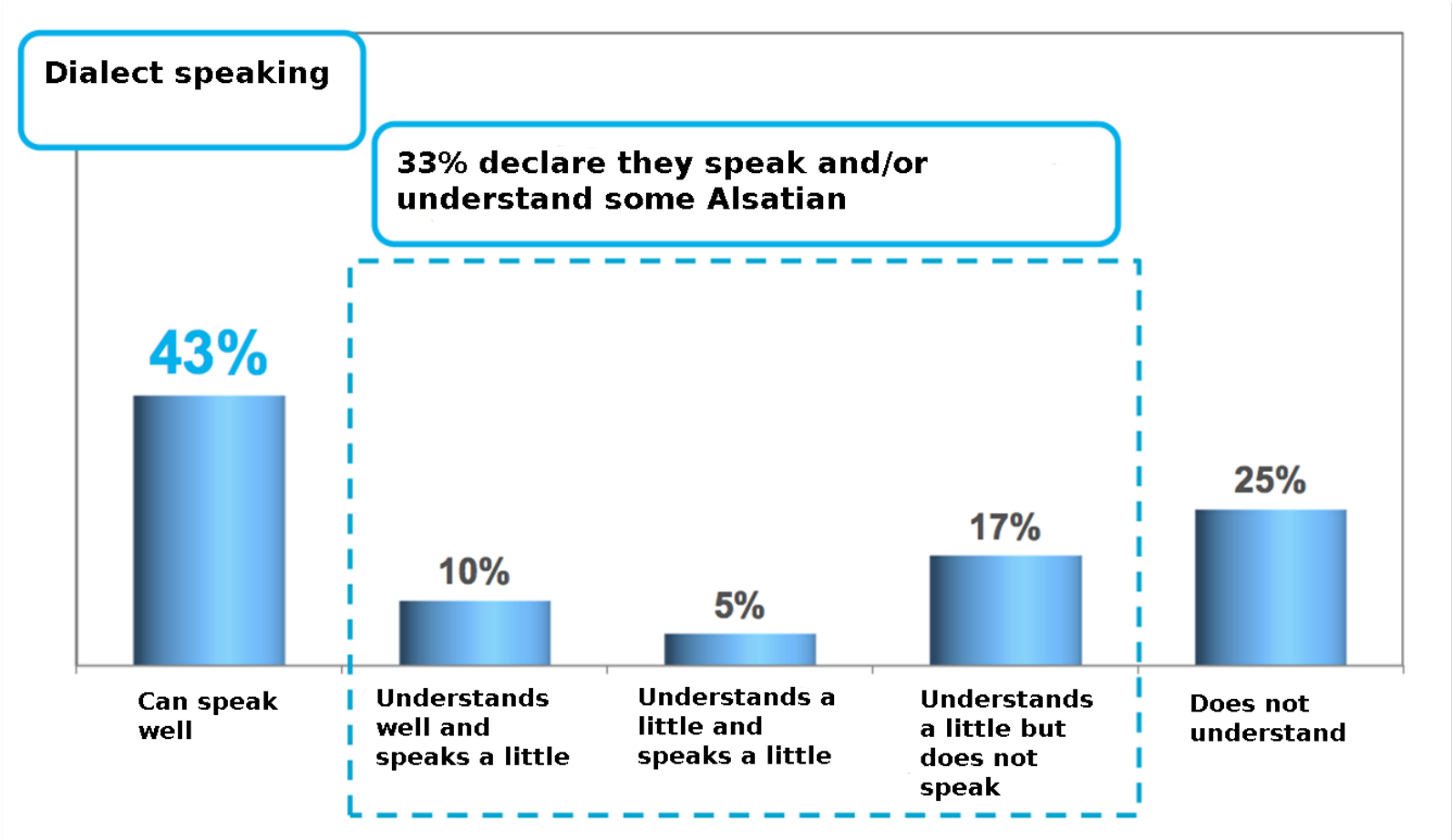


How many people speak Alsatian ?



http://www.insee.fr/fr/insee_regions/alsace/themes/cpar12_1.pdf

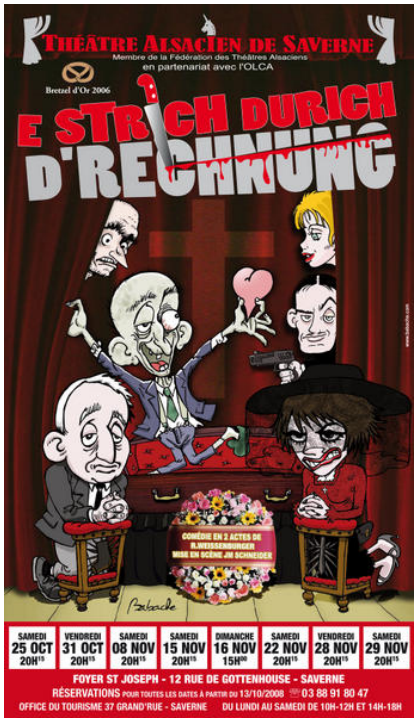
How many people speak Alsatian ?



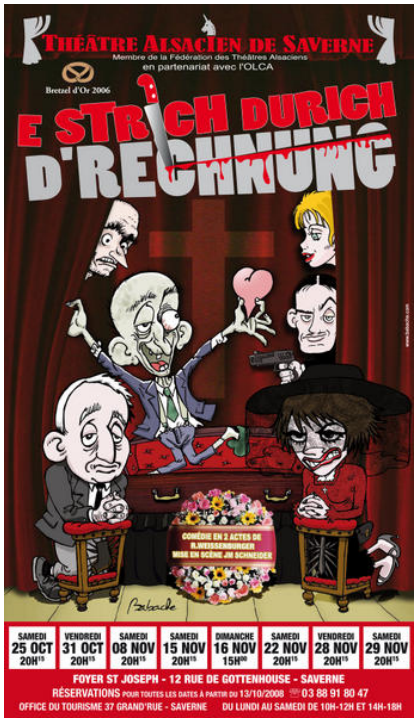
http://www.olcalsace.org/sites/default/files/documents/etude_linguistique_olca_edinstitut.pdf

Writing Alsatian

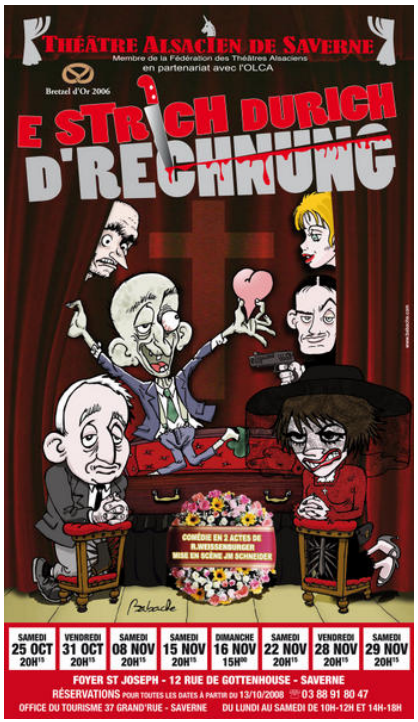
Writing Alsatian

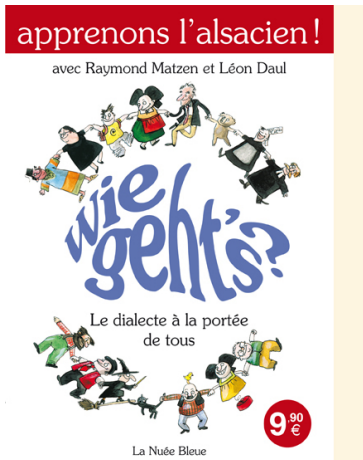


Writing Alsatian

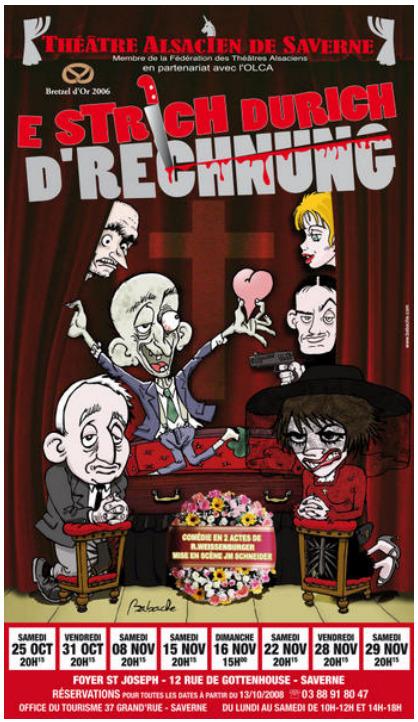


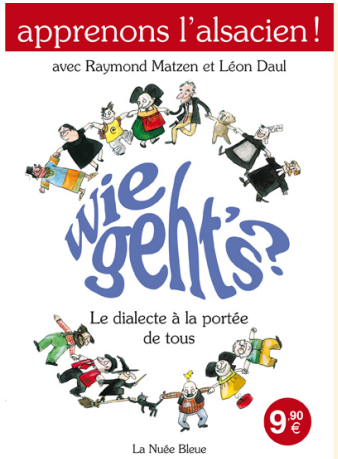
Writing Alsatian





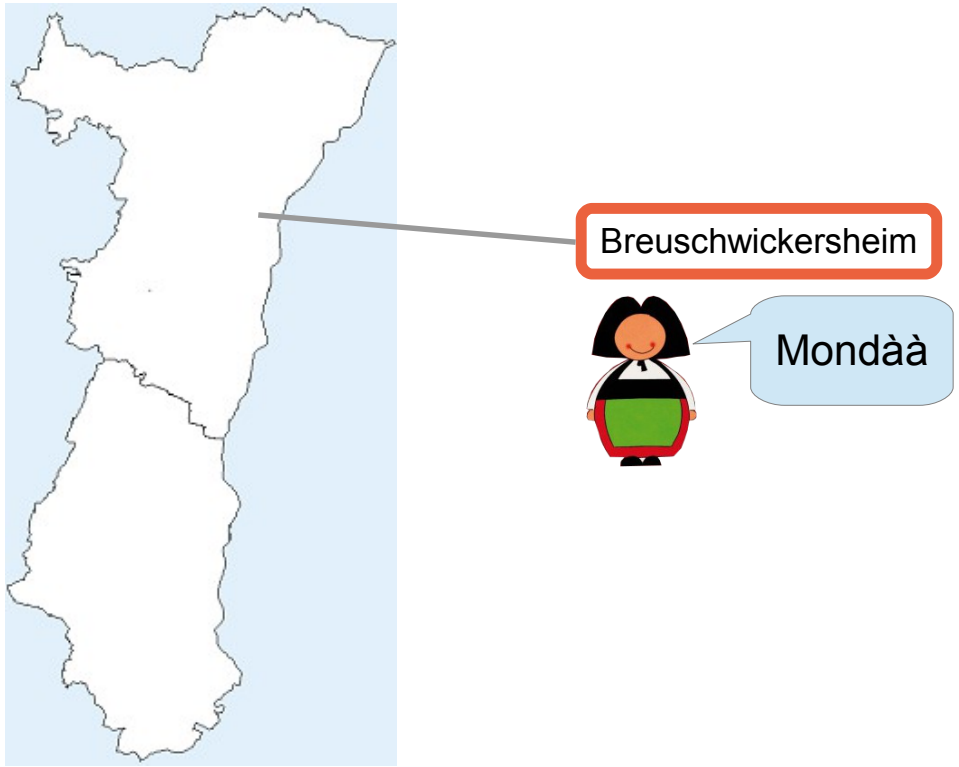
Writing Alsatian



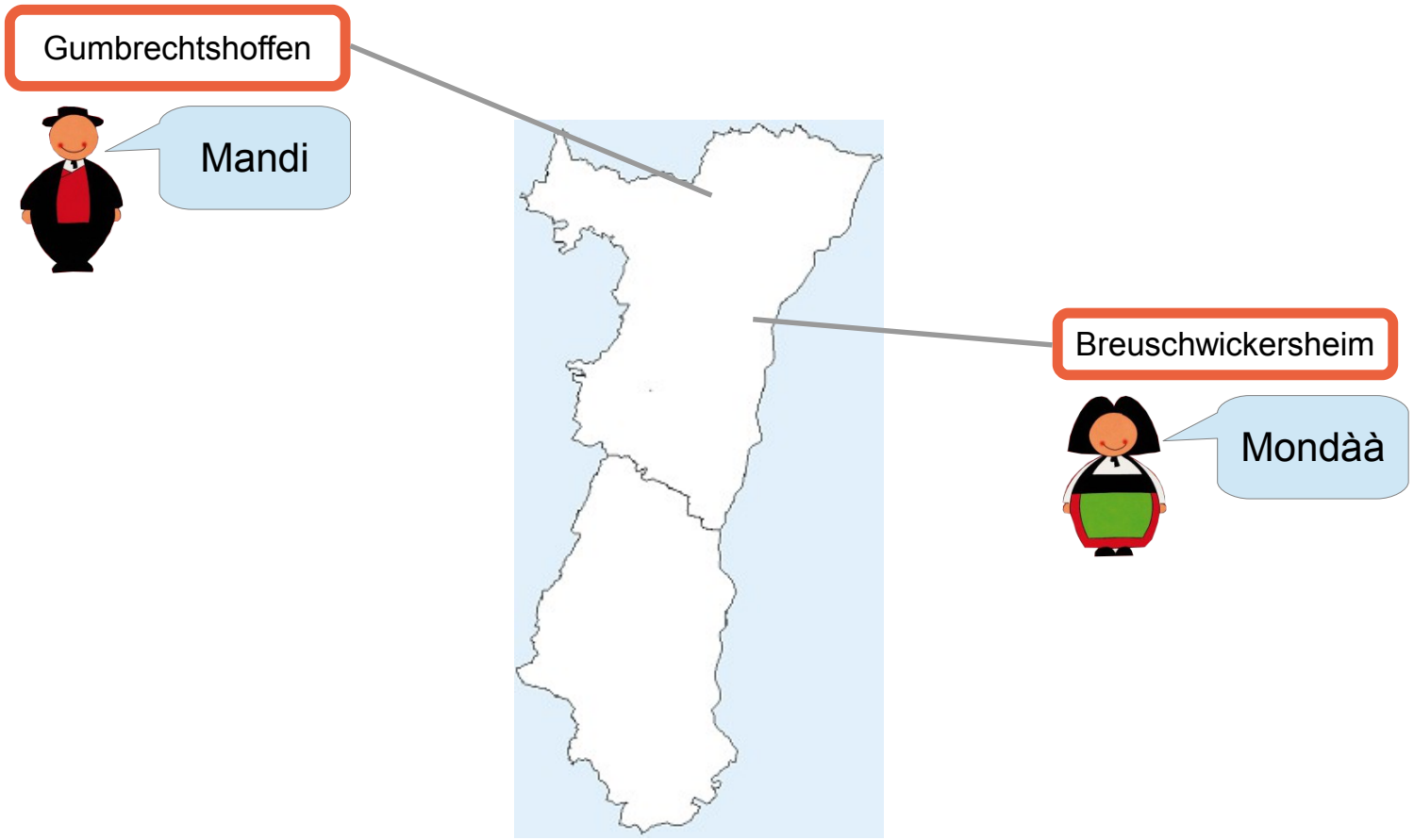


Monday in Alsatian...

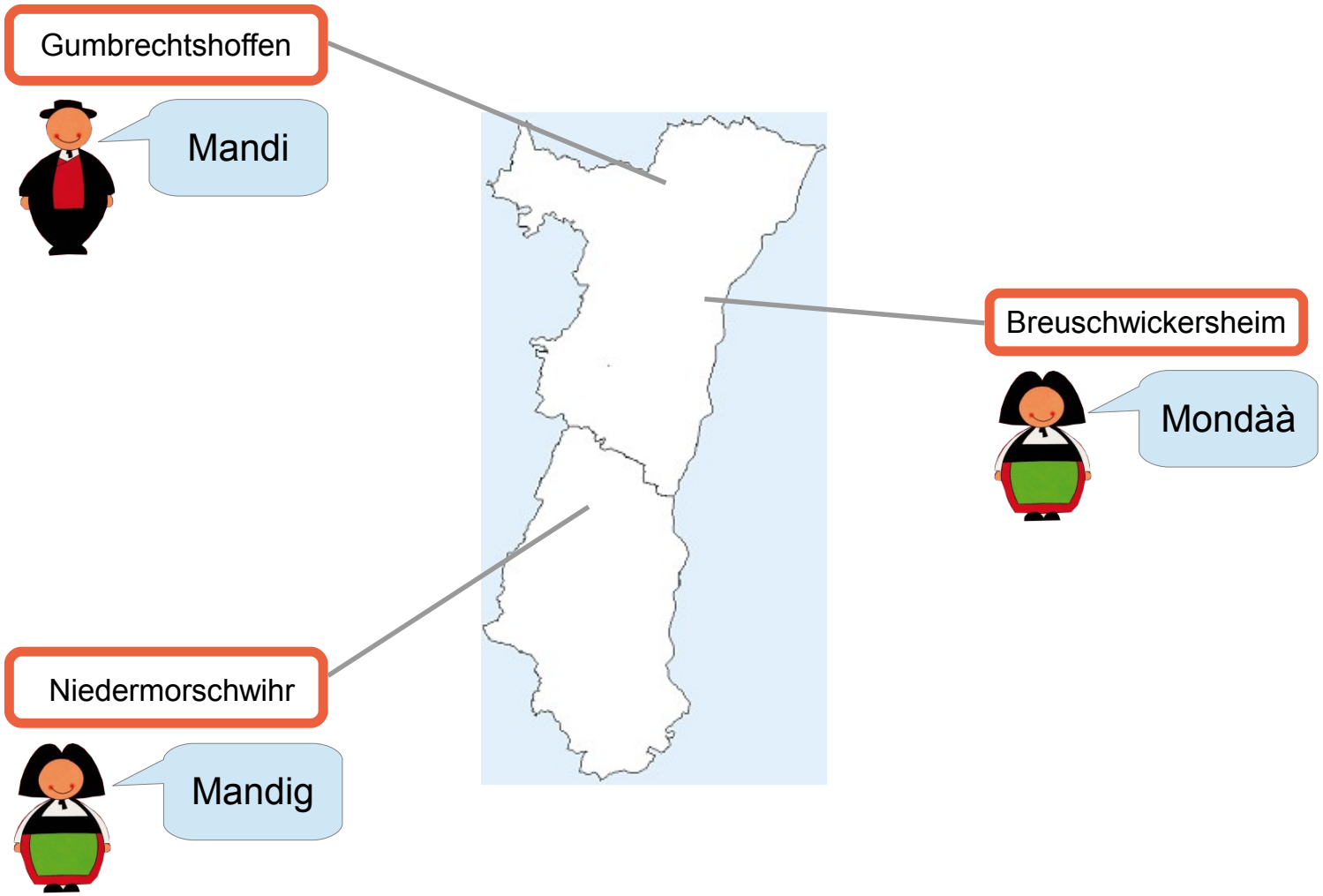
Monday in Alsatian...



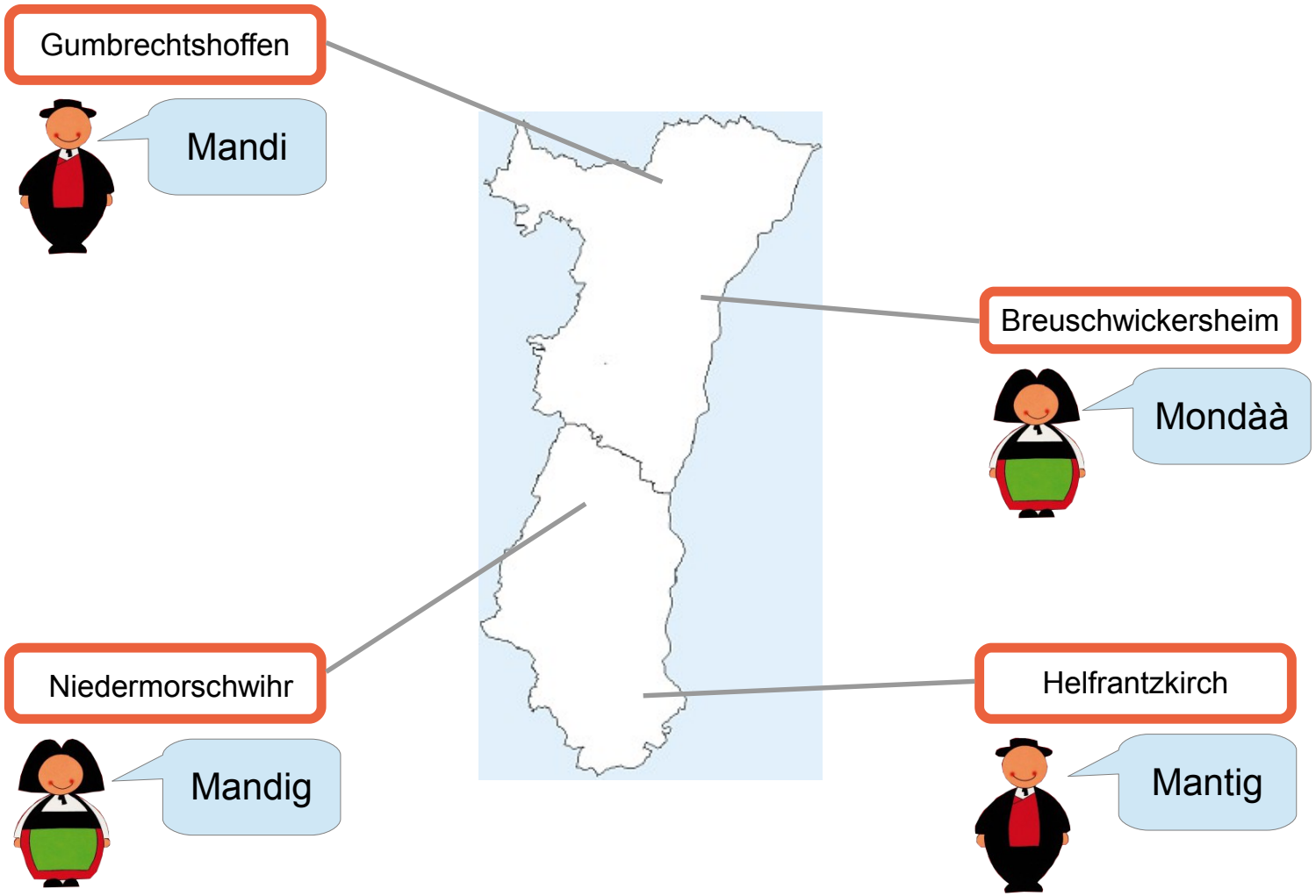
Monday in Alsatian...



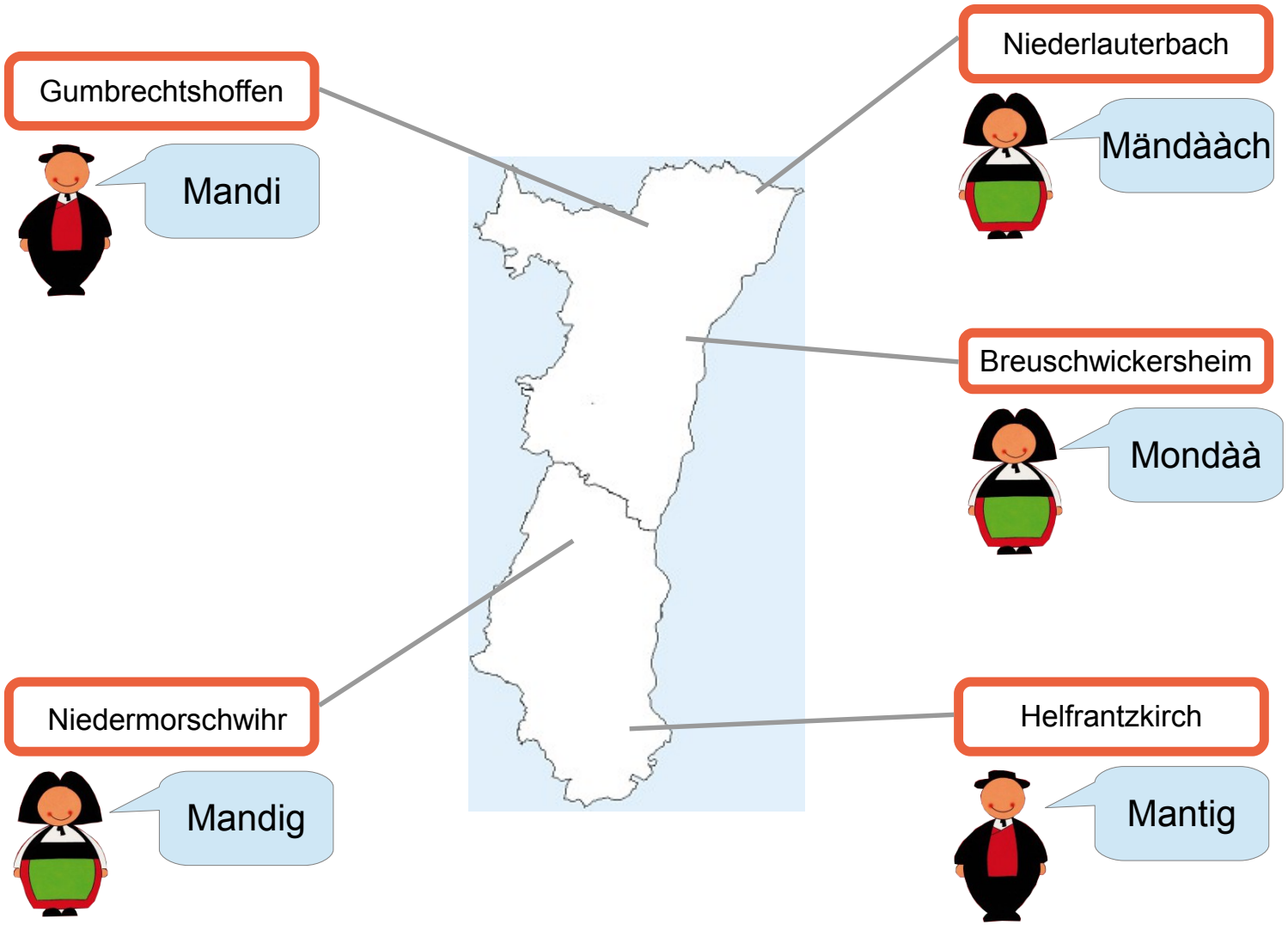
Monday in Alsatian...



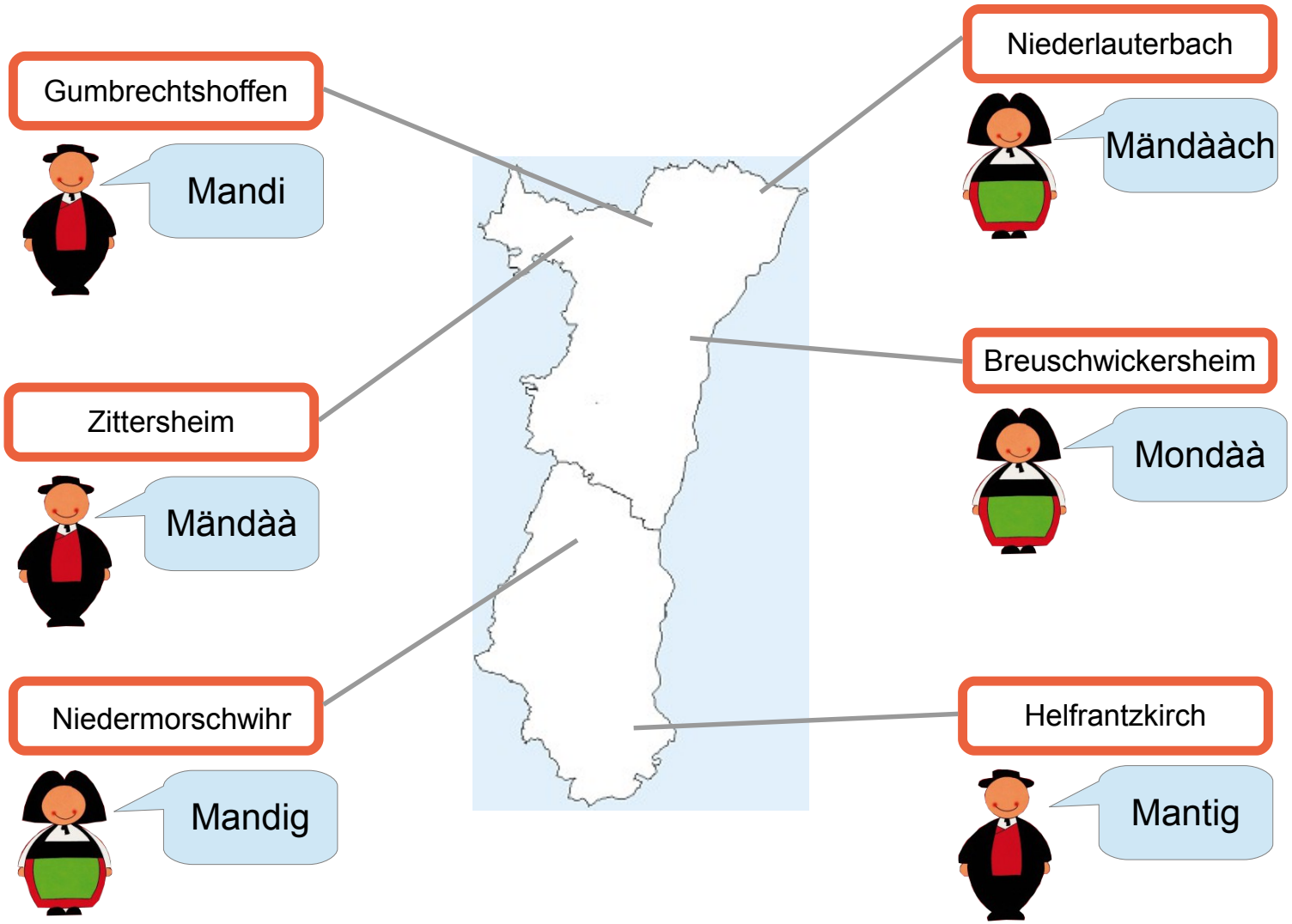
Monday in Alsatian...



Monday in Alsatian...



Monday in Alsatian...



Challenges and objectives

Challenges for Alsatian

- Geolinguistic variants
- Lack of standard spelling
- Low-resourced language

Objectives

- Align variants of the same lexeme found in several lexicons
- Build a unified and aligned resource
- Map the aligned variants to a multilingual linked open data resource : BabelNet

Outline

Source bilingual French-Alsatian lexicons

- OLCA : domain-specific lexicons produced by the OLCA (*Office pour la Langue et la Culture d'Alsace*). These lexicons provide variants for the Bas-Rhin (Lower Rhine : OLCA-67) and Haut-Rhin (Upper Rhine : OLCA-68) Alsatian departments.
- WKT : a lexicon retrieved from a Wiktionary user page ;
- ACPA : a bilingual lexicon authored by André Nisslé from Association Culture et Patrimoine d'Alsace.

Examples from the source lexicons

French	**corbeau**	**jambe(s)**	**grenier**
English	crow	leg	attic
German	Rabe	Bein	Dachboden
ACPA	Kräje Kràbb	Bai Unterschankel	Behna **Behn** Ästrich Dàchbooda
WKT	Grâb Kràpp **Ràmm**	**Bein** **Baan**	**Behn** Behni Bhena Kàscht Späicher Spicher
OLCA	Kràb **Ràmm**	**Bein** Bei **Baan**	

Examples from the source lexicons

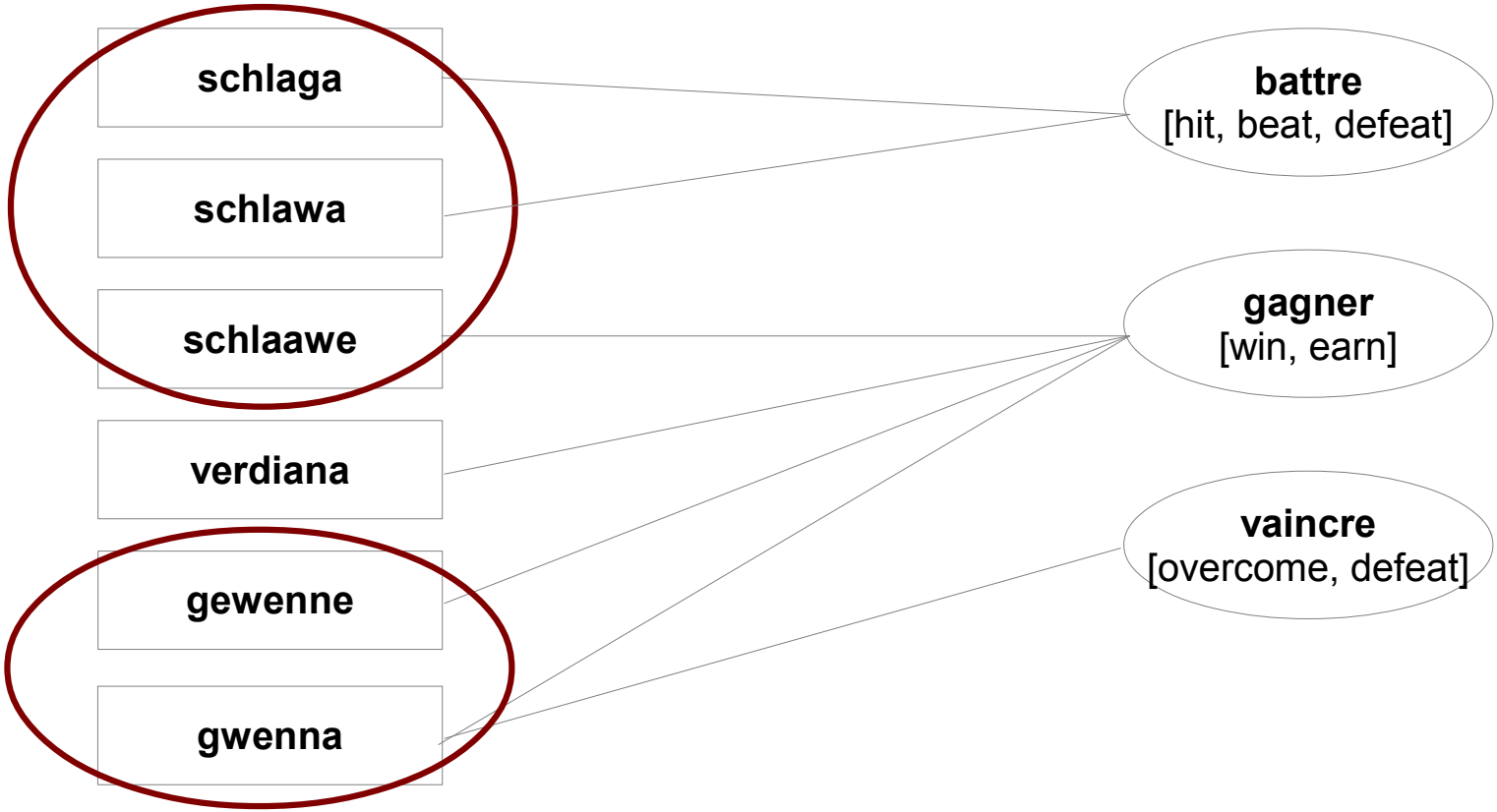
French	**corbeau**	**jambe(s)**	**grenier**
English	crow	leg	attic
German	Rabe	Bein	Dachboden
ACPA	Kräje Kràbb	Bai Unterschankel	Behna **Behn** Ästrich Dàchbooda
WKT	Grâb Kràpp **Ràmm**	**Bein** **Baan**	**Behn** Behni Bhena Kàscht Späicher Spicher
OLCA	Kràb **Ràmm**	**Bein** Bei **Baan**	

Examples from the source lexicons

French	**corbeau**	**jambe(s)**	**grenier**
English	crow	leg	attic
German	Rabe	Bein	Dachboden
ACPA	Kräje Kràbb	Bai Unterschankel	Behna **Behn** Ästrich Dàchbooda
WKT	Grâb Kràpp **Ràmm**	**Bein** **Baan**	**Behn** Behni Bhena Kàscht Späicher Spicher
OLCA	Kràb **Ràmm**	**Bein** Bei **Baan**	

Lexicon alignment

Challenge : double ambiguity



Identifying variants

Edit distance ?

	ε	M	a	n	d	i	g
ε	0	1	2	3	4	5	6
M	1	0	1	2	3	4	5
a	2	1	0	1	2	3	4
n	3	2	1	0	1	2	3
t	4	3	2	1	1	2	3
i	5	4	3	2	2	1	2
g	6	5	4	3	3	2	**1**

It could work...

`http://www.kurzhals.info/static/samples/levenshtein_distance/`

Identifying variants

Edit distance ?

	ε	M	a	n	d	i	g
ε	0	1	2	3	4	5	6
M	1	0	1	2	3	4	5
a	2	1	0	1	2	3	4
n	3	2	1	0	1	2	3
t	4	3	2	1	1	2	3
i	5	4	3	2	2	1	2
g	6	5	4	3	3	2	**1**

	ε	M	o	n	d	à	à
ε	0	1	2	3	4	5	6
M	1	0	1	2	3	4	5
a	2	1	1	2	3	4	5
n	3	2	2	1	2	3	4
t	4	3	3	2	2	3	4
i	5	4	4	3	3	3	4
g	6	5	5	4	4	4	**4**

It could work...

...but only for simple cases

http://www.kurzhals.info/static/samples/levenshtein_distance/

Identifying variants
Double metaphone

- Originally proposed by [Phillips, 2000] for information retrieval in English ;
- Transforms the input string into one or two keys which are identical for words which are pronounced in a similar manner

Alsatian	**French**	**English**	**Key 1**	**Key 2**
Schloofwàga	wagon-lit	sleeping car	XLFVK	XLFVY
Schlofwaawe			XLFVV	XLFVY
Rüejdàà	jour de repos	rest day	RT	/
Rüaijtààg			RTK	RT
beschtadiga	confirmer	confirm	PXTTK	PXTTY
Uffschtànd	insurrection	insurrection	AFXTNT	/

Identifying variants
Double metaphone

- Originally proposed by [Phillips, 2000] for information retrieval in English ;
- Transforms the input string into one or two keys which are identical for words which are pronounced in a similar manner

Alsatian	French	English	Key 1	Key 2
Schloofwàga	wagon-lit	sleeping car	XLFVK	XLFVY
Schlofwaawe			XLFVV	XLFVY
Rüejdàà	jour de repos	rest day	**RT**	/
Rüaijtààg			RTK	**RT**
beschtadiga	confirmer	confirm	PXTTK	PXTTY
Uffschtànd	insurrection	insurrection	AFXTNT	/

Lexicon alignment

1. All entries in the input lexicons are added to a large graph : the nodes correspond to Alsatian words and their French translations ;
2. Alsatian words are connected to their French translations in the lexicons by an edge ;
3. Two Alsatian words are connected by an edge if all of the following conditions are met :
    1. they have the same French translation ;
    2. they share one of their double metaphone keys ;
    3. they have the same part-of-speech.

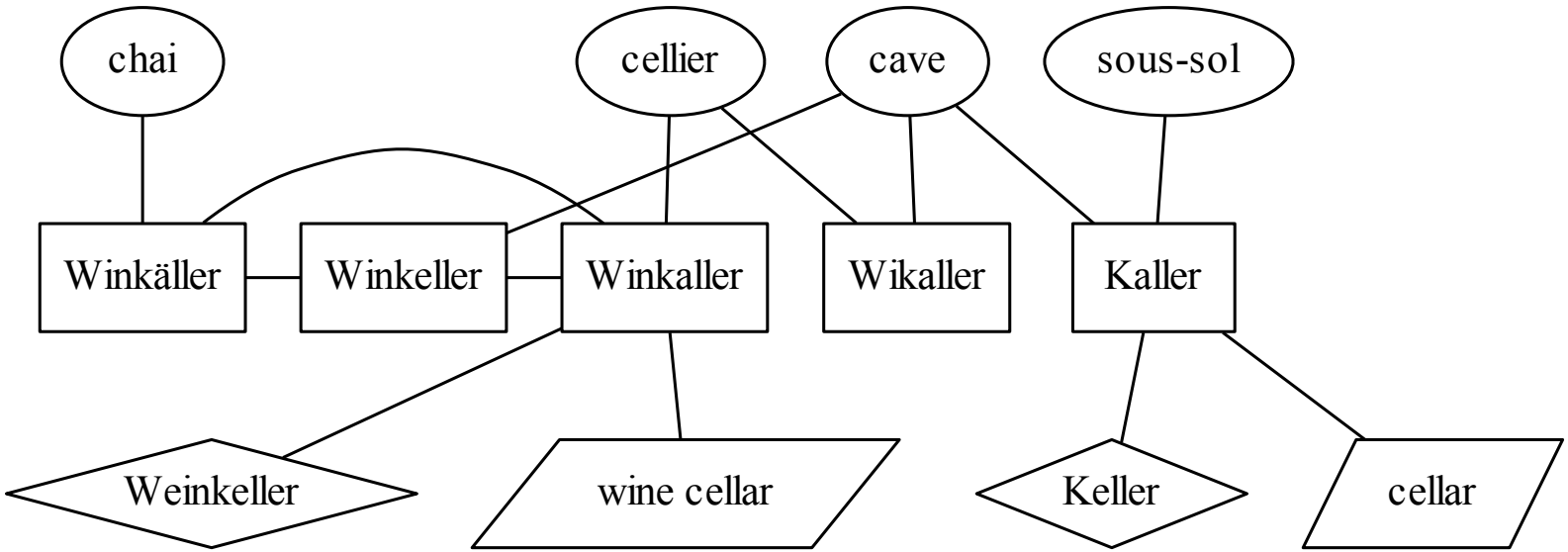
Additional resources

- Synonyms from JeuxDeMots [Lafourcade, 2007], a freely available French lexical network built through crowdsourcing games.
- BabelNet [Navigli and Ponzetto, 2012], a multilingual semantic network, which integrates knowledge from WordNet and Wikipedia.

$\Rightarrow$ add edges between Alsatian words which have synonymous French translations in these resources.

Alignment of Alsatian Variants

Detection of connected components in the graph formed by Alsatian word forms :



Outline

Example synset in BabelNet

Mapping method

- Calculate the cosine similarity between binary bag-of-words representations of Babel synsets and clusters of aligned Alsatian variants.
- Bag-of-words representations :
  1. French lexicalisations for Babel synsets – French translations for Alsatian variant clusters
     ["Winkäller", "Winkeller", "Winkaller"] → ["chai", "cellier", "cave"]
  2. French + German / English lexicalisations for Babel synsets – French translations for Alsatian variant clusters + cognate German / English translations
     ["Winkäller", "Winkeller", "Winkaller"] → ["chai", "cellier", "cave", "Weinkeller", "wine cellar"]

Outline

Evaluation methodology

- 100 manual alignments between the lexicons and BabelNet ;
- based on randomly selected entries from a multilingual French-German-Alsatian-English dictionary [Adolf, 2006] ;
- evaluation in terms of precision, recall, F-measure and the proportion of correct mappings.

Evaluation results

	Lexicon alignments			Mapping to BabeNet		
	P	**R**	**F**	**top 1**	**top 2**	**top 3**
baseline	1.00	0.69	0.82	0.52	0.83	0.88
+ BN FR	0.98	0.71	0.83	0.56	0.85	0.89
+ JDM	1.00	0.71	0.83	0.52	0.80	0.86
+ BN FR & DE	0.98	0.71	0.83	0.72	**0.90**	**0.94**
+ BN FR & EN	0.98	0.71	0.83	0.63	0.83	0.91
+ BN FR, DE & EN	0.98	0.71	0.83	**0.76**	0.87	0.93
+ JDM + BN FR & DE	0.98	0.72	0.83	0.71	**0.90**	0.93
+ JDM + BN FR, DE & EN	0.98	0.72	0.83	0.75	0.87	0.92

Remaining issues

- Different metaphone keys, e.g. "Chilche" - `KLX / XLX` vs. "Kirche" - `KRX`
- Erroneous POS tags
- Coverage of the additional resources (JeuxDeMots and BabelNet)

Conclusion and perspectives

- Method to both align spelling variants of the same Alsatian lexeme found in several lexicons and map the variants to synsets in BabelNet
- Could in principle be applied to many less-resourced languages, as the only needed resource is a bilingual lexicon.
- In the future, provide the aligned lexicon in a standard format and use it in NLP applications.

Merci vielmols !

References


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


Lafourcade, M.
Making people play for Lexical Acquisition.
In *Proceedings of SNLP 2007, 7th Symposium on Natural Language Processing*, Pattaya, Thailande, 2007.


Navigli, R. and Ponzetto, S. P.
BabelNet : the automatic construction, evaluation and application of a wide-coverage multilingual semantic network.
Artificial Intelligence, 193 :217–250, December, 2012.


Phillips, L.
The Double Metaphone Search Algorithm.
C/C++ Users Journal, 2000.