



HAL
open science

Collecting and annotating corpora for three under-resourced languages of France: Methodological issues

Delphine Bernhard, Anne-Laure Ligozat, Myriam Bras, Fanny Martin, Marianne Vergez-Couret, Pascale Erhart, Jean Sibille, Amalia Todirascu, Philippe Boula de Mareüil, Dominique Huck

► To cite this version:

Delphine Bernhard, Anne-Laure Ligozat, Myriam Bras, Fanny Martin, Marianne Vergez-Couret, et al.. Collecting and annotating corpora for three under-resourced languages of France: Methodological issues. *Language Documentation & Conservation*, 2021, 15, pp.316-357. hal-03273196

HAL Id: hal-03273196

<https://hal.science/hal-03273196>

Submitted on 29 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Collecting and annotating corpora for three under-resourced languages of France: Methodological issues

Delphine Bernhard
Université de Strasbourg,
LiLPa, UR 1339

Pascale Erhart
Université de Strasbourg,
LiLPa, UR 1339

Anne-Laure Ligozat
LISN, ENSIIE

Jean Sibille
CLLE, UMR 5263,
Université Toulouse Jean Jaurès, CNRS

Myriam Bras
CLLE, UMR 5263,
Université Toulouse Jean Jaurès, CNRS

Amalia Todirascu
Université de Strasbourg,
LiLPa, UR 1339

Fanny Martin
Habiter Le Monde, EA 4287,
Université de Picardie Jules Verne

Philippe Boula de Mareüil
LISN, CNRS

Marianne Vergez-Couret
FoReLLIS, EA 3816,
Université de Poitiers

Dominique Huck
Université de Strasbourg,
LiLPa, UR 1339

In contrast to French, the vast majority of regional languages of France can be considered as under-resourced. In this article, we present the results of a research project aiming to produce annotated resources for three regional languages of France: Alsatian, Occitan, and Picard. These languages cover three different language families (Germanic and two subfamilies of Romance, Oil and Oc languages) and different sociolinguistic situations. Yet, they all face issues common to many under-resourced languages: lack of human and financial resources and presence of geolinguistic variation. The originality of this project is that it brought together researchers from different fields (sociolinguistics, descriptive linguistics, dialectology, natural language processing, digital humanities) to work together towards the common goal of developing annotated corpora for Alsatian, Occitan, and Picard. This created a favorable and stimulating working environment which could not have been achieved had different research groups worked independently, each on a single language. This article details the annotation process, with a special focus on the delimitation of the tokens and the definition of the part-of-speech tags.

1. Introduction The development of natural language processing (NLP) resources and systems for under-resourced minority languages requires study of these languages from the perspective of computational tools. In comparison to more traditional micro- and macrolinguistic studies, NLP imposes specific constraints related to the limitations and scope of state-of-the-art methods: it is necessary to classify complex phenomena into a limited set of categories or make consistent decisions for large quantities of extremely variable data. This leads to new ways of thinking about the languages under consideration, in order to provide linguistic descriptions, which are both implementable and operable. While well-described languages are the subject of many detailed linguistic studies, under-resourced languages usually lack complete, thorough, and up-to-date descriptions of their morphological and syntactic systems. Furthermore, they are often characterized by graphical variation and the lack of a standard spelling. In many cases, spelling is not normalized and reflects geolinguistic variation. Typical phenomena involve variations in pronunciation, as well as morphological variations, in which inflected forms or derived forms vary by locality, or lexical variations, in which different lexemes are used to refer to the same concept. Moreover, the absence of standard spelling systems leads to interpersonal variation, in which each writer chooses her/his own spelling convention. This phenomenon is sometimes observed for neologisms in dominant languages, before they stabilize; it is, however, very marginal in comparison to what is observed in regional languages.

In this article, we describe the linguistic issues raised during the process of annotating corpora for three regional languages of France: Alsatian, Occitan, and Picard. This work was performed in the context of the RESTAURE¹ project, whose goal was to develop resources and tools for these under-resourced French regional languages. We focus in particular on two important aspects: the delimitation of tokens and the definition of the part-of-speech (POS) tags. We also detail strategies which can be implemented to make the annotation process as efficient as possible, in the context of limited human and financial means: automatic pre-annotation, automatic post-annotation verification, and addition of layers of annotation to produce other resources such as bilingual lexicons or toponym lexicons.

2. State-of-the-art before the project When we submitted the RESTAURE project, we identified the needs for digital resources and tools for Alsatian, Occitan, and Picard. The situation of the different languages before the start of the project is summarised in Table 1. This table shows many deficits and a heterogeneous situation in terms of experiences and needs. It also reflects the state-of-the-art which was established in 2014 in the *Inventaire des ressources linguistiques des langues de France* (Leixa et al. 2014) and which indicates a low volume of linguistic resources for Alsatian (8/10)² and Oïl languages (7/10), including Picard, and a medium volume for Occitan (5/10), which is at the same level as Breton and below Basque (4/10), Catalan (4/10), and French (3/10).

¹<http://restaure.unistra.fr/>

²A rating of 1/10 indicates an excellent language resource base. Conversely, a rating of 10/10 indicates a weak or non-existent base.

Table 1. Tools and resources available before the project.

Resource / tool	Alsatian	Occitan	Picard
Unannotated corpus	∅	BaTelÒc experimental database (Bras & Thomas 2008)	PICARTEXT ³
Annotated corpus	∅	∅	∅
Lexicons and dictionaries	(Bernhard 2014)	Dico d'òc (Congrès Permanent de la Lenga Occitana) under construction	∅
Tokeniser	∅		∅
POS tagging	(Bernhard & Ligozat 2013a; Bernhard & Ligozat 2013b)	(Vergez-Couret 2013; Vergez-Couret & Urieli 2014)	∅
Syntactic analysis	∅	∅	∅
Ranking established by (Leixa et al. 2014)	8	5	7

It should be noted that the notions of corpus, lexicon, or dictionary used in the RESTAURE project refer to resources which can be directly used for work in natural language processing. Thus, texts or lexicons that would be available in unstructured or semi-structured digital formats (web pages, PDF, word processing documents) do not strictly speaking constitute resources that can be directly exploited without preparatory work: extraction and tagging of elements of interest, removal of unnecessary information, transformation to a standard format such as the Text Encoding Initiative, TEI (TEI Consortium 2020). In addition, some existing resources, such as Dico d'òc, contain structured data that can be directly exploited by natural language processing tools but are not necessarily available in their entirety for research because of intellectual property rights. We therefore have a more restrictive definition than that used in the Leixa et al. (2014) report. Furthermore, we are, for the time being, only interested in written resources, while Leixa et al. also include oral resources and tools.

3. Description of the languages All three regional languages presented here are languages of France listed in Bernard Cerquiglini's 1999 report (Cerquiglini 1999), which establishes the list of regional languages of France under the definition of the European Charter for Regional and Minority Languages. A common characteristic is that they are not standardized and present dialectal and orthographic variation. We present each language with its linguistic properties, its history, and its current situation; we then detail the issues with orthography and the consequences for NLP tools.

³<https://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/>

3.1 Alsatian The Alsatian dialects are spoken in northeastern France and are part of the High German (vs. Low German) dialects, which are subdivided into Central German and Upper German. The majority of the Alsatian dialects belong to (Low) Alemannic, an Upper German dialect. High Alemannic German is found in the south of the region, and a small part of the dialect area (northwest and northeast) belongs to Central German Franconian (see Figure 1).



Figure 1. Dialectal domain of Alsace and Moselle (map created by Anne Horrenberger in 2017 to illustrate our publications)

The Alsatian dialects have a long history which dates back to the linguistic changes brought about in the region by the Alemanni and the Franks as early as the 6th century (Huck 2015). The geographic region now called Alsace was progressively integrated into France much later, during the 17th century. Despite this, the middle and lower classes went on using the local German dialects for everyday communication. The situation changed only after World War II, and the use of French as the dominant language of communication in all everyday activities accelerated in the last third of the 20th century. This period marked the gradual decline of the Alsatian dialects: in a study carried out in 2012, only about 43% of the surveyed Alsatian people declared that they still had some knowledge of the dialects (EDinstitut & OLCA 2012).

Throughout their history, the Alsatian dialects have always been used mainly orally, in informal situations. For writing, either French or German was preferred, depending on the territorial affiliation of Alsace. There are however some instances of written Alsatian, from the second half of the 17th century onwards (Wackenheim 1993). Alsatian literature is considered to have begun with the verse comedy *Der Pflingstmontag* by Jean-Georges-Daniel Arnold in 1816.

Since then, the focus of literary production in the Alsatian dialects involved two main text genres: plays (mostly comedies) and poetry (Wackenheim 1994, 1997, 1999, 2003). Some other genres are also represented, but are much rarer: prose poetry, songs, nursery rhymes, tales, translations, and adaptations of works in other languages. Linguistic works focusing on the description of the dialects have also been written: dictionaries, glossaries and grammars. Interestingly, texts in prose are quite rare (with the notable exception of Marie Hart's writings), and French or German are usually preferred in this case.

Interest in Alsatian dialectal languages began to develop in the more general context of the German "romantic" movement in the 19th century which promoted and exalted popular culture (storytelling, oral literature) and which conceived of language as a mirror of the people. The first lexical collections in Alsace were carried out by the folklorist Auguste Stoeber (1808–1884), who also participated in the surveys launched by Jakob and Wilhelm Grimm for the development of their *Deutsches Wörterbuch*. These lexical collections form the basis of Martin and Lienhart's dictionary (*Wörterbuch der elsässischen Mundarten*), which remains at present the most complete but also the most dated reference available (Martin & Lienhart 1899–1907).

Modern dialectology really began at the time when Alsace and part of Lorraine were integrated into the German Empire (1870–1918), with Georg Wenker's dialectal collection, in the form of a correspondence survey for the constitution of a *Sprachatlas des Deutschen Reichs* (1876–1887). After the Second World War, a linguistic atlas of the Alsace region was launched, under the direction of Ernest Beyer (Beyer & Matzen 1969; Bothorel-Witz et al. 1984). The data of this atlas have been digitized⁴ and still serve as a reference today for the work of dialectologists. These atlases are supplemented by linguistic descriptions (Brunner et al. 1985; Huck 1999; Beyer 1963).

⁴<http://ala.u-strasbg.fr>

Like all dialectal areas, the Alsatian dialects are characterized by spatial variation, on the phonological as well as on the lexical level, which is the main characteristic of a dialect. While there are recent proposals for spelling conventions (for example ORTHAL (Zeidler & Crévenat-Werner 2008)), the writing of Alsatian dialects is not strictly standardized and uses are therefore very diverse. Moreover, the Alsatian dialects are not taught at school within the national education system: Standard German is preferred, even in bilingual education. There are some voluntary schools though, where the Alsatian dialects are taught alongside German and French (ABCM-Zweisprachigkeit association), but only in their oral form, since standard German is used for the teaching of writing.

Since the second half of the 20th century, the vast majority of writers and speakers of Alsatian have had a plurilingual repertoire, with French becoming increasingly dominant and driving them to use, in Alsatian, linguistic strategies and calques from French. According to Huck (to appear):

At the turn of the millennium, approximately 500,000 people declared knowing how to speak Alsatian and approximately 300,000 people declared having an active knowledge of German. However, the transmission of dialect speech was declining rapidly insofar as, overall, studies and surveys tend to show that less than 10% of the younger generation (under 18 years old around 2010) still had active knowledge of the dialect, whatever the definition of “active knowledge”. (Huck to appear)

The last survey to date (IFOP 2020) shows that only 5% of the respondents declare “Alsatian” as their main language, while 25% declare themselves “bilingual” (Alsatian-French). 70% declare French as their only language. Despite their lack of institutional support and an undeniable decline in their practice, Alsatian dialectal speeches still seem to be invested with symbolic functions which have led to new uses of “Alsatian”, no longer oral, but written, on the Internet and social networks (Erhart 2020).

The fundamental morphosyntactic characteristics of Alsatian are not greatly affected by this phenomenon. They are common to all Alsatian dialects and are very similar to those of standard German. The linearization of verbal utterances presents significant continuity between Alsatian and German. In continuous linearization, three fields are identifiable: thematic (what is being talked about), phematic (positioning of the speaker), and rhematic (what is said about what is spoken of). The finite form of the verb is found in this last field. In discontinuous linearization (the most frequent), a part of the rhematic field is positioned before the thematic field: the personal form of the verb. This is why, in assertive or interrogative sentences, the finite part of the verb is syntactically in the second position. For the noun phrases, the order ‘determiner’ / (adjective) / ‘nominal base’ is obligatory in Alsatian as in German. Finite verbs are marked with exponents of tense, mood, and person-number, and noun phrases are marked for number, case, and gender. The tense and mood system is much simpler than in standard German and there is only one person-number

marker for the plural persons (Huck to appear; Huck 1999). However, the surface form of the morphemes can present intradialectal variation.

3.2 Occitan Occitan is a Romance language spoken in southern France (except in the Basque and Catalan areas), in several valleys of the Italian Piedmont, and in the Val d’Aran in Spain. The northern limit of the Occitan-speaking area runs a few kilometers to the north of the cities of Bordeaux, Guéret, Montluçon, Valence, and Briançon.

Occitan belongs to the Gallo-Romance group of Romance languages, together with standard French and “langues d’oil”, Francoprovençal, and Catalan. Occitan is closer to Catalan than to French and, according to Bec (1970), forms with Catalan a subgroup called *occitano-roman*.

Occitan has several varieties, organized into dialects. The most widely accepted classification proposed by Bec (1995) includes Auvernhàs (in French: *Auvergnat*), Gascon, Lengadocian (*Languedocien*), Lemosin (*Limousin*), Provençau (*Provençal*), and Vivaroaupenc (*Vivaro-alpin*) (see Figure 2), but these dialects are not homogeneous: they form a continuum with areas of greater or lesser variation. Along the northernmost part of the Occitan area, the so-called Creissent (in French: *Croissant*) forms an area of interference between Occitan and Oil varieties.

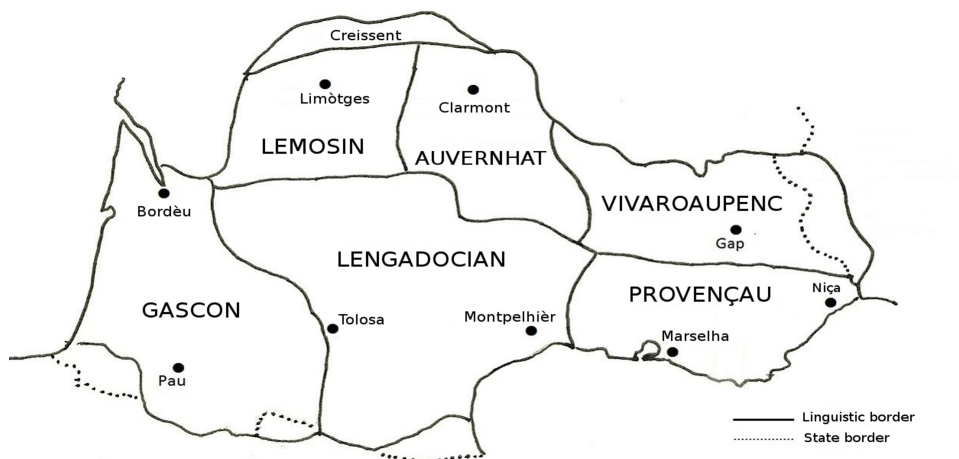


Figure 2. Linguistic areas for Occitan dialects.

The term *lingua occitana* and the adjective *occitanus* appear in Latin texts from the 13th century, but Occitan has often been designated by other names: Provençal, Gascon, Lemosin – these are terms that refer to a part of Occitan territory but which could also be used to designate the language as a whole –, (*la*) *langue romane*, and also *patois*, which is originally a pejorative term but is not seen as pejorative for all speakers.

Occitan is not standardized as a whole, but regional or local standards – which function in a convergent way – are used for teaching, literature, press, and the Internet. However, these standards are generally not in use among native speakers.

The usage of Occitan declined considerably after World War II. Family transmission stopped at the beginning of the 20th century in large towns and after 1945 in the countryside. The number of Occitan speakers nowadays is hard to assess. Around 1920, the linguist Jules Ronjat (Ronjat 1930) estimated that there were about 10 million speakers, while around 1960 the figure was thought to be between 2 and 3 million. According to the data of *Enquête sur l'histoire familiale* related to the 1999 census, the number of active speakers can be calculated to be around 700,000 (Sibille 2010) at the time of the survey. From data concerning the district (*département*) of Hautes-Pyrénées, Bernissan (2012) estimates, by extrapolation, the number of “real speakers” at around 100,000, but we think that this figure is an underestimate due to very restrictive criteria for defining the notion of speaker and to some other distortions (for example, Bernissan only counts speakers who remained in the area, but many villages were deserted because of the rural exodus). Nowadays, the vast majority of native speakers are people born in rural areas before 1945 and the number of neo-speakers is outstripped by the loss of native speakers: thus, the decrease of speakers is not progressive but exponential.

Two sociolinguistic surveys were carried out in 2008 and 2010 in the old Aquitaine and Midi-Pyrénées administrative Regions.⁵ According to these surveys, there were more than 750,000 speakers of Occitan in these two Regions:⁶ 10% of the population in Aquitaine consisted of native and fluent speakers or speakers able to have a simple conversation in Occitan, while in Midi-Pyrénées, the percentage of speakers was estimated to 18% with 4% of native or fluent speakers and 14% of speakers with an average competence.

A new sociolinguistic survey was carried out by the also new *Ofici Public per la Lengua Occitana* (OPLO,⁷ Public Office for Occitan Language) in the also new administrative regions *Nouvelle Aquitaine* and *Occitanie*, and in Val d’Aran in Spain, altogether representing a larger area than the two ex-Regions mentioned above.⁸ The number of Occitan speakers (i.e. speakers declaring to speak Occitan without difficulty or enough to hold a simple conversation) is evaluated to 7%, i.e. approximately 600,000 speakers. Considering the numbers given above, the loss in the last ten years is important. Nonetheless, 92% of respondents come out in favor of preservation and development of Occitan language.

Occitan has been written continuously for at least a thousand years, but with varying fortunes. In the Middle Ages, it was a great language of culture at the origin of European lyric poetry. Throughout the Middle Ages it was also used as a legal and administrative language in competition with Latin. The first literary texts date from the year 1000, the earliest being religious poems. From the late 11th century, troubadour poetry flourished, and its influence reached throughout Europe. After the

⁵Those two surveys were *Résultats synthétiques de l'étude sociolinguistique sur la présence, les pratiques et les perceptions de la langue occitane en Aquitaine 2008* and *Résultats synthétiques de l'étude sociolinguistique 'Présence, pratiques, et perceptions de la langue occitane en région Midi-Pyrénées'* 2010. See Région Midi-Pyrénées (2010).

⁶<https://www.ofici-occitan.eu/oc/los-enjocs/>

⁷<https://www.ofici-occitan.eu/>

⁸<https://www.ofici-occitan.eu/oc/restitucion-de-las-resultas-de-lenquesta-sociolinguistica/>

13th century, Occitan literature entered a period of slow decline, but only the field of literature was affected: Occitan was not yet threatened as a language of everyday life, or even as the language of legal and administrative writing. In the 16th century, with the centralization of power, French became the language of administration and elites. In the 17th century, a period of diglossic bilingualism began and Occitan literature became confined to genres considered minor or lower-class.

In the 19th century, Occitan experienced a literary renaissance driven by the *Felibrige*, an organization founded by Frederic Mistral. The primary focus of the renaissance was on poetry, but throughout the 19th and 20th century Occitan gradually reclaimed all fields of literature. Contemporary literature, by writers including Mistral, Boudou, Max Rouquette, and Manciet, has been translated into other languages. Literary production also includes plays. Although much less widely present in public life than it was before World War II, Occitan is nowadays present in newspapers, on the internet, on the radio and television, and in some state schools and universities. Non-governmental organizations maintain and promote Occitan: these include the *Congrès Permanent de la Lengua Occitana*, the *Felibrige*, the *Institut d'Estudis Occitans*, the network of immersive schools *Calandretas*, and the adult education institutes *Centres de Formacion Professional Occitans*.

As for spelling, medieval conventions were forgotten during the 16th century. In the mid-19th century, the renewal of literature led to the emergence of a new convention named after Frederic Mistral. However, the period from the 16th to the 19th century is characterized by a plethora of non-standard individual spellings, which are all close to French conventions (Sibille 2002). From the mid-20th century on, a new spelling convention called “classical”, based on medieval conventions, has been developed by the grammarian Louis Alibèrt (1935–1937, 1976). The aim is to minimize the dialectal differences, while dialectal particularities remain (Sibille 2002), cf. §3.4 for more details.

Many dialectological studies describe the Occitan area: after the *Atlas Linguistique de la France* (Gilliéron & Edmont 1902–1910), nine regional atlases were produced in the second part of the twentieth century, collecting a large amount of linguistic data (now gathered together by the Thesoc project;⁹ Oliviéri et al. 2017). A wide range of bilingual dictionaries is also available, some of which are now available online via the dico d'Òc application of *Lo Congrès Permanent de la Lengua Occitana*.¹⁰

As a Romance language, Occitan is characterized by several morphosyntactic properties. It is a null subject language with tense, person, and number inflection marks on finite verbs for each person. The morphosyntactic level is also subject to dialectal variation (e.g., verb inflection varies from one dialect to the other). Many dialects (Lengadocian, Gascon, part of Vivaroaupenc, and, in a lesser extent, Lemosin and Auvernhàs) mark number and gender inflection on all components of the noun phrase. Unlike contemporary French, Occitan maintains the use of the preterit (*passé simple*), which contrasts with the perfect tense (*passé composé*), and the use of the imperfect subjunctive, even in oral colloquial speech.

⁹<http://thesaurus.unice.fr/index.html>

¹⁰<https://www.locongres.org/>

3.3 Picard Picard is a langue d’oïl, belonging to the group of Romance languages. The Picard linguistic area includes the current Hauts-de-France region (from the 2014 territorial reform) – namely the former region of Nord-Pas-de-Calais, part of the Picardy region, with the exception of the southeast of the administrative territory – and the Belgian province of Hainaut, where Picard is officially recognized as an endogenous regional language, by a decree of the French Community of Belgium in 1990 (see Figure 3). Recent works (Forlot & Martin 2014; Martin 2015; Martin & Forlot 2016) indicate that the linguistic area is today reduced mainly to the south of the territory.

The Picard language has a long dialectological tradition, which made it possible to establish, in 1957, the layout of the linguistic area, in connection with the surveys previously carried out (Dubois 1957) and the publication of a linguistic and ethnographic Picard atlas (Carton & Lebègue 1989–1998; Carton et al. 1997). For delimiting the linguistic area, we refer to Figure 3, established by René Debrie in 1957, based on the dialectological studies conducted in the 1950s.

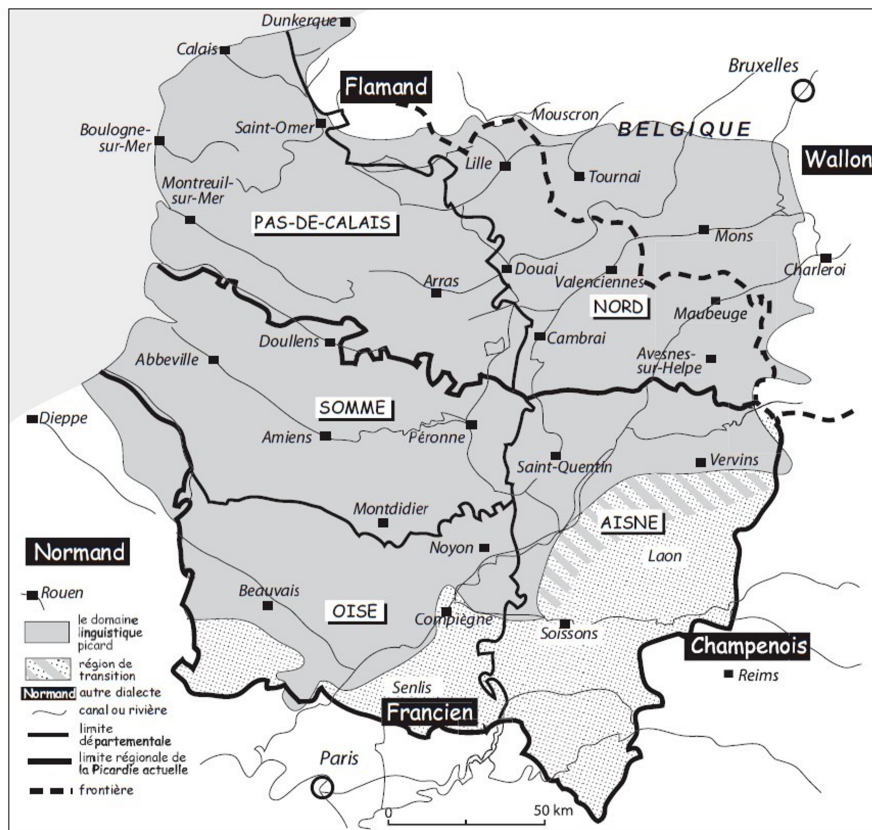


Figure 3. Picard linguistic area. Map established by René Debrie according to Raymond Dubois’s data, and drawn by Joëlle Désiré (Amiens, Université Jules Verne, 1985).

To pursue this reflection on the linguistic area, recent studies (Forlot & Martin 2014; Martin 2015) allow us to state that different focal areas for language practice have emerged in the Picardy region, in particular: west of Abbeville, around Amiens, around Valenciennes, around Lille, the region of Arras, around Beauvais, the western buttresses of the Ardennes massif, and in the region of Tournai in Belgian Hainaut. In the heart of the Picard linguistic area, we can therefore clearly see varieties of Picard. These are defined in relation to each other (graphically, syntactically, and in terms of identity): they make sense for the speakers, but also for the writers who cultivate distinctiveness within the Picard linguistic area. It is also with a view to countering the codification and standardization of Picard that we are seeing the emergence of such focal areas, which we believe cultivate and affirm their differences within this area (Martin 2015).

In the history of the linguistic study of Picard, there were many debates in the 1960s and 1970s about standardization and spelling. However, to date, no standardization has been undertaken for the entire Picard linguistic domain. Moreover, this situation is nowadays considered a conscious freedom of choice, countering standardization. Although Picard does not exist as a standardized language, it has been written for several centuries and the absence of a standard does not in any way impede its presence on the literary scene today.

Grammars and dictionaries exist since the end of the 18th century. While Picard and Occitan are both Romance languages, Picard inflection more closely resembles French. 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 shared with the other *langues d'oïl* is the fact that a subject personal pronoun is obligatory in the verb phrase.

3.4 Orthography In the absence of well-established written traditions and of a single authority for each of them, the regional languages of France do not have unanimously recognized and accepted standards (Caubet et al. 2001). The orthographic normalization (or, if one prefers, normativization) of a language is difficult where several variants exist, which is necessarily the case in a predominantly oral culture. In order for the majority to have access to information, to allow for the greatest intercomprehension, a writing system with the primary function of communication must take into account the diversity of usages. Another function of writing is to give a graphic image to the identity of a language, through notations that become symbolic emblems. Consequently, the transcription solutions proposed vary from one language to another, even among the regional languages of France, the spellings of which are based on the Latin alphabet and are meant to resemble pronunciation (at least to some extent).

The orthographies adopted are more or less phonetic (reflecting a particular local pronunciation) or diasystemic (emphasizing the unity of a set of dialects). Sometimes, the system is hybrid, in the interests of efficiency, noting aspects of pronunciation, which differ from French, while following the orthographic conventions of French.

Writing contributes to the safeguarding of the cultural heritage of a community. Writing is also a necessity for minority languages, to regain public use and occupy new domains: it is all the more important as the chain of family transmission is often interrupted. In at least the early stages of learning, there is a need for textbooks, which, where possible, highlight the homogeneity of the promoted language. On the other hand, the absence of constraints can allow writers a measure of freedom, in computer- and smartphone-mediated communication, as can be observed in the use of social networks by Alsatian dialect speakers (Huck & Erhart 2019; Erhart 2018). This is especially relevant for some speakers of minority languages, who do not recognize their language (that of immediacy, everyday life, and family) in neo-standards considered to be scholarly artefacts.

Even when an agreement is reached on the main principles unifying a minority language, the search for a standardized orthography readily crystallizes petty squabbles. The resulting orthographic conventions are not always consistent, and this may affect the performance of natural language processing tools. This is the case, in particular, for the languages addressed in the RESTAURE project (Alsatian, Occitan, and Picard), the spelling systems of which are briefly presented below.

For Occitan, we chose to work only on texts that use the so-called classical orthography. Based on Alibèrt's work (Alibèrt 1976) on Lengadocian (intermediate between the other dialects), this norm favors origin and history, rather than the link to pronunciation. This results in a spelling with a strong etymological component, complemented by rules with phonological value. Alibèrt's work was later extended by the *Conselh de la Lengua Occitana* (CLO), which specifies additional details and rules for the transcription of different Occitan varieties (Sumien 2007): recommendations for transcribing the sounds /s/, /z/, words in *-atge-*, for example, the use of umlaut, loanwords, etc. The recommended standard admits frequent uses, such as the omission of diacritics to distinguish the Lemosin *e~è* (opposition neutralized as a single archiphoneme /E/), the Provençau plural in *-ei* (and *-eis* before vowels), the *h* (regular reflex of *f*) and *u* (or *v* or *ü* to write [w]) in Gascon, or the dropping of intervocalic consonants in Northern Occitan.

For Picard, a spelling inspired by the Feller-Carton system (Carton 2009), proposes two principles, accompanied by region-dependent rules to resolve cases where these two principles are in contradiction: (1) priority to French orthography, provided that it does not create ambiguity; (2) priority to the phonetics of Picard. In practice, however, there are almost as many systems as writers, who often overvalue the proliferation of micro-dialectal specific properties. Contrary to French, Picard texts can contain a dot which must not be considered as a word boundary but specifies pronunciation peculiarities, e.g., *lon.mint* (for a long time), *erwet.tent* (look), *fin.mes* (women); this phenomenon is also found in Occitan. More often than in French, words can also contain apostrophes, markers of orality – *c'min* (path) – and hyphens – *gardin-neux* (gardeners).

For Alsatian, part of the corpus uses the ORTHAL system (Zeidler & Crévenat-Werner 2008), with orthographic reference to written German (*Schriftsprache*), respecting dialectal particularities by adding diacritics which reflect the pronunciation

of each speaker and which mark some distance to the German standard. As in Picard though, there are almost as many systems as writers, as can be seen in the other part of the corpus, which reflects the variety of the spellings on the one hand and the influence of the contact with the French and the German standards on the graphic choices made by the different writers on the other hand. All this diversity may of course challenge NLP tools.

In contrast to the Occitan corpus, which includes only texts in the classical orthography, there is no consistent use of a given spelling system in the texts in Picard and Alsatian. As stated above, even though spelling systems were proposed for Picard (Feller-Carton) and Alsatian (ORTHAL), they are not used by all authors. Orthographic variation is an interesting issue of our corpus and one of our goals is to check if the NLP tools can deal with this variation. Another possibility for this kind of corpora is to normalize the texts, but this is a challenging and open research question *per se*, which we chose not to address for the time being.

4. Corpus collection For the creation and annotation of the corpus, we aimed to follow a standard methodology based on the four steps identified by Fort (2012):

- Preparatory work: identifying the “participants”, in our case the annotators and experts, selecting the corpus, creating and modifying the annotation guide;
- Pre-campaign: building a sample reference corpus, training the annotators;
- Annotation: running-in period, annotation work and updates in the corpus and annotation guide;
- Finalization: adjudication, review, publication.

The preparatory step was particularly complex in our case for several reasons:

- The corpora had to be collected and selected, taking copyright issues into account, since our goal was to be able to redistribute the corpora. The situation was different for each language: there was no corpus for Alsatian, while for Occitan many texts were not freely redistributable.
- There were no tokenization tools for any of the three languages.
- There were no POS tag annotation guidelines for these languages, though some existed for closely related languages.
- Annotators are scarce since speakers for these languages who are also trained linguists are difficult to recruit.

4.1 Existing textual databases In order to create the corpus for our project, we first considered re-using existing corpora. Due to its large literary production, Occitan had rich potential sources. Furthermore, text databases already existed for Occitan and Picard, which could be used as sources for our corpus:

- For Occitan, the BaTelÒc text base (Bras & Vergez-Couret 2016) contains wide coverage text collections, with written texts exemplifying literature (prose, drama, and poetry) and other genres such as technical texts and newspapers. The texts show dialectal and spelling variation. 3.7 million words have already been gathered. All the texts in the database are encoded according to the XML TEI P5 format (TEI Consortium 2020).
- For Picard, the PICARTEXT text database¹¹ is a large panchronic literary resource, which contains about 8 million tokens from texts ranging from the 17th century to the 21st century. All the texts in PICARTEXT are encoded according to the guidelines of the XML TEI P5.

4.2 Selection of the sources In order to select the sources, we considered several criteria:

- We wanted the corpus to be freely available, so we needed to collect texts which were not subject to copyright restrictions on distribution.
- The texts had to be representative of the variety found in each language: geolinguistic and graphical variation in spelling, texts from different written genres and time periods.
- The texts had to be relevant to our objective to create generic models: for example, creating a POS tagger trained on a corpus containing only 19th century poetry would lead to a POS tagger which would be too specific.

4.2.1 Alsatian The Alsatian corpus is composed of two main sources: WKP – Wikipedia articles from the Alemannic Wikipedia¹² and HRM – chronicles written in an information magazine published by the Haut-Rhin department (southern Alsace) General Council. Given that the Alemannic Wikipedia contains articles written in several dialects from the Alemannic linguistic area, we only used articles which were explicitly categorized as being written in Alsatian.¹³ In addition to these articles, two more specific genres were used for the annotator training phase: an excerpt from a play and several recipes. As shown in Figure 1, there are actually several sub-dialects, all commonly referred to as “Alsatian”. We included only the largest dialectal areas, i.e. Low Alemannic from the north and the south of the region.

4.2.2 Occitan For the first steps in annotating corpora and developing NLP tools, we chose a single spelling and orthography norm. Among the different spelling conventions and orthography norms available (see §3.2 and §3.4), we chose the classical

¹¹<http://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/>

¹²<http://als.wikipedia.org>

¹³The articles in the Alemannic Wikipedia are manually tagged with their sub-dialect by the authors. We did not check whether this categorization was correct, but the annotators did not notice particular issues with this dialectal categorization.

spelling and orthography conventions, which are the most widely used. Although a text database already existed, the Occitan corpus was problematic to constitute, because there are very few freely available texts in BaTeloC and existing freely available texts do not necessarily follow the classical norm which is relatively recent. The solution that we chose was to select short extracts from the texts with copyright issues and to add other texts for which we obtained permission to distribute.

We automatically selected 55 extracts of 60 words maximum from 16 texts representing different authors and different dialects to create the corpus to be annotated. We also added texts from the online Occitan newspaper *Lo Jornalet*¹⁴ with their kind permission. *Lo Jornalet* contains texts mostly in Lengadocian, and some in Gascon (all in Alibèrt's classical norm); we selected several texts from each dialect to complete the Occitan corpus.

4.2.3 Picard We selected a subset of PICARTEXT according to the project criteria: copyright issues, diachronic diversity, variety of dialects (all Picard dialects are represented), and genres.

4.2.4 Resulting corpora Table 2 presents the resulting corpora. The distribution of the sources according to dates is not always representative of the whole time frame due to the limited availability of adequate sources and the small size of the corpora.

4.3 Tokenization Prior to manual annotation, the texts had to be split into sentences and then tokenized, i.e. split into tokens (words, punctuation marks). This can be done either manually or automatically. In either case, language-specific tokenization guidelines have to be defined in order to make consistent tokenization choices (Habert et al. 1998). These guidelines are of particular importance for punctuation marks, which can be considered either as delimiters, indicating token boundaries, or as being integral parts of a word.

Two types of punctuation marks can be distinguished:

- non-ambiguous delimiters such as question marks (?), exclamation marks (!), colons (:), semi-colons (;), brackets ([], {}), or ellipsis (...)
- ambiguous punctuation marks such as the space, the dash (-), the apostrophe (') and the period (.)

Ambiguous punctuation marks raise specific problems for the correct delimitation of tokens and multiword expressions (such as compound nouns or verbs, location names, prepositional phrases, fixed expressions). For example, the apostrophe (') might be used to replace an elided vowel (*s'n*) but also as a token boundary: *s'n'ac-cidint* should be split as *s'n' accident* (in French: 'son accident'). We identify several exceptions when the punctuation mark is not a separator and should not be split.

¹⁴<https://www.jornalet.com/>

Table 2. Annotated corpora: number of tokens (occurrences, including punctuation) and types (i.e. different tokens), as well as information on each sub-corpus.

Lang.	Dialect/ Region	Source	Docs	Genre	Dates	Tokens	Types
Alsatian	Low Alemannic / North and south of the region	WKP	13	encyclopedic	2016	8,435	3,169
	Low Alemannic / South of the region	HRM	6	chronicles	2005 to 2012	3,550	1,336
	Low Alemannic / North of the region	OLCA	1	recipes	2014	364	188
	Low Alemannic / North of the region	archive.org	1	theater	1906	233	136
					total	12,582	4,535
Occitan	Lengadocian	Jornalet	3	newspaper	2016	931	436
	Gascon	Jornalet	5	newspaper	2016	2,268	1,057
	Lengadocian	BaTelÒc	6	poetry, essays, novels	1948 to 2012	3,394	1,708
	Gascon	BaTelÒc	3	novels, tales	1966 & 2000	1,621	769
	Provençau	BaTelÒc	3	novels, essay	2002 to 2008	1,361	648
	Lemosin	BaTelÒc	3	novels	2000 to 2013	1,313	700
	Vivaro-Aupenc	BaTelÒc	1	novel	1990	526	269
	Auvernhàs	BaTelÒc	1	novel	2008	481	241
				total	11,895	5,828	
Picard	Nord, Belgium, Somme, Aisne, Oise	Picartext	18	narrative	1905 to 2015	8,588	2,657
	Pas-de-Calais, Oise	Picartext	5	poetry	1907 to 1989	1,941	753
	Pas-de-Calais, Aisne	Picartext	2	theater	1923 & 2012	875	377
				total	11,404	3,358	

As we presented in (Bernhard et al. 2017), we define tokens as terminal nodes, following (Webster & Kit 1992). A token will not be divided into smaller units in the annotation process and will be considered as the annotation unit. It is necessary in some cases to decompose a single word into several tokens (like *aux* in French that becomes *à + les*). Conversely, some multiword expressions include spaces, hyphens, or apostrophes and are considered as a single token.

4.3.1 Tokenization for Alsatian The tokenizer for Alsatian was developed in Python. A distinction is made between two types of tokens: (1) lexical and grammatical tokens and (2) signs which mark epenthesis, mostly the euphonic <n> and <w>. Alsatian tends to represent these phonetic phenomena graphically and we chose to single out epenthesis as tokens, as is often the case in the context of oral corpus annotation (Benzitoun et al. 2012). Table 3 presents some tokenization examples for Alsatian:

Table 3. Tokenization examples for Alsatian.

Initial text	Proposed tokenization	Phenomenon	Genre and source
<i>zitter'm Ààfàng</i> (‘since the beginning’)	zitter_’m Ààfàng	preposition + article in the dative case	encyclopedia: Alemannic Wikipedia
<i>in’ra Volkssproch</i> (‘in a dialect’)	in_’ra Volkssproch	preposition + article in the dative case	encyclopedia: Alemannic Wikipedia
<i>uf’e’me Schàrebbà</i> (‘on a charabanc’)	uf_’e’me Schàrebbà	preposition + article in the dative case	story: (Sonnendrücker & Kauss 1998)
<i>hàt fànga-n-à drucka</i> (‘has started to print’)	hàt fànga_n-n_à drucka	epenthesis	encyclopedia: Alemannic Wikipedia
<i>mine-n-Anforderunge</i> (‘my demands’)	mine_n-n_Anforderunge	epenthesis	theater: (Stoskopf 1906)
<i>Heere-n-Er</i> (‘Do you hear’)	Heere_n-n_Er	epenthesis	theater: (Redslob 1907)
<i>geb-w-i</i> (‘go I’)	geh_w-w_i	epenthesis	conversation guide (Keck & Daul 2010)

The tokenizer has been evaluated (Bernhard et al. 2017) on a corpus of 2,633 tokens from different genres (Facebook, poetry, theater, Wikipedia, and narrative), manually tokenized by one expert, and obtains consistent F-measures above 0.99.

4.3.2 Tokenization for Occitan For Occitan, tokenization and sentence segmentation are performed by a Perl script. The tokenization program is adapted from the French tokenization program by Tanguy & Hathout (2007). It is based on the recognition of word separators as outlined above. In practice, segmentation issues in French and Occitan are roughly the same where punctuation signs are used ambiguously to

link separate words syntactically, as in *agacha-las* (look at them) or to create a lexical multiword expression, such as *cap-adjudant* (chief warrant officer). Some Occitan specificities need to be taken into account, for instance the interpunct used in some Gascon words, such as *in·hèrn* (hell) to create a bigram (*n* followed by *h*). Another difference is the use of epenthetic consonants, such as *-n-* or *-s-*, as described for Alsatian above. Sentence segmentation follows from tokenization. It is based on the recognition of reliable sentence separators such as the period, exclamation marks, and question marks. The tokenizer for Occitan has been evaluated on a corpus of 2,707 tokens from two genres (novel and newspaper) and two dialects (Lengadocian and Gascon). The evaluation shows good results overall with F-measures above 0.98, even though results are slightly better for Lengadocian (0.99) than Gascon (0.98). Table 4 presents some tokenization examples for Occitan:

Table 4. Tokenization examples for Occitan.

Initial Form	Proposed tokenization	Phenomenon	Source	Dialect
que s'en.hlma	que s'Ꞥen.hlma	enonciative particle + reflexive pronoun + verb	novel: Lavit	Gascon
que s'estanquè	que s'Ꞥestanquè	enonciative particle + reflexive pronoun + verb	jornalet	Gascon
que's morí	queꞤ's morí	enonciative particle + reflexive pronoun + verb	jornalet	Gascon
Fintatz-me	fintatzꞤ-me	imperative + clitic pronoun	novel: Marti	Lengadocian
a n'aquò	a n'Ꞥaquò	epenthesis	novel: Viaule	Lengadocian

4.3.3 Tokenization for Picard To tokenize Picard texts, we identify several ambiguous punctuation marks, based on (Debrie 1983a). Firstly, the apostrophe is used in several contexts:

- I. to mark a vowel ellipsis: *k'min* ('chemin'/‘path’), *facil'mint* ('facilement'/‘easily’), *v'noét* ('venait'/‘came’), *r'vèttiot* ('regardait'/‘looked’), *atein-n'tté* ('attendait'/‘waiting’)

2. to stress a specific pronunciation: *jam'mas* ('jamais'/'never'), *caban'ne* ('cabane'/'hut'), *un'n* ('une'/'a')
3. as the final character of a word: *és'z'* ('les'/'the')

In other cases, the apostrophe is used to delimit two tokens (*Coreed'l'histoire*: 'encore de l'histoire'/'still the story') as in French.

The hyphen can also occur within a word in Picard: *train-nailler* ('flâner'/'wander'), *éton-nant* ('étonnant'/'surprising'), *einsan-n'* ('ensemble'/'together'), *parson-ne* ('personne'/'nobody').

The hyphen may be used as a delimiter: *Est-ce-què* ('est-ce que'/'Is it ...?').

The period (.) is often used between two double consonants or between two vowels, in the middle of a token: *dormt.tent* ('dorment'/'sleep'), *fin.mes* ('femmes'/'women'). As in many other languages, it is also used to delimit sentences or is included in some abbreviations. Thus, the period is ambiguous and specific tokenization rules were proposed: if the period is placed between two nasals (m or n), we replace the period by an underscore, which avoids further splitting.

The space is ambiguous: it can be included into a multiword expression or used to mark specific phenomena such as epenthesis: *lé z éfans* ('les enfants'/'the children').

Table 5 presents some tokenization examples for Picard:

Table 5. Tokenization examples for Picard.

Initial Form	Tokenization	Phenomena	Source
les gardin-neux ('les jardiniers'/'the gardeners')	les_gardin-neux	determiner + noun	(Debrie 1983a)
Est-ce-què ('est-ce que'/'interrogative form')	Est_ce_què	verb + pronoun + conjunction	(Debrie 1983a)
I se proumon.ne ('il se promène'/'he wanders around')	I se proumon.ne	pronoun + pronoun + verb	(Debrie 1983a)
O z avon ('nous avons'/'we have')	O z avon	pronoun + support consonant + auxiliary	(Debrie 1983a)
Té mérit'roès qu'j'el diche à tin père, quand qu'il arvarro ('Tu mériterais que je le dise à ton père, quand il arrivera'/'You deserve that I tell your father, when he arrives')	Té mérit'roès qu'_j'_el diche à tin père_, quand qu'_il arvarro	relative pronoun + personal pronoun + pronoun, conjunction + pronoun	(Debrie 1983a)
Eze z'éfan i s'abiye ('les enfants s'habillent'/'the children are getting dressed')	Eze z'_éfan i s'_abiye	support consonant + noun, personal pronoun + verb	(Debrie 1983a)

The tokenizer for Picard was developed in Perl. We designed several contextual rules to handle ambiguous cases, using specific regular expressions. We first handle the exceptions that should not be split by replacing the punctuation mark with another character (e.g., underscore). For all the other cases, the punctuation mark is used to split the words. Finally, we insert the initial punctuation mark back in place.

In the following example, we check the characters preceding and following the apostrophe, and we temporarily replace the apostrophe with the character “_”.

```
(1) if($ligne=~/(^qu(\'|') (mi|nt|ti|li))/) #nos${
    # qu' is followed by mi, nt, ti, or li;
    # if other letters follow, we replace qu' with qu__
    $ligne=~s/(qu(\'|'))/qu__;/
}
```

The tokenizer is also able to identify multiword expressions, thanks to a lexicon of multiword expressions (including hyphens and spaces), compiled from several Picard dictionaries (Debie 1975, 1981, 1983b, 1985, 1986, 1987), containing 2,390 entries. The multiword expression lexicon contains prepositional phrases (*a travér dech*) or adverbial phrases (*tout ein heüt*), fixed multiword expressions (*pi vlaù qu*), compound nouns (*faiseu-d'-jeux* ‘saltimbanque’), etc. A lexicon of 204 exceptions (tokens containing hyphens or apostrophes) identifies those cases for which the contextual rules do not apply. The exceptions were identified after a study of samples of texts from several genres (theater, poetry, novel, letter).

First, the rules for annotating multiword expressions are applied to avoid further splitting. Then, we apply the list of exceptions to avoid splitting when this is not necessary. These exceptions are annotated with special delimiters. The contextual rules identify the cases when the hyphen or the dot are included into the word, and we annotate them with a special character. We apply the rules handling non-ambiguous punctuation marks and splitting sentences. Finally, we replace the special delimiters with the space, period, or hyphen respectively. The tokenizer for Picard has been evaluated (Bernhard et al. 2017) on a corpus of 4,191 tokens from various genres (letter, theater, novel, poetry). The evaluation corpus is different from the corpus used to develop the rules. The evaluation shows good results for all the genres (the F-measure varies with the genre from 0.94 to 0.98). This evaluation has been performed by comparing the results of the automatic tokenization with the manual annotation of one expert fluent in Picard.

All three tokenizers are freely available (Bernhard 2018; Vergez-Couret 2019; Todirascu 2018).

4.4 Pre-Annotation Pre-annotation has been widely shown to improve the quality and speed of manual annotation. Marcus et al. (1993) in particular showed that, when creating the Penn treebank, manual annotation not only took twice as long as correction, but had an inter-annotator disagreement rate about twice as high and an

error rate about 50% higher. Existing taggers were thus used where possible. The pre-annotation procedures are fully described in (Bernhard, Ligozat, et al. 2018). They differ for each language:

- In Alsatian, the German TreeTagger (Schmid 1994) was used with several adaptations to enhance its performance on Alsatian (see (Bernhard, Ligozat, et al. 2018) for details).
- In Occitan, an existing Occitan tool designed for machine translation (Aper-tium, a free machine translation platform)¹⁵ was used for a pre-annotation, and once a first subcorpus had been manually corrected, successive models of the TALISMANE tagger (Vergez-Couret & Urieli 2014; Vergez-Couret & Urieli 2015) were trained to pre-annotate the whole corpus.
- In Picard, no pre-annotation was used: no tagger existed for Picard. We tested the French TreeTagger and manually assessed the annotation on a few sentences, but the results were considered too poor to be useful.

This step was handled quite differently for each language, due in particular to the existence of taggers and lexicons for some languages. Another factor was the different backgrounds of the researchers in each team, as some had NLP experience while others came from fields such as sociolinguistics, and this led to different ways of handling the task.

5. Design of the annotation guidelines Ideally, the tagsets should be fine-grained in order to take into account all the linguistic phenomena that can be observed. Yet, a fine-grained annotation is time-consuming, and, in our case, there were few annotators for each language.

The choices made for the tagset were based on the following criteria:

- Use of existing tagsets: the objective was to be as close as possible to a standard tagset. We chose the Universal POS tags (Nivre et al. 2016) as a common tagset. We added an EPE tag (which can be projected to PART) in order to annotate epentheses, which are common in the regional languages.
- Decisions based on tagging choices for close languages: for example, for Alsatian, some choices are based on those made for German. It was often necessary to adapt the guidelines since the morphosyntax of the regional languages sometimes differs from closely related languages.
- Common tagset for all three languages: each language was tagged with a specific tagset, but with a projection to a common tagset.

The main difficulties were, to a varying degree depending on the language, the scarcity of reference resources (grammars for example) and the scarcity of specialists

¹⁵<https://www.apertium.org/>

trained in both morpho-syntax and corpus annotation. Nevertheless, each language team increased in proficiency over the course of the project.

In addition to the POS, we annotated information of potential interest for exploitation of the corpora: compound nouns for Picard, since they can indicate neologisms; epenthesis since they are written traces of the oral forms; location names which are quite frequent and can be used to test named entity recognition for these languages; French translations which make it possible to query the corpus in French and create bilingual lexicons, making the annotation more effective.

The tagsets and guideline creation procedures are fully described in (Bernhard, Ligozat, et al. 2018). The tagsets and guidelines are also available (Bernhard, Erhart, et al. 2018; Bras 2018; Martin et al. 2018a). We will focus here on the choices specific to each language.

5.1 Alsatian The POS tagset for Alsatian is very close to UD (Universal Dependencies),¹⁶ except for the following tags: the EPE tag already mentioned, the APPRART tag, which is the tag for contracted forms of a preposition and a determiner, the MOD tag specific to modal verbs, and the FM tag for foreign words. Lemmas, lemmatized French translations for each token, and location names were also annotated.

5.2 Occitan For Occitan, the standard GRACE tagset (Rajman et al. 1997), derived from the MULTEXT (Ide & Véronis 1994) and EAGLES (von Rekowski 1996) tagsets, was previously chosen for the lexicon of inflected forms Loflòc (Vergez-Couret 2016), as it had been used for several similar annotated corpora for French and Catalan. This tagset, which has also been used for the Occitan TALISMANE tagger (see §4.4), was likewise chosen for the annotation phase.

We made slightly different choices from Alsatian for lemma annotation: for example, numbers are given the same numerical lemma, whether the word form is numerical (23) or not (*vint-e-tres*/twenty-three). As for Alsatian, lemmas, lemmatized French translations, and location names were also annotated.

5.3 Picard The POS tagset for Picard is close to UD, but with some refinements, to take the characteristics of Picard into account and obtain a finer-grained categorization for further analysis of this language. The following categories were added: ADJIND (indefinite adjective), ADJPOS (possessive adjective), ADVLOC (adverbial phrase), ADPDET (contracted preposition + determiner), ADPLOC (prepositional phrase), EPE (epenthesis), NOUNCOMP (compound noun), pronouns (PRONDEM: demonstrative pronoun, PRONIND: indefinite pronoun, PRONPERS: personal pronoun, PRONPOSS: possessive pronoun, PRONREL: relative pronoun, PRONINT: interrogative pronoun), verbs (VERBCONJ: finite verb, VERBINF: infinitive verb, VERBPP: verb in the past participle, and VERBPPR: present participle).

In Picard, neologisms are often created by composition of autonomous lexical units. Compound words were thus annotated with specific tags (NOUNCCOMP

¹⁶<https://universaldependencies.org/u/pos/>

for example for common nouns) in an additional column. As was decided in the project, epentheses were given a specific tag, EPE. This phenomenon is quite common in Picard. In the Picard corpus, in contrast to the Alsatian and Occitan corpora, the French translations correspond to inflected forms as opposed to lemmas. The inflected forms are useful to find lemmas in the dictionaries of several variants of Picard.

Some differences in the choices made were motivated by linguistic differences between the languages, different linguistic interests, or the existent resources in each language. Other differences however stem from the fact that the project development was not at the same stage in the three languages at a given point, which made it difficult to make the choices: for example, the Picard corpus was re-annotated several times in order to be consistent with the other corpora in terms of tokenization and tag choices.

6. Corpus annotation

6.1 Manual annotation As mentioned earlier, the manual annotation was based on a pre-annotation for Alsatian and Occitan. This annotation is described in (Bernhard, Ligozat, et al. 2018). Here we focus on the challenges encountered and the solutions developed.

6.1.1 Alsatian Six persons took part in the annotation. The whole corpus was annotated by one person (a student, who is an expert speaker) and this annotation was reviewed and corrected by another annotator, assisted by two experts. Two other annotators annotated some of the documents to measure inter-annotator agreement. Inter-annotator agreement ranged from 84.1% (κ 0.824) to 93.7% (κ 0.930). The main challenges met during the annotation were as follows:

Limitations in the tagset Sticking to a limited list of POS tags such as the Universal tagset, even with the additions we made, can be difficult at times. Some phenomena could have warranted additional categories, according to the expert linguists, e.g.:

- Light verbs (“Funktionsverb”) with reduced semantic contents such as *lon* (‘let’), *màche* (‘do’), *gënn* (‘give’), *bringe* (‘bring’), *nëmmè* (‘take’)
- Pronominal adverbs such as *drunder* (‘under this’), which in German have a separate category in STTS (Stuttgart-Tübingen Tagset) (Schiller et al. 1999), namely PAV (Pronominaladverbien), but are categorized as adverbs in the German UD corpus

For the time being, we chose to stick to a tagset which is as close as possible to the UD tagset and leave these potential extensions for future work. These could in principle be integrated into the tagset, as long as we provide for a mapping to the UD tagset.

Choice between several possible POS Additional issues are related to the annotation process per se: do we annotate according to the token’s typological classification or to its function and position in the sentence? We chose the second option. When we hesitated over a token’s POS, we mainly used two sources of information to make a decision, relying on the close proximity between the Alsatian dialects and German: the guidelines for tagging German corpora with the STTS and the German UD corpus v 2.0.

While this was often helpful, it also highlighted difficult cases even in German. For instance, *als* has at least five different possible tags in UD_German-GSD v. 2.2: ADP (adposition), ADV (adverb), CCONJ (coordinating conjunction), PART (particle), SCONJ (subordinating conjunction). If we take the example of the sequence *mehr als NUM* (‘more than’ followed by a numeral), there are 37 occurrences in UD_German-GSD v. 2.2 and 7 different POS sequences for the annotation of *mehr als*: ADV ADP (24 occ.), ADV CCONJ (5 occ.), ADV ADV (4 occ.), PRON ADP (1 occ.), PRON ADV (1 occ.), PRON CCONJ (1 occ.), and ADV PART (1 occ.). If we focus on *als* only, it has been assigned 4 different tags in this very specific construct (ADP, ADV, CCONJ, PART). Of course, some of these are clearly annotation errors: Wisniewski & Yvon (2018) have studied the annotation errors and inconsistencies in the UD corpora and shown that they negatively impact the performance of tools which were trained on corpora with errors. The two most frequent annotations for *als* in this construct (ADP and CCONJ) also account for respectively 1,367 and 124 annotations of *als* in the whole corpus (out of 1,658 occurrences), which seems to indicate that there is an ambiguity between both tags in some contexts. In such ambiguous cases, we also consulted the STTS guidelines to make an informed decision. These specific instances of *als* are included in the broader “Konjunktionen” (conjunctions) category in STTS, and we followed this classification by annotating them as conjunctions, in contrast to the predominant ADP annotation found for this construct in the German UD corpus. While existing annotation guidelines and annotated corpora for closely related languages are certainly useful for bootstrapping and informing the annotation process, they nevertheless need to be considered with caution and assessed critically with respect to our target languages. These annotations cannot always be transposed directly, and difficult or ambiguous cases in one language are also difficult or ambiguous in the closely related language.

Lexical variation Since we collected texts which represent several geolinguistic variants, some words may be used only in parts of the region and thus unknown for some of the annotators. In these cases, the help of the expert linguists during the adjudication phase made it possible to identify the suitable annotation and translation into French. We also identified cases which were clearly spelling errors, e.g. *Schilfer* instead of *Schilder*. These are marked in the final CONLL-U format with “Typo=Yes”.

Lemmas and translations into French Verbs with separable particles, when they are conjugated, are split into two parts: the finite verb and the particle, which can be separated from one another by several tokens (e.g., *În dam Sinn kummt/VERB eim*

d elsassische Sproch so vor/PART wie na Dornreesla). Here the issue concerns the lemma for the finite verb (lemma with or without the particle?) and the translation into French (translation of the verb with or without the particle?). Since our goal was also to be able to collect bilingual French-Alsatian lexicons from our annotated corpora, we chose to indicate the lemma strictly corresponding to the finite verb without the particle (this is also the solution we observed in the German UD corpus) and the literal gloss for the translation into French. For the time being, there is no indication of the relation between the finite verb and the particle: this could be specified in the future by also annotating the syntactic dependency relations.

Issues of tokenization In some cases, authors chose to merge tokens which are usually independent. These cases are difficult to resolve with automatic tokenization, because they are very variable and author-dependent. We chose to split some of these instances manually in order to be able to annotate the tokens, e.g. *àsses* split to *àss/SCONJ* + *es/PRON*. This issue is typically due to the lack of standardization which not only affects the way words are spelled but also their boundaries and the use of punctuation marks. In the future, we could also consider using specific POS tags for such merged words, as it was done for Swiss German by Hollenstein & Aepli (2014).

6.1.2 Occitan Six occitan speakers took part in the annotation, including four researchers in Occitan linguistics and two students (only in the initial phase). The corpus was annotated in three steps. First, a Lengadocian corpus was automatically tagged by the Apertium tools described in §4.4 and then manually corrected by four annotators. We then trained the Talismane tagger to annotate a bigger corpus that included texts in two dialects, Lengadocian and Gascon (Vergez-Couret & Urieli 2014). This corpus was manually corrected by two annotators. We trained again Talismane in order to pre-annotate a bigger corpus including the six dialects of Occitan. This final corpus was manually corrected by three expert annotators. At this final stage, inter-annotator agreement ranged from 94.7% (κ 0.940) to 95.3% (κ 0.947).

Annotation guidelines The description of the POS tagset with examples of each tag as provided by the guidelines of the lexicon of inflected forms Loflòc had to be enriched to assist the annotators. We organized triple annotations with expert linguists to make annotation decisions that were then recorded in an annotation guideline (Bras 2018). For example, we made decisions on the annotation of temporal adverbials and dates frequently used in Lo Jornalet articles:

- weekdays are always analyzed as common nouns whether they are included in an NP (*lo dimecres d'abans* 'on the previous Wednesday'), or not (*dimècres* 'on Wednesday')
- schedule times can be analyzed as common nouns (*9h* in *abans 9h* 'before 9 o'clock') or NPs (cardinal determiner + common noun) (*9 oras* in *a 9 oras* 'at 9 o'clock')

Here are two examples of tags for dates with the GRACE-Loflòc tagset and the Universal Dependency tags:

(2) *lo dijòus 23 de març*
 Da/DET Nc/NOUN Nk/NUM Sp/ADP Nc/NOUN
 ‘on Tuesday 23 March’

(3) *lo primèr (jorn) de febrièr*
 Da/DET Ao/NUM (Nc/NOUN) Sp/ADP Nc/NOUN
 ‘on the first (day) of February’

Improvements to the tagset As the tagset was already designed for Loflòc, the annotation process did not raise limitation issues, unlike for Alsatian. Nevertheless, annotating tokens in context helped us refine either the tagset itself or the tag attributes.

For example, in Loflòc, the choice was made not to introduce a new tag for the merged categories preposition (Sp/ADP) + determiner (Da/DET) (*del = de + lo*), but to tag them as a sum of tags: Sp+Da/ADP+DET. During the annotation process, we decided to introduce a new tag SpDa, for technical reasons due to the use of automatic NLP tools.

Regarding tag attributes, for personal pronouns (Pp/PRON), we had initially chosen not to include the nominative case in the attributes for the case tagset because Gascon and Lengadocian dialects do not have subject clitic personal pronouns. When we wanted to annotate more peripheral dialects, like Lemosin and Vivaroaupenc, we had to include nominative case.

Choice between several possible POS As for Alsatian, we chose to annotate according to the token’s function in context. For example, when infinitives are used as nouns, as in *lo manjar e lo beure* (lit. ‘the eating and the drinking’), we annotate them as nouns. We also had to handle polysemic closed class words, such as *que* which can be a relative pronoun (Pr/PRON, as in French), a subordinating conjunction (Cs/SCONJ), and also a coordinating conjunction (Cc/CCONJ). We defined criteria to distinguish past participles from their adjectival uses. Another example of polysemic word is *mai*, which can be either an intensive adverbial (Rq/ADV) – *un còp de mai* (‘once more’), *vos vòli pas espaurugar mai* (‘I don’t want to frighten you more’) – or a coordinating conjunction Cc/CCONJ – *2 mai 2 fan 4* (‘2 plus 2 makes 4’).

As far as negative expressions are concerned, we had to describe the different tagging possibilities:

- *pas* is always tagged Rg/ADV, whether it is the only negative marker in a proposition or part of discontinuous negation: *Maria (non) es pas venguda* (‘Mary has not come’)

- *ne* and *non* are tagged Rp/ADV when they occur with *pas* in discontinuous negation (see example above), but *non* is tagged Rg/ADV when it is the only negative marker in the proposition: *Maria non es venguda* ('Mary has not come')
- the adverbials *cap*, *pus*, *plus*, *brica*, *mai*, *fôrça*, *plan* used with *pas* in discontinuous negation are tagged Rq/ADV (intensive adverbial): *I a pas mai de vin* ('there is no more wine')
- the indefinite pronouns *res*, *degun*, *enlòc* also used with *pas* are tagged Pi/PRON: *I a pas degun* ('there is nobody')

Lexical variation and lemmatization We annotated texts in the 6 dialects of Occitan. For cases of cross dialectal variation, for example *castanha/chastanha* ('chestnut'), we decided not to neutralize the variation in the lemma and to choose *castanha* as the lemma for *castanha* and *chastanha* for *chastanha*. The information that *castanha* and *chastanha* are indeed lexical variants could be annotated in a second run. By contrast, for spelling variants, we decided to standardize the lemma, for example the lemma for *çaquelà* ('although') will be the same as that for *ça que la*.

Lemmas translation into French Although both languages are close, when translating lemmas, we faced the classical problem of impossible lemma-to-lemma mapping. We tried to stay as close as possible to the lemma sense with a single lemma. In some cases, however, we had to give long paraphrases, as for the verb *pirenejjar* translated as *faire de la montagne dans les Pyrénées* ('to hike in the Pyrenees mountains').

6.1.3 Picard We defined POS tagging guidelines for Picard. The data was annotated by a student in linguistics. These annotations were corrected by one of the authors of this article who is fluent in Picard. The problems were discussed with the other researcher specialists of Picard from the RESTAURE project. Due to the lack of time and of native speakers of Picard, the corpus has been annotated by a single annotator, corrected by the other experts. Thus, it is not possible to provide an inter-annotator agreement for this part of the corpus.

Choice between several POS The choice of a tag was sometimes difficult: choosing a tag requires good knowledge of the language, and this was not available for all Picard dialects. The procedure described in the guidelines helped to mitigate this difficulty. Sentence context was used to disambiguate between possible interpretations, and text genre was also taken into account: the use of a particular word could be a stylistic effect or a particularity of the author's style.

Identification of the dialect Another issue was to identify the dialect of the text in order to select appropriate external resources which could help the annotators. The dialect was sometimes unknown and had to be inferred from information about the author's life.

Choice of the lemma This issue has been of great concern to Picard specialists in previous years (Brun-Trigaud & Carton 2003). The choice of a lemma was based on the following criteria, from highest to lowest priority:

- Presence of the lemma in a lexicographic resource representative of the text's dialect
- Presence of an occurrence of the lemma in the text
- Presence for an occurrence of the lemma in the PICARTEXT text database

When several choices were possible (which occurred frequently especially for determiners and pronouns), the most frequent lemma in the text was selected. We also decided not to include the apostrophes at the end of a word in the lemma: for example, *p'tchot'* was lemmatized as *p'tchot*. Moreover, we added a final *e* to the lemma when we had observed occurrences with a final *e* in the text or in dictionary of the same dialect. For example, *quinzain'* was lemmatized as *quinzaine*.

Choice of the translation into French The French translation was chosen according to the following criteria, from highest to lowest priority:

- Sentence context
- Conservation where possible of the same morphosyntactic category as in Picard
- Use of reference dictionaries
- Study of other occurrences to infer translation
- Application of graphical variation rules to look for a similar token in dictionaries when it was not possible to find the exact form

Issues of tokenization The tokenization was done automatically using the space as a separator, and then the tokenization was corrected by hand since the tokenizer was developed at the same time as the manual annotation. However, the difficulties encountered during manual annotation helped the development of the tokenizer.

6.2 Post correction We performed a semi-automatic post correction of the corpora, inspired by (Boudin & Hernandez 2012; Wisniewski & Yvon 2018): for each word, we checked if the same word in the same context had been attributed the same tag and lemma. Otherwise, a new tag or lemma was suggested and submitted to manual validation.

This allowed us to correct basic annotation errors which had been missed by the annotators (NOUN instead of PUNCT for a dot for example). For the Picard corpus, this allowed us to correct 23 tags and 9 lemmas. For the Occitan and Alsatian corpora, no such errors were detected.

6.3 Transformation into the UD format For Alsatian, the tagset was already very close to UD and required only a few minor adjustments, such as mapping EPE, which was a tagging choice made specifically for the project, to PART.

For Occitan, a conversion table was created in order to map the GRACE-Loflòc tags to UD POS tags. Some mappings were trivial (common nouns for example had a specific tag in each tagset); others had to take into account specific features of the language: for example, the As tag for possessive adjectives was mapped to ADJ (instead of DET) because in Occitan, possessive adjectives can occur before or after the noun they modify (*my son* for example can be translated by *lo miu filh* ('the mine son') or *lo filh miu* ('the son mine')). Cardinal nouns, adjectives, and pronouns were all mapped to NUM.

For Picard, a conversion table was also created, which mostly maps the more fine-grained tags to UD tags: for example, the tag for possessive adjectives ADJPOSS becomes adjective ADJ.

The major difficulties that we encountered were the following:

- The first one was a technical one: The raw text had not been kept during the annotation process. Since annotated sentences are supposed to be preceded by their raw version in UD v2, we had to re-align raw text and annotated texts, a process complicated by the fact that the annotated versions had been slightly modified during the annotation process (e.g., correction of spelling errors as explained in §6.1.1).
- The second difficulty came from the fact that contractions had not been annotated according to UD v2 guidelines: contractions are supposed to be annotated by adding the tokens composing the contraction, each token having its own annotation. Contractions had previously been annotated as single tokens, and we had to post-process them semi-automatically.
- Moreover, some multiword expressions had been annotated as a single token and had to be split into several tokens.

7. Final corpora The tag distributions for the final corpora are given in Figures 4, 5, and 6. As expected, in all three languages, the most frequent tags are nouns, determiners, verbs, pronouns, and adpositions (plus punctuation marks).

The vocabulary growth for Picard is presented in Figure 7, which gives the number of different tokens (types) for each number of tokens selected from the corpus. The objective was to estimate the vocabulary size in each regional language, compared to its corresponding well-resourced language, by using the translation information of the annotated corpora. In Picard, the French translations are not lemmatized, so they are compared to the token wordforms. The vocabulary grows faster in Picard than in French, which was expected since there is more graphical and geographical variation in Picard. The same phenomenon can be observed for Alsatian, as Figure 8 shows, and to a lesser extent for Occitan, as Figure 9 shows. This can be explained by the fact that the Occitan texts in the corpus all follow the same orthographic norm.

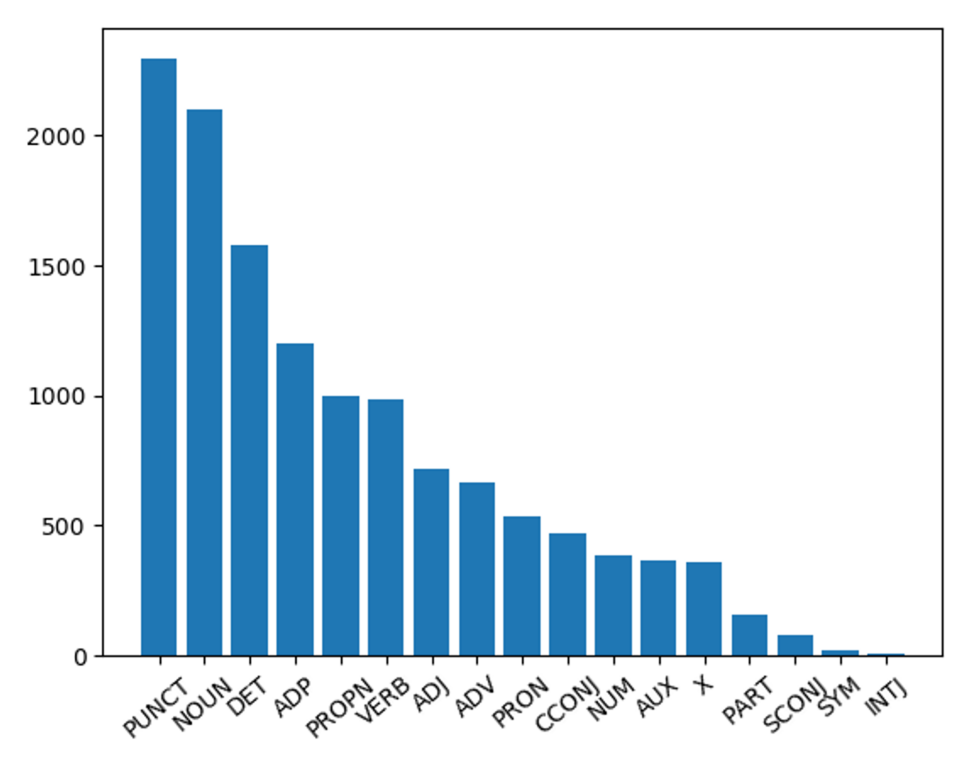


Figure 4. Tag distribution for the Alsatian corpus.

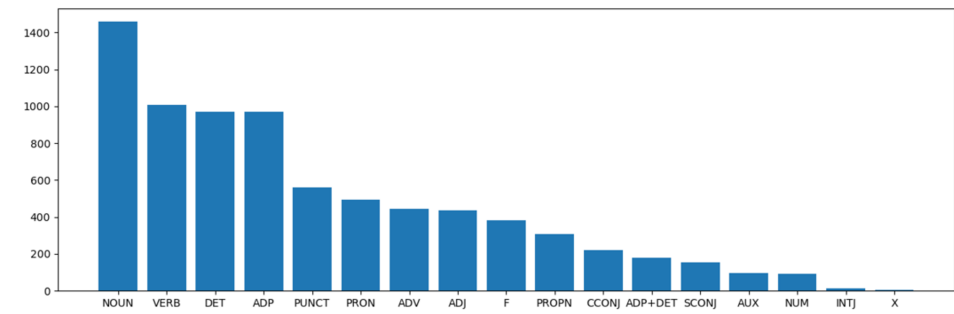


Figure 5. Tag distribution for the Occitan corpus.

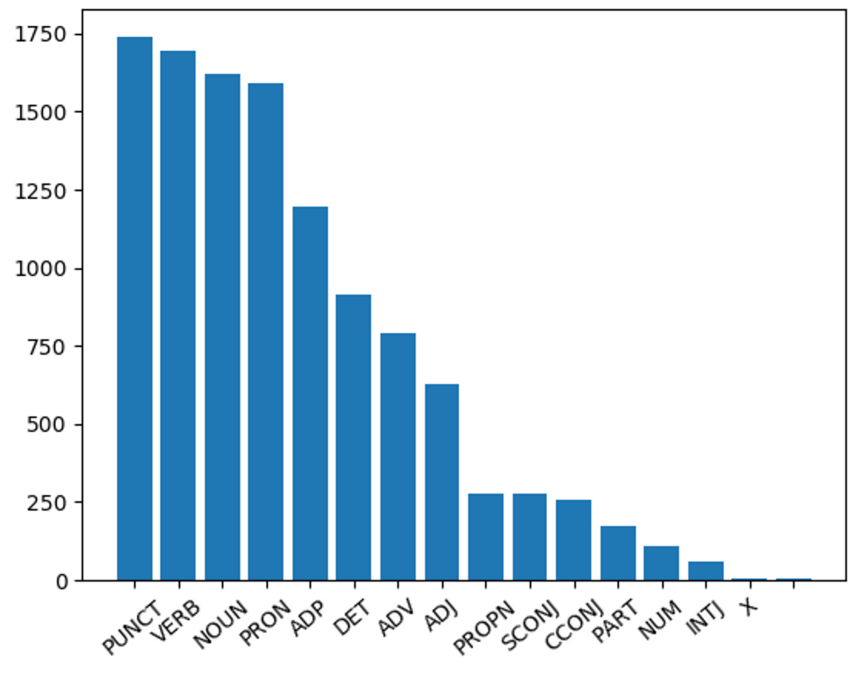


Figure 6. Tag distribution for the Picard corpus.

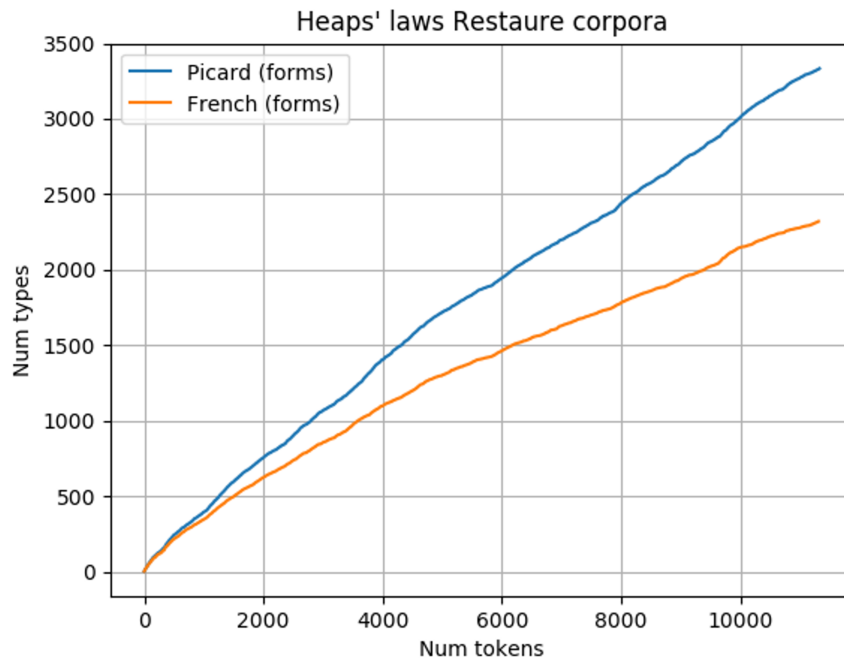


Figure 7. Vocabulary growth for the Picard corpus and its French translation

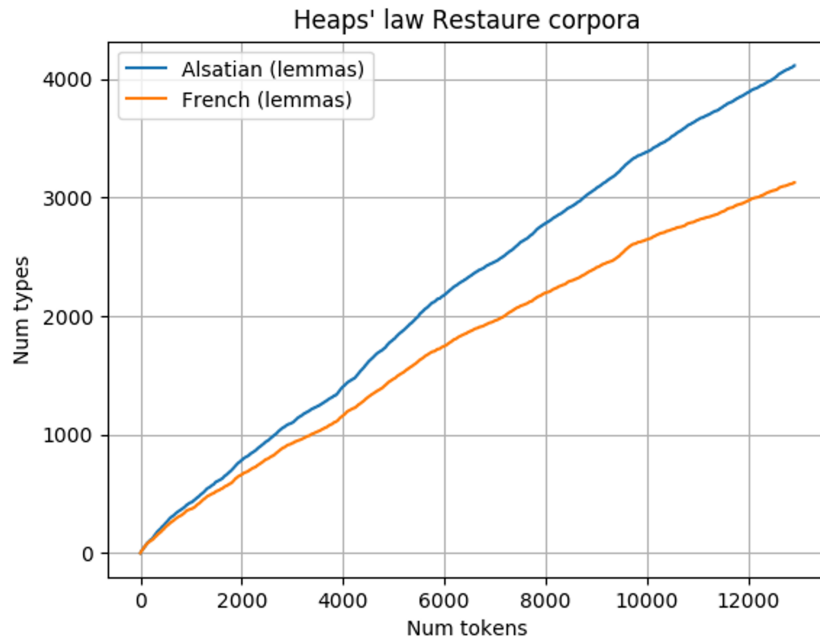


Figure 8. Vocabulary growth for the Alsatian corpus and its French translation (lemmas)

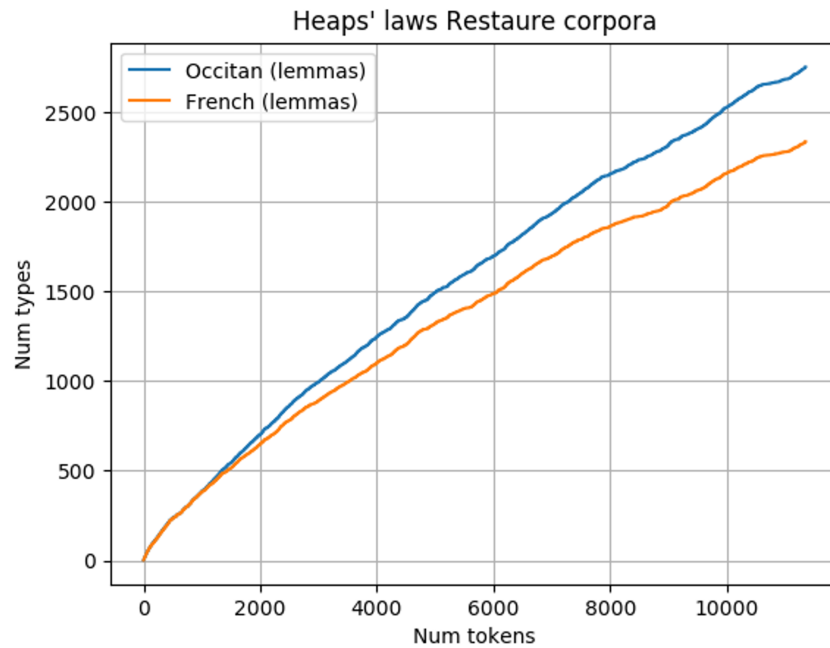


Figure 9. Vocabulary growth for the Occitan corpus and its French translations (lemmas)

8. Related work Standard methodologies for text annotation have been proposed, for example by Fort (2012), and part-of-speech annotation is generally considered as a standard annotation task: part-of-speech annotation has a very long tradition in linguistics and NLP, and many corpora and guidelines exist.

Yet, for under-resourced languages, the task may be more complex due to several issues: the lack of available corpora, the difficulty of recruiting expert annotators when the number of speakers is low, the substantial geographical and graphical variation in the texts, and the possible absence of a full description of the language. Several proposals have been made in response to these issues. Crowdsourcing for example can be used for annotation (Millour et al. 2017; Hovy et al. 2014). Yet, crowdsourcing requires a suitable platform and dissemination to possible speakers, who in the case of Picard, for example, are rare. In order to optimize annotation efficiency, Garrette & Baldridge (2013) compare what they call type and token supervision. Type supervision consists in using tag dictionaries (lists of words with their possible POS tags) as constraints in a semi-supervised learning model, while token supervision consists in manually annotating a set of full sentences with POS tags. They obtain a 71%–78% accuracy for Kinyarwanda, Malagasy, and English.

Another way to alleviate some of these issues is to base the annotation, and in particular the tagset, on similar work in order to benefit from the experience accumulated by other researchers.

The UD guidelines for part-of-speech annotation are very interesting in this respect since they aim at developing a unified tagset suitable for many languages. The tagset is quite small, which can make the annotation process easier. Moreover, these guidelines have already been used to create corpora for several under-resourced languages, e.g. Ainu (Senuma & Aizawa 2018) or Amharic (Seyoum et al. 2018).

Whether they use the UD tagset or not, work on under-resourced languages often faces the very same issues as the ones underlined in this paper, but the solutions proposed can be different.

In order to deal with the lack of a standardized spelling, Seyoum et al. (2018) only collect source texts from grammar books and exclude data collected using online sources, which display more variety in spelling. Moreover, spelling errors were manually corrected. For the Ainu language, the original texts were transcribed in a modern orthographic system (Senuma & Aizawa 2018). Jarrar et al. (2017) describe the process of developing an annotated corpus for the Palestinian Arabic Dialect, which is characterized by a high level of spelling variation. They extended existing guidelines for a Conventional Orthography for Dialectal Arabic (CODA) in order to rewrite each word in the corpus, thus providing two layers for the representations of word forms: raw word (Unicode Arabic script and Buckwalter transliteration) and surface word (Unicode CODA and Buckwalter transliteration of CODA). Normalizing word forms is a delicate issue: in some cases, the normalization can target a given standard. But this seemingly simple approach is not always easy to apply. In these cases, particular guidelines have to be developed. For instance, Krasselt et al. (2015) describe rules for manually normalizing historical German texts, where the norm consists of a modern German wordform, when available. The guidelines for Arabic

dialect orthography by Habash et al. (2018), named CODA*, are an improvement over CODA and aim at transcribing the particularities of each dialect while taking their similarity with modern standard Arabic into account.

In our project, we did not address this issue: we provide no additional layer with a standardized form. However, we added the French gloss, which could, in principle, be used to identify orthographic variants.

9. Conclusion Soria et al. (2013) put forward some recommendations for the development of language resources and technologies for regional and minority languages: connect and cooperate, use standards, document resources and technologies, reuse and recycle, crowdsource resources, be open, share and sustain, cooperate to focused development. We have striven to respect these principles. Firstly, our experience shows that projects of this kind require many different types of skills and it may be difficult to gather the ideal team for one specific language. The work in our project was carried out in a collaborative manner, since the different teams involved had different levels of expertise in the computational work necessary for this project (pre-processing, combination of tools, etc.). The cooperation between linguists, computational linguists, and computer scientists, who share their experience and knowledge, creates a favorable context for the development of resources and tools for under-resourced languages. This work led to the first freely available annotated corpora for Alsatian, Occitan, and Picard. Secondly, the produced corpora are provided in the CONLL-U format, making them easily reusable for further NLP research. They are fully documented and distributed under a *Creative Commons Attribution Share Alike 4.0 International* license (Bernhard et al. 2019; Bras et al. 2018; Martin et al. 2018b). Finally, we adopted a pragmatic approach for the annotation by reusing and adapting existing tagsets and tools (tokenizers) and performing pre-annotation with tools available for closely related languages.

The newly created annotated corpora for Alsatian, Occitan, and Picard offer interesting perspectives. They will be complemented in the future with the annotation of dependency relations and will then be complete for integration into the Universal Dependencies set of corpora. They also make it possible to develop new resources (bilingual lexicons) and tools (tokenizers, POS taggers, lemmatizers) which could then be used for other projects or to annotate other resources (e.g. to speed up the manual annotation process).

To go beyond the purely scientific scope of this work, we hope that we have been able to help bridge some of the growing digital divide between majority languages and the minority languages discussed in this article. Indeed, the development of digital language resources and NLP technologies is essential to make less-resourced languages fully usable on digital media. This allows for greater and stronger linguistic and cultural diversity in the digital world. In our view, this is also a matter of simple respect for the principle of equality between humans and the languages they speak.

References

- Alibèrt, Loís. 1935–1937. *Gramatica occitana segon los parlars lengadocians*. Toulouse: Societat d’Estudis Occitans.
- Alibèrt, Loís. 1976. *Gramatica occitana segon los parlars lengadocians*. 2nd edn. Montpellier: Centre d’Estudis Occitans.
- Bec, Pierre. 1970. *Manuel pratique de philologie romane*, vol. 1. Paris: Picard.
- Bec, Pierre. 1995. *La langue occitane*. 6th edn. Paris: Presses Universitaires de France.
- Benzitoun, Christophe, Karën Fort, & Benoît Sagot. 2012. TCOF-POS: Un corpus libre de français parlé annoté en morphosyntaxe. In Antoniadis, Georges, Hervé Blanchon, & Gilles Sérasset (eds.), *Actes de la Conférence Conjointe JEP-TALN-RECITAL 2012*, Grenoble, vol. 2, 99–112. (<https://hal.archives-ouvertes.fr/hal-007-09187>)
- Bernhard, Delphine. 2014. Adding dialectal lexicalisations to linked open data resources: The example of Alsatian. In Pretorius, Laurette, Claudia Soria, & Paolo Baroni (eds.), *Proceedings of the Workshop on Collaboration and Computing for Under Resourced Languages in the Linked Open Data Era (CCURL 2014)*, Reykjavik, 23–29. (<http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014-Workshop-CCURL2014-Proceedings.pdf>)
- Bernhard, Delphine. 2018. Tokeniser for the Alsatian Dialects. *Zenodo*. doi:10.5281/zenodo.2454993
- Bernhard, Delphine, Pascale Erhart, Dominique Huck, & Lucie Steiblé. 2018. Part-of-speech annotation guidelines for the Alsatian dialects. *Zenodo*. doi:10.5281/zenodo.1171925
- Bernhard, Delphine, Pascale Erhart, Dominique Huck, & Lucie Steiblé. 2019. Annotated corpus for the Alsatian dialects [Data set]. *Zenodo*. doi:10.5281/zenodo.2536041
- Bernhard, Delphine & Anne-Laure Ligozat. 2013a. Hassle-free POS-tagging for the Alsatian dialects. In Zampieri, Marcos & Sacha Diwersy (eds.), *Non-standard data sources in corpus based-research (ZSM Studien)*, 85–92. Düren: Shaker.
- Bernhard, Delphine & Anne-Laure Ligozat. 2013b. Es esch fäscht wie Ditsch, oder net? Étiquetage morphosyntaxique de l’alsacien en passant par l’allemand. In Morin, Emmanuel & Yannick Estève (eds.), *Actes de TALARE 2013: Traitement Automatique des Langues Régionales de France et d’Europe*, Les Sables-d’Olonne, 209–220.
- Bernhard, Delphine, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steiblé, Pascale Erhart, Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynés, Sophie Rosset, Jean Sibille, & Thomas Lavergne. 2018. Corpora with part-of-speech annotations for three regional languages of France: Alsatian, Occitan and Picard. In Calzolari, Nicoletta, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, & Takenobu Tokunaga (eds.), *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan, 3917–3924.
- Bernhard, Delphine, Amalia Todirascu, Fanny Martin, Pascale Erhart, Lucie Steiblé, Dominique Huck, & Christophe Rey. 2017. Problèmes de tokénisation pour deux

- langues régionales de France, l'alsacien et le picard. In *Actes de DiLiTAL 2017 : Diversité Linguistique et TAL*, Orléans, France, 14–23.
- Bernissan, Fabrice. 2012. Combien de locuteurs compte l'occitan en 2012?. *Revue de Linguistique Romane* 303–304. 467–512.
- Beyer, Ernest. 1963. *La flexion du groupe nominal en alsacien: Etude descriptive et historique avec 60 cartes*. Paris: Les Belles-Lettres.
- Beyer, Ernest & Raymond Matzen. 1969. *Atlas linguistique et ethnographique de l'Alsace*, vol. 1. Paris: Éditions du C.N.R.S.
- Bothorel-Witz, Arlette, Marthe Philipp, & Sylviane Spindler. 1984. *Atlas linguistique et ethnographique de l'Alsace*, vol. 2. Paris: Éditions du CNRS.
- Boudin, Florian & Nicolas Hernandez. 2012. Détection et correction automatique d'erreurs d'annotation morpho-syntaxique du French TreeBank. In Antoniadis, Georges, Hervé Blanchon, & Gilles Sérasset (eds.), *Actes de la Conférence Conjointe JEP-TALN-RECITAL 2012*, Grenoble, vol. 2, 281–291. (<http://www.aclweb.org/anthology/F12-2021>)
- Bras Myriam. 2018. Part of speech annotation guidelines for the Occitan language. *Zenodo*. doi:10.5281/zenodo.1173113
- Bras, Myriam, Louise Esher, Jean Sibille, & Marianne Vergez-Couret. 2018. Annotated corpus for Occitan [Data set]. *Zenodo*. doi:10.5281/zenodo.1182949
- Bras, Myriam & Jean Thomas. 2008. Batelòc: Cap a una basa informatizada de tèxtes occitans. In Rieger, Angelica & Domergue Sumien (eds.), *IXème Congrès de l'Association Internationale d'Etudes Occitanes*, Aachen, 661–670. (<https://hal.archives-ouvertes.fr/hal-00986409>)
- Bras, Myriam & Marianne Vergez-Couret. 2016. BaTelÒc: A text base for the Occitan language. In Ferreira, Vera & Peter Bouda (eds.), *Language documentation and conservation in Europe*, 133–149. Honolulu: University of Hawai'i Press. (<http://scholarpace.manoa.hawaii.edu/handle/10125/24675>)
- Brunner, Jean-Jacques, Arlette Bothorel-Witz, & Marthe Philipp. 1985. Parlers alsaciens. In *Encyclopédie de l'Alsace*, vol. 10, 5838–5853. Strasbourg: Editions Publitotal.
- Brun-Trigaud, Guylaine & Fernand Carton. 2003. Lemmes, supra-lemmes: Dilemmes ... problèmes d'indexation de l'Atlas linguistique picard et de l'Atlas linguistique du Centre. In Bouvier, Jean-Claude, Jacques Gourc, & François Pic (eds.), *Sempre los camps auràn segadas resurgantas: Mélanges offerts à Xavier Ravier*, 63–72. Toulouse: CNRS – Université Toulouse-Le Mirail. (<https://hal.archives-ouvertes.fr/hal-01360721>)
- Carton, Fernand. 2009. Pourquoi et pour qui on transcrit ? Les graphies du picard moderne. *La Linguistique* 45(1). 113–123.
- Carton, Fernand & Maurice Lebègue. 1989–1998. *Atlas linguistique et ethnographique du Picard*. Paris: Éditions du CNRS.
- Carton, Fernand, Maurice Lebègue, Jacques Chaurand, & Denise Poulet. 1997. *Atlas linguistique et ethnographique picard: Le temps, la maison, l'homme, animaux et plantes sauvages, morphologie*. Paris: Éditions du CNRS.

- Caubet, Dominique, Salem Chaker, & Jean Sibille. 2001. *Codification des langues de France*. Paris: L'Harmattan.
- Cerquiglioni, Bernard. 1999. *Les langues de la France*. (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.)
- Debrie, René. 1975. *Lexique picard des parlers ouest-amiénois*. Amiens: Centre d'Études Picardes.
- Debrie, René. 1981. *Lexique picard du Vimeu*. Amiens: Centre d'Études Picardes.
- Debrie, René. 1983a. *Eche pikar bèl é rade*. Paris: Éditions Disques Omnivox.
- Debrie, René. 1983b. *Lexique picard des parlers est-amiénois*. Amiens: Centre d'Études Picardes.
- Debrie, René. 1985. *Lexique picard du Ponthieu*. Amiens: Centre d'Études Picardes.
- Debrie, René. 1986. *Lexique picard des parlers du Santerre*. Amiens: Centre d'Études Picardes.
- Debrie, René. 1987. *Lexique picard du Vermandois*. Amiens: Centre d'Études Picardes.
- Dubois, Raymond. 1957. *Le domaine picard: Délimitation et carte systématique, dressée pour servir à l'inventaire général du picard et autres travaux de géographie linguistique*. Arras: Archives du Pas-de-Calais.
- EDInstitut & Office pour la Langue et Culture d'Alsace (OLCA). 2012. *Etude sur le dialecte alsacien*. (https://www.olcalsace.org/sites/default/files/documents/etude_linguistique_olca_edinstitut.pdf)
- Erhart, Pascale. 2018. Les émissions en dialecte de France 3 Alsace : des programmes hors normes pour des parlers hors normes?. *Les Cahiers du GEPE* 10. (<http://www.cahiersdugepe.fr/index.php?id=3201>)
- Erhart, Pascale. 2020. Von der 'Mundart' zur 'Fingerart': Was bedeutet es heute, Elsässisch zu sprechen bzw. zu schreiben?. *IDS Sprachreport* 36(1). 6–13. (<https://pub.ids-mannheim.de/laufend/sprachreport/sr20.html>)
- Forlot, Gilles & Fanny Martin. 2014. Entre invisibilité et (auto) occultation: Les paradoxes des pratiques langagières minoritaires en Picardie. In Djordjevic, Ksenija (ed.), *Les minorités invisibles: Diversité et complexité (ethno) sociolinguistiques*, 77–87. Limoges: Éditions Lambert-Lucas.
- Fort, Karèn. 2012. *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. Paris: Université Paris 13. (Doctoral dissertation.)
- Garrette, Dan & Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of HLT-NAACL*, Atlanta, 138–147. (<https://www.aclweb.org/anthology/N13-1014.pdf>)
- Gilliéron, Jules & Edmond Edmont. 1902–1910. *Atlas linguistique de la France*. Paris: Honoré Champion.
- Habash, Nizar, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghrouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al-Shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, & Hind Saddiki. 2018. Unified guidelines and resources for Arabic dialect orthography. In Calzolari, Nicoletta, Khalid

- Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, H el ene Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, & Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, 3628–3637. Paris: European Languages Research Association. (<https://www.aclweb.org/anthology/L18-1574.pdf>)
- Habert, Beno t, Gilles Adda, Martine Adda-Decker, Philippe Boula de Mar euil, St ephane Ferrari, Olivier Ferret, Gabriel Illouz, & Patrick Paroubek. 1998. The need for tokenization evaluation. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC 1998)*, Granada, vol. 1, 427–431.
- Hollenstein, Nora & No emi Aepli. 2014. Compilation of a Swiss German dialect corpus and its application to PoS tagging. In Zampieri, Marcos, Liling Tan, Nikola Ljube i c, & J org Tiedemann (eds.), *Proceedings of VarDial: Applying NLP Tools to Similar Languages, Varieties and Dialects*, Dublin, 85–94. (<https://www.aclweb.org/anthology/W14-5310/>)
- Hovy, Dirk, Barbara Plank, & Anders S ogaard. 2014. Experiments with crowdsourced re-annotation of a POS tagging data set. In Toutanova, Kristina & Hua Wu (eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, vol. 2, 377–382. (<https://www.aclweb.org/anthology/P14-2062>)
- Huck, Dominique. to appear. Dialectal speech in Alsace / Alsatian. In Boas, Hans, Ana Deumert, Mark L. Loudon, & Peter Maitz (eds.), *Varieties of German worldwide*, vol. 1. Cambridge: Cambridge University Press.
- Huck, Dominique. 1999. Les dialectes en Alsace – l’allemand standard. In Dominique, Huck, Arlette Laugel, & Maurice Laugner (eds.), *L’ l ve dialectophone en Alsace et ses langues, L’enseignement de l’allemand aux enfants dialectophones   l’ cole primaire. De la description contrastive dialectes / allemand   une approche m thodologique. Manuel   l’usage des ma tre*, 15–71. Strasbourg: Oberlin.
- Huck, Dominique. 2015. *Une histoire des langues de l’Alsace*. Strasbourg: La Nu e Bleue.
- Huck, Dominique & Pascale Erhart. 2019. Frankreich – Das Elsass. In Beyer, Rachel & Albrecht Plewnia (eds.), *Handbuch des Deutschen in West- und Mitteleuropa: Sprachminderheiten und Mehrsprachigkeits-konstellationen*, 155–192. T bingen: Narr.
- Ide, Nancy & Jean V ronis. 1994. MULTTEXT: Multilingual text tools and corpora. In *Proceedings of the 15th Conference on Computational linguistics*, Kyoto, vol. 1, 588–592. Stroudsburg, PA: Association for Computational Linguistics.
- Institut d’ tudes Opinion et Marketing en France et   l’international (IFOP), January 7. 2020. Enqu te sur la question r gionale en Alsace. (<https://www.ifop.com/publication/enquete-sur-la-question-regionale-en-alsace/>) (Accessed 2020-06-02.)
- Jarrar, Mustafa, Nizar Habash, Faeq Alrimawi, Diyam Akra, & Nasser Zalmout. 2017. Curras: An annotated corpus for the Palestinian Arabic dialect. *Language Resources & Evaluation* 51. 745–775. doi:10.1007/s10579-016-9370-7
- Keck, B n dicte & L on Daul. 2010. *L’alsacien pour les nuls* (Pour les nuls ( d. de poche)). Paris: First  ditions.


- Krasselt, Julia, Marcel Bollmann, Stefanie Dipper, & Florian Petran. 2015. *Guidelines für die Normalisierung historischer deutscher Texte / Guidelines for normalizing historical German texts* (Bochumer Linguistische Arbeitsberichte 15). Bochum: Ruhr-Universität Bochum. (<https://digitalcollection.zhaw.ch/handle/11475/4003>)
- Leixa, Jérémy, Valérie Mapelli, Khalid Choukri. 2014. *Inventaire des ressources linguistiques des langues de France* (ELDA/DGLFLF-2013A). Paris: ELDA/DGLFLF.
- Marcus, Mitchell P., Mary Ann Marcinkiewicz, & Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics* 19(2). 313–330.
- Martin, Ernst & Hans Lienhart. 1974 [1899–1907]. *Wörterbuch der elsässischen Mundarten*, vols. 1–2. Berlin: Walter de Gruyter. (<https://woerterbuch-netz.de/?sigle=ElsWB#0>)
- Martin, Fanny. 2015. *Espaces et lieux de la langue en Picardie au XXIème siècle. Approche complexe de la structuration des répertoires linguistiques en situations ordinaires. Enquête en Picardie*. Amiens: Université de Picardie Jules Verne. (Doctoral dissertation.)
- Martin, Fanny & Gilles Forlot. 2016. Hétérogénéité linguistique et poids des idéologies sur les pratiques linguistiques en Picardie. In Boudreau, Annette & Laurence Arrighi (eds.), *La construction discursive du locuteur francophone en milieu minoritaire: Problématiques, méthodes et enjeux*, 193–210. Québec: Presses de l'Université Laval.
- Martin, Fanny, Christophe Rey, & Philippe Reynés. 2018a. Part-of-speech annotation guidelines for Picard. *Zenodo*. doi:10.5281/zenodo.1173428
- Martin, Fanny, Christophe Rey, & Philippe Reynés. 2018b. Annotated corpus for Picard [Data set]. *Zenodo*. doi:10.5281/zenodo.1485988
- Millour, Alice, Karèn Fort, Delphine Bernhard, & Lucie Steiblé. 2017. Vers une solution légère de production de données pour le TAL: création d'un tagger de l'alsacien par crowdsourcing bénévole. In Eshkol, Iris & Jean-Yves Antoine (eds.), *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles*, Orléans, 139–154. (<https://www.aclweb.org/anthology/2017.jeptalnrecital-long.10/>)
- Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, & Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Morena, Jan Odjik, Stelios Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, 1659–1666. (http://www.lrec-conf.org/proceedings/lrec2016/pdf/348_Paper.pdf) (Accessed 2016-06-14.)
- Olivieri, Michèle, Sylvain Casagrande, Guylaine Brun-Trigaud, & Pierre-Aurélien Georges. 2017. Le Thesaurus occitan dans tous ses états. *Revue française de linguistique appliquée* 22(1). 89–102.

- Rajman, Martin, Josette Lecomte, & Patrick Paroubek. 1997. *Format de description lexicale pour le français, partie 2: Description morpho-syntaxique*. (Rapport Grace Gtr-3-2.1)
- Redslob, Robert. 1907. *D'r Schlitterhannes: Elsaessisches Bauerndrama in zwei Akten*. Strassburg: E. van Hauten.
- Région Midi-Pyrénées. 2010. *Résultats synthétiques de l'étude sociolinguistique "Présence, pratiques, et perceptions de la langue occitane en région Midi-Pyrénées"*. Toulouse: Région Midi-Pyrénées. (<http://www.midipyrenees.fr/IMG/pdf/EnqueteOccitan.pdf>)
- Ronjat, Jules. 1930. *Grammaire istorique [sic] des parlers provençaux modernes*, vol. 1. Montpellier: Société des Langues Romanes.
- Schiller, Anne, Simone Teufel, Christine Stöckert, & Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Stuttgart: Universität Stuttgart.
- Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, 44–49. (<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.2-8.1139>)
- Senuma, Hajime & Akiko Aizawa. 2018. Universal Dependencies for Ainu. In Calzolari, Nicoletta, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, & Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, 2354–2358. Paris: European Language Resources Association (ELRA). (<https://www.aclweb.org/anthology/L18-1373.pdf>)
- Seyoum, Binyam Ephrem, Yusuke Miyao, & Baye Yimam Mekonnen. 2018. Universal Dependencies for Amharic. In Calzolari, Nicoletta, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, & Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, 2216–2222. Paris: European Language Resources Association (ELRA). (<https://www.aclweb.org/anthology/L18-1350.pdf>)
- Sibille, Jean. 2002. Écrire l'occitan: essai de présentation et de synthèse. In Caubet, Dominique, Salem Chaker, & Jean Sibille (eds.), *Codification des langues de France*, 18–37. Paris: L'Harmattan.
- Sibille, Jean. 2010. Les langues autochtones de France métropolitaine. Pratiques et savoirs. In Gruaz, Claude & Christine Jacquet-Pfau (eds.), *Autour du mot: pratiques et compétences, séminaire du centre du français moderne*, vol. 2, 69–85. Limoges: Lambert-Lucas.
- Sonnendrücker, Paul & Alain Kauss. 1998. *Kochersberg: récits en dialecte avec version française*. Strasbourg: BF Éditions.
- Soria, Claudia, Joseph Mariani, & Carlo Zoli. 2013. Dwarfs sitting on the giants' shoulders – How LTs for regional and minority languages can benefit from piggy-


- backing major languages. In *Proceedings of XVII FEL Conference*, Ottawa, 73–79. (<https://hal.archives-ouvertes.fr/hal-01840828>)
- Stoskopf, Gustave. 1906. *D'r Hoflieferant. Elsaessische Komædie in 3 Aufzuegen*. Strassburg: Schlesier und Schweikhardt.
- Sumien, Domergue. 2007. Preconizacions del Conselh de la Lengua Occitana. *Linguistica Occitana* 6. 1–157.
- Tanguy, Ludovic & Nabil Hathout. 2007. *Perl pour les linguistes: Programmes en Perl pour exploiter les données langagières*. Stanmore: Hermes Science.
- Text Encoding Initiative (TEI) Consortium. 2020. TEI P5: Guidelines for electronic text encoding and interchange. *Zenodo*. doi:10.5281/zenodo.3667251
- Todirascu, Amalia. 2018. Tokeniser for Picard. *Zenodo*. doi:10.5281/zenodo.1493642
- Vergez-Couret, Marianne. 2013. Tagging Occitan using French and Castilian Tree Tagger. In *Proceedings of the 6th Language and Technology Conference*, Poznan, 78–82. (<https://hal.archives-ouvertes.fr/hal-00986426/document>)
- Vergez-Couret, Marianne. 2016. *Description du lexique Loflòc*. CLLE-ERSS. (Research report.) (<https://hal.archives-ouvertes.fr/hal-01338774>)
- Vergez-Couret, Marianne. 2019. Tokenization for Occitan (Gascon and Lengadocian). *Zenodo*. doi:10.5281/zenodo.2533873
- Vergez-Couret, Marianne & Assaf Urieli. 2014. Pos-tagging different varieties of Occitan with single-dialect resources. In Zampieri, Marcos, Liling Tan, Nikola Ljubešić, & Jörg Tiedemann (eds.), *Proceedings of VarDial: Applying NLP Tools to Similar Languages, Varieties and Dialects*, Dublin, 21–29. (<https://www.aclweb.org/anthology/W14-5303/>)
- Vergez-Couret, Marianne & Assaf Urieli. 2015. Analyse morphosyntaxique de l'occitan languedocien: l'amitié entre un petit languedocien et un gros catalan. In *Actes de TALARE 2015: Traitement Automatique des Langues Régionales de France et d'Europe*, Caen. (<https://hal.archives-ouvertes.fr/hal-01214566>) (Accessed June 28, 2016.)
- von Rekowski, Ursula. 1996. *Elm-fr: Specifications for French morphosyntax, lexicon specification and classification guidelines*. (EAGLES document.)
- Wackenheim, Auguste. 1993. *La littérature dialectale alsacienne: une anthologie illustrée*, vol. 1. Paris: Prat Éditions.
- Wackenheim, Auguste. 1994. *La littérature dialectale alsacienne: une anthologie illustrée*, vol. 2. Paris: Prat Éditions.
- Wackenheim, Auguste. 1997. *La littérature dialectale alsacienne: une anthologie illustrée*, vol. 3. Paris: Prat Éditions.
- Wackenheim, Auguste. 1999. *La littérature dialectale alsacienne: une anthologie illustrée*, vol. 4. Paris: Prat Éditions.
- Wackenheim, Auguste. 2003. *La littérature dialectale alsacienne: une anthologie illustrée*, vol. 5. Paris: Prat Éditions.
- Webster, Jonathan J. & Chunyu Kit. 1992. Tokenization as the initial phase in NLP. In *Proceedings of the 14th conference on Computational linguistics*, Nantes, vol. 4, 1106–1110. Stroudsburg, PA: Association for Computational Linguistics. (<http://dl.acm.org/citation.cfm?id=992434>) (Accessed 2016-03-18.)

- Wisniewski, Guillaume & François Yvon. 2018. Divergences entre annotations dans le projet Universal Dependencies et leur impact sur l'évaluation des performance d'étiquetage morpho-syntaxique. In Sébillot, Pascale & Vincent Claveau (eds.), *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles*, Rennes, vol. 1, 567–576. (<https://hal.archives-ouvertes.fr/hal-01793784>)
- Zeidler, Edgar & Danielle Crévenat-Werner. 2008. *Orthographe alsacienne: Bien écrire l'alsacien de Wissembourg à Ferrette*. Colmar: J. Do Bentzinger.


Delphine Bernhard
dbernhard@unistra.fr

 orcid.org/0000-0001-7857-5873


Anne-Laure Ligozat

 orcid.org/0000-0002-2188-3426


Fanny Martin

 orcid.org/0000-0003-3179-8388


Marianne Vergez-Couret

 orcid.org/0000-0002-0483-0525


Pascale Erhart

 orcid.org/0000-0002-5674-8320

Amalia Todirascu

 orcid.org/0000-0002-3092-3549

Philippe Boula de Mareüil

 orcid.org/0000-0002-8213-2693