

# Adding Dialectal Lexicalisations to Linked Open Data Resources

The Example of Alsatian

Delphine Bernhard

May 26 2014

CCURL 2014

Workshop



# Outline

- 1 The Alsatian Dialect(s)
- 2 Alignment of dialectal variants
- 3 Mapping to BabelNet Synsets
- 4 Evaluation

# The Alsace Region



Source : Wikipedia

# Some clichés about Alsace

Storik



Flâmmeküeché



Köjelhopf



Minschter



Fächwarkhüss



Brattschtall

Sürkrüt



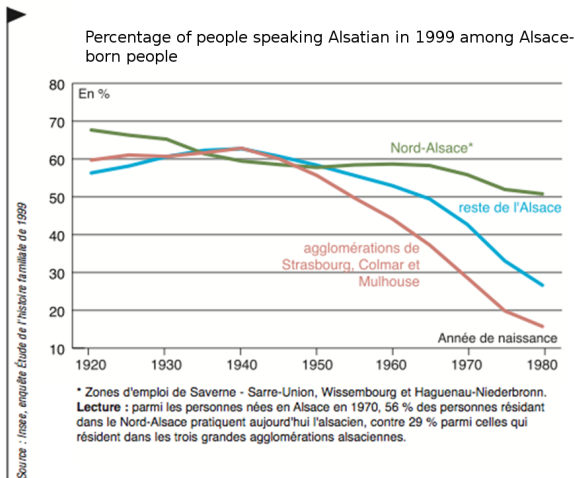
Source : Wikipedia

# Linguistic Situation in Alsace

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

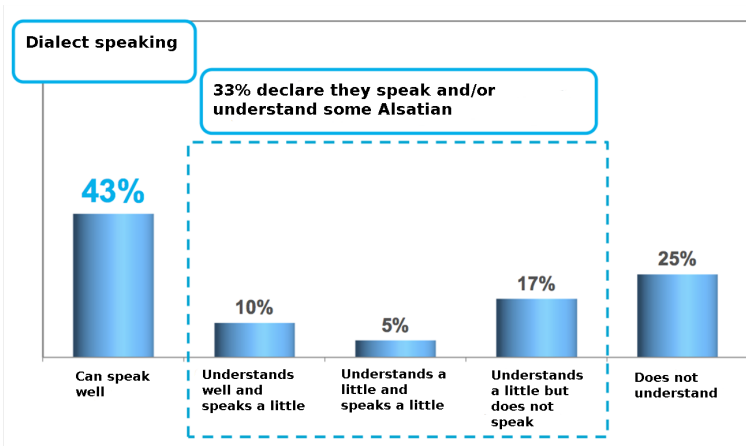


# How many people speak Alsatian ?



[http://www.insee.fr/fr/insee\\_regions/alsace/themes/cpar12\\_1.pdf](http://www.insee.fr/fr/insee_regions/alsace/themes/cpar12_1.pdf)

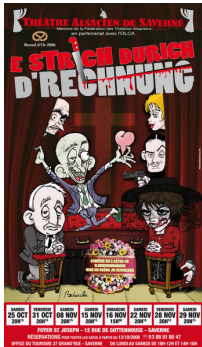
# How many people speak Alsatian ?



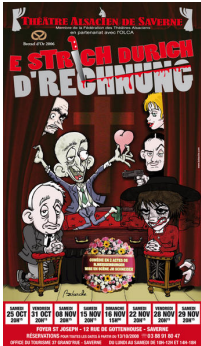
[http://www.olcalsace.org/sites/default/files/documents/etude\\_linguistique\\_olca\\_edinstitut.pdf](http://www.olcalsace.org/sites/default/files/documents/etude_linguistique_olca_edinstitut.pdf)

# Writing Alsatian

# Writing Alsatian



# Writing Alsatian

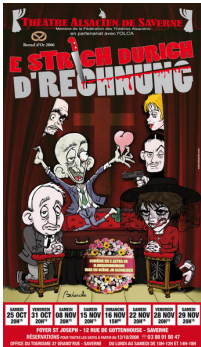


# Writing Alsatian

## 's MIMI ÛN DE LEO GEHN ÌN D'SCHUEL

Mimi et Léo  
vont à l'école

Mimi und Leo  
gehen zur Schule



apprenons l'alsacien!

avec Raymond Matzen et Léon Daul



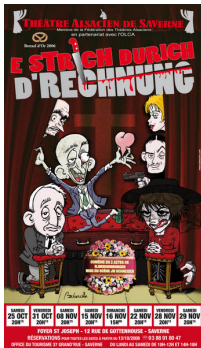
C'est chic  
de parler alsacien...  
...et surtout  
si facile!

avec Raymond Matzen et Léon Daul  
les deux auteurs spécialistes de la transmission du dialecte  
DIALOGUES • VOCABULAIRE DE BASE • GRAMMAIRE SIMPLIFIÉE  
• 20 fiches et 100 schémas de la vie quotidienne  
• 1 000 mots usuels  
• 200 proverbes et dictons  
• deux lexiques (français-alsacien et alsacien-français)  
• conjugaisons, déclinaisons et règles grammaticales essentielles

La meilleure  
des méthodes pour  
apprendre le dialecte



# Writing Alsatian



apprenons l'alsacien !

avec Raymond Matzen et Léon Daul



C'est chic  
de parler alsacien...  
...et surtout  
si facile !



avec Raymond Matzen et Léon Daul  
les deux meilleurs spécialistes de la transmission du dialecte  
DIALECTES - VOCABULAIRE DE BASE - GRAMMAIRE SIMPLIFIÉE :  
• 20 leçons et 100 schémas de la vie quotidienne  
• 1 000 mots usuels  
• 200 exercices et dictées  
• deux lexiques (français-alsacien et alsacien-français)  
• conjugaisons, déclinaisons et règles grammaticales essentielles

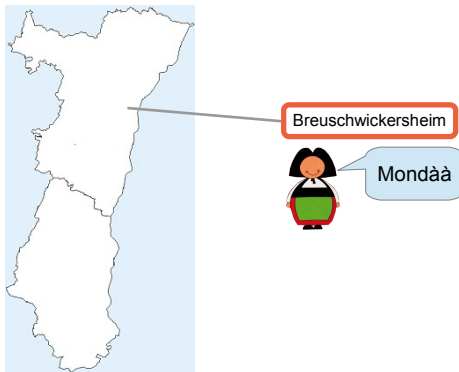
La meilleure  
des méthodes pour  
apprendre le dialecte



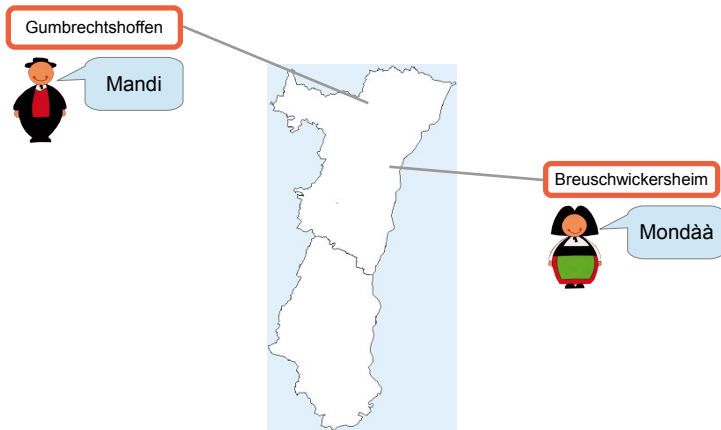
# Monday in Alsatian...



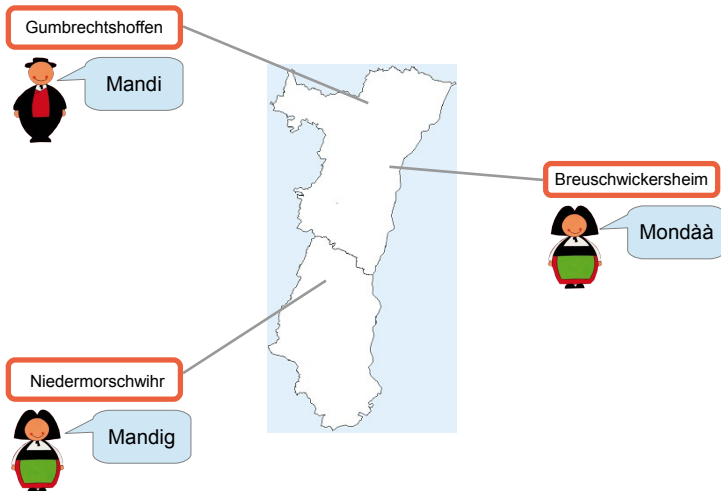
# Monday in Alsatian...



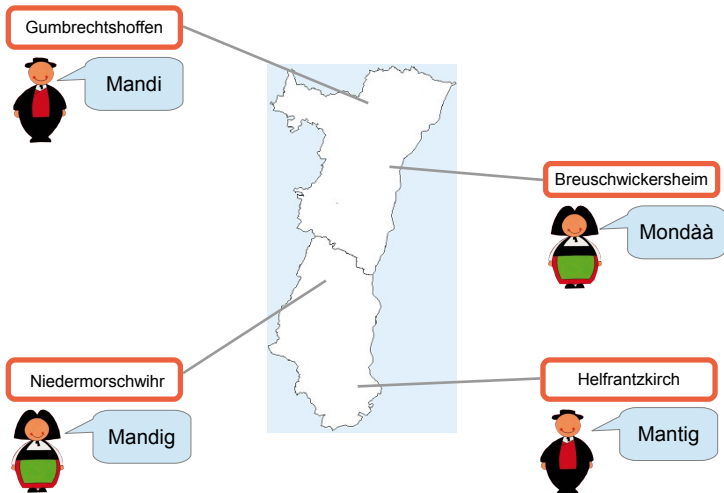
# Monday in Alsatian...



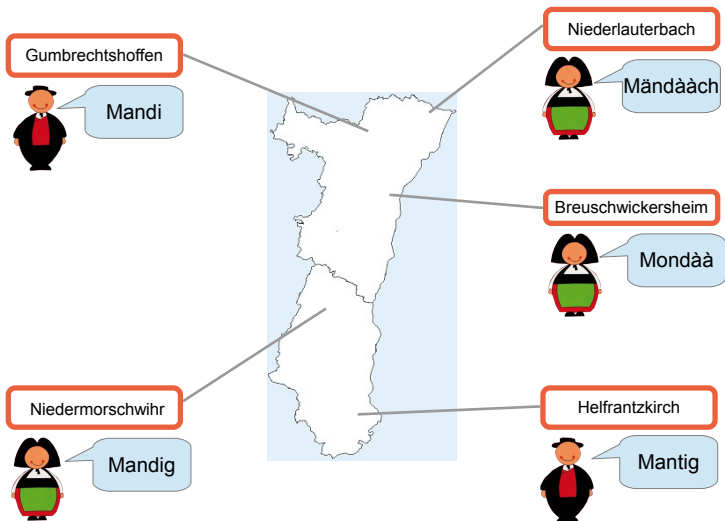
# Monday in Alsatian...



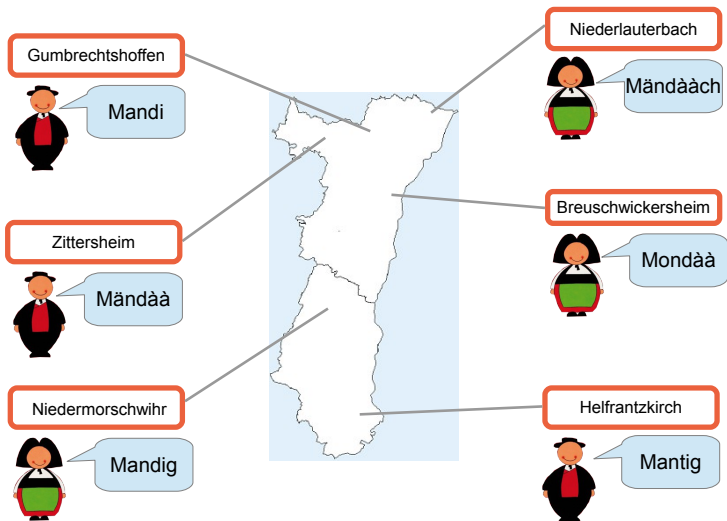
# Monday in Alsatian...



# 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

- 1 The Alsatian Dialect(s)
- 2 Alignment of dialectal variants
- 3 Mapping to BabelNet Synsets
- 4 Evaluation

# 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

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

# Examples from the source lexicons

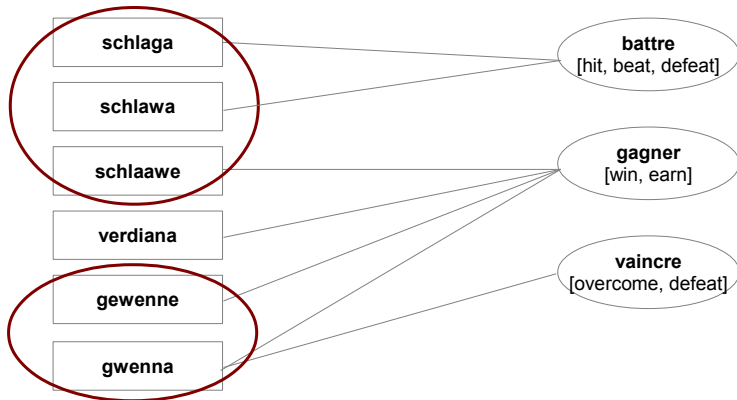
<b>French</b>	<b>corbeau</b>	<b>jambe(s)</b>	<b>grenier</b>
<b>English</b>	crow	leg	attic
<b>German</b>	Rabe	Bein	Dachboden
<b>ACPA</b>	Kräje Kràbb	Bai Unterschankel	Behna <b>Behn</b> Ästrich Dàchbooda
<b>WKT</b>	Grâb Kràpp Ràmm	<b>Bein</b> <b>Baan</b>	<b>Behn</b> Behni Bhena Kàscht Späicher Spicher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

# Examples from the source lexicons

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

# Lexicon alignment

## Challenge : double ambiguity



# Identifying variants

Edit distance ?

	<b>ε M a n d i g</b>						
<b>ε</b>	0	1	2	3	4	5	6
<b>M</b>	1	0	1	2	3	4	5
<b>a</b>	2	1	0	1	2	3	4
<b>n</b>	3	2	1	0	1	2	3
<b>t</b>	4	3	2	1	1	2	3
<b>i</b>	5	4	3	2	2	1	2
<b>g</b>	6	5	4	3	3	2	<b>1</b>

It could work...

[http://www.kurzhangs.info/static/samples/levenshtein\\_distance/](http://www.kurzhangs.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	<b>1</b>

	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	<b>4</b>

It could work...

...but only for simple cases

[http://www.kurzhaus.info/static/samples/levenshtein\\_distance/](http://www.kurzhaus.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	XLFBVK	XLFBVY
Schlofwaawe			XLFBVV	XLFBVY
Rüejdàà	jour de repos	rest day	RT	/
Rüaijtààg			RTK	RT
beschadiga	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
beschtdiga	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.

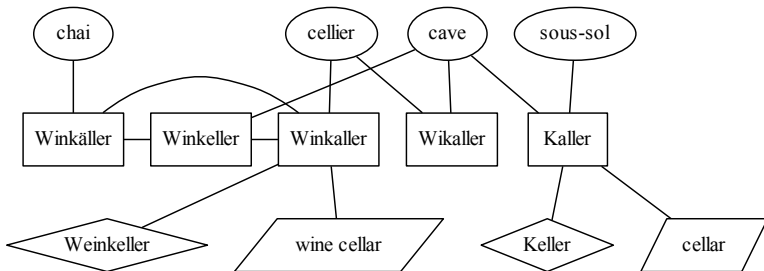
# Additional resources

- 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.

⇒ 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

- 1 The Alsatian Dialect(s)
- 2 Alignment of dialectal variants
- 3 Mapping to BabelNet Synsets**
- 4 Evaluation

# Example synset in BabelNet

<b>Meaning:</b> cellar <sup>3</sup> • <b>ID:</b> bn:00017041n • <b>Type:</b> Concept	
<b>Senses:</b>	
  cellar <sup>3</sup> , wine cellar <sup>1</sup>	
  Wine cellar,  قو (مخزن),  <b>Cave à vin</b> ,  Weinkeller,  Cantina,  Cava (bodega)	
  Wine closet, Winecellar, Wine cellars, Wine Cellar, Wine room,  Cave à vin réfrigérée, Cave a vin, Chai (Cellier), Armoire à vin, Stockage du vin, <b>Cave À Vin</b> ,  Bodeguero, Bodega de vino	
  wine cellar,  cellier, cave à vin,  Weinkeller,  cantina,  bodega	
  أقنية النبيذ, قو النبيذ,  葡萄酒窖,  caves à vin, <b>cave à vin</b> ,  weinkeller, größten weinkeller,  κελάρια κρασιού, κάβα,  מרתפי יין,  शराब cellars, शराब तहखाने,  cantine, cantina,  ワインセラー,  винных погребов, винный погреб,  bodegas, bodega	
  مخزن النبيذ,  酒窖,  <b>cave à vin</b> ,  weinkeller,  κάβα,  יקב,  शराब तहखाने,  cantina,  ワイン貯蔵室,  винный погреб,  bodega	

# 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

- 1 The Alsatian Dialect(s)
- 2 Alignment of dialectal variants
- 3 Mapping to BabelNet Synsets
- 4 Evaluation**

# 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	<b>0.90</b>	<b>0.94</b>
+ 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	<b>0.76</b>	0.87	0.93
+ JDM + BN FR & DE	0.98	0.72	0.83	0.71	<b>0.90</b>	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, Thaïlande, 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.