Towards a Corsican Basic Language Resource Kit

Laurent Kevers, Stella Retali Medori

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2726–2735
Marseille, 11–16 May 2020

Towards a Corsican Basic Language Resource Kit

Laurent Kevers, Stella Retali-Medori
UMR CNRS 6240 LISA, Università di Corsica - Pasquale Paoli
Avenue Jean Nicoli, 20250 Corte, France
{kevers_l, medori_e}@univ-corse.fr

Abstract

The current situation regarding the existence of natural language processing (NLP) resources and tools for Corsican reveals their virtual non-existence. Our inventory contains only a few rare digital resources, lexical or corpus databases, requiring adaptation work. Our objective is to use the *Banque de Données Langue Corse* project (BDLC) to improve the availability of resources and tools for the Corsican language and, in the long term, provide a complete Basic Language Ressource Kit (BLARK). We have defined a roadmap setting out the actions to be undertaken: the collection of corpora and the setting up of a consultation interface (concordancer), and of a language detection tool, an electronic dictionary and a part-of-speech tagger. The first achievements regarding these topics have already been reached and are presented in this article. Some elements are also available on our project page (`http://bdlc.univ-corse.fr/tal/`).

Keywords: less-resourced languages, Corsican, corpora, lexical resources, language identification, POS tagging

We will first contextualize our research and present some general information about the Corsican language (1.1., its digital presence (1.2.) and the BDLC project (1.3.), as previously done in Kevers et al. (2019) and Kevers and Retali-Medori (2020). Not without having drawn up a brief state of the art (2.1.), we will present the achievements concerning the NLP resources and tools for Corsican (2.2.). In particular, we will address corpora collection (2.3.), the set up of a consultation interface in the form of a concordancer (2.4.) and of a language detection tool (2.5.), the electronic dictionary building (2.6.) and POS tagging (2.7.). These developments are guided by a roadmap that we have defined (Kevers et al., 2019) and that is largely in line with the steps proposed by Ceberio Berger et al. (2018).

1. The Corsican language and its digital presence

1.1. Corsican language

Corsican is a Latin language and is part of the Italo-Romance domain. It has known various contacts and linguistic influences. Due to the Pisan domination (9th-13th centuries), it is particularly linked to medieval Tuscan which constitutes its superstratum. Corsican has borrowed from other Italo-Romance and even Romance varieties as well as from Germanic and Arabic languages.

From a dialectal point of view, four or even five areas are identifiable (Dalbera-Stefanaggi, 2002; Dalbera-Stefanaggi, 2007). The extreme southern area even crosses the borders of Corsica, extending into Gallura, in the north of Sardinia. However, these five areas constitute a *continuum* and do not prevent interunderstanding between speakers, even with those of the central and southern varieties of Italy.

The genetic and historical relation between Corsican and Tuscan languages led the latter to be the writing language of the islanders from the Middle Ages until the emergence of a conscious writing in Corsican language in the 19th century. The spelling of Corsican is therefore, with some adaptations, based on the Italian graphic system (Retali-Medori, 2015). However, despite the implementation of a polynomic approach (Marcellesi, 1984) that encompasses all dialectal variants, the writing of the language is not standardised, which is an obstacle for its automatic processing.

1.2. Digital presence

Nowadays, Corsican is, with French, part of a diglossic language environment, and its use is declining. The development of tools is necessary for its preservation, enhancement, transmission and promotion[1]. Public policy encourages the exploitation of new technologies to this end. Several tools and resources exist for the learning and linguistic description of Corsican dialects, but their inclusion in the digital humanities domain remains insufficient[2].

In particular, sites and applications dedicated to translation, lexicon and syntax contain little data in comparison with the richness and complexity of the language. On the other hand, this wealth is found on databases such as the *Banque de Données Langue Corse* (BDLC) and *Infcor*[3] (*Banca di dati di a lingua corsa*). The latter was conceived in an associative context by ADECEC, an organisation active in the field of Corsican language and culture. It has been developed and was made available to the general public as a website as well as a smartphone application wich offer numerous lexical records and multi-criteria query modes[4]. Regarding the BDLC, it is a tool designed in a scientific context, associated with the production of a linguistic atlas, the NALC.

[1] Recommendation of the UNESCO (Brenzinger et al., 2003)

[2] Corsican language can be found in various forms on the web: interfaces in Corsican language (*Qwant* and *Google* search engines, *Facebook* social network), websites or application like *Google Translate*, *Wikipedia* or *Wiktionary*, smartphone apps for tourism, education or translation, online educational content such as Corsican language learning websites, blogs about Corsican language and culture.

[3] `http://infcor.adecec.net`

[4] Altough all the lexical data would require a revision of the information related to variation, meanings, morphology and etymology, the material contained in this database is essential and could help in the development of new tools.

1.3. The BDLC project

The *Nouvel Atlas Linguistique et ethnographique de la Corse* (NALC) was initiated by the CNRS in 1975 and transmitted to Marie-José Dalbera-Stefanaggi in 1981 at the opening of the University of Corsica. In 1986, in response to a request from the Regional Assembly of Corsica, the *Banque de Données Langue Corse*[5] (BDLC) was created and was naturally linked to the atlas[6].

The NALC-BDLC hosts linguistic data related to Corsican know-how and cultural traditions throughout the island. During field surveys with local speakers, these data are collected through thematic questionnaires[7] made up of French wordlists[8]: for example *la vigne* ("the vine"), *le cep* ("the vine stock"), *tailler la vigne* ("pruning the vine"), *le tonneau* ("the barrel"). Based on a question such as *Cumu si dici "tailler la vigne" in corsu?* ("How do we say 'to prune the vine' in Corsican language?"[9]), the corresponding translations into Corsican are recorded and a semi-directed interview entirely in Corsican, and relating to practices, is undertaken in order to collect ethnotexts (testimonies). This data is then processed, linguistically analysed and put online on the website `http://bdlc.univ-corse.fr`.

If this database constitutes a real asset for the development of NLP applied to Corsican, one of the major difficulties comes from the rich variation characterising the Corsican language. According to the objects, significant lexical variations are documented: for example, 25 lemmas were collected to describe the act of pricking the vine out. These lemmas will in turn be affected by variable transcriptions, particularly as a result of the non-standardisation of the Corsican language and as a production carried out by various transcribers during the 30 years of the project existence. The different writing choices meet objectives such as:

- keep the dialectal richness, for example to name "the jar", according to the pronunciation in the localities surveyed, we will find the forms *cerra* and *gerra* from the same etymology;
- in some cases, explicitly indicate the placement of the tonic accent as well as the opening of the vowels, for example in *tróvula* rather than *trovula* / "bowl", *còmpulu* / *compulu* / "shelter", *pèrgula* / *pergula* / "arbour", *tépidu* / *tepidu* / "tepid" (proparoxytones vowels) or *durmìa* / *durmia* / "he was sleeping" (hiatus accentuation);

2. Development of NLP resources and tools for Corsican

2.1. State of the art

To our knowledge, there are very few resources and tools designed for Natural Language Processing (NLP) in Corsican. The ELDA 2014 report on linguistic resources dedicated to the languages of France (Leixa et al., 2014) lists 93 resources for Corsican. More than a third of these are recordings and transcriptions from the BDLC project. The rest is made up of various documents: blogs, scientific papers, institutional sites, newspapers sites, etc. There are also some lexicons, including *Infcor* or the Corsican version of the *Wiktionary*. In addition to this inventory, there are a few other contributions, including the *Speaking Atlas of the Regional Languages of France* (Boula de Mareüil et al., 2018), the *W2C* corpus (Majlis and Zabokrtský, 2012), the *BabelNet* semantic network (Navigli and Ponzetto, 2012), which offers a number of units in Corsican, and the Corsican resources created within the *Crúbadán Project* by Scannell (2007). Except for the last three, the majority of available resources are not directly usable for NLP.

Corsican therefore falls into the category of less-resourced languages. These languages are an active area of research. The *TALN* conference hosted several events dedicated to this issue, including the workshop *Traitement automatique des langues minoritaires et des petites langues* (Streiter, 2003), as well as the *TALaRE* workshops (Morin and Estève, 2013; Vergez-Couret et al., 2015). Similarly, the *LREC* conference hosted multiple workshops, the most recent of which are *SaLTMiL* (Alegria et al., 2010; De Pauw et al., 2012) and *CCURL* (Pretorius et al., 2014; Soria et al., 2016; Soria et al., 2018). Recently, the *TAL* journal also published a thematic issue on the subject (Bernhard and Soria, 2018). However, the place of the Corsican language in these publications is almost non-existent.

2.2. Initiatives for Corsican

Given this observation, we have decided to work to improve the situation of the Corsican language with regard to its place in the digital world, and more particularly in the field of Natural Language Processing. To achieve this objective and start tooling up the Corsican language, we rely on the BDLC project. In the following sections, we detail the progress of these different points and present our first achievements.

Our objective is to gradually build up the set of resources and tools that can form a Basic Language Resource Toolkit (BLARK), as presented by Krauwer (2003). This toolkit can then be used as a basis for the development of more ambitious applications. We follow the recommendations of Soria et al. (2013), particularly in adopting an open approach, using standards, allowing sharing, reuse and cooperation. In order to progress as quickly as possible, and in the hope that it will lead to emulation and collaboration, we wish to make our achievements available to the scientific community. Obviously, this approach can only be considered in accordance with copyright rules and other legal

[5] `http://bdlc.univ-corse.fr`

[6] A synthesis of the project history is presented by (Dalbera-Stefanaggi and Retali-Medori, 2015). The project has been directed since 2015 by S. Retali-Medori. A semi-popularisation collection entitled *Detti è Usi di paesi, matériaux et analyses extraits de la Banque de Données Langue Corse* was also created in 2006 around the NALC.

[7] The topics covered in the BDLC are: farming, agriculture, people, homes and daily life, nature, villages or cities and beliefs.

[8] The questionnaires were created at the beginning of the project through preliminary recordings made on the island about different technical or cultural topics. From their transcripts, the list of words, also named the *responsaire* by M. J. Dalbera-Stefanaggi (Dalbera-Stefanaggi, 1992, p.397), was established.

[9] The question is expressed in Corsican, but the term to be translated is in French.

constraints regarding the resources and tools that we will use.

The progress of the various currently ongoing works does not yet allow us to distribute a complete kit, or even fully developed resources. We set up a page dedicated to our project (http://bdlc.univ-corse.fr/tal/), which aims at progressively making the resources and tools available. This paper gave us the opportunity to upload the first elements of the future resource kit. We plan to expand this repository regularly in the coming months.

2.3. Corpus

We already set out the general problem of corpus building in Kevers and Retali-Medori (2020). We reproduce below the essential elements before presenting our new results.

Collecting corpora is generally among the top priorities for processing less-resourced languages. In addition to documenting the language, there are many uses for corpora, starting with comparing the intuition and linguistic knowledge of language specialists with large datasets. Corpora can also be useful for building lexical resources, for creating automatic processing tools, especially through machine learning, or even in the educational field.

This task faces two main obstacles: the availability of documents, preferably in a digital form, and their legal terms of use. Apart from the question of the existence of the documents, the first difficulty is essentially technical. The first step is to identify existing resources and process them according to their nature. Printed documents will have to be digitised. If they are already in a digital format, conversion operations[10] or even "harvesting"[11] may be necessary.

The second difficulty lies in respecting the rights that apply to this content. Indeed, the copyright laws do not generally allow their free and complete use, even for research purposes. This obstacle constitutes a real limitation for research in general, and for the digital development of less-resourced languages in particular, and has therefore been highlighted on many occasions, including by Zayed et al. (2016): *One of the big obstacles for the current research is the lack of large-scale freely-licensed heterogeneous corpora in multiple languages, which can be redistributed in the form of entire documents. [...] due to the restrictive license of the content, many corpora cannot be re-distributed because of the risk of copyright infringement*. The task of automatic corpora building from the web[12] is particularly affected by this problem. Tools proposed for this purpose, such as BootCaT[13] (Baroni and Bernardini, 2004) or Sketch Engine[14] (Kilgarriff et al., 2014), will be difficult to use if we intend to redistribute the resources and tools created from these corpora.

However, Directive 2019/790/EU of the European Parliament and of the Council on copyright and related rights in the Digital Single Market[15], adopted on 17 April 2019, should improve the situation. This text introduces new exceptions to copyright, in particular *for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access* (article 3, paragraph 1). It should be noted that this directive must be transposed into the national laws of the Member States in order to be implemented, which should be the case by 2021.

In the meantime, we do not know any corpus in Corsican that is ready to use. By this we mean a set of documents, in a structured format, for which we have easy access to plain text (without formatting, tags or other layout elements), and guarantees a legally sustainable use, mainly with regard to copyright issues. However, the exception of the Universal Declaration of Human Rights[16] should be noted, but it consists of an isolated text. Although *Wikipedia* can be used under a CC BY-SA 3.0 license, and is accessible as XML dumps, it requires cleaning work to isolate article pages and remove the different *wiki* tags from the text.

We therefore made a special effort to collect, clean, formalise in a standard format (XML TEI P5), and finally release a number of corpora. The currently available ressources are detailed in table 1[17]. In addition to the Corsican *Wikipedia* corpus, our data also comprises the Corsican translation of the *Bible*. This document is available on the Internet as a bilingual PDF (French-Corsican)[18]. We have obtained permission from the author to release our XML TEI version under a Creative Commons license. A third corpus consists of a set of articles from the journal *A Piazzetta*[19], which its publisher kindly shared with us, again under a Creative Commons license. This list will obviously evolve over time.

Corpus	Size (words)	License
Wikipedia co	919 382	CC BY-SA 3.0
Bible	770 560	CC BY-SA-NC 4.0
A Piazzetta	504 225	CC BY-SA-NC 4.0
TOTAL	**2 194 167**	

Table 1: Summary of the available corpora

2.4. Concordancer

One of our objectives, as part of the development of this first set of resources and tools for Corsican, is to set up an online corpora consultation interface, that allows the expression of queries based on linguistic criteria (complex patterns using lemmas, grammatical categories, etc.) and to get the results in the form of concordances. This type

[10] Such as switching from PDF to text format.

[11] For content published in the form of websites.

[12] An ACL *Special Interest Group* (SIG) is dedicated to this domain under the name of *Web AS Corpus* (SIGWAC - https://www.sigwac.org.uk/).

[13] https://bootcat.dipintra.it/

[14] https://www.sketchengine.eu/

[15] http://data.europa.eu/eli/dir/2019/790/oj

[16] Published by the United Nations (https://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=coi), but can also be obtained from the Unicode consortium website (http://unicode.org/udhr/) or from the NLTK data set (http://www.nltk.org/nltk_data/).

[17] The word count was estimated using *wc* under *Linux*.

[18] See the website https://www.dico-bible-corse.fr/.

[19] https://www.apiazzetta.com/

of interface is interesting for linguists to carry out research, explore data, check and illustrate with examples the linguistic theories they wish to defend. Filtering content according to metadata can allow for contrastive studies, which is particularly relevant in the case of a language with many synchronic and diachronic variations. Finally, from an educational point of view, it is also a tool that enables students to carry out practical work as part of their language learning.

This type of interface having already been developed by others, we initiated a collaboration with the Cental (Center for Natural Language Processing, UCLouvain[20]) in order to benefit from their concordancer. The data injected into the system was extracted from the BDLC.

We do not currently have the lemmatised versions of the texts that would enable querying on linguistic criteria. However, it seemed interesting to us to set up the tool with the unlemmatised data, even if it limits queries to the graphical forms contained in the corpus. We plan to update data later, when their lemmatised version is available. Access to the interface is provided from our project page.

2.5. Language identification tool

2.5.1. Introduction

Language detection is a well-known problem for Natural Language Processing. First, it is interesting to group documents by language in order to select only those written in some languages. Moreover, linguistic processing is more accurate if it is adapted to the language being processed. The Language identification component is important from the beginning of the tooling up of a language because it allows to build the best possible quality corpora, useful for the development of future tools.

Assigning a language to an entire document is a task that is generally well handled and for which very good results, close to perfection, are obtained. However, performance depends on the number of languages supported by the tool, the availability of learning resources in sufficient quantity and quality, and the proximity between the different languages covered. Detecting sections of different languages within the same document is also a slightly more complex task. It involves identifying changes from one language to another, which can be tricky, especially if the parts of the text does not correspond to document divisions such as paragraphs or sentences, or when the sections are very short (one or two words for example).

Unfortunately, we cannot draw up an extensive state of the art in this paper. However, a very detailed study can be found in (Jauhiainen et al., 2018).

2.5.2. Development for Corsican

As a first step, we focused on language identification for entire documents. We leave the identification of parts within a document for further research.

While many studies and a number of tools have been published, few of them have been tested on Corsican. There is indeed no difficulty in finding research papers or software for the detection of major languages, such as English, French, Spanish, Italian or German. On the contrary, it is much more difficult to find support for less widely used languages. The challenge is therefore to test different approaches and adapt them so that they can be applied to the Corsican language.

Due to the large number of publications, we had to select a few tools that we considered representative of the different methods and that offered, if possible, an implementation usable without major modification and a free license. The most frequent approaches are the use of characteristic *stopwords* in different languages, a long known method since Ingle (1976), as well as the use of *n-grams* of characters built from a learning corpus, as proposed amongst others by Cavnar and Trenkle (1994).

Our work consisted in gathering the necessary resources to test the different tools on Corsican as well as on a set of 8 other languages: English, French, Italian, Spanish, Portuguese, German, Dutch and Romanian. This choice was mainly driven by the availability of linguistic resources (word lists and learning corpora required for training). However, we made sure to select a number of important languages (such as English for example) as well as languages that could be perceived as "similar" to Corsican by the different processing methods (Italian or Romanian for example).

For the 8 languages other than Corsican, the learning corpora were created from the data available in the multilingual sentence database *Tatoeba*[21] (see table 2)[22].

Language	Words	Sentences	Weight (Ko)
English	8 945 380	1 165 661	46 691
Italian	4 197 656	711 983	25 093
German	3 639 723	458 646	22 701
French	2 895 933	381 935	16 878
Portuguese	2 212 496	318 355	12 786
Spanish	2 086 565	297 825	12 122
Dutch	580 185	86 496	3 251
Romanian	117 013	17 892	701

Table 2: Learning corpora (except Corsican)

For Corsican, we had to build a corpus from scratch by collecting a series of documents from, among other sources, *Wikipedia*, part of the Corsican translation of the *Bible*, educational content, articles from newspapers or blogs available online, etc. This set contains a total of 3 161 036 words, spreads over 7 904 documents, which places the Corsican corpus in the average of the other languages represented in this test.

The tools being tested are the following: *CueLanguage*[23] (stopwords), *Libre Office* version of *LibTextCat*[24] (n-grams), *Langid.py*[25] (n-grams), *Langdetect*[26] (n-grams), *FastText*[27] (n-grams) and *Ldig*[28] (maximal substring). Two

[20] https://cental.uclouvain.be

[21] https://tatoeba.org/eng/

[22] The word count was estimated using *wc* under *Linux*.

[23] https://github.com/jdf/cue.language

[24] https://github.com/LibreOffice/libexttextcat

[25] https://github.com/saffsd/langid.py

[26] https://github.com/shuyo/language-detection

[27] https://github.com/facebookresearch/fastText

[28] https://github.com/shuyo/ldig

additional methods were implemented by us to provide a baseline (one based on the distribution of letters or groups of two letters, the other on the stopword method).
The evaluation corpus was obtained for Corsican by subtracting 10% of the collected corpus, the rest being used for the learning process. For the 8 other languages, we used the parallel corpus of European Parliament acts proposed by Koehn (2004). For each language, we selected 100 of the largest documents from the sessions of September and October 2009. With a few exceptions, the documents are identical for each language.
The results obtained are shown in the table 3. The tool that offers the best performance is *Ldig*. This tool is particularly effective for small documents. It is nevertheless quite slow during the learning phase. As this is done only occasionally, this does not constitute an element of exclusion in our view.

Method	8 languages	Corsican	Total
MyLetterDistrib	99.62	93.04	98.89
MyStopWords	99.62	93.56	98.95
CueLanguage	99.50	84.41	97.82
LibreTextCat	100	95.62	99.51
Langid.py	98.75	95.23	98.36
Langdetect	100	96.65	99.63
FastText	100	95.49	99.50
Ldig	**100**	**98.58**	**99.84**

Table 3: Evaluation results for the language identification test (accuracy in %).

Based on these results, we have decided to make available, on our project page, a demonstration based on the *Ldig* software. Given the copyright implications, we cannot distribute neither the final model nor the Corsican corpus used for this evaluation. However, as mentioned in the section 2.3., we are currently working on making some part of this corpus, as well as other documents, available. When a set of about 3 million words will be reached, we are considering re-training a new model that could then be released.

2.6. Electronic dictionary

2.6.1. General approach

In the context of developing lexical resources for NLP, we have adopted a method that goes somewhat against the usual approach used for the BDLC and for the creation of linguistic atlases in general. While the latter aims at specifying as precisely as possible the use of different lexical units, and this in different geographical areas, our approach consists first and foremost in maximising the inventory of forms, even if they are somewhat different from what could be observed during the field surveys. In this way, the resource will be as broad and robust as possible for the analysis of texts potentially from a wide variety of styles and sources (literature, press articles, blogs, texts from social networks, etc.).
At a first step, this resource should not be seen as a finished product, but as an intermediate tool for the eventual creation of a real electronic dictionary that can be used for NLP. However, since its format is identical to the one of this type of dictionary, there is nothing to prevent it from being used as such in the meantime.
The use of this intermediate resource on carefully selected corpora can provide us with very interesting information. First, it will make it possible to distinguish the forms that are actually observed in the texts from those that are not[29], which will lead ultimately to the final version of the dictionary. Furthermore, data about the use of language extracted from the corpora can be compared with data collected during the field surveys carried out for atlases. While it is clear that the nature of linguistic information is not the same (the records made from speakers selected for their linguistic skills are compared with writings whose level is potentially very variable), the fact remains that the corpora textual content constitutes a proven use of the language. The conclusions to be drawn from this type of analysis may of course be diverse, ranging from educational recommendations to correct future speakers/writers on some expressions considered irrelevant, or even erroneous and therefore undesirable, to the identification of new or poorly known terms that could then be considered as integrated into the language in view of their observed frequency.
From a technical point of view, the data are organised according to the LADL dictionary format (Gross, 1989; Courtois, 1990; Silberztein, 1993) and processed using the Unitex software (Paumier, 2016). The entries are saved in text files in the following format: `form,lemma.grammatical_semantic_codes:flexional_codes/comment`.

2.6.2. The BDLC as a starting point

As explained in the section 1.3., field surveys have, for several decades, made it possible to gather in the BDLC a lot of lexical information. We therefore first built an initial version of the dictionary from an export of this database. In order to make it a coherent resource, a work of standardisation had to be carried out. We decided to follow recommendations of the Universal Dependencies project (Nivre et al., 2016), especially when it comes to choosing grammatical codes[30].
Currently, the dictionary has 21 108 forms[31], of which 18 079 are simple forms (referring to 10 248 lemmas) and 3 029 are compound forms (referring to 2 250 lemmas). The distribution between the different grammatical categories is shown in table 4.
When this dictionary is used to analyse our corpus of ethnotexts, also extracted from the BDLC - which is about 160 000 forms, just under 15 000 of which are unique - about 49% of occurrences are recognised. For these terms, several competing analyses may coexist (lexical ambiguity)

[29] This information may be included in an other version of the dictionary, either by simply removing the unconfirmed forms or by using a semantic code.

[30] In order to best comply with the UD recommendations, some items of the dictionary will probably have to be removed, such as expressions (LOC code which stands for *locution* in French.)

[31] This count includes some external additions regarding verbs. We have manually added the 763 forms related to verbs *esse* ("to be"), *avè* ("to have"), *andà* ("to go"), *dà* ("to say"), *fà* ("to make") and *stà* ("to be", state).

Category	Count
NOUN	14 928
VERB	3 481
ADJ	1 534
LOC	781
ADV	159
NUM	67
PRON	65
INTJ	93

Table 4: Dictionary, detailed count by grammatical category

and the correct analysis may be missing (due to the incompleteness of the dictionary). It should also be noted that processing the most frequent unrecognised forms would quickly improve coverage: the first 20 of these elements cover no less than 31% of the total unknown forms.
The work needed to improve this first dictionary has not yet been started, as it is planned to be carried out as part of the lemmatising task (see section 2.7.). This process should constitute a recurring contribution to the dictionary over the course of the project. We are also considering the possibility of integrating data from outside the BDLC.
As this resource is still under development, and the legal aspects of the underlying data we are using have yet to be processed, it is not currently possible for us to disseminate it. The objective, however, is to make it available through the project page when it becomes possible.

2.6.3. Enhancements for verbal forms

With regard to verbs, the challenge is not only to gather a number of infinitive forms, but also to have all the verbal paradigms. If we consider all the simple tenses, this represents almost 50 forms, which will largely increase when we also take into account the dialectal variations that may exist in verbal morphology in the Corsican language. In order to make the collection of a first set of verbal data, we choose to build automatic inflection transducers. It is a formalism that makes it possible to manipulate a canonical form in order to produce a series of morphological variants to which it is possible to attach the relevant flexional codes. This approach, already used by Steiblé and Bernhard (2016) for Alsatian, allows to automatically generate all the conjugated forms for any infinitive that has been attached to a verbal class[32]. For each defined class, an inflection grammar describing its morphological characteristics was constructed. This sort of grammar is represented in the form of a graph such as that shown in figure 1.
The aim of this approach is to provide a first resource for recognising the most regular forms of verbs. Its ambition is not to cover all the phenomena that can appear in different parts of the island, sometimes in a very local way[33]. Missing forms can be integrated later, among other things during lemmatisation projects (see section 2.7.) during which the electronic dictionary will be updated according to the content of the analysed corpora. Similarly, generated forms that are not observed may also be discarded.
The defined verbal classes are shown in table 5 and follow the classification established by Medori (1999).

Class	Ending of infinitive / past participle
V1	'à'
V2	'è', not implemented (only 7 verbs)
V3	'ì' and 'isce' or 'isca'
V4e	'e' with pp in 'utu'
V4a	'a' with pp in 'utu'
V5e	'e' with pp in 'itu'
V5a	'a' with pp in 'itu'

Table 5: Verbal classes defined for regular verbs

In order to maximise the coverage of the resource, we also consulted various other publications, such as the learning grammars of Romani (2000) or Comiti (2012), as well as the website *Cunghjugatori corsu* [34]. The consultation of these different sources revealed many differences. At this stage, we do not give any opinion on the relevance of these. As explained above, we will be able to carry out analyses of this kind when the lexical data will be confronted with corpora.
Semantic codes "Dn" (northern dialects) and "Ds" (southern dialects) were also used. These codes are not intended to define a very precise dialectal area but rather to provide information on an area in which the form is *a priori* present. The assignment of more precise information is very difficult because field studies have shown that the forms used can vary within a given geographical area, even to the extent of differences between neighbouring villages.
A list of 312 verbs with reference to the verbal class has been defined. This allowed the automatic inflection to be performed and a dictionary of verbal forms containing (40 526) elements to be generated.

Class	Count
V1	204
V3	39
V4e	20
V4a	20
V5e	14
V5a	15

Table 6: Distribution of the 312 verbs by class

The automatic inflection process, including the writing of inflection grammars, and their application to the list of verbs was performed using Unitex (Paumier, 2016). A grammar example is given in figure 1.

[32]This approach is particularly well suited for regular verbs.

[33]For example, Medori (1999, p. 229) notices the existence, in some villages of Cap Corse, of forms receiving the affixing of an enclitic: *un autre fait qui caractérise la conjugaison de l'imparfait dans le nord du Cap Corse: l'apposition systématique (ainsi qu'à d'autres temps et modes que l'imparfait de l'indicatif), d'un 'tu' enclitique à la seconde personne du singulier. J'ai observé ce phénomène à Luri et à Morsiglia: par exemple cantavatu "tu chantais" pour cantava ou cantavi [...]*

[34]http://aiaccinu.free.fr/conjugueur-corse/

After the step of automatic inflection, we added an additional treatment. This one is dedicated to the recognition and processing of a vowel shift phenomenon: apophony. It is a mutation of the vowel that occurs when there is a shift of the tonic accent to another syllable than the one where it was originally located (in relation to the form at the third person singular of the present tense). When the syllable that loses the accent contains the vowel 'e' this one turns into an 'i'. When it is an 'o', it will become a 'u'. The criteria for the occurrence of this phenomenon are therefore primarily phonetic, and it is quite difficult to detect it with great precision by using solely the orthographic transcriptions of words. The processing operation therefore aims to propose forms for which vowel alternation is *possible*. Given our approach to creating an intermediate resource that will be confronted with real data coming from corpora, this approximation is acceptable. Of the 40 526 automatically inflected verbal forms, 4 681 have been marked as potentially relevant for vowel alternation, and 1 860 eventually gave rise to a modified form (the original form being in any case preserved). Adding these forms to the initial list increases its number of lines to 42 386.

When this updated dictionary of verbs is used with the one extracted from the BDLC (section 2.6.2.) to analyse the same corpus of ethnotexts as before, 3 240 additional forms are recognised, which increased the coverage by 2% to 51%.

2.7. Part-of-speech tagging

Morphosyntactic analysis is a basic tool that is very frequently used as a pre-processing step before more complex NLP tasks. The development of such a tool generally involves making a manual or semi-automatic annotation of part-of-speech (*POS*) tags on a rather large learning corpus. This annotated corpus generally allows, through a supervised learning method, to obtain an annotation model that will be quite efficient.

The development of learning corpora is a substantial effort, especially in the case of less-resourced languages, for which it is generally necessary to start from scratch. Our intention is to work on it, and to this end we have started defining and developing a lemmatisation procedure (Kevers et al., 2019). This process, in addition to enriching lexical resources, will allow us to build a learning corpus for Corsican. This resource will eventually allow us to train the necessary models using tools such as the *Tree Tagger* (Schmid, 1994), *Stanford Part-Of-Speech Tagger* (Toutanova et al., 2003) or *Talismane* (Urieli, 2013). The latter is a particularly interesting alternative because it proposes a hybrid approach combining rule-based methodology to supervised artificial learning. Its effectiveness has been demonstrated by Vergez-Couret and Urieli (2014) for the analysis of less-resourced languages, particularly Occitan and its dialectal variations, despite the small size of the learning corpus (2 500 tokens). The lack of data is, in the case of Talismane, partly counterbalanced by the intensive use of a lexicon, as well as by the use of some rules.

However, there is another possibility to set up a POS tagger. The so-called *transfer* approaches make it possible to take advantage of the proximity of well-resourced languages by relying on their resources and tools. This strategy was used by Hana et al. (2011) who describe a tagger for the old Czech. The same approach has also been adopted by Bernhard and Ligozat (2013) regarding Alsatian dialects, which have many similarities with standard German, as well as by Vergez-Couret (2013), this time for Occitan from French and Castilian. As far as Corsican is concerned, its proximity to Italian could make it possible to set up this type of approach.

3. Conclusions and future work

The status of less-resourced language is no longer to be demonstrated for the Corsican language. We therefore undertook to gradually gather and build the resources and tools that will eventually constitute a Basic Language Resource Kit (BLARK). Practically speaking, the first steps we focused on are: the collection of corpora, the setting up of a concordancer, a language detection tool integrating Corsican, as well as the construction of a first version of an electronic dictionary.

Of course, there is still some work to be done before we have a complete resource kit. As for corpora, our intention is to continue the collection effort undertaken so far (a little more than 2 million words are available at present), in order to have a larger set of documents under appropriate licenses. A finer annotation of the sections that are not in Corsican could also be interesting. These corpora will allow us, among other things, to re-train the language detection tool, and therefore be able to make the underlying model available. In the future, the concordancer should receive lemmatised texts, which would make it possible to use requests involving linguistic criteria (grammatical codes, lemmas, etc.). To this end, we have begun to define a methodology and procedure for the lemmatisation of texts. This effort will also provide documents for training a POS tagger, as well as enrich our electronic dictionary. The latter could also be upgraded by extending the automatic inflection method to a larger list of regular verbs and by using external data sources. In any case, it will have to be refined and tested against corpora to result in a reliable and usable resource for NLP.

Depending on legal and copyright constraints, we hope to be able to make most of the resources and tools available to the scientific community. To this end, we have created a project page which will be used to disseminate our results: `http://bdlc.univ-corse.fr/tal/`.

4. Acknowledgements

We would like to thank Marie-José Dalbera-Stefanaggi who kindly agreed to take note of our work and share with us her knowledge on the Corsican language. Thanks to Mr. Christian Dubois (translator of the Bible into Corsican) and Mr. Petru Paulu de Casabianca (online newspaper *A Piazzetta*) for making their corpora available. Many thanks to Hubert Naets (Cental/UCLouvain, Belgium), for his work on the concordancer. Finally, this text would probably be less readable without Julia Medori's proofreading and corrections. This work was carried out thanks to a post-doctoral fellowship and the CPER fund allocated by the *Collectivité de Corse*.

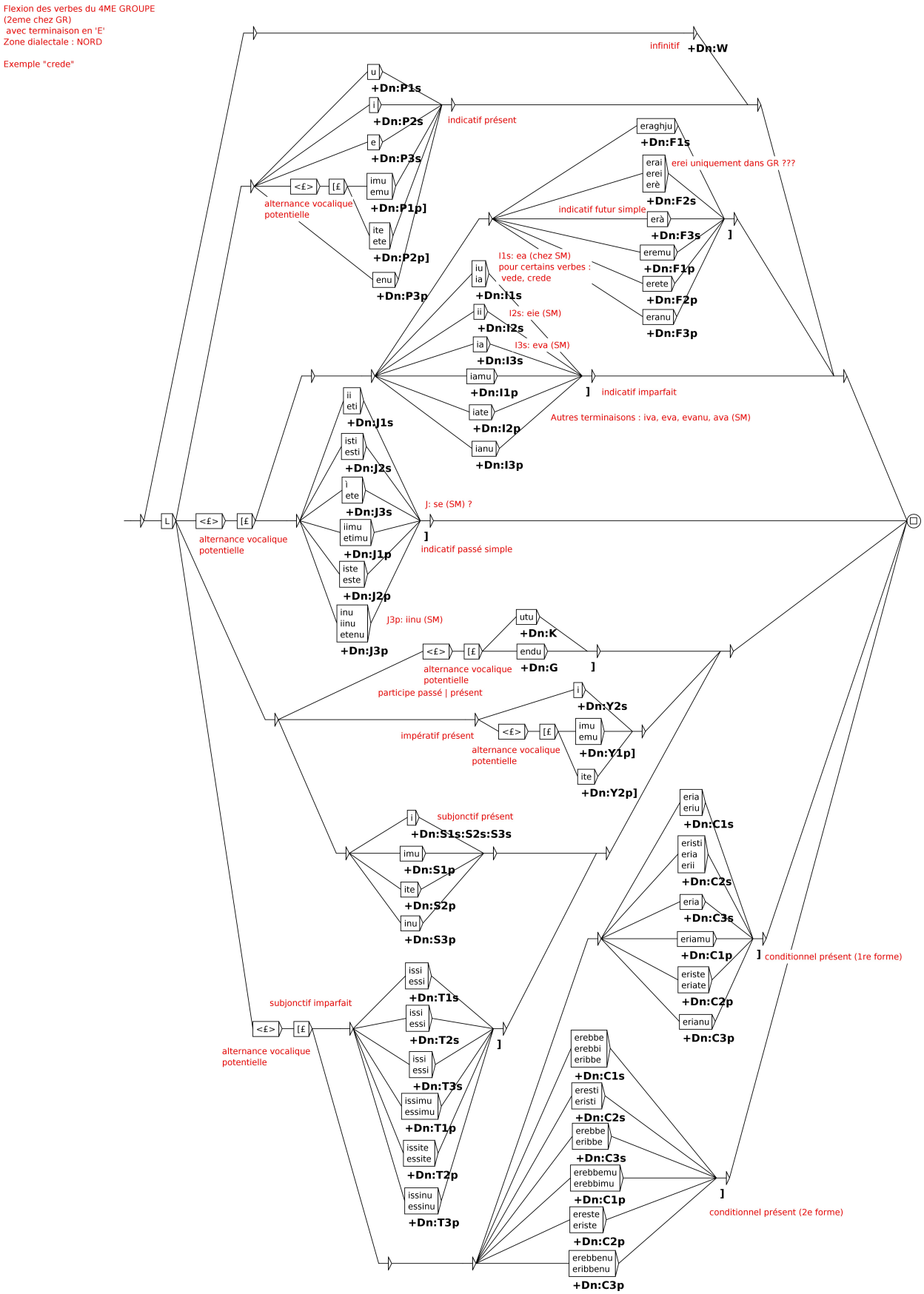

Figure 1: Inflection grammar for the “V4e” verbal class

5. Bibliographical References

Inaki Alegria, et al., editors. (2010). *7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC2010 Workshop)*, May.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC 2004)*.

Bernhard, D. and Ligozat, A.-L. (2013). Es esch fàscht wie Ditsch, oder net? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand. In *TALaRE, Traitement Automatique des Langues Régionales de France et d'Europe. Atelier TALN 2013*, pages 209–220, Les Sables d'Olonne, France, June.

Bernhard, D. and Soria, C. (2018). Traitement automatique des langues peu dotées. *Traitement Automatique des Langues*, 59(3).

Boula de Mareüil, P., Vernier, F., and Rilliard, A. (2018). A Speaking Atlas of the Regional Languages of France. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May.

Brenzinger, M., Dwyer, A. M., de Graaf, T., Grinevald, C., Krauss, M., Miyaoka, O., Ostler, N., Sakiyama, O., Villalón, M. E., Yamamoto, A. Y., and Zepeda, O. (2003). *Language Vitality and Endangerment. Document submitted by the Ad Hoc Expert Group on Endangered Languages to the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages*. United Nations Educational, Scientific and Cultural Organization, Paris. Available at `https://ich.unesco.org/doc/src/00120-EN.pdf`.

Cavnar, W. B. and Trenkle, J. M. (1994). N-Gram-Based Text Categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

Ceberio Berger, K., Gurrutxaga Hernaiz, A., Baroni, P., Hicks, D., Kruse, E., Quochi, V., Russo, I., Salonen, T., Sarhimaa, A., and Soria, C. (2018). *Digital Language Survival Kit. The DLDP Recommendations to Improve Digital Vitality*. The Digital Language Diversity Project. Available at `http://www.dldp.eu/sites/default/files/documents/DLDP_Digital-Language-Survival-Kit.pdf`.

Comiti, J.-M. (2012). *A practica è a grammatica*. Albiana, Ajaccio; Corte, December.

Courtois, B. (1990). Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, 87:11–22.

Dalbera-Stefanaggi, M.-J. and Retali-Medori, S. (2015). Trente ans de dialectologie corse : autour du programme Nouvel Atlas Linguistique et Ethnographique de la Corse et Banque de Données Langue Corse. In Stella Retali-Medori, editor, *Actes du colloque Tribune des chercheurs, études en linguistique*, volume 6 of *Corse d'hier et de demain - Nouvelle série*, pages 17–25, Bastia, France, June. Société des Sciences Historiques et Naturelles de la Corse.

Dalbera-Stefanaggi, M.-J. (1992). Le Nouvel Atlas Linguistique de la Corse et son articulation sur une base de données. In *Atlanti Linguistici italiani e romanzi: esperienze a confronto. Atti del Congresso Internazionale (Palermo 3-7 Ottobre 1990)*, pages 395–402, Palermo. Centro di Studi Filologici e Linguistici Siciliani.

Dalbera-Stefanaggi, M.-J. (2002). *La langue corse*. Number 3641 in Que sais-je? PUF, Paris, June.

Dalbera-Stefanaggi, M.-J. (2007). *Nouvel atlas linguistique et ethnographique de la Corse : Volume 1, Aréologie phonétique, édition revue et corrigée*. Comité des travaux historiques et scientifiques - CTHS, Ajaccio : Paris, Alain Piazzola edition, December.

Guy De Pauw, et al., editors. (2012). *8th SaLTMIL & AfLaT 2012 Workshop, Language Technology for Normalisation of Less-Resourced Languages (LREC 2012 Workshop)*.

Gross, M. (1989). La construction de dictionnaires électroniques. *Annales de Télécommunications*, 44:4–19.

Hana, J., Feldman, A., and Aharodnik, K. (2011). A low-budget tagger for Old Czech. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 10–18, June.

Ingle, N. C. (1976). A language identification table. *The Incorporated Linguist*, 15(4).

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., and Lindén, K. (2018). Automatic Language Identification in Texts: A Survey. *arXiv:1804.08186 [cs]*, April. arXiv: 1804.08186.

Kevers, L. and Retali-Medori, S. (2020). Copyright in the context of tooling up corsican and other less-resourced languages. In *Proceedings of the International Conference on Language Technologies for All (LT4All), Enabling Linguistic Diversity and Multilingualism Worldwide*, Paris, France.

Kevers, L., Guéniot, F., Tognotti, A. G., and Retali-Medori, S. (2019). Outiller une langue peu dotée grâce au TALN : l'exemple du corse et de la BDLC. In *Actes de la 26e conférence sur le Traitement automatique des langues naturelles (TALN)*, Toulouse, France, July.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1):7–36, July.

Koehn, P. (2004). EuroParl: A parallel corpus for statistical machine translation. 5, November.

Krauwer, S. (2003). The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of the 2003 International Workshop Speech and Computer (SPECOM 2003)*, pages 8–15.

Leixa, J., Mapelli, V., and Choukri, K. (2014). *Inventaire des ressources linguistiques des langues de France*. ELDA, September. Available at `http://portal.elda.org/en/projects/archived-projects/review-existing-lrs-france/`.

Majlis, M. and Zabokrtský, Z. (2012). Language Richness of the Web. In *Proceedings of the Eighth International*

Conference on Language Resources and Evaluation, Istanbul, Turkey, May.

Marcellesi, J.-B. (1984). La définition des langues en domaine roman : les enseignements à tirer de la situation corse. In *Actes du Congrès de Linguistique et de Philologie Romanes 5*, pages 307–314, Aix-en-Provence.

Medori, S. (1999). *Les parlers du cap corse : une approche microdialectologique*. Phd thesis, Université de Corse, Corte.

Emmanuel Morin et al., editors. (2013). *TALaRE 2013 - Traitement Automatique des Langues Régionales de France et d'Europe. Atelier TALN 2013*, Les Sables d'Olonne, France.

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

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Paumier, S. (2016). *Unitex 3.1 User Manual*. Université Paris-Est Marne-la-Vallée, March. Accessible à l'adresse http://releases.unitexgramlab.org/3.1/man/Unitex-GramLab-3.1-usermanual-en.pdf.

Laurette Pretorius, et al., editors. (2014). *CCURL 2014 : Collaboration and Computing for Under - Resourced Languages in the Linked Open Data Era, LREC 2014 Workshop*.

Retali-Medori, S. (2015). La documentation corse. In Maria Iliescu et al., editors, *Anthologies, textes, corpus et sources des langues romanes*, number 7 in Manuals of Romance Linguistics, pages 558–564. De Gruyter, Tübingen.

Romani, G. (2000). *Grammaire corse pour le collège et l'école*. G. Romani, Ajaccio.

Scannell, K. P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In Cédrick Fairon, et al., editors, *Proceedings of the 3rd Web as Corpus Workshop*, volume 4 of *Cahiers du Cental*, Louvain-la-Neuve, Belgium.

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

Silberztein, M. D. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Masson, Paris.

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

Claudia Soria, et al., editors. (2016). *CCURL 2016 : Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity (LREC 2016 Workshop)*, May.

Claudia Soria, et al., editors. (2018). *CCURL 2018 : 3rd Workshop on Collaboration and Computing for Under-Resourced Languages, Sustaining knowledge diversity in the digital age (LREC 2018 Workshop)*.

Steiblé, L. and Bernhard, D. (2016). Vers un lexique ouvert des formes fléchies de l'alsacien : génération de flexions pour les verbes. In *Actes de la 23ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, volume 2, pages 547–554, Paris, France.

Olivier Streiter, editor. (2003). *Traitement automatique des langues minoritaires et des petites langues, Actes du Workshop TALN 2003*, Batz-sur-Mer. ATALA.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich Part-of-speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, pages 173–180. Association for Computational Linguistics.

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

Vergez-Couret, M. and Urieli, A. (2014). Pos-tagging different varieties of Occitan with single-dialect resources. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 21–29, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Marianne Vergez-Couret, et al., editors. (2015). *TALaRE 2015 - Traitement Automatique des Langues Régionales de France et d'Europe. Atelier de TALN 2015*, Caen, France.

Vergez-Couret, M. (2013). Tagging Occitan using French and Castillan Tree Tagger. In *Less Resourced Languages, new technologies, new challenges and opportunities*, page 6, Poznan, Poland, December.

Zayed, O., Habernal, I., and Gurevych, I. (2016). C4corpus: Multilingual Web-size Corpus with Free License. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May.