Language Report French

Gilles Adda, Ioana Vasilescu, François Yvon

Chapter 16
Language Report French

Gilles Adda, Ioana Vasilescu, and François Yvon

Abstract This chapter presents a survey of the current state of technologies for the automatic processing of the French language. It is based on a thorough analysis of existing tools and resources for French, and also provides an accurate presentation of the domain and its main stakeholders (Adda et al. 2022). The chapter documents the presence of French on the internet and describes in broad terms the existing technologies for the French language. It also spells out general conclusions and formulates recommendations for progress towards deep language understanding for French.

1 The French Language

French is typologically a Romance language, closely related to other languages whose origin is Latin (e. g., Italian, Spanish, Portuguese, Romanian). French inherited Gaulish features from the Celtic dialects spoken by ethnic groups that previously populated the territory conquered by the Romans, and was later influenced by Germanic dialects as a consequence of the invasions that marked the fall of the Roman Empire. Modern French uses the Latin alphabet and has retained many Latin linguistic features. For instance, French is a nominative-accusative and article-based language (SVO) that greatly simplified the nominal and verbal declensions. French developed a large vocalic system including 12 oral and 4 nasal vowels.

With 128 million "native and real speakers" worldwide and an estimate of close to 300 million speakers overall (Collectif 2019), French appears only as the 16th most spoken native language, but as the 6th most spoken language in the world, after English, Chinese Mandarin, Spanish, Hindi and Russian. French is an official language in close to 30 countries, most notably in Europe (France: 65m speakers, Belgium: 7m speakers, Switzerland: 3m speakers, and Luxembourg), Africa, Canada and Haiti. In Europe, it is estimated that 129 million people speak French making it the 3rd most spoken second language, after English and German. French-speaking

Gilles Adda Ioana Vasilescu François Yvon
Université Paris-Saclay, CNRS, LISN, France, gilles.adda@limsi.fr, ioana.vasilescu@limsi.fr, francois.yvon@limsi.fr

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_16

countries together constitute *La Francophonie*, with the *Organisation Internationale de la Francophonie* coordinating policies between 88 associated states and entities.

Collectif (2019) notes that in 2018 French occupies the fourth place on the internet behind English, Chinese and Spanish, with a comfortable lead over the next set of languages. Pimienta (2022) observes that although French remains in fourth place on the internet in 2022, the gap to the following languages has considerably narrowed. The presence of French on the internet derives from its role as an international language: French is an official language of the EU and one of the three working languages of the European Commission. French is also a working language at the Organisation for Economic Co-operation and Development, and at the United Nations. French is also one of the three official languages of the European Patent Office and one of the four working languages of the African Union.

2 Technologies and Resources for French

Looking at the technology landscape for French, most state-of-the-art tools and applications rely almost exclusively on generic machine learning technologies, a major change with respect to the previous survey (Mariani et al. 2012): the most important ingredients for system building are data and, to a lesser extent, compute resources. We will, therefore, focus on the most critical language resources and give a general overview of the various technologies derived from them.

Large-scale, general purpose lexica for French associating lemmas or word forms to morpho-syntactic information are widely available. There is no official French National Corpus that would contain a representative subset of the language, balanced across periods, genres and domains, as may exist for other languages. However, sizable corpora (up to billions of tokens) of mixed genres are accessible and searchable.

The CommonCrawl project aggregates Web data that is orders of magnitude larger than these resources; and it is updated on a regular basis. Using French subsets of CommonCrawl, it has been possible to train large language models (LMs): FlauBERT uses a corpus of 12B running words, while CamemBERT uses the 22B words OSCAR. Other large LMs for French are available for research and commercial use; they help to boost the state-of-the-art for multiple NLP tasks.

Large-scale annotated (segmented in sentences, speakers and turns, transcribed) speech databases, containing thousands of hours of recordings are available for several genres. Such resources have enabled advanced technologies for French (transcription, synthesis, NLU). However, the collection of large sets of recordings remains a pressing issue to widen the applicability of these technologies, an objective addressed by Mozilla's Common Voice[1] or the Voice Lab project.[2]

Basic NLP tools were already well covered in 2012 and they have benefited from improvements in machine learning. Open source industrial strength tokenizers, lem-

[1] https://commonvoice.mozilla.org/fr

[2] http://www.levoicelab.org

matizers and POS taggers for French are available. We note, however, that no recent systematic performance comparisons exist for these tasks; most of these tools process "generic" French and too little exists for specific sublanguages.

Having moved to fully neural, the availability of Machine Translation systems for French mostly depends on the availability of parallel corpora. Good resources exist for French, especially when matched with an English translation.

As for most social science and humanities domains, the digital revolution has created new avenues for language analysis. Such methodological changes are also happening for French and impact all linguistic domains, with the creation of corpora, tools and methods. Regarding corpora, both written and spoken varieties of French are well covered, although for historical reasons written sources are more common.

Owing to its role as an international language and the comparatively large size and advanced development of French-speaking markets, French is relatively well covered by international LT services: French-English has been one of the earliest translation pairs on the Web, and French versions of Siri, Amazon Echo and Google Home have been available for years. The development of LTs for French far exceeds the activity observed in France or other French-speaking countries.

Institutional support to LTs is mostly operated by the ANR (the French National Research Agency), albeit with a lack of continuous funding; large variability in funding over the years is not favourable to planning. The French research community is nonetheless active, with a dozen significant academic clusters all over France, as well as Belgium, Canada and Switzerland, covering the full spectrum of NLP. This research has greatly benefited from the development of the Jean Zay platform, an open high-performance computing infrastructure tailored to AI applications.

3 Recommendations and Next Steps

Many open-domain French corpora are the result of uncoordinated initiatives and consequently only partially cover the needs of domain-specific applications. This state of affairs results in 1. a lack of visibility of tools and data that are only known to restricted sub-communities, and 2. a waste of resources, as existing datasets are underused, or even duplicated, when other pressing needs remain unsatisfied. A first recommendation is thus to institutionalise clearer policies for the archiving of LRs for French, when they are produced by public research projects.

A second recommendation, aimed to increase the diversity and size of existing corpora, is to open the large datasets produced by public administration and institutions (e. g., in health, culture, media, justice or education) which are hard to access. Policies are needed to amplify the actions of the European CEF/ELRC programme to incentivize the development of open repositories with clear access rules.

Applications that involve social network data (e. g., opinion mining, fake news and hate speech detection) require specific actions, as they are often associated with delicate legal issues (related to proprietary rights or personal information) that limit their dissemination and exploitation. To reduce the dependency on current data poli-

cies of content holders, a third recommendation would be to secure access to sensitive data for research purposes and to facilitate the dissemination of publicly produced databases and models (e. g., using privacy-preserving techniques).

Recommendation four is the definition of a strategic roadmap for identifying, building, curating, annotating and securing resources for language varieties or domains that are critical for research, industry or for the administration in each French-speaking country, based on a precise analysis of the gaps in the existing datasets (some were alluded to above). This roadmap should also identify cases where resources can be transferred from English through MT.

Recommendation five aims to ensure, through recurrent funding, that evaluation campaigns specifically targeting French for a large number of applications are organized on a regular basis and widely advertised, so that systems are evaluated under real world conditions, so as to document their biases, defects and harmful impacts.

The final recommendation is to increase the support for research on themes such as fair and explainable deep learning for large language models, deep language analysis algorithms and technologies, multimodal resources for the study of language acquisition through interactions and grounding, and the study of pathological language processing. This multidisciplinary research should involve all relevant communities.

References

Adda, Gilles, Annelies Braffort, Ioana Vasilescu, and François Yvon (2022). *Deliverable D1.14 Report on the French Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-french.pdf.

Collectif (2019). *La langue française dans le monde*. OIF/Gallimard.

Mariani, Joseph, Patrick Paroubek, Gil Francopoulo, Aurélien Max, François Yvon, and Pierre Zweigenbaum (2012). *La langue française à l' Ère du numérique – The French Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/french.

Pimienta, Daniel (2022). "La place du français sur Internet". In: *La langue française dans le monde 2022*. OIF/Gallimard, pp. 26–27. https://www.francophonie.org/sites/default/files/2022-03/Synthese_La_langue_francaise_dans_le_monde_2022.pdf.