12

Language Documentation & Conservation Special Publication No. 9 (January 2016):
Language Documentation and Conservation in Europe
ed. by Vera Ferreira and Peter Bouda, pp. 133–149
http://nflrc.hawaii.edu/ldc/
http://hdl.handle.net/10125/24675

BaTelÒc: A Text Base for the Occitan Language[1]

Myriam Bras and Marianne Vergez-Couret
CLLE-ERSS, UMR 5263 Université Toulouse Jean Jaurès & CNRS

Language Documentation, as defined by Himmelmann (2006), aims at compiling and preserving linguistic data for studies in linguistics, literature, history, ethnology, sociology. This initiative is vital for endangered languages such as Occitan, a romance language spoken in southern France and in several valleys of Spain and Italy. The documentation of a language concerns all its modalities, covering spoken and written language, various registers and so on. Nowadays, Occitan documentation mostly consists of data from linguistic atlases, virtual libraries from the modern to the contemporary period, and text bases for the Middle Ages. BaTelÒc is a text base for modern and contemporary periods. With the aim of creating a wide coverage of text collections, BaTelÒc gathers not only written literary texts (prose, drama and poetry) but also other genres such as technical texts and newspapers. Enough material is already available to foresee a text base of hundreds of millions of words. BaTelÒc not only aims at documenting Occitan, it is also designed to provide tools to explore texts (different criteria for corpus selection, concordance tools and more complex enquiries with regular expressions). As for linguistic analysis, the second step is to enrich the corpora with annotations. Natural Language Processing of endangered languages such as Occitan is very challenging. It is not possible to transpose existing models for resource-rich languages directly, partly because of the spelling, dialectal variations, and lack of standardization. With BaTelÒc we aim at providing corpora and lexicons for the development of basic natural language processing tools, namely OCR and a Part-of-Speech tagger based on tools initially designed for machine translation and which take variation into account.

1. INTRODUCTION. Occitan is a Romance language, spoken in southern France and in several valleys of Spain and Italy. The number of speakers is hard to estimate: According to several studies, it can be evaluated between 600,000 to 2,000,000 (Martel 2007, Sibille 2010). Occitan is not a unitary language, it has several varieties. The most accepted classification of Occitan dialects was suggested by Bec (1995) and includes Auvernhàs, Gascon, Lengadocian, Lemosin, Provençau, and Vivaro-aupenc.

Occitan is not standardized as a whole. Nevertheless, it is written since the Middle Ages and has a very important literary tradition. Its literature has been translated to other languages (Mistral, Boudou, Rouquette, Manciet, etc.). Although much less socialized than it was before the Second World War, Occitan is now present in newspapers, on the internet, on

[1] This project was carried out with the help of the *Région Midi-Pyrénées*, which has funded, together with the University of Toulouse-Le Mirail, a two-year post-doctoral fellowship devoted to BaTelÒc.

ISBN: 978-0-9856211-5-5

the radio and television, and in some public schools and universities. Non-governamental organizations maintain and spread Occitan: the *Felibrige*, the *Institut d'Estudis Occitans*, the associative network of imersive schools *Calandreta*, the linguistic training institute for adults – *Centre de Formacion Professionala Occitan*. However, Occitan has no official status in France.

In this paper, we present a text base for the Occitan language, called BaTelÒc. Nowadays, Occitan documentation mostly consists in dialectological data (several regional linguistic atlases are gathered in the THESOC database searchable online[2]), digitized lexicographic data (few bilingual dictionaries are searchable online[3]), virtual libraries (books in PDF format) from the modern (Bibliotheca Tholosana Occitana,[4] 16th–18th) to the contemporary period (CIEL d'Òc,[5] 19th–21st), and machine-readable texts of the Middle Ages (Concordance of Medieval Occitan (Ricketts et al. 2001) and Linguistic Corpus of Old Gascon (Field 2013)). In addition, the CIRDOC (International Center for Occitan Documentation) is developing a multimedia library, Occitanica[6] which offers access to a multiplicity of sources: written texts, images, virtual exhibitions, documentary films, sound records, etc. The BaTelÒc project aims at complementing those resources with machine-readable texts for modern and contemporary periods (see Bras 2006, Bras & Thomas 2011 for a description of the text base experimental version). It aims at developing a wide range of text collections by gathering written literary texts (prose, drama and poetry) and others genres such as technical texts and newspapers, and also by embracing dialectal and spelling variations.

This resource will be relevant for the linguistic description of Occitan in lexicology, morphology, syntax, semantics, and discourse studies. It could also provide data for lexicographic projects (Bras & Thomas 2007) and for studies in literature, anthropology, ethnology, history, etc. And last but not least, the text base is also meant for teachers and new speakers by giving real language uses in learning competencies.

In this paper, we first discuss the aims of BaTelÒc (Section 2): As Occitan is an endangered language, building a data base of Occitan texts can be regarded as a language documentation process; at the same time, as there is a big amount of written texts, the text base can be built in a corpus linguistic perspective. We then present the first version of BaTelÒc (Section 3) and end with some perspectives (Section 4).

2. Text Bases within the fields of Language Documentation and Corpus Linguistics.

2.1. Language Documentation and Corpus Linguistics. Language Documentation, as defined by Himmelmann (2006), aims at compiling and preserving linguistic data for studies in linguistics, literature, history, ethnology, sociology. It should include all language modalities, covering spoken and written language, various registers, for the study of the language as social practice and cognitive faculty. The special situation of endangered languages, namely the unsustainable small and declining speaker base, influences greatly the objectives: to collect on an opportunistic basis all the possible varieties, first and foremost spoken data (because if writings remain, spoken words fly away!). Language

[2] http://thesaurus.unice.fr/index.html (10 December, 2015).
[3] http://www.locongres.org (10 December, 2015).
[4] http://www.bibliotheca-tholosana.fr/bth/pageStatique.seam (10 December, 2015).
[5] http://www.cieldoc.com/fr.htm (10 December, 2015).
[6] http://occitanica.eu (10 December, 2015).

Documentation offers well-documented material which might serve for linguistic analysis. From this material, linguists can extract a coherent corpus for their own specific studies (Cox 2011).

Corpus Linguistics has been mostly developed in response to the problems caused by introspective methods. It consists in studying languages based on attested examples. It mostly concerns well-studied languages and is based on specific goals of linguistic studies. This generally requires the development of big corpora (millions of words), "sampled texts, written or oral, in machine readable form" (McEnery et al. 2006: 4, cited by Mosel 2013), to be representative of the language variety under study.

2.2. TEXT BASES. Text bases can be seen as the result (an online production) of both language documentation and corpus linguistics projects. Data in text bases must then consist of marked-up, organized, and well-documented machine-readable authentic texts. Those data can therefore be gathered as corpus. Text bases are available online with a range of tools that allow the user to select his/her own corpus and interrogate it (concordancer, frequencies, ...). The efforts to create text bases range from well-studied languages such as English, French, Portuguese, etc., to vulnerable languages[7] such as Basque or even endangered languages such as Picard.

The *British National Corpus* (BNC) is probably the most famous general corpus. It has been created by the industrial/academic BNC Consortium (led by Oxford University Press and composed of the following publishers: Addison-Wesley Longman and Larousse Kingfisher Chambers, and academic research centers: Oxford University Computing Services (OUCS), the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University, and the British Library's Research and Innovation Centre). The corpus is designed to represent a wide range of modern British English, from written (newspapers, letters and memoranda, school and university essays) to spoken data (everyday conversations, formal business conversations, radio material).[8] Other corpora have been designed as the BNC to allow comparisons across genre and provide a reliable basis for contrastive language studies, namely the *American National Corpus*, the *Korean National Corpus*, and the *Polish National Corpus*. There are several ways to access BNC online, for instance through the access provided by the Brigham Young University (the BYC-BNC).[9] The Brigham Young University also provides other corpora with the same exploring tools (see Table 1 below), such as for example the *Corpus of Contemporary American English*.[10]

Another well-known text base is *The Bank of English* (BoE), extracted from the 2.5-billion-word *Collins Corpus*. It is a balanced corpus of current English regarding genres and diatopic variation. It includes a wide range of written (websites, newspapers, magazines and books) and spoken data (everyday conversations, radio and TV material). The text base also considers diatopic variation including material mainly from UK, US, Australia, and Canada but also from India, New Zealand, South Africa, and Ireland. Moreover, BoE is constantly updated with new material. This base was designed to support the development of COBUILD dictionary by giving real-life examples, collocations, and frequencies.

As for French, *Frantext*[11] was designed to support the creation of *Trésor de la langue française*. Frantext has been developed by ATILF laboratory (Analyse et Traitement

[7] The classification chosen here is the one established by UNESCO.
[8] http://www.natcorp.ox.ac.uk (10 December, 2015).
[9] http://corpus.byu.edu/bnc (10 December, 2015).
[10] http://corpus.byu.edu (10 December, 2015).
[11] http://www.frantext.fr (10 December, 2015).

Informatique de la Langue Française). It is a reference text base in French linguistics and literature. It includes a collection of nearly 4,800 texts comprising 286 million words (literary, scientific, and technical) from the 12th to the 21st century. Unlike the two corpora described above, Frantext is not a balanced corpus. Almost all the text bases offer the possibility to select texts to work on. But in the case of Frantext, this step is achieved separately. Exploring the text base is then done in two steps: first building a corpus and then searching for concordances. The main advantage is that once the corpus is defined, it may be used for the entire work session.

Considering Catalan, an example of a more recent official language, the *Corpus Textual Informatitzat de la Llengua Catalana*[12] was created mostly to support the development of a lexicographic project (*Diccionari del Català Contemporani*) by the Institut d'Estudis Catalans. It only includes written texts (for instance literary, newspaper, technical, and scientific materials).

However, not all text bases are designed for lexicographic purposes and many other rich-resourced languages have their text bases online as for instance the *Reference Corpus of Contemporary Portuguese (CRPC)*[13] developed by the Centro de Linguística da Universidade de Lisboa. It contains written (literary, newspaper, technical, scientific, didactic, etc.) and spoken material (formal and informal conversation) and considers diatopic variation as well (Europe, Brazil, Angola, Cape Verde, Guinea-Bissau, Mozambique, São Tomé and Principe, Goa, Macau, and East-Timor). Only the written part of the corpus is available online.

Text bases are being also created for vulnerable languages such as Basque and endangered languages such as Picard. The *XX. Mendeko Euskararen Corpus Estatistikoa* has been developed by the Real Academia de la Lengua Vasca. It includes 6,351 texts organized by historical period, dialects (Bizkaiera, Gipuzkera, Zuberera, Lapuertera-Nafarrera, Unified Basque) and genre (administrative, literary, scientific, newspaper writing). *Picartext*[14] has been developed by CERCLL-LESCLaP (Laboratoire d'Etudes Sociolinguistiques sur les Contacts de Langues et la Politique Linguistique, EA 4283). It includes 300 written texts organized by genre (literature, letters, songs, comics, and dictionary) and diatopic features (by reference to the place of the authors – different French and Belgian districts).

As far as technical aspects of text bases are concerned, the *Text Encoding Initiative* seems to be chosen by all bases as a means to provide and classify texts. The different bases also provide tools for analyzing languages and/or help lexicographic work, for instance collocation, frequency, and concordance checks to observe different word usages. Most of them are enhanced with linguistic annotations, Part-Of-Speech tagging (POS), and lemmatization when relevant. Text bases of under-resourced languages (as for example Basque and Picard) normally do not provide such kind of annotation. Texts and tools are made available online to a wide range of users: academic researchers, general public, learners, new speakers, etc. Table 1 offers a synthesis for all the text bases presented above.

2.3. BUILDING A TEXT BASE FOR THE OCCITAN LANGUAGE. We are developing a text base for Occitan, called BaTelÒc. As Occitan is an endangered language, building a data base of Occitan texts can be regarded as a language documentation process: Our

[12] http://ctilc.iec.cat (10 December, 2015).
[13] http://www.clul.ul.pt/en/resources/183-reference-corpus-of-contemporary-portuguese-crpc (10 December, 2015).
[14] http://www.u-picardie.fr/LESCLaP/PICARTEXT/Public/index.php (10 December, 2015).

Text Bases	BYC-BNC	BoE	Frantext	CTILC	CRPC	XX MECE	Picartext
Language	English	English	French	Catalan	Portuguese	Basque	Picard
Language Status	Official language	Official language	Official language	Recent official language	Official language	No official status at country level	No official status
UNESCO Atlas of the World's Languages in Danger						Vulnerable language	Endangered language
Corpus Type	Balanced corpus	Balanced corpus	Collection of texts	Balanced corpus	General corpus	Diversified corpus	Diversified corpus
Content	Written and oral British English	Written and oral Diatopic variations	Written	Written	Written and Oral Diatopic variations	Written	Written
Period	1980-1993	2001-2005	1100-2012	1832-1988	1850-2006	1900-1987	1700-2000
No. words	100 million	650 million	270 million	50 million	300 million	4,6 million	8 million
Markup	XML TEI	?	XML TEI	?	?	XML TEI	XML TEI
Define your own corpus	by genre	by genre, domain, country, period	by author, title, genre and period	One corpus	by genre and country	by genre, period and dialect	by genre, period and region of the author
Annotations provided	POS	Lemma POS	Lemma POS	Lemma POS	Lemma POS	Lemma	
Tools	Concordancer Frequencies Collocations	Concordancer Frequencies Collocations	Concordancer (Regular expressions) Frequencies Collocations	Concordancer	Concordancer Frequencies	Concordancer Frequencies	Concordancer (Regular expressions)

TABLE 1: Examples of Text Bases

goal is to gather written literary texts (prose, drama, and poetry) and others genres, such as technical texts and newspapers, of modern and contemporary periods (from the 16th to the 21st century) and to embrace dialectal and spelling variations. There is enough material available to foresee a text base of hundreds of millions of words. The texts are well-documented with several types of metadata (genre, author's name, year of author's birth, dialect, year of publication, ...). The text base provides tools to allow feature-oriented selection of texts that can, thus, be gathered as a study corpus.

	Typical traditional corpus	BaTelÒc Text Base	Language Documentation corpus
Language status	Well-studied languages	Endangered languages[15]	Less-studied languages Endangered languages
Content	Oral and written data	Written data	Recordings, transcriptions, translations
Resources gathered by	Team of native speakers	Teams of non-native speakers	One non-native speaker
Size	Millions of words	More than one million words	Less than one million words
Availability of data	Huge amount of data growing every day	Big amount of data	Few amount of data
Goals	Linguistic research	Preservation, linguistic and interdisciplinary research	Preservation, linguistic and interdisciplinary research
Compilation	Representative sample of one variety under study	All the possible opportunities	All the possible opportunities

TABLE 2: BaTelÒc Text Base between language documentation corpus and typical traditional corpus, inspired by Mosel (2013)

3. THE BATELÒC TEXT BASE. For now, three million words have already been compiled within the text base. The texts are organized in nine genres (novel, poetry, memoir, short-story, fairy-tale, technical text, essay, song, treaty) and in five dialects (Gascon, Lengadocian, Lemosin, Provençau, and Auvernhàs). From a technical standpoint, BaTelÒc consists of texts encoded according to the international standard for sharing files on the internet, i.e. XML Text Encoding Initiative TEI P5, to ensure an accurate dissemination and the reusability of texts in the base and to provide high performance tools to select corpus (see Section 3.1).

BaTelÒc is designed to provide tools for linguistics studies, be it corpus linguistics or descriptive linguistics. The text base may be explored using a search engine that includes a concordancer to extract forms (word, part of word or sequence of words). It also includes more complex enquiries through the use of regular expressions (see Section 3.2).

[15] According to UNESCO Atlas of the World's Languages in Danger, http://www.unesco.org/culture/languages-atlas/index.php?hl=fr&page=atlasmap (10 December, 2015).

3.1. BUILDING THE TEXT BASE. All the texts in the base are encoded according to XML TEI P5 format. XML is a computer language that defines a set of rules for encoding text segments using markups (recognizable by the use of angle brackets). This format is both human-readable and machine-readable. Every document is decomposed in two main parts. The head contains all the metadata about the document (author's name, year of author's birth, year of publication, dialect, spelling, ...). The body contains the whole document. Markups assign to text segments characteristics such as text formatting (bold, italic, ...) and document structure (paragraph, title, subtitle, ...). XML format is not a unitary format for all documents. It provides a markup syntax to be defined depending on needs and document types. Within BaTelÒc, we follow the Text Encoding Initiative P5 markup schema, which is a norm to encode digitized texts created for librarians and research in order to emphasize simplicity, generality, and usability of digitized texts and to ease search within texts. Figure 1 gives an extract of a BaTelÒc text in XML format.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!–<!DOCTYPE TEI SYSTEM "TeiP5.dtd">–>
<TEI>
    <teiHeader>
                        ..... Un centenat de linhas de metadonadas .....
     </teiHeader>
    <text>
        <front>
                <doctitle><hi rend="M">l'estilò negre de la pluma d'aur</hi> </doctitle>
        </front>
        <body>
                <div type="preface">
                <p>Sèm plan astrucs ! Òc ben, plan astrucs ! Ai l'estilò negre de la pluma d'aur, l'estilò de
totas las istòrias, de totas las jòias e de totas las lagremas, de tots los espèrs e de totas las ràbias. </p
                        ..... Aqui i a tot lo tèxte ondrat de gavitèls .....
        </body>
    </text>
</TEI>
```

FIGURE 1: Extract of a BaTelÒc text in XML format

Several steps are needed in order to get a XML file from a rtf or doc file (see Figure 2 below): Manual pre-treatment to clean-up and mark-up for example the title of sections and sub-sections. The most important part of body markup is done automatically with a Perl program, for instance the markup of paragraphs, dialogs. All the metadata are saved in an Access database which generates the head of the file. Both elements together constitute the entire xml file.

3.2. BUILDING TOOLS TO SEARCH THE TEXT BASE. The exploration of the text base is done is two separated steps. For beginners, a "discovery corpus" is preselected as a default setting and it is searchable from the homepage. For advanced users, the basic consultation mode starts with building his/her own specific study corpus, based on various criteria. Once the corpus is defined, it may be used for the entire session. Tools are then provided to search for concordances.

3.2.1. CORPUS DEFINITION. Tools offer the possibility to select texts according to various criteria: author's name, title of the book, year of publication, genres, dialects, spelling

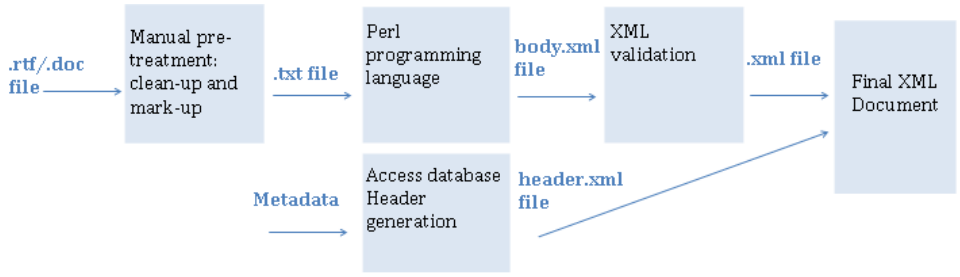

FIGURE 2: Follow up processing to build a xml file

norms. This allows the definition and adaption of the corpus to specific research goals (for instance, a corpus organized according to genre for descriptive linguistic purposes, a corpus organized by time periods in order to study the diachronic evolution of Occitan, a corpus organized by dialects to study diatopic variation or a corpus organized by authors to study diastratic and diaphasic variation). Figure 3 shows an example where all the novels (*roman*) of the text base are selected.

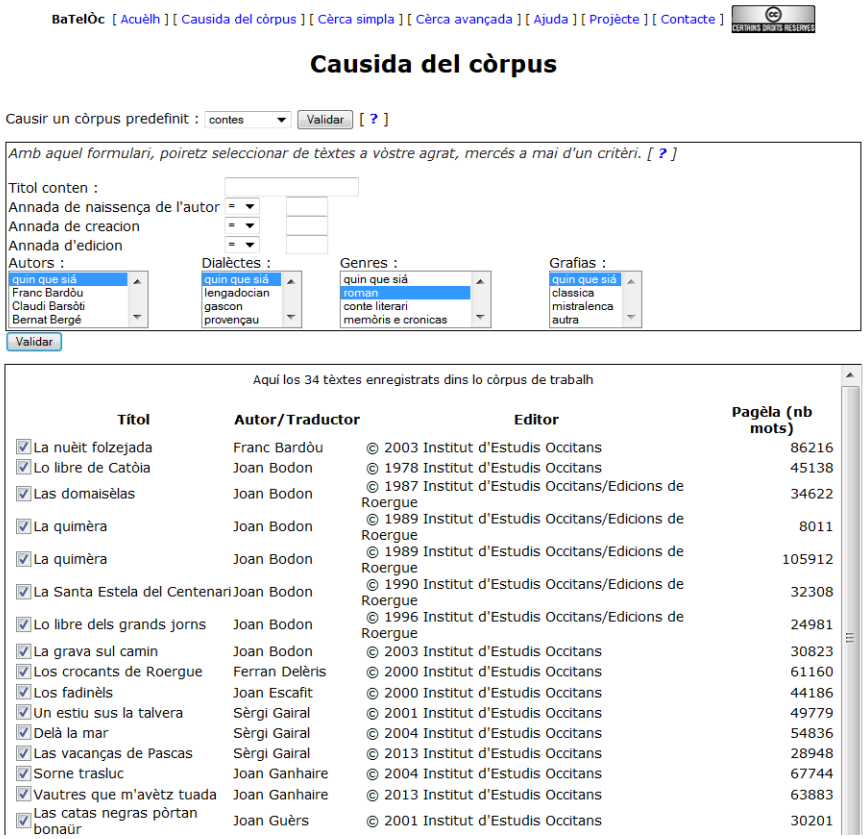

FIGURE 3: Building a corpus of novels

3.2.2. CONCORDANCE SEARCH. BaTelÒc provides tools to search for concordances of words or forms (for instance the word *çaquela* "however"). The results showing forms in context may be displayed in a narrow context (or KWIC), as in Figure 4, or in a larger context as in Figure 5.

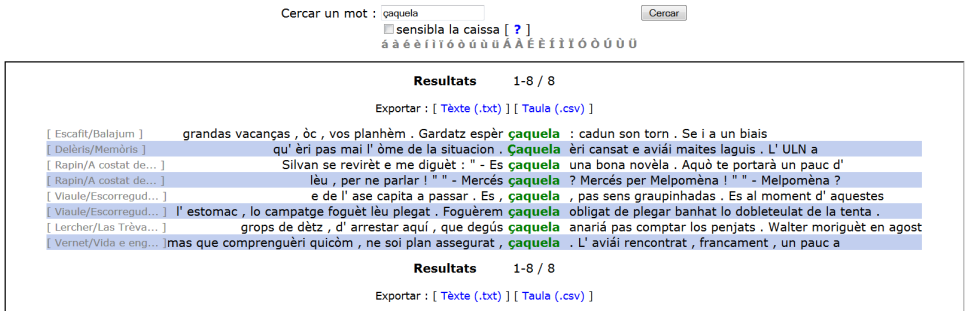

FIGURE 4: Searching for the form *çaquela* in a narrow context

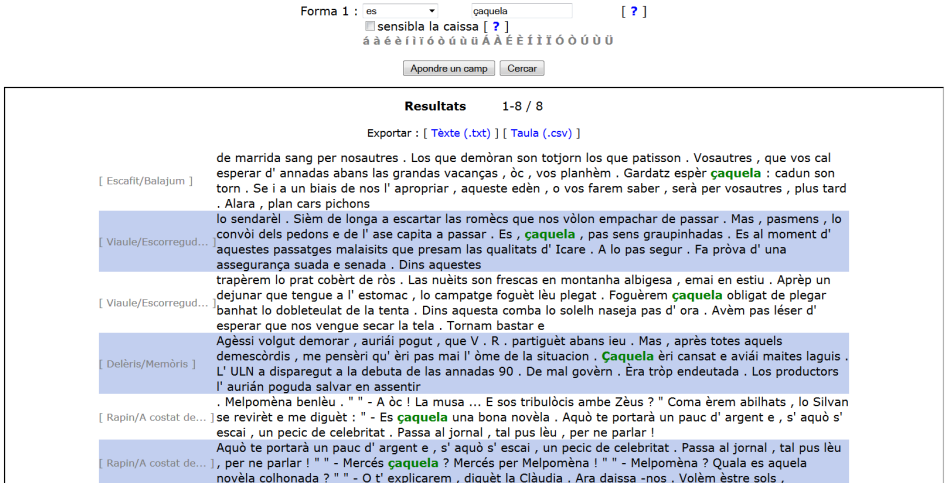

FIGURE 5: Searching for the form *çaquela* in a larger context

BaTelÒc also provides tools to look for two or three forms in a larger context. Figure 6 shows an example of *costat* "side" and *autre* "other" as search forms within an interval between 1 and 9 forms after the first form.

BaTelÒc search engine also includes more complex enquiries using regular expressions. Regular expressions are special text strings that allow more complex searches, i.e. through regular expressions it is possible to look for many different forms within one request. The disjunction (|) is used to look for alternative forms. For example, the following request (*es|ei*) (*aquò|çò*) allows you to search for *es aquò*, *es çò*, *ei aquò*, *ei çò* – "it is that" (see Figure 7).

The functions "starts with" (*comença per*) and "ends by" (*s'acaba per*) can be used to look for regular forms inside of words such as prefixes and suffixes. Figure 8 presents an example of search for the suffix *–òt* (note that not all the found forms are suffixed words,

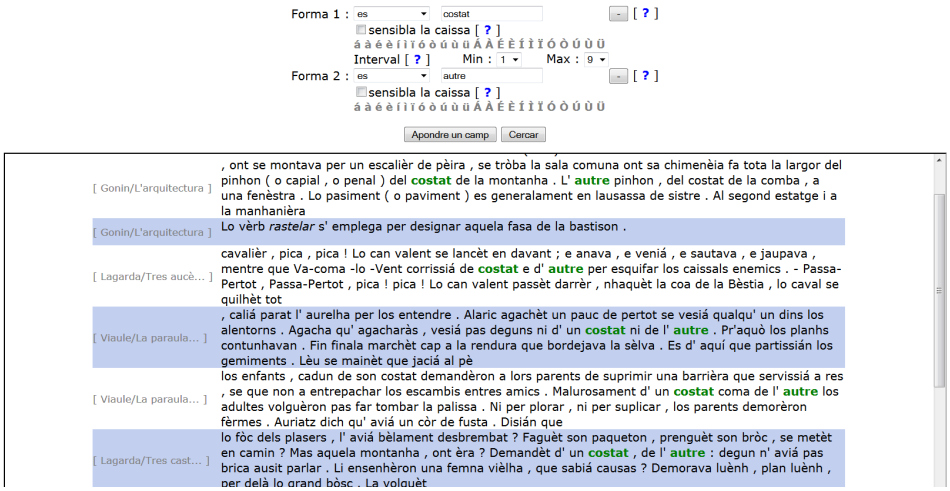

FIGURE 6: Searching for one form (*costat*) followed by another form (*autre*)

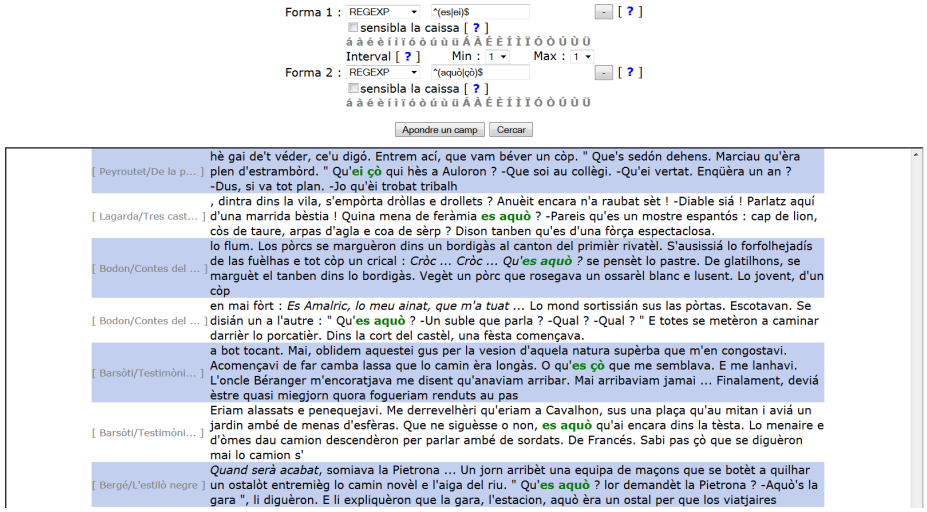

FIGURE 7: Searching for sequences of words *es aquò*, *ei aquò*, *es çò*, *ei çò*

as for example *pòt* "can" which is a finite verb). To look for all possible alternate forms (masculine, feminine, singular, plural), we can use the REGEXP function and the request may be formulated as follows: *..òta?s?$* (the two first dots aim at avoiding the very frequent verb form *pòt*) as is shown in Figure 9.

3.3. ENRICHING BATELÒC WITH LINGUISTIC ANNOTATIONS. Text base development logically involves enrichment of the texts with linguistic annotations. But Natural Language Processing (NLP) of endangered language such as Occitan is very challenging. It is not possible to transpose directly existing models for rich-resourced languages, partly because of the spelling and dialectal variations. We aim at providing corpora and lexi-

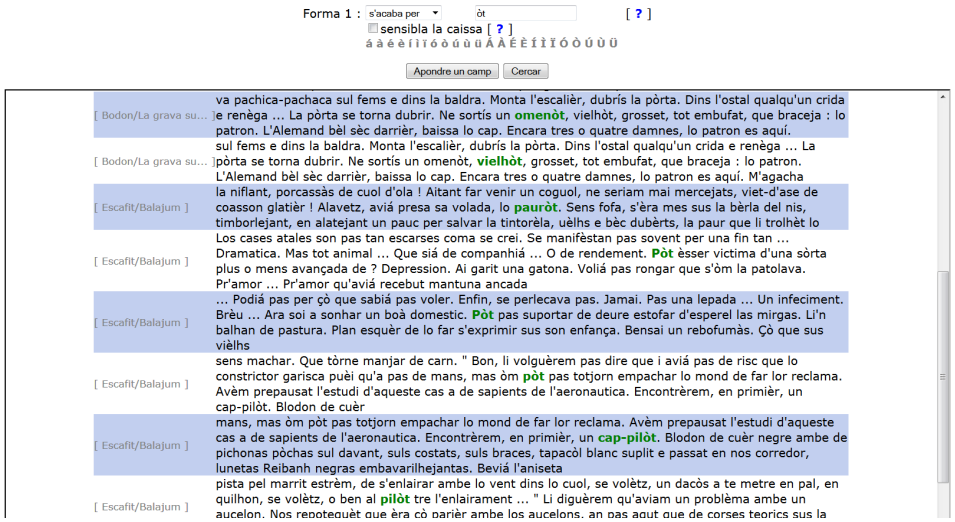

FIGURE 8: Searching for words ending with *–òt*

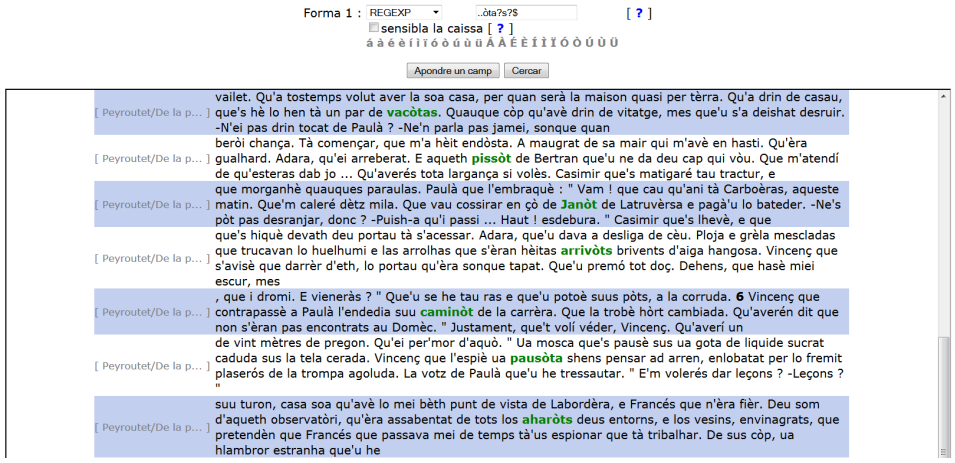

FIGURE 9: Searching for words ending with *òt*, *òta*, *òts*, *òtas*

cons in order to develop basic natural language processing tools, namely OCR (Urieli & Vergez-Couret 2013) and a Part-of-Speech (POS) tagger.[16]

In the foreseeable future, we aim at enriching one part of the text base with POS and lemma annotations. Within the framework of the text base, lemmatization and POS will allow new request possibilities, for example the search of all finite verb forms or inflected forms of adjectives or the disambiguation of homographs such as *poder* (common noun "power") and *poder* (verb "can"). One of the major difficulties to overcome is connected to the strong variation in written Occitan. The existence of numerous spellings is one of its

[16] A Part-of-Speech tagger marks-up words in a text as corresponding to a particular part of speech with additional morphosyntactic information.

causes. The spelling used during the Middle Ages is called the "troubadour spelling".This spelling disappeared gradually with the decline of the literary production. Since the 19th century, one can distinguish two major types of spellings: The Mistral's spelling (created in the Provence and influenced by the French spelling) and the Gaston Febus's spelling, which is used in Biarn. The last one appeared during the 20th century. It is a unified spelling, known as "classical spelling", inspired by the "troubadour spelling", and diffused in all Occitan territories (Sibille 2007). Another reason for the variation is related to the dialectal complexity of the language. The classical spelling naturally integrates the geolinguistic varieties (for instance *lo filh* vs. *eth hilh* "the son", *luna* vs. *lua* "moon" or *cabra* vs. *craba* "goat"). Variations can also be seen as the result of the normalization process which is now in progress (for instance, differences in the spelling of conjugated verbs: *avian* vs. *avián* "they had"). Moreover, there are also phonological intra-dialectal variations reflected in the spelling (for example *contes* vs. *condes* "tales"). Because of spelling and dialectal variations in Occitan, it is difficult to simply apply the existing systems of POS tagging to create annotated corpora and large coverage lexicons as it is currently done with rich-resourced standardized languages.

In order to create morphosyntactic annotated corpora for Occitan, we first used a POS Tagger for Occitan available in the Apertium chain (Forcada et al. 2011). Apertium originally proposes open source systems for automatic translation, generally for related language pairs. Armentano I Oller (2008) developed a translation system for the Occitan/Catalan pair, which includes a POS tagger for Occitan. It is based on the use of one lexicon containing 36,500 entries. From our experience with this POS Tagger, three main difficulties can be raised:

a) If a word is not in the lexicon, no tag is proposed for this word. This results in strong performance variation depending on the texts used, especially because all possible spelling forms for all dialects are not included in the lexicon. Armentano I Oller announces an accuracy of 0.8 of correct tags for a text in Lengadocian (on an evaluation corpus of 600 tags). We made the same experiment on a text in Gascon (on an evaluation corpus of 1000 tags). We reached an accuracy of 0.6 of correct tags. Indeed, there were more words that could not be tagged in Gascon (19%) when compared to Languedocien (13%) (Vergez-Couret 2013).

b) The lexicon includes indiscriminately forms from various dialects. It should be evaluated if it is better to have one larger lexicon for all dialects or, on the contrary, one lexicon for each dialect.

c) The only way to improve the current performances of Apertium is to enrich the lexicon which can be a very time-consuming task.

To cope with this kind of problem for under-resourced languages, some researchers develop a system requiring a minimum of lexical resources. The main circumvention strategy is to use existing systems for a rich-resourced etymologically close language. Hana et al. (2011), for instance, use the proximity between Old Czech and Modern Czech (two successive states of the language) and Bernhard & Ligozat (2013) exploit the similarities between Alsatian and German (Alsatian being considered as a German dialect). We proposed to adapt the latter method for Occitan, using an etymologically related language – Castillan (Vergez-Couret 2013). It should be mentioned that the relation between these two languages is looser than the one described above. The method consists in running

existing morphosyntactic tools, here Tree Tagger for Castillan, on Occitan texts with a pre-transposition of the most frequent words in Castillan first. This method only requires the construction of a bilingual lexicon (Occitan/Castillan) of the 300 most frequent words (based on a corpus) in Occitan[17] and their translation into Castillan. We chose Tree Tagger software (hereafter TT) (Schmid 1994) trained for Castillan. The TT's performance relies on the use of a large coverage lexicon. But unlike Apertium, TT predicts tags, using the probability of POS tag sequences calculated on a manually tagged training corpus. We assumed that probabilities of POS tag sequences would be fairly similar between Occitan and Castillan.

- Transposition: This step consists in transposing the Occitan words which are in the lexicon in Castillan (see Table 3). Only the bold words have been translated. The other ones remain in their original form.
- POS tagging with TT: We used a simplified tag set (see Table 4) inspired by the one used in Frantext. POS tagging was done twice, first on the original text and then on the transposed text. Table 3 gives an example of tags for original texts and for transposed texts. The symbol ✓means known words and ✗ unknown words by TT. Correct tags are greyed out.
- Running TT on original text: Some words are graphically similar in each pair of languages. For example, the feminine singular definite article is *la* both in Occitan and Castillan (see Table 3). TT will consider them as known words (✓) and will use the available information about this word. Nonetheless, it does not insure that similar words in the two languages will have the same POS.
- Running TT on transposed text: In the transposed text, similar words and transposed words are considered known by TT.

The precision of Apertium on original text in Occitan is rather low, 0.65. As we explained above, Apertium only assigns tags to known words. As a consequence, the performances correlate significantly with the number of unknown words (19% for our evaluation corpus). The Spanish TT precision for the original text is unsurprisingly low, less than 0.5. After transposition, the precision reaches up to 0.8. These experiments show that the methodology first used between a language and one of its dialects (Bernhard & Ligozat, 2013) is exportable with similar results for pairs of languages, less close even if etymologically related.

Our approach is resource-free and gives a precision of 0.8 but this system is more robust to deal with variations than Apertium. Nevertheless, improvements are required to raise the precision up to 0.95, as usually expected for POS tagging. The strategy we retained was, therefore, to use the two systems described above to foster the very long and fastidious process of creating a gold standard for POS annotation in Occitan. The corrected annotations were used for training more traditional supervised learning systems such as Tree Tagger or Talismane (Urieli & Tanguy 2013), see Vergez-Couret & Urieli (2015) for more details. Along with this comes the building of lexicons.

[17] For the first experiment we used the Gascon dialect.

Original text		Tag	Trans_ca		Tag
Dab	✗	Np	**Con**	✓	Pp
la	✓	D	La	✓	D
complicitat	✗	S	Complicitat	✗	S
de	✓	Pp	De	✓	Pp
la	✓	D	La	✓	D
lua	✗	S	Lua	✗	S
que[18]			Que		
vau	✗	A	**Voy**	✓	V
poder	✓	Vi	Poder	✓	Vi
,		Cm	,		Cm
adara	✗	A	**Ahora**	✓	Adv
,		Cm	,		Cm
tirar	✓	Vi	Tirar	✓	Vi
camin	✗	S	Camin	✗	S
.		<sent>	.		<sent>

TABLE 3: Extract from original texts in Occitan and transposed texts in Spanish tagged with Spanish TT

Dca	Cardinal number as article
Pe	Enunciative particle
V	Finite verb (except Vi, Vpp, Vps)
Vi	Infinite verb
Vpp	Present participle
Vps	Past participle
Inj	Interjection
Np	Proper noun
S	Common noun
P	Pronoun
Pp	Preposition
Pr	Relative pronoun
<sent>	End of sentence
Cm	Comma

TABLE 4: Corresponding tag set

	Original text	Transposed text
Apertium	0.65	
Spanish TT	0.46	0.80

TABLE 5: TT precision

[18] We deleted all the enunciative particles because there is no equivalent in Castillan and the tags for them would have been inevitably wrong. While these particles play a role at the enunciative level, the following propositions are still grammatical.

As far as the lexicon is concerned, available lexical resources for Occitan are mostly dictionaries, in paper and electronic format. But those resources do not conform to the standardized norm (*Text Encoding Initiative, Lexical Markup Framework*) and are not currently operable for NLP systems the way they are. We are constructing a lexicon, based on the entries of the Laus dictionnaries (Laus 2001, 2005) and a list of finite verbal forms provided by Lo Congrès Permanent de la Lenga Occitana (Verbòc app). The lexicon currently includes around 700,000 inflected forms. But, as we have mentioned above, all the spelling, dialectal, and intra-dialectal variations are not included in the lexicon. It will be necessary to find strategies to enrich the lexicon, for example by using morphological similarities (Baroni et al. 2012). Other types of strategies to enrich the lexicon using existing resources for a related rich-resourced language would also help. For example, Scherrer & Sagot (2013) acquire German/Palatin (a German dialect) cognate pairs with unsupervised automatic learning methods. Such cognates could be used in our case for pairing dialectal and spelling variations.

4. PERSPECTIVES. The text base is now available online.[19] The next step will be to design new tools for calculating frequencies and to gradually increase the amount of texts. The strategy we want to retain is to increase the range of genres, domains, dialects, neither following the lead of Frantext which pays greater attention to literature, nor the example of English balanced corpus, but rather gathering as many various data as possible to satisfy as many BaTelÒc users as possible. BaTelÒc consists in a two-fold effort: to document the language and to provide data for linguistic studies. We hope that providing linguistic data will help the research field of Occitan linguistics to gain new researchers, which is another way of playing a role in the conservation of the Occitan language.

REFERENCES

Armentano I Oller, Carme. 2008. Traduction automatique occitan-catalan et occitan-espagnol: difficultés affrontées et résultats atteints. *IXème Congrès International de l'AIEO*, Aix-La-Chapelle.

Baroni, Marco, Johannes Matiasek & Harald Trost. 2012. Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the 6th Workshop of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, 48–57. Philadelphia: Association for Computational Linguistics.

Bec, Pierre. 1995. *La langue occitane*. Que sais-je n°1059. Paris: PUF.

Bernhard, Delphine & Anne-Laure Ligozat. 2013. Es esch fàscht wie Ditsh, oder net? Étiquetage morpho-syntaxique de l'alsacien en passant par l'allemand. In *Actes de TALARE 2013: Traitement Automatique des Langues Régionales de France et d'Europe*, 209–220.

Bras, Myriam. 2006. Le projet TELOC: construction d'une base textuelle occitane. *Langues et Cité: bulletin de l'observation des pratiques linguistiques* 10. 9.

Bras, Myriam & Jean Thomas. 2007. Diccionaris, *corpora*, e basas de donadas textualas. *Linguistica Occitana* 5. http://superlexic.com/revistadoc/wp-content/uploads/2013/07/Linguistica-occitana-5-BrasThomas.pdf (10 December, 2015).

[19] BaTelÒc can be consulted on the REDAC website: http://redac.univ-tlse2.fr (20 January 2016). REDAC gathers the linguistic resources developed at CLLE-ERSS research unit and makes them available for download or browsing.

Bras, Myriam & Jean Thomas. 2011. Batelòc : cap a una basa informatisada de tèxtes occitans. In Angelica Rieger (ed.), *L'Occitanie invitée de l'Euregio. Liège 1981 – Aix-la-Chapelle 2008 Bilan et perspectives*, Actes du IXème Congrès International de l'AIEO, 661–670. Achen: Shaker Verlag.

Cox, Christopher. 2011. Corpus linguistics and language documentation: challenges for collaboration. In John Newman, Harald Baayen & Sally Rice (eds.), *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*, 239–264. Amsterdam & New York: Rodopi.

Field, Thomas. 2013. *The Linguistic Corpus of Old Gascon*. Database for linguistic research on Southwestern France.

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez & Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation* 25(2). 127–144.

Hana, Jirka, Anna Feldman & Katsiaryna Aharodnik. 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 Humanitie (LaTeCH'11)*, 10–18. https://www.aclweb.org/anthology/W/W11/W11-15.pdf (10 December, 2015).

Himmelmann, Nikolaus P. 2006. Language documentation: what is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*, 1–30. Berlin & New York: Mouton de Gruyter.

Laus, Cristian 2001. *Dictionnaire Occitan-Français (Languedocien Central)*. Puylaurens: IEO.

Laus, Cristian. 2005. *Dictionnaire Français-Occitan*. Puylaurens: IEO.

Martel, Philippe. 2007. Qui parle occitan? *Langues et Cité: bulletin de l'observation des pratiques linguistiques* 10. 3.

McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based language studies, and advanced resource book*. Abington & Oxon: Routledge.

Mosel, Ulrike. 2013. Language Documentation and Corpus Linguistics. Paper presented at *International Conference on Endangered Languages in Europe 2013*, Minde, Portugal.

Ricketts, Peter, Alan Reed, F. R. P. Akehurst, John Hathaway & Cornelis van der Horst. 2001. *Concordance de l'occitan médiéval (COM)*. Turnhout: Brepols Publishers.

Sauzet, Patrick & Josiane Ubaud. 1995. *Le verbe occitan. Lo vèrb occitan*. Aix-en-Provence: Édisud.

Scherrer, Yves & Benoît Sagot. 2013. Étiquetage morphosyntaxique de langues non dotées à partir de ressources pour une langue étymologiquement proche. In *Actes de TALARE 2013: Traitement Automatique des Langues Régionales de France et d'Europe*, 195–208. http://www.taln2013.org/actes/www/TALARE-2013/actes/talare-2013-long-002.pdf (10 December, 2015).

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

Sibille, Jean. 2006. L'Occitan, qu'es aquò. *Langues et Cité: bulletin de l'observation des pratiques linguistiques* 10. 2.

Sibille, Jean. 2010. Les langues autochtones de France métropolitaine. Pratiques et savoirs. In Claude Gruaz & Christine Jacquet-Pfau (eds.), *Autour du mot: pratiques et compétences. Séminaire du Centre du français moderne, Tome II, 2006-2009*, 69–85. Limoges: Lambert-Lucas.

Urieli, Assaf & Ludovic Tanguy. 2013. L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In *Actes de la conférence Traitement Automatique des Langues Naturelles (TALN 2013)*, Les Sables d'Olonne, France. http://aclweb.org/anthology/F/F13/F13-1014.pdf (10 December 2015).

Urieli, Assaf & Marianne Vergez-Couret. 2013. Jochre, océrisation par apprentissage automatique : étude comparée sur le yiddish et l'occitan. *Atelier TALaRE (Traitement Automatique des Langues Régionales d'Europe), Conférence TALN (Traitement Automatique des Langues Naturelles)*. http://w3.erss.univ-tlse2.fr/textes/pagespersos/urieli/UrieliVergezCouret2013TALARE-Jochre.pdf (10 December, 2015).

Vergez-Couret, Marianne. 2013. Tagging Occitan using French and Castillan Tree Tagger. In *Proceedings of Less-Resourced Languages Workshop, Language & Technology Conference*, Poznan, Poland. https://hal.archives-ouvertes.fr/hal-00986426/document (10 December, 2015).

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, France.

Myriam Bras
myriam.bras@univ-tlse2.fr

Marianne Vergez-Couret
vergez@univ-tlse2.fr