Resources and Methods for the Automatic Recognition of Place Names in Alsatian

Delphine Bernhard, Pierre Magistry, Anne-Laure Ligozat, Sophie Rosset

Resources and Methods for the Automatic Recognition of Place Names in Alsatian

Delphine Bernhard[1], Pierre Magistry[2],
Anne-Laure Ligozat[3] and Sophie Rosset[2]

[1]LiLPa - EA 1339, Université de Strasbourg, France
[2]LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
[3]LIMSI, CNRS, ENSIIE, Université Paris-Saclay, F-91405 Orsay France
E-mail: dbernhard@unistra.fr,
{annlor,magistry,sophie.rosset}@limsi.fr

Abstract

This article describes annotated resources (corpus, lexicons) for the automatic recognition of place names in the Alsatian dialects. The two main issues are related to their non-standardized orthography, leading to spelling variants, and the scarcity of available resources. We also present automatic methods using recurrent neural networks for the identification of place names that take these aspects into account.

1 Introduction

The detection of real-world entities is important for many text understanding applications. The main purpose of named entity recognition (NER) is to identify such pieces of information as names of persons, places and organisations. Detecting named entities in text relies on resources, mainly lexicons, whatever the approach used to develop the system (rule-based or statistical).

In this article, we present resources and methods for the automatic identification of place names in the High German Alsatian dialects, spoken in North-Eastern France. The Alsatian dialects are an heritage of the linguistic changes brought in the region by the Alemanni and the Franks as early as the 6th century [12]. The geographic region now called Alsace started to be progressively integrated into France during the 17th century. However, this did not have a real impact on everyday language use, especially for the middle and lower classes. Major changes to this situation occurred only after World War II, and in particular in the last third of the 20th century, when French also became the language of communication in all everyday activities. This period is also characterized by the gradual decline of the Alsatian dialects.

The Alsatian dialects have always been used mainly orally, with either French or German being used as the languages of choice for writing, depending on the time period. The earliest writings in Alsatian can be traced back to the second half of the 17th century [20], but the real beginnings of Alsatian literature are attributed to the comedy in verse published by Jean-Georges-Daniel Arnold in 1816 (*Der Pfingstmontag*). Since then, there has been an ongoing –albeit not very numerous– literary and cultural production with a focus on two main text genres : theater plays (mostly comedies) and poetry [21, 22, 23, 24]. Other genres are also represented: poetic prose, songs, nursery rhymes, tales, translations and adaptations of works in other languages. In addition to this literary production, there have also been efforts to provide linguistic descriptions in the form of dictionaries, glossaries and grammars. However, it should be noted that texts in prose are rarer, with the exception of a few authors such as Marie Hart: either French or German are used in this case.

While French is clearly dominant nowadays in the public space, place names are one of the cases where Alsatian is still present in everyday life, even to non dialect speakers. An increasing number of villages and cities choose to have French-Alsatian bilingual town entrance and street signs.[1] Each location in Alsace has two names: an "official" name and a "popular" name [16]. Official names are written and stem from the German language for 95% of them, though often with spelling deviations with respect to the norm [16]. Popular names are oral and often correspond to the dialectal pronunciation of the official name. There may however exist form differences, e.g. the popular name of the village *Schwindratzheim* is *Schwìngelse* [16]. Moreover, the pronunciation may differ depending on the dialectal variant in use, e.g. *Wissembourg* is locally pronounced *Waisseburch*, but *Wisseburi* or *Wissaburg* elsewhere in the region [16].

Repositories of place names in Alsatian are very scarce and have to be collected, improved and categorized according to a well-defined typology. Various typologies of named entities exist and they differ by their semantic coverage or categorical representations [7]. The MUC-6 project [9] was the first to propose a definition of NEs as proper nouns that refer to specific semantic classes: Person, Location, Organisation, etc. New typologies were proposed based on this first definition, aiming at more fine grained classes [5, 18]. More recently, a new typology was defined within the QUAERO project [10] with the objective of being more general than Sekine's typology [18] while having wide semantic coverage. In this work, we used the location type definition as defined in QUAERO's typology. As shown in Figure 1 (left-hand side), the different location categories are: administrative locations with geopolitical definitions (cities, countries, states...), physical locations (geonyms, hydronyms, astronyms), odonyms (streets, squares...), facilities (train stations, universities...) and addresses (physical or electronic).

The main contributions of this article are as follows (1) we present lexicons of Alsatian place names which have been manually categorized according to the

[1] See for instance `https://commons.wikimedia.org/wiki/File:Mulhouse_entr\%C3\%A9e_agglom\%C3\%A9ration.JPG`

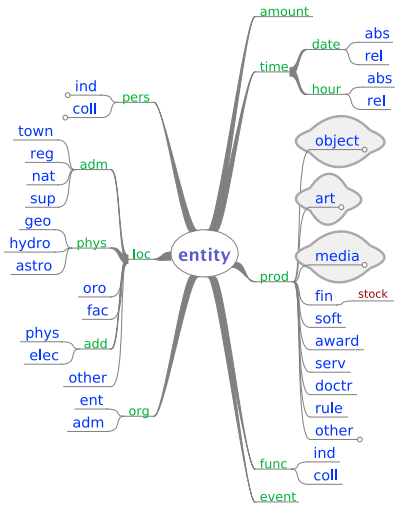

Figure 1: Quaero Typology [17].

QUAERO named entity types (Section 2.1) ; (2) we describe the first corpus manually annotated with place names for the Alsatian dialects (Section 2.2) ; (3) we propose and experiment with methods to automatically identify locations in Alsatian corpora (Section 3).

2 Description of the Resources

2.1 Lexicons

The lexicons we collected consist of bilingual lists of place names in French and Alsatian, manually categorized with respect to the QUAERO location types. The sources used for the collection include printed and online material:

- Alsadico: a printed French-Alsatian dictionary [14];
- Elsàsser: a website which includes, among others, a list of Alsatian place names with their phonetic transcription and translation into French [13];
- WikiAls: the Alemannic Wikipedia (`als.wikipedia.org`). Wikipedia pages corresponding to locations were collected thanks to specific categories (e.g. [[Kategorie:Ort (Unterelsass)]]) and the dialect tag of the page;
- WikiFr: French Wikipedia (`fr.wikipedia.org`). Relevant pages were identified thanks to their category. Pages are written in French but often contain the Alsatian place names in the text, for example “La ville est appelée *Mìlhüsa* en alsacien” (The city is called Mìlhüsa in Alsatian). These instances were identified thanks to regular expressions detecting the presence of the expression "(en) [[alsacien]]" in the context.

Printed material, which was available only on paper, has been re-typed to obtain digital resources, while online material has been collected automatically and manually corrected afterwards. It should be noted that both Alsadico and Elsàsser are copyrighted and hence cannot be freely redistributed.[2] We nevertheless used these resources in order to perform some linguistic analyses and assess the quality and coverage of the Wikipedia-derived datasets. The categorization into the QUAERO entity types has been performed by a first annotator and then reviewed and corrected by a second annotator. Table 1 shows an extract from the lexicon. The translation into French makes it possible to retrieve spelling variants.

Alsatian	**French**	**Type**	**Sources**
Algelse	Algolsheim	<loc.adm.town>	Elsàsser
Àlgesle	Algolsheim	<loc.adm.town>	AlsaDico
Algolse	Algolsheim	<loc.adm.town>	WikiAls
Elsass	Alsace	<loc.adm.reg>	WikiFr
Elsàss	Alsace	<loc.adm.reg>	WikiAls

Table 1: Extract from the lexicon.

Table 2 details the contents of the lexicons. The overlap between the lexicons is rather low : on average, an Alsatian spelling variant is found in 1.18 lexicons : 2,484 spellings are found in one lexicon only, 376 in two, 66 in three and only 3 are present in all four lexicons (*Kurzehüse* ; *Molse* ; *Sundhüse*).

Type	**AlsaDico**	**Elsàsser**	**WikiAls**	**WikiFr**	**Corpus**
loc.fac – Facility	0	9	2	1	2
loc.phys.astro – Astronym	0	0	0	0	1
loc.phys.geo – Geonym	0	55	16	6	60
loc.phys.hydro – Hydronym	0	22	22	1	9
loc.adm.nat – Country	0	0	10	0	50
loc.adm.reg – Region	3	1	89	1	119
loc.adm.sup – Supranational	0	0	0	0	3
loc.adm.town – City	1,184	1,038	891	94	154
loc.oro – Odonym	0	0	0	1	3
Total	1,187	1,125	1,030	104	401

Table 2: Types of locations in the resources. For the corpus, the figures correspond to the number of occurrences.

2.2 Corpus

The corpus is composed of two main sources of Alsatian documents: Wikipedia articles from the Alemannic Wikipedia and chronicles from an information magazine

[2]The non-copyrighted lexicons are scheduled to be released on the project's website `http://restaure.unistra.fr/` under a CC BY-SA license.

published by the Haut-Rhin department (southern Alsace) General Council. The Wikipedia articles are written in different Alemannic geolinguistic variants found in the Alsace region, while the chronicles are written in Low Alemannic from the southern part of the region. In addition, two more specific genres were used for the annotator training phase: one excerpt from a theater play and some recipes. The two main sources meet the requirements set within our project whose overall goal is to provide resources and tools for language learning and documentation: (i) freely redistributable,[3] (ii) contemporary forms of the Alsatian dialects, (iii) texts in prose with a wide potential audience.[4] Texts in prose were favoured over poems or theatre plays because they are easier material both for the development of NLP tools and for language learners. Texts from other genres and time periods will be considered in the future.

The annotated corpus contains 21 documents and 12,570 tokens (including punctuation). Location annotation was performed following a BIO (Begin-Inside-Out) notation: the first token of a location is annotated with a B-LOC tag, the following ones with I-LOC, and non location tokens with an O tag.

Overall, 401 occurrences of place names have been annotated in the corpus, corresponding to 207 different place names. As can be seen in Table 2, location types found in the corpus are much more varied than those found in the lexicons. While names of cities are largely predominant in the lexicons, the corpus also contains many geonyms, countries and regions. Only 40 tokens were annotated with a I-LOC tag: most location entities are expressed by a single token.

2.3 Spelling Variation in the Resources

Spelling variation is an important issue in the Alsatian dialects, since there is no widely acknowledged and standardised orthography. In the lexicons, a location name has on average 1.92 Alsatian graphical variants. The maximum number of spelling variants is 14, for the town Mulhouse (see Figure 2). In the corpus, a location name has on average 1.38 Alsatian graphical variants. Interestingly, the maximum number of spelling variants is also found for the town Mulhouse, with six different spellings (see Figure 2).

Mulhouse (in the lexicons): Mïehlhüse ; Mielhüse ; Mielhüusa ; Mièlhuuse ; Mihlhüsa ; Milhüsa ; Mìlhüsa ; Milhüse ; Mïlhüse ; Milhüüsa ; Milhüüse ; Mìlhüüse ; Mìlüüse ; Mílhüse
Mulhouse (in the corpus): Mìhlhüse ; Milhüsa ; Milhüse ; Milhüüsa ; Milhüüse ; Mulhouse

Figure 2: Spelling variants for the town “Mulhouse”

[3] We secured the authorization of the publisher for the chronicles.

[4] The Alemannic Wikipedia is freely available on the Web, while the Haut-Rhin information magazine is distributed in all homes in the department.

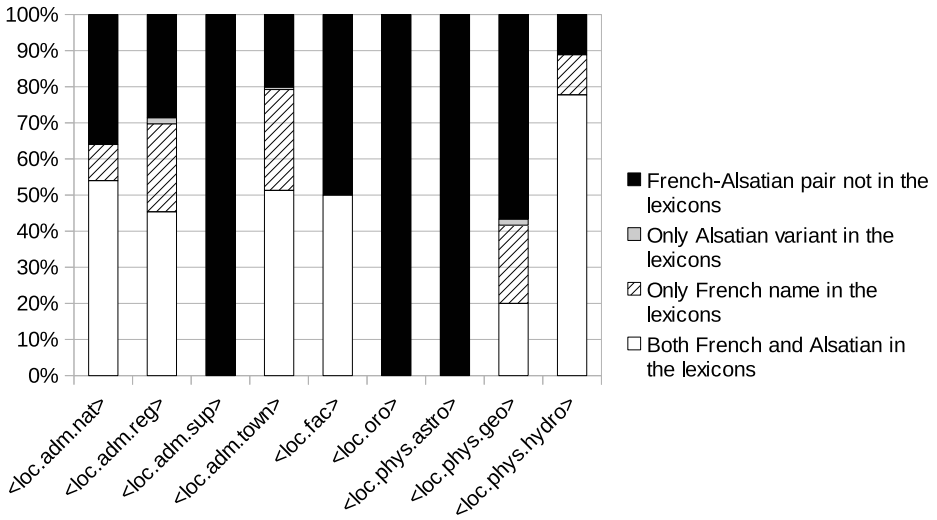

Figure 3: Overlap between the corpus and the lexicons.

We also measured the overlap between the corpus and the lexicons for each entity type (see Figure 3). Overall, from the 401 occurrences in the corpus, 180 French-Alsatian pairs are found in the lexicons, 91 French location names are found in the lexicon but not with the same Alsatian spelling variant, 4 Alsatian spelling variants are found in the lexicons but not with the same French location name and 126 French-Alsatian pairs are totally absent from the lexicons (neither the French location name, nor the Alsatian spelling variant is found in the lexicons). As could be expected from the composition of the lexicons, the coverage is highly dependent on the entity type: supranational entities, odonyms and astronyms are completely absent from the lexicons.

To sum up, there are two main issues: (i) the low coverage of the lexicons with respect to the place names found in the corpus and (ii) the large amount of spelling variants found in the corpus.

3 Location Recognition and Linking

Since the location lexicons lack coverage, we chose to use word embeddings to inform our location recognition and linking methods. Word embeddings are semantic representations of words in a low dimension vector space, learnt in an unsupervised way from large corpora.

We first tested a baseline system for location named entity recognition. The system performs named entity recognition for locations and is based on a bidirectional LSTM (Long Short Term Memory, which is a type of recurrent neural network) and word representations built with fastText [1]. fastText is a library for learning word representations using character *n*-grams of words, which improves

the representation of rare words. The architecture of this system is the biLSTM-CRF as proposed in [15]. This model predicts a sequence of tags for any sequence of observations (in our case, a sentence) based on multiple sources of information. To select a tag for a word, it combines contextual information from surrounding words and characters (modelled as vectors) and tagging decisions to be made about the other tokens of the sentence. The model uses the BIO tagged corpus. Results are given in Table 3.

System	**Recall**	**Precision**	**F-measure**
LSTM (1doc vs. all)	.53	.60	.56
LSTM (80/20)	.61	.79	.69

Table 3: Named entity recognition for locations

We used cross-validation on documents to evaluate our system under two conditions:

- first, the corpus of 21 annotated documents was divided into 21 train/test sets, with one document being the test corpus and all the others the training corpus (1 doc vs. all condition).
- we also tested cross-validation on sentences (80/20 condition): 80% of the corpus was used to train the model, and 20% to test it.

To build the word embeddings, we rely only on the WikiAls data. The missing vectors for the words from our annotated documents which are absent from the WikiAls corpus are generated in a second step using the method included in fastText to avoid having Out-of-Vocabulary tokens. This vector space is then used for the initialization of the embedding layer of the biLSTM-CRF. This system achieves a .56 F-measure with the 1 doc vs. all condition and .69 F-measure with the 80/20 condition, which is a little easier since parts of the train and test corpora come from the same documents.

The second system performs named entity linking: when given a new location name recognized by the first system, it links it to an existing location cluster by selecting the most similar locations in the embeddings. We tested the following method: for each location in the annotated corpus, we selected the 10 most similar words in the embeddings, and checked if these similar words were associated to the same lexicon entry. For example, for the location “Stroßburg” in the text, one of the most similar word is “Stroßburi”, and both are Alsatian variants for “Strasbourg” in the lexicon. The goal is to assess whether, if the text location was absent from the lexicon, it could still be linked to the correct lexicon entry. For the 207 distinct location entities from the annotated corpus, 61 can be linked to the correct entry. Graphical variants are found among the most similar words, such as “Milhüüse” for “Milhüüsa”, “Schwitz” for “Schwiz”, “Tännchel” for “Taennchel” or “Sawere” for "Zàwere”.

4 Related Work

Most NER systems rely on information on the words to be annotated: forms of the word, presence in a specific lexicon, part-of-speech tags etc. All this information is obviously affected by surface form variations [19, 2]. Spelling variation is indeed a well-documented issue for the identification of place names in historical texts [3], given that the standardization of orthography is often quite recent.

Concerning specifically location detection in historical data, Borin et al. [2] proposed a knowledge- and rule-based approach which aims specifically at handling the variation in 19th century Swedish literature. Their best system reaches an F-measure of 86.4%. For the Arabic language, which also presents high variation, a knowledge and rule-based approach is described in [19]. The presented system reached an F-measure of 85.9%. All these works consider only one class for the location type. The QUAERO typology, on which our work is based, was used on French old press data [8] and on Swiss old press data in French [6] which also contains a lot of variation. In the latter, the authors compared various systems and for the location type the results ranged between 48% and 69% of F-measure depending on the system tested, which is similar to what we obtained in this work. Most current models for NER consider it as a supervised sequential classification problem where each sentence is a sequence [4, 11, 15]. In order to categorize words, the model can rely on orthographic information, captured by character-based representations, and distributional information, captured by word embeddings. Recently a method to represent such information which includes character level information was proposed [1]. This approach is often considered as being robust against variation. Our hypothesis was that this model may be useful for our purpose.

5 Conclusion and Perspectives

We have described the first corpus manually annotated with place names for the Alsatian dialects as well as lexicons collected from different sources. The automatic methods proposed to identify place names face two main issues: low coverage of the lexicons and spelling variants.

The corpora and lexicons correspond to contemporary Alsatian and the texts in the corpus belong to two particular text genres (encyclopaedia and chronicle). The next step will be to assess whether the methods and resources can be applied to older texts and different text genres, in particular theatre plays and tales.

Acknowledgements

We would like to thank Clément Dorffer, Elisa Feuerstein, Mario Luis Figueroa Miranda and Gwendoline Hollner for their participation in the collection and annotation of the datasets. This work was supported by the French "Agence Nationale de la Recherche" (ANR) (project no.: ANR-14-CE24-0003).

References

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[2] Lars Borin, Dimitrios Kokkinakis, and Leif-Jöran Olsson. Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature. In *Proceedings of LaTeCH 2007*, pages 1–8, Prague, Czech Republic, June 2007.

[3] James O. Butler, Christopher E. Donaldson, Joanna E. Taylor, and Ian N. Gregory. Alts, Abbreviations, and AKAs: historical onomastic variation and automated named entity recognition. *Journal of Map & Geography Libraries*, 13(1):58–81, 2017.

[4] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.

[5] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The Automatic Content Extraction (ACE) Program, Tasks, Data, and Evaluation. In *Proceedings of LREC-2004*, Lisbon, Portugal, May 2004.

[6] Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. Diachronic Evaluation of NER Systems on Old Newspapers. In *Proc. of KONVENS 2016*, pages 97–107, Bochum, Germany, 2016.

[7] Maud Ehrmann, Damien Nouvel, and Sophie Rosset. Named Entity Resources - Overview and Outlook. In *Proceedings of LREC 2016*, may 2016.

[8] Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum, and Ludovic Quintard. Extended Named Entities Annotation on OCRed Documents: From Corpus Constitution to Evaluation Campaign. In *Proceedings of LREC'12*, Istanbul, Turkey, may 2012.

[9] Ralph Grishman and Beth Sundheim. Design of the muc-6 evaluation. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, 1995.

[10] Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. Proposal for an Extension of Traditional Named Entities: From Guidelines to Evaluation, an Overview. In *Proceedings of LAW-V, ACL*, pages 92–100, Portland, OR, June 2011.

[11] James Hammerton. Named entity recognition with long short-term memory. In *Proceedings of HLT-NAACL 2003-Volume 4*, pages 172–175, 2003.

[12] Dominique Huck. *Une histoire des langues de l'Alsace*. la Nuée bleue, Strasbourg, 2015.

[13] Marc Hug. Toponymes d'Alsace. Online, `http://elsasser.free.fr/NomCommu/ecrantot.html`, 2007.

[14] Edmond Jung. *L'alsadico : 22 000 mots et expressions français-alsacien*. La Nuée bleue, Strasbourg, 2006.

[15] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural Architectures for Named Entity Recognition. In *Proceedings of NAACL-HLT*, pages 260–270, June 2016.

[16] Michel Paul Urban. *La grande encyclopédie des lieux d'Alsace*. La Nuée Bleue, Strasbourg, 2nd edition, 2010.

[17] Sophie Rosset, Cyril Grouin, and Pierre Zweigenbaum. Entités nommées structurées : guide d'annotation Quaero. Technical report, 2011. Notes et Documents LIMSI N° : 2011-04.

[18] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended Named Entity Hierarchy. In *Proceedings of LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain*, 2002.

[19] Khaled Shaalan and Hafsa Raza. NERA: Named entity recognition for Arabic. *Journal of the Association for Information Science and Technology*, 60(8):1652–1663, 2009.

[20] Auguste Wackenheim. *Tome 1. Du XVIIe au XIXe siècle – Les anonymes, les précurseurs, les fondateurs*. La littérature dialectale alsacienne: une anthologie illustrée. Prat-éditions, Paris, 1993.

[21] Auguste Wackenheim. *Tome 2. L'âge d'or du XIXe siècle: la fin de l'Empire, la Restauration, le Second Empire*. La littérature dialectale alsacienne: une anthologie illustrée. Prat-éditions, Paris, 1994.

[22] Auguste Wackenheim. *Tome 3. La période allemande, 1870-1918*. La littérature dialectale alsacienne: une anthologie illustrée. Prat-éditions, Paris, 1997.

[23] Auguste Wackenheim. *Tome 4. D'une guerre mondiale à l'autre: 1918 - 1945*. La littérature dialectale alsacienne: une anthologie illustrée. Prat-éditions, Paris, 1999.

[24] Auguste Wackenheim. *Tome 5. De 1945 à la fin du XXe siècle*. La littérature dialectale alsacienne: une anthologie illustrée. Prat-éditions, Paris, 2003.