Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian

Delphine Bernhard

Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian

Delphine Bernhard
LiLPa - Linguistique, Langues, Parole
EA 1339, Université de Strasbourg
dbernhard@unistra.fr

Abstract

This article presents a method to align bilingual lexicons in a resource-poor dialect, namely Alsatian. One issue with Alsatian is that there is no standard and widely-acknowledged spelling convention and a lexeme may therefore have several different written variants. Our proposed method makes use of the double metaphone algorithm adapted to Alsatian in order to bridge the gap between different spellings. Once variant citation forms of the same lexeme have been aligned, they are mapped to BabelNet, a multilingual semantic network (Navigli and Ponzetto, 2012). The mapping relies on the French translations and on cognates for Alsatian words in the English and German languages.

Keywords: lexicon alignment, spelling variants, Alsatian

1. Introduction

Linked Open Data Resources have recently emerged as a new way to represent linguistic knowledge in many languages, by linking resources represented using standard formats. In practice, many of these resources are based either on existing word nets or on collaboratively built encyclopaedias or dictionaries such as Wikipedia or Wiktionary. As a consequence, not all languages are covered and even automatic approaches which acquire knowledge from e.g. Wikipedia or Wiktionary are not always usable because of the lack of information available for under-resourced languages.

In this article, we focus on a dialect, namely Alsatian, and propose to make use of resources which are more easily exploited and readily available, i.e. bilingual lexicons, to provide additional lexicalisations to existing linguistic linked open resources.

The Alsatian dialects are spoken in the Alsace region, located in the North-East of France. They belong to the Franconian and Alemannic language families (Huck et al., 2007). According to a recent study, 43% of the Alsatian population still speak the regional dialect (OLCA / EDinstitut, 2012). However, the proportion of Alsatian speakers is decreasing regularly since the 1960s, to the benefit of the French language. Moreover, the Alsatian dialects are mostly oral and there is no standard written norm.

There have been some initiatives aimed at defining spelling conventions. The ORTHAL system (Zeidler and Crévenat-Werner, 2008) refers to standard German spelling while allowing the transcription of phenomena which are specific to the Alsatian dialects. The GRAPHAL-GERIPA system (Hudlett and Groupe d'Etudes et de Recherches Interdisciplinaires sur le Plurilinguisme en Alsace et en Europe, 2003) defines a set of rules to go from sound to grapheme. However, it is difficult to estimate the actual dissemination and use of these systems. Moreover, they accommodate for the various geolinguistic variants encountered in Alsace and thus do not guarantee a unique spelling for the citation form of a given lexeme.[1]

To sum up, Alsatian dialects pose several important challenges for NLP:

- There is no standard and widely acknowledged spelling convention ;
- The Alsatian dialect is actually a continuum of dialects, with geographic lexical and pronunciation variants ;
- There are no large amounts of digital text corpora available.

In this article, we present a first step towards building digital lexical resources for the Alsatian dialects which consists in (i) aligning several bilingual French-Alsatian lexicons and (ii) mapping the Alsatian words to BabelNet, a multilingual semantic network which is connected to the Linguistic Linked Open Data cloud (Navigli and Ponzetto, 2012).

The proposed method relies on the following observations:

- The spelling conventions adopted in the French-Alsatian lexicons are very variable, and thus an Alsatian lexeme may have a different citation form in each lexicon, and even several different citation forms in a given lexicon, to accommodate for geolinguistic variants. Also, many of the Alsatian words are similar to their translation into standard German and even sometimes English.
- Different lexicon authors may choose different translations into French for a given Alsatian lexeme. This complicates the alignment, which cannot only rely on a simple mapping using French lemmas.

[1] We use *lexeme* in the sense given by Bauer (2003): "A lexeme is a dictionary word, an abstract unit of vocabulary. It is realised (...) by word-forms, in such a way that the word-form represents the lexeme and any inflectional endings (...) that are required. (...) The citation form of a lexeme is that word-form belonging to the lexeme which is conventionally chosen to name the lexeme in dictionaries and the like."

We address these issues as follows:

- We propose a variant of the double metaphone algorithm adapted to the Alsatian dialects, in order to identify spelling variants. The algorithm also tackles standard German and English spelling in order to find cognates;
- We use external resources to obtain information about synonyms in the French language and translations into German and English.

The article is organised as follows: in the following section we review previous work on the identification of spelling variants and the alignment of lexical resources. Section 3 details the lexical resources used in our work. We present our alignment and mapping method in Section 4 and the evaluation results in Section 5.

2. State of the Art

2.1. Identification of Spelling Variants

Non-standard writing is an issue when dealing with different kinds of texts, e.g. data from the Web, in particular Web 2.0, historical texts and languages which are mainly oral and thus non-written.

A first family of methods target normalisation, i.e. transforming a minority variant to a given standard. Scherrer (2008) uses orthographic Levenshtein distance and trained stochastic transducers in order to build a bilingual lexicon for a Swiss German dialect and standard German. Hulden et al. (2011) present two methods which automatically learn transformations from a dialectal form to the standard form using a limited parallel corpus for the Basque language and the Lapurdian Basque dialect. The first method relies on an existing tool, `lexdiff` (Almeida et al., 2010), which detects spelling differences. The spelling differences identified are used to obtain replacement rules which are compiled as transducers. The second method is inspired by ILP (Inductive Logic Programming) and tries to select the best set of replacement rules, using both positive and negative examples. Salloum and Habash (2011) describe a rule-based method to generate paraphrases of dialectal Arabic in standard Arabic. The paraphrases are used for Arabic-English statistical machine translation. For historical language variants, Porta et al. (2013) propose a method to map historical word forms to their modern counterparts. The approach is based on a Levenshtein transducer and a linguistic transducer implementing sound change rewrite rules.

In a different vein, Dasigi and Diab (2011) present a clustering algorithm which aims at grouping orthographic dialectal variants. They experiment with several word similarity measures and conclude that string similarity metrics perform better for this task than contextual similarity metrics.

Our work is closest to Dasigi and Diab (2011), in that we cluster dialectal variants and do not resort to normalisation. We preferred this approach as normalisation is not applicable in our case. Fist, there is no consensus on the writing norm for Alsatian dialects and it is thus difficult to decide which form should prevail. Moreover, even though Alsatian is closely related to German, there are a number of lexical and syntactic differences which have to be taken into account. Added to that, considering German as the standard for Alsatian is a very sensitive sociolinguistic issue, which has implications reaching deeper than purely linguistic considerations. Given all these reasons, our proposed method does not attempt to normalise writing variants but preserves their diversity by considering clusters of variants as lexicon entries.

2.2. Alignment of Lexical Resources

The main objective of our work is not only to identify spelling variants of the same Alsatian lexeme, but also to align entries stemming from different bilingual lexicons and map the alignments to a semantic network.

A lot of work has been devoted recently to the alignment of collaborative resources, such as Wikipedia, and classical lexical knowledge bases, such as WordNet.

Niemann and Gurevych (2011) detail a method for aligning senses in WordNet and Wikipedia, which was later employed for creating the UBY lexical-semantic resource (Gurevych et al., 2012). The method relies on a machine learning method which classifies alignments as valid or non-valid. The similarity of aligned sense candidates is computed based on a bag-of-word representation of the senses and then provided to the classifier. For the UBY resource, cross-lingual word sense alignments are induced in the same manner, by first automatically translating the textual representations of the senses.

Navigli and Ponzetto (2012) propose a method to relate Wikipedia pages to WordNet senses used for building the BabelNet resource. The method applies several different strategies sequentially. In particular, it re-uses a technique used for Word Sense Disambiguation which consists in defining a disambiguation context for each Wikipedia page and WordNet sense. The disambiguation context is a set of words obtained from information provided in the resources (e.g. labels, links, redirections and categories in Wikipedia ; synonyms, hypernyms / hyponyms, glosses in WordNet). A similarity score can then be computed based on this context.

When there is no lexical resource in one language, automatic translation of resources in another language is often the best option, in terms of construction costs. In this case, an existing resource is extended with lexicalisations in another language.

The WOLF (Wordnet Libre du Français) has been built by Sagot and Fišer (2008) using the Princeton WordNet and several multilingual resources. The main assumptions underlying their approach are that different senses of an ambiguous word in one language often correspond to different translations in another language and words which are translated by the same word in another language often have similar meanings. They enforce these ideas by collecting a multilingual lexicon with 5 languages from a parallel corpus and by assigning the most likely synset to each lexicon entry, relying on the intersections between the synsets associated to each non-French word in the lexicon in the Princeton WordNet or in wordnets from the BalkaNet project. Hanoka and Sagot (2012) have extended the WOLF resource using a new approach relying on a large

synonymy and translation graph built from Wikipedia and Wiktionary. The graph is queried with literals from synset-aligned multilingual wordnets to get the best translation candidate, based both on translation and back-translation relations.

In our work, we also apply the idea of extending an existing lexical-semantic resource with lexicalisations from another language, namely Alsatian. We use French as a pivot language to obtain a mapping between Alsatian variants and BabelNet. We also exploit the cognacy between Alsatian, German and English in order to enrich the feature vectors.

3. Resources

In this section, we detail the resources used in our work.

3.1. Bilingual French-Alsatian Lexicons

We have retrieved three bilingual French-Alsatian lexicons available on the Web:

- OLCA: the lexicons produced by the OLCA (*Office pour la Langue et la Culture d'Alsace*)[2]. These lexicons are domain-specific (beer, shopping, football, medicine, weather, nature, fishing, pharmacy, vine) and provide variants for the Bas-Rhin (Lower Rhine) and Haut-Rhin (Upper Rhine) Alsatian departments. In the rest of the article, these two variants are identified as OLCA-67 (for Bas-Rhin) and OLCA-68 (for Haut-Rhin);
- WKT: a lexicon retrieved from a Wiktionary user page;[3]
- ACPA: a bilingual lexicon authored by André Nisslé.[4]

These lexicons, though machine-readable, are not available in a standard format. They have been preprocessed with specific parsers to extract French-Alsatian word pairs. When available, information about part-of-speech is kept.[5] Otherwise, we used two heuristics for guessing the part-of-speech : (i) apply the French TreeTagger (Schmid, 1994) to obtain a category for French single words[6] ; (b) for nouns, check the presence of a determiner next to the Alsatian form.

Table 1 lists the number of French entries in the lexicons after preprocessing. The table shows that the coverage of the different parts-of-speech is uneven, and that the lexicons mostly focus on nouns, verbs and adjectives.

The lexicons follow different graphical conventions as exemplified by Table 2, which lists the translations found in the lexicons for several lexemes. Many translations in Table 2 are actually graphical variants of the same Alsatian lexeme (e.g. "Kràb " and "Kràpp"). However, these graphical variants can be very dissimilar if we only consider the characters used.

	OLCA-67	OLCA-68	WKT	ACPA
adjective	194	195	122	1,898
adverb	16	16	49	295
determiner	0	0	20	15
noun	2,628	2,617	1,049	15,770
past participle	45	46	59	476
pronoun	1	1	38	47
verb	276	276	292	3,017
unknown	671	676	393	2,015
TOTAL	3,831	3,827	2,022	23,533

Table 1: Number of French entries in the French-Alsatian 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**	

Table 2: Example translations found in the lexicons. Identical variants found in at least two lexicons are in bold format.

In addition to the bilingual lexicons, we also used two semantic networks: JeuxDeMots and BabelNet.

3.2. JeuxDeMots

JeuxDeMots (Lafourcade, 2007) is a freely available French lexical network built through crowdsourcing games.[7] We used the version dated November 30, 2013,[8] which contains 171,029 occurrences of the synonymy relation (though the network also contains many other types of relations, e.g. association, domain, hypernymy, hyponymy, etc.).

3.3. BabelNet

BabelNet (Navigli and Ponzetto, 2012) is a multilingual semantic network, which integrates knowledge from Word-

[2] http://www.olcalsace.org/

[3] Available from the user page of Laurent Bouvier: http://fr.wiktionary.org/wiki/Utilisateur:Laurent_Bouvier/alsacien-fran%C3%A7ais

[4] http://culture.alsace.pagesperso-orange.fr/dictionnaire_alsacien.htm

[5] We used the following list of POS categories: verb, adjective, adverb, preposition, phrase, conjunction, pronoun, interjection, proper noun, past participle, determiner abbreviation, noun (feminine, masculine, neutral, plural).

[6] We use the TreeTaggerWrapper by Laurent Pointal available at http://perso.limsi.fr/pointal/dev:treetaggerwrapper.

[7] The games can be played on the following website: http://www.jeuxdemots.org

[8] Available from http://www.lirmm.fr/~lafourcade/JDM-LEXICALNET-FR

Net and Wikipedia. BabelNet is composed of Babel synsets, which are concepts with lexicalisations in several languages. The multilingual lexicalisations were obtained either thanks to Wikipedia's inter-language links or to Machine Translation. We used BabelNet version 2.0.[9]

4. Method

In this section, we present our method for aligning the lexicons. It relies on a variant of the double metaphone algorithm, adapted to Alsatian dialects.

4.1. Double Metaphone for Alsatian Dialects

Given the absence of a widely spread writing convention, as well as differences due to geolinguistic variants, it is not possible to align lexicon entries based on their written forms only using classical string similarity measures (consider for instance "Grâb" and "Kràbb" from Table 3). In order to cater for these differences, we have developed a double metaphone algorithm for Alsatian dialects. Double metaphone (Phillips, 2000) was originally proposed for information retrieval, in order to find names spelled differently than the search string, but referring to the same entity. Double metaphone belongs to the class of phonetic encoding algorithms, as it transforms the input string into a key which is identical for words which are pronounced in a similar manner. For instance, for the three given names "Stephan", "Steven" and "Stefan" the resulting key is `STFN`. In order to take ambiguities into account, double metaphone actually returns two keys in some cases. Double metaphone has for instance been used for Web 2.0 text normalisation (Mosquera et al., 2012).
The double metaphone transformations for Alsatian were written based on an analysis of our input lexicons.[10] We also took standard German into account, in order to obtain identical keys for German and Alsatian cognates. Table 3 gives some examples of the double metaphone keys obtained for several Alsatian and German words.

4.2. Lexicon Alignment

Our first objective is to be able to align entries across several bilingual Alsatian-French lexicons. In a first step, all entries in the input lexicons are added to a large graph. The nodes correspond to Alsatian words and their French translations. Alsatian words are connected to their French translations in the lexicons by an edge. Moreover, 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.[11]

[9] Available from `http://www.babelnet.org/download.jsp`

[10] Our implementation of Double Metaphone for Alsatian dialects is based on an existing Python module for English `http://www.atomodo.com/code/double-metaphone/metaphone.py/view`.

[11] Adjectives and past participles are considered as the same category.

We also use information obtained from the resources detailed in Section 3 in order to relax condition 1.

French Synonyms The JeuxDeMots synonyms list is used to connect two Alsatian words which have synonymous French translations in this resource.

BabelNet French Senses BabelNet French senses are used in the same way as the JeuxDeMots synonyms, to connect Alsatian words which have French translations belonging to the same sense.

4.2.1. Alignment of Alsatian Variants

Alsatian variants corresponding to the same lexeme are retrieved by detecting connected components in the subgraph containing only Alsatian words.
Figure 1 shows a portion of the initial graph. The translations into French, German and English are also shown. In the subgraph formed by the Alsatian words, there are three connected components: (1) ["Winkäller", "Winkeller", "Winkaller"], (2) ["Wikaller"] and (3) ["Kaller"]. The words "Winkäller", "Winkeller" and "Winkaller" are therefore aligned and considered as variants of the same lexeme.

4.3. Mapping to BabelNet Synsets

Our second objective is to map aligned Alsatian variants to BabelNet synsets. For instance, taking the example of Figure 1, the cluster formed by ["Winkäller", "Winkeller", "Winkaller"] should be mapped to the synset with ID `bn:00017041n` (see Figure 2).

Figure 2: Synset `bn:00017041n` in BabelNet's online search interface.

The mapping is achieved by calculating the cosine similarity between binary bag-of-words representations of Babel synsets and aligned Alsatian variants.
In the simplest case, the representation used for Babel synsets consists of their French lexicalisations. Alsatian variants are represented by their French translations: in the example of Figure 1, the cluster formed by ["Winkäller", "Winkeller", "Winkaller"] will be represented by the French words ["chai", "cellier", "cave"].
The bag-of-words representations can be extended by leveraging the translations available in BabelNet. The use of multilingual features has been shown to have a positive effect on the task of word sense disambiguation (Banea and Mihalcea, 2011). However, in looking for translations into English and German for Alsatian lexemes we have to avoid ambiguity. This issue has been addressed in work on the acquisition of bilingual dictionaries for a language pair using a third language as a pivot : in our case, French is the

Word	French translation	English translation	Metaphone key 1	Metaphone 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	/
Iwereinsschtimmung	concordance	agreement	AFRNXTMNK	AVRNXTMNK
bestätigen	confirmer	confirm	PXTTK	/
Aufstand	insurrection	insurrection	AFXTNT	/
Übereinstimmung	concordance	agreement	APRNXTMNK	AVRNXTMNK

Table 3: Example metaphone keys. Alsatian words are in the upper part of the table, while German examples are detailed in the lower part of the table.

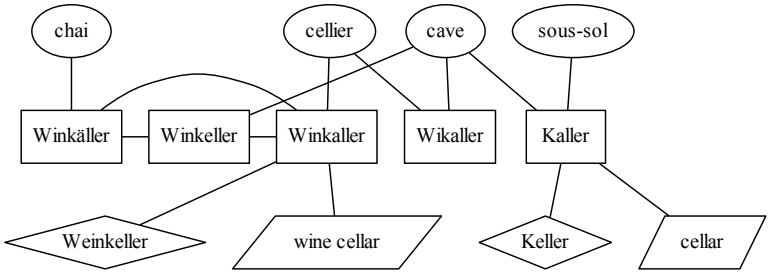

Figure 1: Simplified view of a subgraph. French words are in ellipses, Alsatian words in boxes, German words in diamonds and English words in parallelograms.

pivot language, Alsatian the source language and German and English the target languages. Several methods have been proposed, relying mostly either on the structure of the available bilingual lexicons or on distributional similarity (Tanaka and Umemura, 1994; Saralegi et al., 2011). In our particular case, we exploit the closeness between Alsatian and German, and, to a lesser degree, English. Starting from the French translations, German and/or English translations are added to the bag-of-words representations of Alsatian words if they share one of their double metaphone keys. This constraint performs a sort of disambiguation and ensures that only valid translations are selected. Thus, in the example of Figure 1, the German word "Weinkeller" and English word "wine cellar" will be added to the bag-of-words.

5. Evaluation of the Aligned Lexicon

5.1. Evaluation Methodology

In order to evaluate our method, we manually produced 100 ground-truth alignments between the lexicons and BabelNet. To this aim, we randomly selected entries from a multilingual French-German-Alsatian-English dictionary (Adolf, 2006). This dictionary presents several advantages for the evaluation: several spelling variants are usually proposed for each Alsatian entry, translations into French, German and English are provided, thus facilitating the mapping to BabelNet and, finally, the dictionary focuses on Alsatian lexemes which are very similar to corresponding German and English words.

To produce our evaluation dataset, we excluded BabelNet mappings with no translations into French and chose to limit ourselves to at most two Babel synsets. In case of a tie, the mapping to BabelNet is considered as correct if at least one of the Babel synsets is correct.

The alignment of variants is evaluated in terms of precision, recall and F-measure. For each French word in the evaluation dataset, we count the intersection between its Alsatian variants in the gold standard and in the automatic alignments as true positives (TP). Automatically aligned variants which are not in the gold standard are considered as false positives (FP), while those in the gold standard which are not in the alignments are considered as false negatives (FN). Then, precision (P), recall (R) and F-measure (F) are computed as follows :

$$P = \frac{TP}{TP + FP} \quad ; \quad R = \frac{TP}{TP + FN} \quad ; \quad R = \frac{2 \cdot P \cdot R}{P + R}$$

The mapping to Babelnet is evaluated in terms of the proportion of correct mappings. Since Babel synsets can be ranked according to cosine similarity, we consider the top 1, 2 and 3 mappings and judge the mapping as correct if one relevant Babel synset is found among the top 1, 2 or 3.

	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

Table 4: Evaluation results

5.2. Results

The evaluation results for different settings are detailed in Table 4. The baseline corresponds to a setting which does not make use of any external resource. + JDM entails that the JeuxDeMots synonyms have been used. + BN entails that BabelNet has been used, with lexicalisations in French (FR), German (DE) or English (DE).

Overall, the results for the alignment of variants are stable: the use of external resources leads to a slight drop in precision which is compensated by a a slight rise of recall. Also, recall is always lower than precision.

For the mapping to Babel synsets, the use of translations into German and, to a lesser degree, English, lead to clear improvements, in particular for pushing relevant Babel synsets to the first rank. The synonyms provided by JDM actually have a detrimental effect on the performance, most certainly because the synonym sets in this resource are different from those in BabelNet.

5.3. Discussion

The lower recall obtained for the alignment of variants is mainly due to the constraint which demands identical metaphone keys. In some cases, variants have different keys (e.g. "Chilche" - `KLX / XLX` and "Kirche" - `KRX`). This also raises a more fundamental question: can these variants still be considered as alternatives for the same lexeme, or do they form a new lexeme? In our construction of the gold-standard, we grouped variants as found in the multilingual dictionary, even though they might be rather different in some cases. In addition to the metaphone keys, more classical string similarity measures could be used to align variants, as it is done for cognate identification (Inkpen et al., 2005). These measures could help improving recall.

Some errors are also due to problems in retrieving part-of-speech tags for ambiguous dictionary entries. As one of the alignment conditions requires identical parts-of-speech, such entries are not considered as variants.

As shown by the results, adding multilingual features helps improving the mapping to Babel synsets. For the time being, German and English translations are selected based on their metaphone keys, which leads to missing translations for some features vectors. In future work, this could be improved by using additional bilingual lexicons, not necessarily limited to the translations available in BabelNet. Also, the inverse consultation method proposed in the context of pivot based bilingual dictionary construction could be put to use in order to add translations which are not necessarily cognates of the Alsatian variants (Tanaka and Umemura, 1994). However, since there is no monolingual corpus for the Alsatian dialects, methods based on distributional similarity are excluded.

Finally, the method is able to rank Babel synsets, but not to decide which of the synsets are accurate. A threshold for the cosine similarity could be learned, in order to obtain mappings only to relevant synsets.

6. Conclusion and Perspectives

We have presented a method to both align spelling variants of the same Alsatian lexeme found in several lexicons and map the variants to synsets in BabelNet. The alignment of the variants relies on the double metaphone algorithm while the mapping uses multilingual (German and English) features in its best performing setting. The mapping to BabelNet gives access to different kinds of additional information: definitions and glosses, translations into other languages, images, etc. All these could be used to produce language games or didactic resources for Alsatian. Moreover, this method could in principle be applied to many less-resourced languages, as the only needed resource is a bilingual lexicon.

In the future, we plan to provide the aligned lexicon in a standard format, to allow its use as Linked Open Data. SKOS for instance allows for several alternative lexical labels with no preferred label.[12] However, the absence of normalization is an issue for many NLP applications which could use the lexicon, in particular lemmatization. This will require finding solutions for this pervasive problem.

Acknowledgements

This work was supported by a grant from the scientific council of the Université de Strasbourg. We thank the reviewers for their insightful comments.

7. References

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

[12] See `http://www.w3.org/TR/skos-reference/#L1606`

Almeida, J. J., Santos, A., and Simões, A. (2010). Bigorna – A Toolkit for Orthography Migration Challenges. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Banea, C. and Mihalcea, R. (2011). Word sense disambiguation with multilingual features. In *Proceedings of the Ninth International Conference on Computational Semantics*, page 25–34.

Bauer, L. (2003). *Introducing Linguistic Morphology*. Georgetown University Press. 2nd edition.

Dasigi, P. and Diab, M. (2011). CODACT: Towards Identifying Orthographic Variants in Dialectal Arabic. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 318–326, Chiang Mai, Thailand.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY–A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 580–590, Avignon, France.

Hanoka, V. and Sagot, B. (2012). Wordnet creation and extension made simple: A multilingual lexicon-based approach using wiki resources. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012)*.

Huck, D., Bothorel-Witz, A., and Geiger-Jaillet, A. (2007). L'Alsace et ses langues. Éléments de description d'une situation sociolinguistique en zone frontalière. *Aspects of Multilingualism in European Border Regions: Insights and Views from Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol*, page 13–100.

Hudlett, A. and Groupe d'Etudes et de Recherches Interdisciplinaires sur le Plurilinguisme en Alsace et en Europe. (2003). *Charte de la graphie harmonisée des parlers alsaciens: système graphique GRAPHAL - GERIPA*. Centre de Recherche sur l'Europe littéraire (C.R.E.L.), Mulhouse, France.

Hulden, M., Alegria, I., Etxeberria, I., and Maritxalar, M. (2011). Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, page 39–48, Edinburgh, Scotland, July.

Inkpen, D., Frunza, O., and Kondrak, G. (2005). Automatic identification of cognates and false friends in French and English. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, page 251–257.

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

Mosquera, A., Lloret, E., and Moreda, P. (2012). Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalisation. In *Proceedings of the Worshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*.

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

Niemann, E. and Gurevych, I. (2011). The people's web meets linguistic knowledge: Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, page 205–214.

OLCA / EDinstitut. (2012). Etude sur le dialecte alsacien. Online, visited Feb 11, 2014: `https://www.olcalsace.org/sites/default/files/documents/etude_linguistique_olca_edinstitut.pdf`.

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

Porta, J., Sancho, J.-L., and Gómez, J. (2013). Edit Transducers for Spelling Variation in Old Spanish. In *Proceedings of the Workshop on Computational Historical Linguistics at NoDaLiDa 2013*, volume 87 of *Linköping Electronic Conference Proceedings*, page 70–79.

Sagot, B. and Fišer, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *Actes de TALN 2008-Traitement Automatique des Langues Naturelles*.

Salloum, W. and Habash, N. (2011). Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, page 10–21.

Saralegi, X., Manterola, I., and San Vicente, I. (2011). Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, page 846–856.

Scherrer, Y. (2008). Transducteurs à fenêtre glissante pour l'induction lexicale. In *Actes de RECITAL 2008*, Avignon.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Tanaka, K. and Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, page 297–303.

Zeidler, E. and Crévenat-Werner, D. (2008). *Orthographe alsacienne: bien écrire l'alsacien de Wissembourg à Ferrette*. J. Do Bentzinger, Colmar, France.