The ParCoLab Parallel Corpus and its Extension to Four Regional Languages of France

Dejan Stosic, Saša Marjanović, Delphine Bernhard, Xavier Bach, Myriam Bras, Laurent Kevers, Stella Retali Medori, Marianne Vergez-Couret, Carole Werner

The ParCoLab Parallel Corpus and its Extension to Four Regional Languages of France

Dejan Stosic[1], Saša Marjanović[2], Delphine Bernhard[3]
Xavier Bach[1], Myriam Bras[1], Laurent Kevers[4]
Stella Medori[4], Marianne Vergez-Couret[5], Carole Werner[3]
[1] Université de Toulouse, CLLE UMR 5263, F-31000 Toulouse, France
[2] University of Belgrade, Faculty of Philology, 11000 Belgrade, Serbia
[3] Université de Strasbourg, LiLPa UR 1339, F-67000 Strasbourg, France
[4] Università di Corsica Pasquale Paoli, LISA UMR 6240, F-20250 Corte, France
[5] Université de Poitiers, FoReLLIS UR 15076, F-86000 Poitiers, France
dejan.stosic@univ-tlse2.fr, sasa.marjanovic@fil.bg.ac.rs, dbernhard@unistra.fr,
xavier.bach@univ-tlse2.fr, bras@univ-tlse2.fr, laurent@kevers.org, medori_e@univ-corse.fr
marianne.vergez.couret@univ-poitiers.fr, wernerc@unistra.fr

Abstract

Parallel corpora are still scarce for most of the world's language pairs. The situation is by no means different for regional languages of France. In addition, adequate web interfaces facilitate and encourage the use of parallel corpora by target users, such as language learners and teachers, as well as linguists. In this paper, we describe ParCoLab, a parallel corpus and a web platform for querying the corpus. From its onset, ParCoLab has been geared towards lower-resource languages, with an initial corpus in Serbian, along with French and English (later Spanish). We focus here on the extension of ParCoLab with a parallel corpus for four regional languages of France: Alsatian, Corsican, Occitan and Poitevin-Saintongeais. In particular, we detail criteria for choosing texts and issues related to their collection. The new parallel corpus contains more than 20k tokens per regional language.

Keywords: parallel corpora, low-resource languages, web interfaces

1. Introduction and Objectives

Parallel corpora are not only a prerequisite for developing machine translation tools but also helpful resources for linguists as well as language learners, teachers lexicographers and translators. Yet, parallel corpora are scarce for most of the world's language pairs (Haddow et al., 2022). The situation is by no means different for regional languages of France. These languages are very poorly equipped with digital resources and tools, compared to French (Leixa et al., 2014). While the situation has somewhat improved in the last years thanks to several projects, the disparity between French and the other languages of France is still very important. Recently, Joshi et al. (2020) established a 6-level taxonomy –from 0 to 5– of the world's languages based on the number of existing language resources: while French is classified at level 5 (best possible level, indicating a resource-rich language), Occitan and Corsican only make it to level 1, while Alsatian is still classified at level 0 (worst level).

For these languages, translations are usually available only as non-digitized books. Web crawling is not a relevant solution, since there is little to no use of these languages in the public space, including official web pages and documents. Moreover, web-crawled corpora, while sometimes including low-resource languages, are often of poor quality (Kreutzer et al., 2022; Kevers, 2022). Crowdsourced translations available e.g. in the Tatoeba[1] collection are few in number: 13K sentences for Occitan, with only 24 sentences for Corsican.[2] The FLORES-200 (NLLB Team et al., 2022) test set for machine translation includes Occitan, Basque and Catalan, which are all three considered as regional languages of France, but none of the other 70+ regional languages of France.[3]

In this paper, we focus on four regional languages of France and their integration into the ParCoLab parallel corpus and platform.[4] From its onset, this resource and tool were geared towards lower-resource languages, with an initial corpus including texts in Serbian, along with French and English. The corpus has since been extended to Occitan with a pilot study in 2018 then to Spanish in 2020 and, since 2022, to three other regional languages of France: Alsatian, Corsican and Poitevin-Saintongeais.

[1] https://tatoeba.org/.
[2] These figures have been checked on September 6, 2023.
[3] https://www.culture.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France/Agir-pour-les-langues/Promouvoir-les-langues-de-France.
[4] http://parcolab.univ-tlse2.fr

LREC-COLING 2024, pages 16014–16023
20-25 May, 2024.

The main contributions of this article are as follows:

- We make a thorough and up-to-date description of the ParCoLab corpus and platform used in this work : genesis and objectives, contents, querying possibilities and ongoing work;
- We present three configurations for hosting regional languages of France in the ParCoLab platform;
- We present a unique and novel parallel corpus for four regional language of France, with human translations and manually-checked alignments, which contains more than 20k tokens per regional language;
- We detail criteria for choosing and including regional texts in the corpus, based on the specific constraints of extremely-low-resource languages and issues related to the collection of parallel corpora ;
- The new corpora are made available online on ParCoLab, to facilitate and encourage use by target users such as language learners and teachers, as well as linguists.

2. Description of ParCoLab

The ParCoLab platform used in this work was established in 2015 with the financial support of the Research Valorisation Unit at the University of Toulouse Jean Jaurès (Stosic et al., 2019).

Its primary aim is to facilitate the web-based presentation and retrieval of texts from the underlying parallel corpus. As such, the ParCoLab platform functions as a searchable textual database, housing parallel texts in multiple languages. In the subsequent subsection (2.1), we provide a general presentation of the ParCoLab corpus, with a focus on the Serbian, French, English and Spanish languages. Section 2.2 introduces its contents, while in section 2.3, we delve into the query capabilities of the ParCoLab platform. Section 2.4 summarises the ongoing and prospective research endeavours.

2.1. General Presentation

The development of the ParCoLab corpus began long before the platform was established, tracing its origins back to 2007 (Stosic et al., 2019). This corpus owes its creation to a fruitful collaborative effort involving the CLLE research unit (CNRS and the University of Toulouse-Jean Jaurès, France), working in conjunction with the Department of Romance Studies at the Faculty of Philology, University of Belgrade in Serbia. The ParCoLab corpus initially comprised texts in French, Serbian, and English, providing support for these three languages within the ParCoLab platform (Miletic et al., 2017). Subsequently, between 2017 and 2018, the platform underwent significant enhancements, driven by a grant from the Délégation générale à la langue française et aux langues de France (DGLFLF) under the French Ministry of Culture. These improvements were strategically aimed at expanding the platform's language coverage. Notably, in 2018, a pilot study was carried out to test the feasibility of extending the platform to regional languages by aligning existing texts with Occitan translations (Stosic and Bras, 2018; Bras, 2023). In 2020, Spanish was introduced into the corpus (Terzić et al., 2020). Since 2022, the corpus has been enriched with texts in Alsatian, Corsican and Poitevin-Saintongeais, thanks to financial support from the French National Research Agency (ANR) granted to the DIVITAL project.[5] Section 3 details the process of adding regional languages of France to ParCoLab.

The corpus database is hosted on the servers of the French Huma-Num consortium.[6] It is not available for download but can be freely accessed through the user-friendly ParCoLab query platform at the following address: http://parcolab.univ-tlse2.fr/.

The primary objective of the ParCoLab project is to create a high-quality and highly reliable multilingual parallel corpus (cf. Miletic et al. (2017); Marjanović et al. (2018); Stosic et al. (2019); Terzić et al. (2020)). Its primary purpose is to be a valuable resource primarily used by human users. These users can employ it for developing theoretical language descriptions, particularly in the realm of contrastive linguistics and empirical research in translation studies (Miletic et al., 2017; Stosic et al., 2019).

Simultaneously, the ParCoLab platform aims to serve as a practical tool for various translation tasks, language learning, educational language content creation, and the development of monolingual and bilingual dictionaries (cf. Marjanović et al. (2018)). At the beginning, it was especially useful for Serbian, a low-resource language, as it fostered the development of Natural Language Processing (NLP) resources (see Miletic et al. (2017); Miletic (2018)), which are also openly accessible to other researchers working with the Serbian language. Furthermore, although a secondary aim, the ParCoLab corpus has the potential for broader applications in the field of NLP. These include ter-

[5] https://divital.gitpages.huma-num.fr/
[6] https://www.huma-num.fr/

Data type	Serbian	French	English	Spanish	Total	% of the corpus
Spoken data	3,618,910	4,031,595	3,745,383	3,369,585	14,765,473	**31.3%**
Written data	9,176,893	11,506,632	7,717,803	3,959,741	32,361,069	**68.7%**
Total	**12,795,803**	**15,538,227**	**11,463,186**	**7,329,326**	**47,126,542**	
% of the corpus	**27.2%**	**33.0%**	**24.3%**	**15.5%**		

Table 1: Token Distribution between Written and Spoken Language Data.

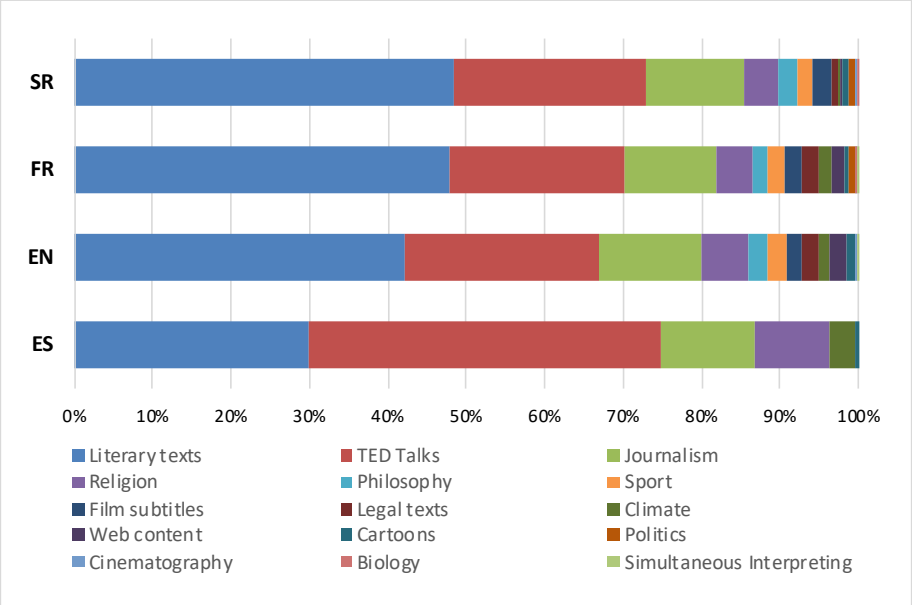

Figure 1: Token distribution by language and text domains.

minology extraction and the development of tools for computer-assisted and automatic translation.

2.2. Content

The content of the ParCoLab parallel corpus is influenced by the initial language selection (Miletic et al., 2017; Stosic et al., 2019). A particular challenge arises when aligning texts in two widely spoken global languages, such as French and English, with texts in Serbian, a less commonly used language on the international stage. Serbian, despite being spoken in several countries, does not hold the status of an official language in major supranational organizations like the European Union. Consequently, there is a limited pool of Serbian texts available that also have French and English versions. Thus, the primary consideration when selecting texts for alignment is their availability, as well as the quality of translation (Miletic et al., 2017). The initial phase focused exclusively on classical literary works with their high-quality translations (Marjanović et al., 2018; Stosic et al., 2019).

The alignment of the original texts with their translations is executed using an algorithm integrated into the ParCoLab platform, without the need for external resources (cf. Marjanović et al. (2018); Stosic et al. (2019)). This algorithm follows a systematic approach, establishing one-to-one alignments initially at the chapter level, then progressing to the paragraph level, and ultimately fine-tuning at the sentence level. The ParCoLab editing platform offers the flexibility to align three languages simultaneously, allowing for language order adjustments. Any errors flagged by the tool are manually corrected, ensuring the utmost reliability of the corpus alignments.

However, during the initial phase, older public domain literary works were primarily aligned, significantly influencing the linguistic composition of the ParCoLab corpus (Miletic et al., 2017; Marjanović et al., 2018; Stosic et al., 2019). This prompted a need for diversification. Despite some efforts made in this regard, literary works continued to dominate, constituting 88.7% of the corpus. In the second phase, the corpus was expanded, incorporating what is referred to as 'spoken language data' (Terzić et al., 2020). Locating authentic spoken content in multiple languages, especially when considering Serbian, posed a challenge. To address this, a decision was made to include transcripts of TED talks with translations in the ParCoLab corpus. These talks offer accessible and diverse content. To account for potential

variations in translation quality compared to classic literary works, transcripts of feature films and dubbed animated films were also integrated into ParCoLab. They were manually transcribed and, given time and resource constraints, these transcripts cover fewer tokens than TED talks. This approach helped achieve a more balanced distribution of written and 'spoken' language data, standing at 55% and 45%, respectively. Starting from 2020, the textual database has seen an expansion with the inclusion of texts from diverse domains, leading to a resurgence of written texts (see Table 1).

The proportion of spoken language data has now decreased to 31% (cf. Terzić et al. (2020)), with 29% of tokens originating from TED talks, and the remaining tokens corresponding to transcripts of films, cartoons, and simultaneous interpreting. Literary content continues to dominate, accounting for 44% of all tokens. Other newly added domains include journalism (12%), religion (6%), philosophy, sports and legislation (2% each), while all remaining categories (climatology, web content, politics, cinematography, biology) together constitute less than 5% of all tokens in the corpus. A detailed token distribution by language and text domains is shown in Figure 1.

2.3. Querying Possibilities in ParCoLab

Queries are carried out using a user-friendly interface that is both visually intuitive and functionally straightforward (cf. Marjanović et al. (2018); Stosic et al. (2019)). This interface is powered by the ElasticSearch[7] search engine, which is well-adapted for querying data in NoSQL databases. It offers versatile querying options, enabling users to perform single-language or multi-language queries involving up to three languages at once. Irrespective of the query's complexity, results are always presented in parallel across three languages. Users can formulate queries for individual words, multi-word expressions, phrases with one or more wildcard characters, words that start or end with specific character sequences etc. Additionally, the interface supports the use of regular expressions and Boolean operators.

All texts are stored in XML format files in accordance with TEI P5 guidelines (TEI Consortium, 2023)[8] (cf. (Miletic et al., 2017; Marjanović et al., 2018; Stosic et al., 2019)). These XML files contain standardized metadata, including title, subtitle, author, translator, publisher, publication place and date, creation date, source, language of the text, language of the original work, domain, text genre and form, token count, and more. This extensive pool of metadata confers upon users the capability to construct finely-tuned queries based on specific metadata criteria. Notably, the interface currently affords users the option to tailor their searches by selecting the source language, constraining queries to particular authors, specific works, or textual domains, thus offering the possibility to choose the subcorpus they want to work on.

Additionally, a subset, though relatively small, of the corpus boasts multiple layers of annotation, including lemmas, parts of speech (POS), morphosyntactic descriptions (MSDs), and syntactic relations (Miletic et al., 2017; Stosic et al., 2019). This applies to a novel with a total of 233K tokens in English, which is complemented by French and Serbian translations, a Serbian literary ParCoTrain-Synt subcorpus comprising approximately 150K tokens (Miletic, 2018), and a compact ParCoJour corpus of news texts in Serbian (Terzić, 2019). These texts serve as valuable resources for both training and evaluating tools dedicated to text annotation in Serbian. The annotation process was conducted using the French Treebank for French (Abeillé et al., 2003), the Penn Treebank for English (Marcus et al., 1993), and the Serbian Treebank for Serbian (Miletic, 2018). Within the query interface, users have the possibility to explore annotated texts by transitioning from the Expression query mode to the Feature mode (Stosic et al., 2019). In Feature mode, queries extend beyond specific word forms and can encompass lemmas, POS, MSDs, and syntactic relations within sentences. To streamline Feature mode queries, users can select the appropriate POS from a dropdown menu, with the interface offering corresponding MSDs and syntactic relations based on the chosen POS. An illustrative representation of the query interface is presented in Figure 2.

2.4. Ongoing Work

Since 2007, this ongoing project has aimed to enhance the ParCoLab database and platform both qualitatively and quantitatively.

Firstly, efforts are being made to achieve a balanced representation of languages within the database. This involves the inclusion of translated texts for languages that were previously underrepresented. Simultaneously, ongoing work is focused on ensuring that all languages are more comprehensively represented with original texts and an increased incorporation of translations of these original texts. This expanded scope aims to facilitate cross-lingual searches to a greater extent.

Secondly, the diversification of the textual database remains a priority. In the near future, there are plans to introduce aligned texts of inter-

[7] https://www.elastic.co/elasticsearch/
[8] https://tei-c.org/guidelines/p5/

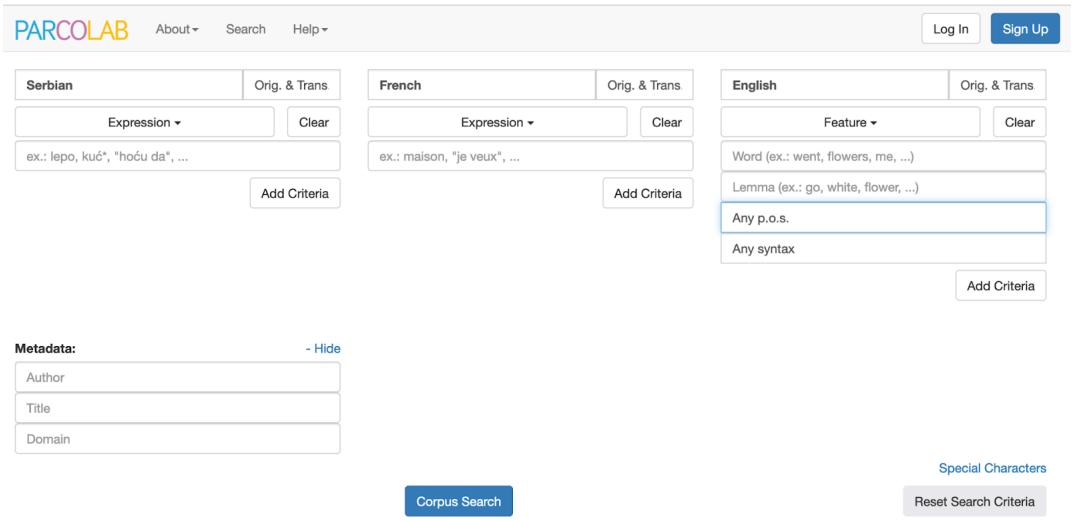

Figure 2: The ParCoLab query interface.

national multilateral conventions in four languages (Serbian, French, English, and Spanish), as well as translated encyclopedic content from French to Serbian.

Thirdly, further steps are being taken towards multi-layered text annotation across all languages, with the intention of enhancing the query capabilities of the texts, aligning with the principles outlined in the preceding subsection.

Moreover, there are intentions to enhance the user interface, allowing users to conduct queries based on an expanded set of textual metadata parameters, including textual genre, form, translator, year of publication or creation, and more.

Lastly, the platform has initiated the inclusion of texts in regional languages of France, with anticipated growth in both quantity and scope, as elaborated upon in Section 3.

3. Regional Languages of France in ParCoLab: the DIVITAL Parallel Corpus

The ParCoLab platform provides a ready to use solution to facilitate access to parallel corpora for low-resource languages such as regional languages of France.

Corpora for regional languages of France can be added to the platform in various configurations that we describe in section 3.1. We then present a parallel corpus in four regional languages plus French (section 3.2). We detail the inclusion criteria for the texts to be collected in Section 3.2.1. The text collection process is described in Section 3.2.2. Finally, we describe the alignment process in Section 3.2.3 and we detail the contents of the regional corpora in Section 3.2.4.

3.1. Hosting Regional Languages

The ParCoLab platform hosts four different regional languages of France along three configurations:

1. The language already has its own textual database gathering texts originally written in this language: the platform offers a nice solution to provide access to these texts in other languages. For example, this is the case for the literary work of the Occitan author Joan Bodon: his Occitan texts are part of the Occitan text database BaTelÒc[9] while these texts aligned with their French translation are displayed in ParCoLab, in a complementary strategy;

2. The language does not have a textual database yet: ParCoLab offers a ready to use solution to provide access to the texts written in this language, be they aligned with translations or not; this is the case for Poitevin-Saintongeais, for which a great deal of texts were gathered in the aim of building a text database, (TELPOS project by Liliane Jagueneau), or for Alsatian;

3. There exist translations in regional languages of widespread texts (international literature works from Kipling, Stevenson, Doyle, Saint-Exupéry, Giono, for example, or international conventions): here again ParColab offers a nice solution to provide access to these translations in less spread languages.

Table 2 shows the number of tokens in the original texts in French, English, Occitan and Poitevin-Saintongeais (in diagonals) and the number of tokens in the target languages translated from these

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

Original=>	French	English	Occitan	PoitSaint
French	5,615,329	5,074,728	67,976	86,508
English	4,784,316	5,136,216		
Occitan	44,660	225,937	61,916	
PoitSaint	20,456			267,573
Alsatian	38,580	2,020		
Corsican	20,266	2,009		

Table 2: French, English and regional languages in the corpus (in number of words). PoitSaint corresponds to Poitevin-Saintongeais.

languages (in rows) already available in ParCoLab. The table shows that, for Poitevin-Saintongeais, ParCoLab hosts a majority of monolingual, non translated yet, documents (see Configuration 2 above), whereas all the Occitan texts in ParCoLab are associated with at least one translation in another language (see Configuration 1 above).

3.2. A Four Regional Language Parallel Corpus

In this section, we present our parallel corpus in four regional languages of France, Alsatian, Corsican, Occitan and Poitevin-Saintongeais, plus French, that belongs to the third configuration mentioned in section 3.1. This corpus, named the "DIVITAL parallel corpus", represents a joint effort to improve the visibility of these languages and to offer the possibility to compare them with other languages for cross-linguistic investigations. This corpus is unique, in the sense that very few texts are available in all four languages, with bilingual French-regional language texts being much more common. It follows the long established tradition of collecting translations of the same story in several languages and dialects. Moreover, this contributes to promoting France's written and literary heritage in languages other than French.

3.2.1. Survey and Inclusion Criteria

As a first step, we carried out a survey of texts translated from or to our target languages. During this survey, we assessed the following desired properties:

1. Availability of translations in at least two of the target regional languages and/or several languages. For instance, *Le Petit Prince* (*The Little Prince*) by Antoine de Saint-Exupéry is one of the most translated texts in the world, with 462 translations listed on the `http://www.petit-prince.at/` website.

2. Copyright status, to ensure open access (Collister, 2022).

3. Level of difficulty in translating the text to other languages.

4. Possible use for manual annotation in parts-of-speech and dependency relationships, which will be done in a second phase of the project.

5. Relevance of the contents to contemporary issues, e.g., the environment, human rights, status of regional languages. For instance, *L'homme qui plantait des arbres* (*The Man Who Planted Trees*) by Jean Giono conveys ecological and humanist messages that are very modern.

6. Reusability of content in pedagogical contexts, e.g. *The Little Prince* can be used in schools for language comparison activities.

7. Use in existing dialectal language documentation works, e.g. *The Parable of the Prodigal Son* (Coquebert de Montbret and de Labouderie, 1831; Favre, 1879), *The Decameron* (Papanti, 1875) or *The North Wind and the Sun* (de Mareüil et al., 2018).

The decision for inclusion in the final corpus was based on a combination of these criteria and was discussed between the project's participants.

3.2.2. Text Collection and Translation

Once we had identified the texts we wished to include in our corpus (see Table 3), we collected the translations. We distinguish between three different cases:

1. Texts in the public domain: these texts can be directly included in our corpus.

2. Existing published translations: in this case, we strive to make agreements with publishers. In the case of difficulties in identifying the copyright holders or publishing house and negotiating with them, we display only limited contexts in ParCoLab, in accordance with the right to quote.[10]

3. Texts with no translation in some or all of the target languages: these texts are translated specifically for the project, with a transfer of property rights from the translator, either by members of the project (co-authors of this paper) or by external translators. For some of the languages, recruiting translators can be difficult: this is particularly the case for

[10] Following Directive 2019/790 of the European Parliament and the Council of 17 April 2019 on copyright and related rights in the Digital Single Market, amending Directives 96/9/EC and 2001/29/EC, and Article L122-5 of the French Intellectual Property Code, we only use these texts for text and data mining within scientific research projects.

Title	Author	Publication date	Domain	Genre	Included part	Tokens (French)
Le Petit Prince (The Little Prince)	Antoine de Saint-Exupéry	1943	Literary	Novella	First 2 chapters	1,150
L'homme qui plantait des arbres (The Man Who Planted Trees)	Jean Giono	1953	Literary	Allegorical tale	Whole	3,800
Lettres de mon moulin (Letters from My Windmill)	Alphonse Daudet	1869	Literary	Short story	2 stories	4,100
Contes du lundi (Monday Tales)	Alphonse Daudet	1873	Literary	Short story	2 stories	3,300
Déclaration Universelle des Droits de l'Homme (Universal Declaration of Human Rights)	UN	1948	Legal	Official charter	Whole	2,070
Décameron (Decameron)	Boccace	1349-1353	Literary	Short story	1 story	300
Pierre et le loup (Peter and the Wolf)	Sergueï Prokofiev	1936	Literary	Symphonic tale	Whole	780
Le fils prodigue (Parable of the Prodigal Son)	Luke	2nd century	Religion	Parable	Whole	510
La bise et le soleil (The North Wind and the Sun)	Esope	6th century BC	Literary	Fable	Whole	130
Chronicles on French regional languages	Michel Feltin-Palas	2021-2023	Journalism	Column	4 chronicles	4,050
					TOTAL	**20,190**

Table 3: DIVITAL parallel corpus contents.

Poitevin-Saintongeais. In this language, we rely on regional language clubs, whose members are particularly committed to initiatives like ours, which help to keep their language alive and promote it.

The issues mentioned make it difficult to collect some of the texts and result in lengthy delays (negotiations, contacts with publishers and translators, translation delivery time). Translations available only in printed form have to be digitised and OCRised. This is no research work *per se*, but is nevertheless mandatory to obtain resources which can be exploited. In addition to the time-consuming labour associated with these tasks, expenses are necessary for the payment of translators.

3.2.3. Alignment

Alignments are performed using the sentence alignment tool available within ParCoLab (see Section 2.2). All the alignments are manually checked and corrected after initial automatic processing, which identifies the document structure (chapter, section...) and segments it into sentences. Some texts are also aligned outside of the ParCoLab platform, using standalone sentence alignment tools. In particular, we use the InterText tool (Vondricka, 2014) which provides an interface to align documents and correct the resulting alignment. These are then automatically transformed to the TEI XML format suitable for ParCoLab using custom Python scripts.

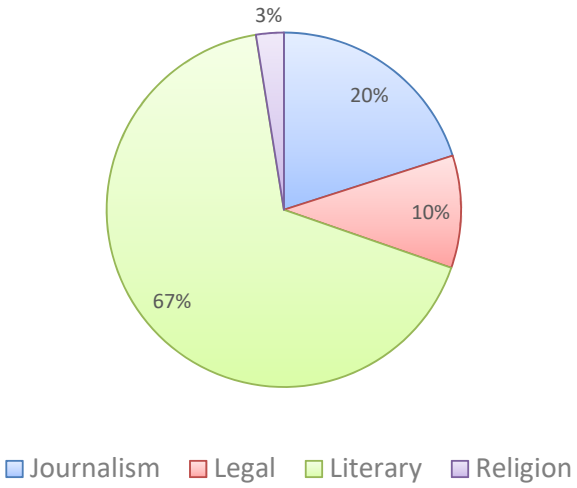

Figure 3: Domains in the DIVITAL parallel corpus.

3.2.4. Corpus Details

Planned corpus contents for the DIVITAL parallel corpus are detailed in Table 3: the corpora are complete for Alsatian, Occitan and French. Corpus acquisition is still ongoing for Corsican and Poitevin-Saintongeais, but several texts listed in Table 3 are already available on ParCoLab. The texts listed in Table 3 are, for the time being, essentially limited to the literary domain, with one legal text, one religious text and some newspaper chronicles (see Figure 3). With 20k tokens per language, including French, we will provide a new 100k token five-lingual parallel corpus.

4. Conclusion and Perspectives

We have presented the ParCoLab corpus: originally comprising texts in Serbian, French and English, and later, Occitan, then Spanish, it is currently being extended with translations for three other regional languages of France. The associated online web platform provides a user-friendly interface for accessing the corpus and carrying out complex queries. It makes the corpus easily usable by language professionals and learners.

Collecting corpora for Alsatian, Corsican, Occitan and Poitevin-Saintongeais is however a time and money consuming process, given the lack of easily accessible digital resources. Whenever possible, we strive to collect texts which can be freely shared under the principles of open science, so as to foster further research. In particular, the corpus can be used to develop NLP tools (machine translation) and resources (multilingual lexicons). The alignment to higher-resource languages such as French or English can be useful for annotation projection across languages (Agić et al., 2016; Akbik and Vollgraf, 2018).

In the future, we plan to manually annotate parts of the regional language texts following the Universal Dependencies (UD) framework (Nivre et al., 2016). This will provide additional layers of annotations for querying the corpus in the platform and performing contrastive linguistics studies.

5. Limitations

(1) The corpus collected so far for regional languages of France lacks diversity with respect to some criteria: author gender, representation of language variants, genres, time periods, source language. (2) Some of the texts included in the corpus are still under copyright and are not shareable: these texts are only made available through the ParCoLab platform, with a limitation in the size of the viewable context, under the principles of Fair Use.

6. Ethics Statement

The professional translators were paid in accordance to current rates for translations in France.

7. Acknowledgements

This work has been carried out within the framework of the ANR-21-CE27-0004 DIVITAL project supported by the French National Research Agency. We would like to thank all those who contributed to translating or aligning the texts: Nathanaël Beiner, Michel Cardineau, Adrien Fernique, Michel Gautier, l'atelier Patrimoine de Vouillé section parlanjhe. We would also like to thank the following authors and publishers for allowing us to use their content: Eric Chaplain (Editions des régionalismes), Michel Feltin-Palas, Nicolas Martin-Minaret, Dr. Walter Sauer (Edition Tintenfaß), IEO Edicions, Letras d'Òc, Vent Terral.

8. Bibliographical References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. *Treebanks: Building and using parsed corpora*, pages 165–187.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual Projection for Parsing Truly Low-Resource Languages. *Transactions of the Association for Computational Linguistics*, 4(0):301–312.

Alan Akbik and Roland Vollgraf. 2018. ZAP: An Open-Source Multilingual Annotation Projection Framework. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Myriam Bras. 2023. Nouvelles perspectives pour la linguistique occitane à partir de la base textuelle BaTelòc. In Annie Rialland and Michela Russo, editors, *Les Langues Régionales de France : Nouvelles Approches, Nouvelles Méthodologies, Revitalisation*, pages 121–142. Société De Linguistique De Paris.

Lauren B. Collister. 2022. Copyright and Sharing Linguistic Data. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister, editors, *The Open Handbook of Linguistic Data Management*. The MIT Press.

Eugène Coquebert de Montbret and Jean de Labouderie, editors. 1831. *Mélanges sur les langues, dialectes et patois: renfermant, entre autres, une collection de versions de la Parabole de l'enfant prodigue en cent idiomes ou patois différens*. Almanach du commerce : Delaunay, Paris, France.

Philippe Boula de Mareüil, Frédéric Vernier, and Albert Rilliard. 2018. A Speaking Atlas of the Regional Languages of France. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, page 6, Miyazaki, Japan.

Léopold Favre. 1879. *Parabole de l'enfant prodigue en divers dialectes ; patois de la France, avec une introduction sur la formation des dialectes et patois de la France, par L. Favre*. L. Favre, Niort.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation. *Computational Linguistics*, 48(3):673–732.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Laurent Kevers. 2022. L'identification de langue, un outil au service du corse et de l'évaluation des ressources linguistiques. *Revue TAL : traitement automatique des langues*, 62(3):13–37.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Jérémy Leixa, Valérie Mapelli, and Khalid Choukri. 2014. Inventaire des ressources linguistiques des langues de France. Technical Report ELDA-DGLFLF-2013A, ELDA/DGLFLF, Paris.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Saša Marjanović, Dejan Stosic, and Aleksandra Miletic. 2018. A sample French-Serbian dictionary entry based on the ParCoLab parallel corpus. In *The XVIII EURALEX International Congress: Lexicography in Global Contexts.*, pages 423–435. Znanstvena založba Filozofske fakultete Univerze v Ljubljani/Ljubljana.

Aleksandra Miletic. 2018. *Un treebank pour le serbe: constitution et exploitations*. Ph.D. thesis, Université Toulouse le Mirail-Toulouse II.

Aleksandra Miletic, Dejan Stosic, and Saša Marjanović. 2017. ParCoLab: A parallel corpus for Serbian, French and English. In *Text, Speech, and Dialogue: 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings 20*, pages 156–164. Springer.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal

dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Giovanni Papanti. 1875. *I parlari italiani in Certaldo alla festa del v centenario di Messer Giovanni Boccacci ; omaggio di Giovanni Papanti*. Livorno, Tipi di F. Vigo.

Dejan Stosic and Myriam Bras. 2018. Une plateforme de constitution et de diffusion de corpus parallèles pour les langues de france. Research report.

Dejan Stosic, Saša Marjanović, and Aleksandra Miletic. 2019. Corpus parallèle ParCoLab et lexicographie bilingue français-serbe: recherches et applications. *Serbica*.

TEI Consortium. 2023. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 4.6.0. Last updated on 4th April 2023, revision f18deffba.

Dusica Terzić. 2019. Parsing des textes journalistiques en serbe à l'aide du logiciel Talismane. In *Traitement Automatique des Langues Naturelles (TALN)-PFIA 2019*, pages 591–604. ATALA.

Dušica Terzić, Saša Marjanović, Dejan Stosic, and Aleksandra Miletic. 2020. Diversification of Serbian-French-English-Spanish Parallel Corpus ParCoLab with Spoken Language Data. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 61–70. Springer.

Pavel Vondricka. 2014. Aligning parallel texts with InterText. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1875–1879.