Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steible, Pascale Erhart, Nabil Hathout, Dominique Huck, et al.

Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard

Delphine Bernhard[1], Anne-Laure Ligozat[2], Fanny Martin[3], Myriam Bras[4], Pierre Magistry[5], Marianne Vergez-Couret[6], Lucie Steiblé[1], Pascale Erhart[1], Nabil Hathout[4], Dominique Huck[1], Christophe Rey[7], Philippe Reynés[3], Sophie Rosset[5], Jean Sibille[4], Thomas Lavergne[8]

[1]LiLPa, Université de Strasbourg, France
[2]LIMSI, CNRS, ENSIIE, Université Paris-Saclay, F-91405 Orsay, France
[3]Laboratoire Habiter le Monde - HM - EA 4278, Université de Picardie Jules Verne, Amiens, France
[4]CLLE, Université de Toulouse, CNRS, UT2J, France
[5]LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
[6]Queen's University, Belfast
[7]LT2D (Lexiques, Textes, Discours, Dictionnaires), Université de Cergy-Pontoise, IUF
[8]LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, F-91405 Orsay, France

[1]{dbernhard,lucie.steible,pascale.erhart,huck}@unistra.fr
[2,5,8]{annlor,magistry,sophie.rosset,thomas.lavergne}@limsi.fr
[3]{fanny.martin,philippe.reynes}@u-picardie.fr
[4]{myriam.bras,nabil.hathout,jean.sibille}@univ-tlse2.fr
[6]m.vergez-couret@qub.ac.uk
[7]christophe.rey@u-cergy.fr

Abstract

This article describes the creation of corpora with part-of-speech annotations for three regional languages of France: Alsatian, Occitan and Picard. These manual annotations were performed in the context of the RESTAURE project, whose goal is to develop resources and tools for these under-resourced French regional languages. The article presents the tagsets used in the annotation process as well as the resulting annotated corpora.

Keywords: corpus, annotation, part-of-speech, Alsatian, Occitan, Picard

1. Introduction

Since the constitutional amendment (Article 75-1) published in 2008, regional languages are officially part of the heritage of France, although the only official language in France is French (Article 2). As such, it is essential to implement modern means for their preservation and transmission, relying particularly on digital technologies. Regional languages of France can be considered as low-resourced, that is to say there are no or only few electronic resources (corpora, lexicons, dictionaries) and tools. All languages with little resources have in common that their computerisation has a low financial profitability which does not compensate for considerable development costs. However, endowing these languages with electronic resources and tools is a major concern for their dissemination, protection and teaching (including for new speakers). Automatic tools help ensure data collection (scanning), storage in standardized formats, categorization and retrieval. Moreover, the availability of digital data and automatic processing tools can transform the attitude of speakers towards regional languages, in particular increase the use of written material that often remains marginal. In a broader perspective, it is the diversity of world languages which would be better preserved and the amount of data available to researchers in human and social sciences (linguistics, sociology, anthropology, literature, history, ...) would increase (Soria et al., 2013).

The overall objective of the RESTAURE[1] project is to provide computational resources and processing tools for three regional languages of France: Alsatian, Occitan and Picard. The three of them belong to the languages of France listed in Cerquiglini's 1999 report (Cerquiglini, 1999), which inventories the regional languages of France within the meaning of the European Charter for Regional or Minority Languages. They have no official status in France and as such, have suffered from a lack of institutional support until recently. The goal of this project is to bring these languages to the front and foster NLP research on these languages. Processing natural language is complex and the development of NLP tools requires significant resources, both human and financial. This explains the lack of such tools for regional languages of France.

In this work, we present the methodology used to create corpora annotated with part-of-speech information for the three regional languages considered in RESTAURE. The main challenge for writing the annotation guidelines was the lack of comprehensive grammatical descriptions, encompassing all the dialectal variants found in our corpora. In addition, the annotation triggered further discussion on tokenisation issues and the use of POS tagsets and taggers developed for closely related languages. The whole process was made possible thanks to the close cooperation between

[1]http://restaure.unistra.fr/

linguists and NLP specialists, as well as the parallel and collaborative work on three different languages facing similar challenges.

2. Description of Alsatian, Picard and Occitan

In this section, we briefly describe the three French regional languages considered in the project, in particular with respect to their morpho-syntactic properties.

2.1. Alsatian

Alsatian is spoken in North-Eastern France and is part of the High German dialects, which are subdivided into Central German and Upper German. The majority of the Alsatian dialects belongs to (Low) Alemannic, an Upper German dialect. A small part of the dialectal space (North-West) belongs to Central German Rhine Franconian. Like all dialectal, phonological and, partially, lexical spaces, the Alsatian dialects are characterized by *spatial variation*, which is the main characteristic of a dialect. Since the second half of the 20th century, the writers and speakers of Alsatian have had, in their vast majority, a plurilingual repertoire, with French taking more and more importance and driving them to use, in Alsatian, linguistic strategies and *calques* (loan translations) from French. The fundamental morphosyntactic characteristics are little affected by this phenomenon. They are common to all Alsatian dialects and are very similar to those of standard German. Finite verbs are marked with morphemes of tense, mode and person-number and noun phrases are marked with number, case and gender. The tense and mode system is much simpler to standard German and there is only one person number marker for the plural persons (Huck, to appear). However, the surface form of the morphemes can present intradialectal variations. The Alsatian dialects are used mainly orally and written production is limited to some literary works (poetry, theater plays), linguistic descriptions (dictionaries, lexicons), small contributions to otherwise French publications (chronicles in newspapers) and online texts (Wikipedia, social networks). What is more, several spelling conventions have been proposed, but none of them can be considered as a widely accepted and used standard.

2.2. Occitan

Occitan, or Oc language, is spoken in a large area in the south of France, in several valleys of Italy and the Aran valley in Spain. Occitan is not a unitary language, it has several varieties, organized in 6 large dialects (Auvernhàs, Gascon, Lengadocian, Lemosin, Provençau, and Vivaro-aupenc). It is a Romance language: as such, it shares many morpho-syntactic properties with other Romance languages (e.g., number and genre inflection marks on all the items of the noun phrase ; tense, person, number inflection marks on finite verbs). It is much closer to Catalan than to French: it is for example a null subject language as all the other Romance languages except French and oïl languages as Picard, Francoprovençal, Rheto-Romance languages and North-Italian dialects. Unlike French and Picard, Occitan has different verb inflection marks for each person. The morpho-syntactic level is also affected by variation across dialects (e.g., verbal inflection varies from one dialect to the other). As far as spelling is concerned, Occitan is not standardized as a whole but has two major spelling standards: the classical system, inspired from the troubadour's medieval spelling, and another system, closer to French conventions (Sibille, 2002).

2.3. Picard

The linguistic area of Picard includes the Hauts-de-France administrative region, and the Hainaut province in Belgium. Picard is an oïl language, which also belongs to the larger Romance language group. It differs from French with respect to several aspects. Word order can be different: for example, *il o foait keud assé* in Picard translates to *il a fait assez chaud* (*it has been quite hot*) –in French, the adverb is placed before the adjective, while in Picard the adjective is placed before the adverb. Even if Picard and Occitan are both Romance languages, Picard is closer to French concerning inflection. As in French and the other langues d'oïl, gender and number are mainly marked in Picard by means of determiners at the level of the noun phrase. Another device is shared with the other langues d'oïl: in the verbal phrase, a subject personal pronoun must be used in order to express personal rank in a deflective way. Picard does not have a unitary standardized spelling system and Picard texts can contain a dot which must not be considered as a word boundary, e.g., *lon.mint* (*for a long time)*, *erwet.tent* (*look*), *fin.mes* (*women*). More often than in French, words can also contain apostrophes – *c'min* (*path*) – and hyphens – *gardin-neux* (*gardeners*).

3. Elaboration of the Tagsets

As we wanted to exploit the proximity to better-resourced languages, an important issue is that of the tagset, i.e. the list of part-of-speech categories used for the manual and, afterwards, automatic annotation. One solution would have been to use the tagsets from annotated corpora and part-of-speech annotation tools for closely-related languages, such as the German TreeTagger or Stanford Tagger for Alsatian. However, these tagsets are usually very detailed: the German TreeTagger and StanfordTagger use a set of 54 tags,[2] the French TreeTagger identifies 33 tags[3] etc. Such level of detail is not necessarily needed and entails several drawbacks: higher cost for training annotators and reduced performance for the part-of-speech (POS) taggers, which have to discriminate between very similar categories. Furthermore, we wanted to create corpora annotated with the same standard tagset if possible, in order to facilitate the diffusion of the corpora, to be able to compare our experiments with state-of-the-art work and to enable comparison between the languages of the project. We thus chose to base the tagsets for Alsatian, Occitan and Picard on the universal POS tags defined in the context of the Universal Dependencies project (Nivre et al., 2016).[4] A first issue was to

[2] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/stts_guide.pdf

[3] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html

[4] http://universaldependencies.org/u/pos/index.html

Tag	Full name
ADJ	adjective
ADP	adposition
ADP+DET	**preposition-determiner contraction**
ADV	adverb
AUX	auxiliary
CCONJ	coordinating conjunction
DET	determiner
EPE	**epenthesis**
INTJ	interjection
MOD	**modal verb**
NOUN	common noun
NUM	numeral
PART	particle
PRON	pronoun
PROPN	proper noun
PUNCT	punctuation
SCONJ	subordinating conjunction
SYM	symbol
VERB	verb
X	other

Table 1: Common tagset. The tags which are not part of the Universal POS tags are in bold format.

evaluate this tagset with respect to our languages and check that it suits our needs.

3.1. Unified Tagset

The Universal POS tags version 2 contains 17 core part-of-speech categories. Yet, we introduced three additional tags: ADP+DET for contractions of a preposition and a determiner; MOD for modal verbs; and EPE for epenthesis.[5] Epenthesis is found in our corpora because oral phenomena are often preserved. These tags were added for ease of annotation –we could have used features with existing tags instead, but it was easier to have only one level of annotation– but can be projected to Universal POS tags: ADP+DET can be split into ADP and DET, MOD becomes AUX, and EPE becomes X. The resulting tagset is presented in Table 1.

3.2. Annotation Guidelines

For guiding the manual annotation, we tried to follow the Universal POS tag documentation as much as possible, but also considered the choices made for closely-related languages. Annotation guidelines were created for all three languages.

Alsatian The documentation for the Universal POS tags was the foundation for the guidelines (Bernhard et al., 2018). The descriptions of the STTS tags for German (Schiller et al., 1999) were consulted for some decisions. Alsatian grammars were also used (Jenny and Richert, 1984; Jung, 1983). Universal POS tags documentation was mostly followed, except for the three added tags mentioned before, and the FM tag for foreign words (see Table 2). Yet, some choices were distinct from Universal POS tags: for example, verb particles separated from their verb were tagged as PART, like in the STTS guidelines (*Er nùtzt/VERB mich üss/PART*, en: *He exploits me*).

[5] Addition of one or several letters to comply with the phonotactics of the language, for example in Alsatian: *fànga␣-n-/EPE ␣à drucka* (en: *begin to print*).

Tag	Full name
ADJ	adjective
ADP	adposition
ADV	adverb
APPART	preposition-determiner contraction
AUX	auxiliary
CONJ	coordinating conjunction
DET	determiner
EPE	epenthesis
FM	*foreign words*
INTJ	interjection
MOD	modal verb
NOUN	common noun
NUM	numeral
PART	particle
PRON	pronoun
PROPN	proper noun
PUNCT	punctuation
SCONJ	subordinating conjunction
SYM	symbol
VERB	verb
X	other

Table 2: Alsatian tagset. The FM tag is not part of the common tagset.

Tag level 1	Tags level 2	Full name
A	Af, Ao, Ak, Ai, As	adjective
C	Cc, Cs	conjunction
D	Da, Dd, Di, Ds, Dt, Dr, Dk, Dp	determiner
F		punctuation
I		interjection
N	Nc, Np, Nk	noun
P	Pp, Pd, Pi, Ps, Pt, Pr, Px, Pk	pronoun
R	Rg, Rx, Rp, Rq	adverb
S	Sp, Spda, Sd	preposition
V	Vm, Va	verb
X		residual

Table 3: Occitan tagset

Occitan The description of morpho-syntactic tags in the lexicon of inflected forms LOFLOC (Vergez-Couret, 2016) was adapted and extended to create annotation guidelines (Bras, 2018). The standard GRACE tagset (Rajman et al., 1997), which comes from the MULTEXT (Ide and Véronis, 1994) and EAGLES (von Rekowski, 1996) tagsets, was chosen, as it has been used for several similar annotated corpora for French and Catalan. We created a conversion script to project the GRACE tags onto Universal POS tags. Table 3 shows the two levels of this tagset (38 tags); the complete description of the tags is given in the guidelines (Bras, 2018).

Tag	Full name
ADJ	adjective
ADJIND	indefinite adjective
ADJPOSS	possessive adjective
ADP	preposition
ADPDET	preposition-determiner contraction or partitive
ADPLOC	prepositional locution
ADV	adverbs
CCONJ	coordinating conjunction
SCONJ	subordinating conjunction
DET	determiner
EPE	epenthesis
INTJ	interjection
NOUN	common noun
NOUNCCOMP	composed common noun
PROPN	proper noun
NUM	cardinal numbers
PRONDEM	demonstrative pronoun
PRONIND	indefinite pronoun
PRONPERS	personal pronoun
PRONPOSS	possessive pronoun
PRONREL	relative pronoun
PRONINT	interrogative pronoun
PART	particle or other function
PUNCT	punctuation
SYM	symbol
VERBINF	infinitive verb
VERBCONJ	conjugated verb
VERBPP	past participle verb
VERBPPR	present participle verb
X	other: loan word, typo, abbreviation...

Table 4: Picard tagset. The tags not found in our common tagset are in italics. The SYM tag was not encountered in the corpus

Picard The annotation guidelines are based on the Universal POS tags and the French Treebank annotation guidelines (Abeillé and Clément, 2003). A specific documentation was nevertheless created, to take into account the specifics of the Picard language, as well as to collect the possible issues and research topics (Martin et al., 2018). Some adjustments were made with respect to the Universal POS tags, namely subcategories, to obtain a better description of the Picard language (see Table 4). A conversion script was created to generate Universal POS tags from these more specific tags.

4. Constitution and Annotation of the Corpora

4.1. Corpus Selection

The first step was to collect texts with rights to distribute without restriction, so that the corpora could be made available. For some texts, we scanned printed texts and performed OCR.[6] One of the main challenges was to obtain resources which represent a large variety of textual genres and geolinguistic variants.

Alsatian The annotated corpus is composed of two main sources: WKP – Wikipedia articles from the Alemannic Wikipedia[7] and HRM – chronicles written in an information magazine published by the Haut-Rhin department (southern Alsace) General Council. In addition, two more specific genres were used for the annotator training phase: one excerpt from a theater play and some recipes. Given that the Alemannic Wikipedia contains articles written in several dialects from the Alemannic linguistic area, we only used articles which were specifically categorized as being written in Alsatian.

Occitan The RESTAURE project led to the finalization of the BaTelÒc text base (Bras and Vergez-Couret, 2016).[8] BaTelÒc is a wide coverage text collection, with written texts of literature (prose, drama and poetry) and other genres such as technical texts and newspapers, and embraces dialectal and spelling variations. 3.7 million words have already been gathered. All the texts in the base are encoded according to XML TEI P5 format. As the texts contained in BaTelÒc pose copyright issues, we selected 55 extracts of 60 words maximum from 17 texts from different authors to create the annotated corpus. We also added 8 texts from the online Occitan newspaper Lo Jornalet [9] with their kind permission. Lo Jornalet contains texts mostly in Lengadocian, and some in Gascon (all in Alibert's classical norm); we selected several texts from each dialect to complete the Occitan corpus. Finally, we also included one text from the Ciel d'òc online virtual library.[10]

Picard We benefited from the textual resources already collected within the PICARTEXT project,[11] a large literary resource. This text database is panchronic and has in its current version, which is still evolving, about 8 million tokens, taken from literature and ranging from the 17th century to the 21st century. The PICARTEXT base was tagged in XML according to the guidelines of the TEI P5. One of our objectives was to enrich this first textual database with literary texts of various genres (poetry, theater, tales, short stories, novels, etc.), of various time periods and taking into account the different varieties of Picard. We selected a subset of 32 texts according to the project criteria: diachronic diversity, variety of dialects and genres.

Table 5 provides the sizes of each corpus and sub-corpus.

4.2. Corpus Preparation

The selected texts were specifically prepared for the manual annotation process.

Alsatian Given the proximity of Alsatian to Standard German, and in order to facilitate the annotation work, the texts were pre-tagged using the TreeTagger (Schmid, 1994) for German and available lexicons for the Alsatian

[6] A specific work was performed on OCR for the three languages, but this work is out of the scope of this paper.

[7] http://als.wikipedia.org

[8] http://redac.univ-tlse2.fr/bateloc/

[9] https://www.jornalet.com/

[10] http://www.cieldoc.com/

[11] http://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/

Lang.	Source	Tokens	Types
Alsatian	WKP (13 doc.)	8,432	3,129
	HRM (6 doc.)	3,542	1,345
	recipes (1 doc.)	364	203
	theater (1 doc.)	232	140
	total (annotated)	12,570	4,497
Occitan	Jornalet (Lengadocian)	927	261
	Jornalet (Gascon)	2403	339
	BaTelÒc (Lengadocian)	5802	1455
	Ciel d'òc (Lengadocian)	538	219
	BaTelÒc (Gascon)	1630	514
	BaTelÒc (Provençau)	1359	390
	BaTelÒc (Lemosin)	462	170
	BaTelÒc (Vivaro-Aupin)	501	145
	BaTelÒc (Auvernhàs)	1386	296
	total (annotated)	15,008	3788
Picard	narrative	8,372	2,066
	poetry	1,924	585
	theater	862	304
	total (annotated)	11,158	2,564

Table 5: Reference corpus

dialects. In details, the following pre-processing steps were performed:

1. Tokenisation using a custom tokenizer for Alsatian (Bernhard et al., 2017).

2. Annotation with the TreeTagger for German in order to identify "unknown words", i.e. words which do not belong to the German TreeTagger lexicon and hence are typically Alsatian.

3. Automatic creation of a custom TreeTagger lexicon by looking up these unknown words in available Alsatian lexicons. Since there is a great amount of spelling variation in written Alsatian, we perform approximate lookup using a variant of the Double Metaphone phonetic algorithm adapted to Alsatian (Bernhard, 2014). This allows us to retrieve POS category information for the words even if they do not appear with exactly the same spelling in the lexicon (e.g. *'Sùnneblüem'* and *'Sunneblüem'*). In order to increase lexicon coverage, we also perform approximate lookup in German lexicons.

4. Transformation of Alsatian spellings for closed class words into their German equivalent in the texts using a custom correspondence dictionary (e.g., Alsatian *nìt* corresponds to Standard German *nicht*). We have shown in previous work that this improves the performance of the German TreeTagger when used for tagging Alsatian (Bernhard and Ligozat, 2013).

5. Second annotation with the German TreeTagger performed on these transformed texts. We provide TreeTagger with the custom lexicon containing suggested categories for unknown words.

6. Transformation of the result of this second annotation into the input format requested by the manual annotation tool, using a correspondence table between our POS tags and the German TreeTagger POS tags.

Occitan Pre-processing was also used:

1. We first used the POS tagger of the APERTIUM translation platform used in the Occitan/Spanish and Occitan/Catalan translators (Armentano I Oller, 2008) to tag an initial corpus in one dialect (Lengadocian). A specific tokeniser for Occitan and a specific inflectional lexicon (LOFLOC) were created to adapt the tagger to our needs, and APERTIUM tags were converted to GRACE tags with a specific script.

2. This first tagger's outputs were manually corrected on a subcorpus.

3. Then, a supervised machine learning tagger, Talismane (Urieli, 2013), was trained on the corrected corpus.

4. The rest of the corpus was annotated with Talismane.

5. Talismane outputs were manually corrected and used for further annotation (Bras and Vergez-Couret, 2014).

Picard We randomly extracted 30 lines excerpts from each of the selected texts. The first issue was tokenisation, which prompted the development of a tokenisation script for Picard. Bernhard et al. (2017) detail the specific issues of tokenisation for Picard, as well as the choices made. In contrast to Alsatian and Occitan, the Picard corpus was not pre-annotated.

4.3. Annotation Methodology

Alsatian The pre-tagged texts were manually corrected with the Analog tool (Lay and Pincemin, 2010). In addition to the POS tags, the annotators were also requested to provide a gloss (translation into French), the lemma, grammatical properties for verbs, nouns and adjectives, as well as location named entities but these further pieces of information will not be discussed in this paper. Overall, 6 persons took part in the annotation. One annotator (A1) annotated all the 21 documents. In order to measure inter-annotator agreement, two annotators (A2 and A3) annotated respectively 6 and 5 of the 21 documents. Finally, annotator A4 made the final decisions and corrections (adjudication) for all the 21 documents. A4 was helped by two experts in the Alsatian dialects (A3 and A6) to solve difficult issues and was also provided with correction proposals by annotator A5 for 7 of the documents annotated by A1. In addition, online resources were consulted for the adjudication: dictionaries (*Wörterbuch der elsässischen Mundarten* (Martin and Lienhart, 1899 1907)[12] ; *DWDS – Digitales Worterbuch der deutschen Sprache*[13]) and the Universal Dependencies version 2.0 German corpus accessed through the search interface by the University of Turku.[14]

[12] http://woerterbuchnetz.de/ElsWB/

[13] https://www.dwds.de/

[14] http://bionlp-www.utu.fi/dep_search

	A1		A4		# Tokens
	%	κ	%	κ	
A1			92.8	0.920	12,570
A2	84.1	0.824	84.6	0.830	2,135
A3	93.0	0.922	93.7	0.930	2,638

Table 6: Inter-annotator agreement for the Alsatian corpus.

	Beginning		Middle	
	%	κ	%	κ
O1 - O2	88.9	0.873	95.3	0.947
O1 - O3	93.8	0.929	94.7	0.940
O2 - O3	88.1	0.864	94.7	0.940
O1 - O2 - O3	87.0	0.889	92.3	0.942
#Tokens	370		169	

Table 7: Inter-annotator agreement for the Occitan corpus .

We measured the inter-annotator agreement (IAA) between annotators A1 to A3 and the final adjudication by A4. Agreement was measured in terms of percentage agreement and the Kappa coefficient κ (Cohen, 1960), computed with the `irr` R package.[15] Table 6 details these IAA values.
Overall, the IAA observed for A1 and A3 was better than for A2. Lower agreement is usually observed for rarer POS tags (e.g. INTJ, SYM, SCONJ). It should be noted that the documents annotated by A2 belong mostly to the training corpus (4 out of 6 texts) and these annotations were performed early in the annotation process. The annotation guide was substantially modified after this first annotation phase by introducing the APPRART, MOD and FM categories. Also, the EPE POS tag was introduced very late in the annotation process and was annotated as X before. In the future, some of the verifications performed by A4 could be automated (e.g. detect missing tags), and integrated to corpus pre-processing and annotation guidelines.

Occitan A first Lengadocian corpus was tagged with the first tagger described in 4.2. and then manually corrected by four annotators. We then trained the Talismane tagger to annotate a bigger corpus including texts in two dialects, Lengadocian and Gascon (Vergez-Couret and Urieli, 2014). This corpus was manually corrected by two annotators. Then we trained Talismane again in order to be able to pre-annotate the corpora including the 6 dialects of the Occitan language. The annotations were corrected by three annotators (O1, O2 and O3) using the Analog tool in order to get the final corpus. Inter-annotator agreement was measured at two different time points: at the beginning and at the middle of the manual annotation phase (see Table 7). The agreement has improved over the time, in particular for annotator pairs O1 - O2 and O2 - O3.

Picard The Picard corpus was annotated in a csv format. The texts were manually tokenised, and each token was annotated with its part-of-speech, its lemma and its French translation in context. Three Picard speakers worked on the manual annotation. A first manual annotation of 20 texts was performed by annotator P1, and then discussed and modified if necessary with annotator P2. The issues detected in this first step were then discussed with the research team in order to adapt the tagset and the guidelines. The remaining annotations were made by P2. All annotations were reviewed by annotators P2 and P3. The annotation phase took place during a period of about 11 months, and each step required additional research: tokenisation requires morphological and grammatical studies taking into account the Picard variety since tokenisation rules differ according to the language variety. The issues of lemmas and translations are interdependent. Moreover, Picard dictionaries are not comprehensive, and a translation for a text in a particular Picard variety could be found in a dictionary of another Picard variety. Finding the right translation thus often required searching in all available dictionaries. Since P2 performed most of the annotation, we measured intra-annotator agreement for several versions of the corpus.[16] For the June 2016 and January 2017 versions, $\kappa = 0.922$ with a percent agreement of 92.9%, the differences being mostly caused by typos and changes in the tagset. Between the January 2017 and the July 2017 version, $\kappa = 0.784$ with a percent agreement of 80.2%. These lower figures are explained mostly by the more important changes in the tagset, the figures for stable tags remaining higher than 0.9 (0.967 for the ADP tag, 1.000 for the PROPN tag...). Finally, a verification script was applied to the corpus to check its coherence: if a word is labeled with different tags, whose distribution is very uneven, the contexts (i.e. preceding and following tag) are compared and if a same context leads to different tags, the annotation is checked manually.

4.4. Resulting Resources and Dissemination

Token	French	Lemma	Tag	*English*
Spàrichle	asperge	Spàrichel	NOUN	asparagus
ìn	dans	ìn	ADP	in
e	une	e	DET	a
Sìbb	passoire	Sìbb	NOUN	sieve
üss	de	üss	ADP	of
Metàll	métal	Metàll	NOUN	metal
màche	mettre	màche	VERB	put

Table 8: Annotation example for Alsatian (some additional annotations are not presented) for the sentence.

Token	Lemma	Tag level 1	Tag level 2	*English*
Los	lo	D	Da	the
cavals	caval	N	Nc	horses
èran	èsser	V	Vm	were
luènh	luènh	R	Rg	far away
.	.	F		

Table 9: Annotation example for Occitan.

[15] `https://cran.r-project.org/package=irr`. Authors: Matthias Gamer, Jim Lemon, Ian Fellows Puspendra Singh.

[16] For one text of the corpus only, because changes were made in the text excerpts and tokenization, which makes it difficult to perform a completely automatic evaluation.

Token	Tag	Lemma	French	*English*
I	PRONPERS	i	il	he
avot	VERBCONJ	avoir	avait	had
fauqu'	VERBPP	fauquer	coupé	cut
chés	DET	euch	les	the
projecteurs	NOUN	projecteur	projecteurs	spotlights
qu'	PRONREL	qui	qui	that
is	PRONPERS	i	ils	they
illuminottent	VERBCONJ	illuminoter	éclairaient	lit
eul	DET	euch	la	the
fosse	NOUN	fosse	fosse	pit
.	PUNCT	.	.	.

Table 10: Annotation example for Picard

Table 8 shows an annotation example for Alsatian, Table 9 for Occitan, and Table 10 for Picard. Each line represents a token, and the colums contain the different annotations (POS tag, lemma, French translation). The columns with the English translations are not available in the corpora and are provided for the sake of readability.

The annotation guidelines and the corpora are available for all three languages on the Zenodo platform, in the RESTAURE project community (see Section 9. for the corpus list).[17]

5. Related Work

Creating annotated corpora for under-resourced languages presents several difficulties. First, it requires assembling a large textual corpus, which can be a challenge for these languages which have few electronic resources. Work on part-of-speech tagging for under-resourced languages is often based on parallel corpora, following (Yarowsky et al., 2001), but there are no such existing electronic corpora for the three languages considered.

Then, a tagset has to be created or adapted to the language, which requires linguistic expertise. Finally, the annotation also requires annotators with linguistic expertise. Crowdsourcing can be used for part-of-speech annotation (Hovy et al., 2014), and was even used for Alsatian (Millour et al., 2017). Yet, crowdsourcing necessitates an adapted platform, and communication to possible speakers, who for example in the case of Picard, are rare. A possible direction for POS tagging could be to create a minimum tag dictionary for the most frequent word types, such as used by (Garrette and Baldridge, 2013). This kind of approach still requires a test corpus to evaluate the tagger; and the performance remains low compared to more resourced languages.

6. Conclusion

We have presented our methodology for producing corpora with POS annotations for three regional languages of France, namely Alsatian, Occitan and Picard. The tagsets are based on an extended version of the Universal POS tags, with some language-specific additions to account for particular linguistic phenomena. The annotation guidelines as well as the manually annotated corpora are freely available. We plan to use these corpora to develop part-of-speech taggers accommodating the spatial variation encountered in the three languages.

[17] https://zenodo.org/communities/restaure/

7. Acknowledgements

This work was supported by the French "Agence Nationale de la Recherche" (ANR) through the RESTAURE project (no.: ANR-14-CE24-0003). We are grateful to the annotators: Clément Dorffer, Gwendoline Hollner, Aurélie Abadie, Louise Esher, Sébastien Gonzales. We would also like to thank those who kindly permitted us to include their texts in our corpora: Yves Bisch, Conseil départemental du Haut-Rhin, OLCA, Laurenç Gosset and l'Institut d'Estudis Occitans, Tricio Dupuy and Ciel d'Òc, Sèrgi Viaule, Eric Chaplain and Editions des Régionalismes, Maurici Romieu and Reclams, Benaset Dazeas and Lo Congrès Permanent de la lenga occitana.

8. Bibliographical References

Abeillé, A. and Clément, L. (2003). Annotation morphosyntaxique. Les mots simples - Les mots composés. Technical report, LLF, Université Paris 7.

Armentano I Oller, C. (2008). Traduction automatique occitan-catalan et occitan-espagnol: difficultés affrontées et résultats atteints. In *IXème Congrès International de l'Association Internationale d'Etudes Occitanes*, Aachen.

Bernhard, D. and Ligozat, A.-L. (2013). Hassle-free POS-Tagging for the Alsatian Dialects. In Marcos Zampieri et al., editors, *Non-Standard Data Sources in Corpus Based-Research*, ZSM Studien, pages 85–92. Shaker. Volume 5.

Bernhard, D., Todirascu, A., Martin, F., Erhart, P., Steiblé, L., Huck, D., and Rey, C. (2017). Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard. In *DiLiTAL 2017*, Actes de l'atelier " Diversité Linguistique et TAL ", pages 14–23, Orléans, France, Jun.

Bernhard, D., Erhart, P., Huck, D., and Steiblé, L., (2018). *Part-of-Speech Annotation Guidelines for the Alsatian Dialects*. DOI: 10.5281/zenodo.1171925.

Bernhard, D. (2014). Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian. In *Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014)*, pages 23–29, Reykjavík, Iceland, May.

Bras, M. and Vergez-Couret, M. (2014). Annotation morphosyntaxique d'un corpus de textes occitans: l'expérience de BaTelÒc. In *XIème Congrès de l'Association Internationale d'Etudes Occitanes*, Lerida, Spain, June.

Bras, M. and Vergez-Couret, M. (2016). BaTelÒc: A text base for the Occitan language. In Vera Ferreira et al., editors, *Language Documentation and Conservation in Europe*, pages 133–149. Honolulu: University of Hawaï Press.

Bras, M., (2018). *Part-of-Speech Annotation Guidelines for the Occitan Language*, February. DOI: 10.5281/zenodo.1173113.

Cerquiglini, B. (1999). Les langues de la France. Technical report, Rapport au Ministre de l'Éducation Nationale, de la Recherche et de la Technologie et à la Ministre de la Culture et de la Communication.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Garrette, D. and Baldridge, J. (2013). Learning a Part-of-Speech Tagger from Two Hours of Annotation. In *HLT-NAACL*, pages 138–147.

Hovy, D., Plank, B., and Søgaard, A. (2014). Experiments with crowdsourced re-annotation of a POS tagging data set. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 377–382.

Huck, D. (to appear). Dialectal speech in Alsace / Alsatian. In Hans Boas, et al., editors, *Varieties of German Worldwide*, volume 1. Cambridge University Press, Cambridge.

Ide, N. and Véronis, J. (1994). Multext (Multilingual Tools and Corpora). In *14th Conference on Computational Linguistics (COLING'94), Kyoto, Japan*.

Jenny, A. and Richert, D. (1984). *Précis pratique de grammaire alsacienne: en référence principalement au parler de Strasbourg*. Librairie Istra.

Jung, E. (1983). *Grammaire de l'alsacien: dialecte de Strasbourg avec indications historiques*. Oberlin.

Lay, M.-H. and Pincemin, B. (2010). Pour une exploration humaniste des textes: AnaLog. In *Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, pages 1045–1056, Sapienza University of Rome.

Martin, E. and Lienhart, H. (1899-1907). *Wörterbuch der elsässischen Mundarten*, volume 1-2. KJ Trubner.

Martin, F., Rey, C., and Reynés, P., (2018). *Part-of-Speech Annotation Guidelines for Picard*. DOI: 10.5281/zenodo.1173428.

Millour, A., Fort, K., Bernhard, D., and Steiblé, L. (2017). Vers une solution légère de production de données pour le TAL: création d'un tagger de l'alsacien par crowdsourcing bénévole. In *Actes de Traitement Automatique des Langues Naturelles (TALN)*.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, may.

Rajman, M., Lecomte, J., and Paroubek, P. (1997). Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique. Technical report, EPFL & INaLF. GRACE GTR-3-2.1.

Schiller, A., Teufel, S., Stöckert, C., and Thielen, C. (1999). Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, Universität Stuttgart & Universität Tübingen.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49.

Sibille, J. (2002). Ecrire l'occitan : essai de présentation et de synthèse. In Dominique Caubet, et al., editors, *Codification des langues de France*, pages 17–37. L'Harmattan, Paris, France.

Soria, C., Mariani, J., and Zoli, C. (2013). Dwarfs sitting on the giants' shoulders—how LTs for regional and minority languages can benefit from piggybacking major languages. In *Proceedings of XVII FEL Conference*, pages 73–79.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, Université de Toulouse II le Mirail.

Vergez-Couret, M. and Urieli, A. (2014). Pos-tagging different varieties of Occitan with single-dialect resources. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages Varieties and Dialects*. Association for Computational Linguistics and Dublin City University.

Vergez-Couret, M. (2016). Description du lexique Loflòc. Research report, CLLE-ERSS, Apr.

von Rekowski, U. (1996). ELM-FR: A typed French incarnation of the EAGLES-TS – Definition of Lexical Specification and Classification Guidelines. Technical report, GSI-Erli.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8.

9. Language Resource References

Bernhard, D., Erhart, P., Huck, D., and Steiblé, L. (2018). Annotated Corpus for the Alsatian Dialects. DOI: 10.5281/zenodo.1170129.

Bras, M., Esher, L., Sibille, J., and Vergez-Couret, M. (2018). Annotated Corpus for Occitan. DOI: 10.5281/zenodo.1182949.

Martin, F., Rey, C., and Reynés, P. (2018). Annotated Corpus for Picard. DOI: 10.5281/zenodo.1172576.