European Language Equality - Report on the French Language

Gilles Adda, Annelies Braffort, Ioana Vasilescu, François Yvon

EUROPEAN LANGUAGE EQUALITY

D1.14

Report on the French Language

Authors	Gilles Adda, Annelies Braffort, Ioana Vasilescu, François Yvon
Dissemination level	Public
Date	28-02-2022

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D1.14
Deliverable title	Report on the French Language
Type	Report
Number of pages	42
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP1: European Language Equality – Status Quo in 2020/2021
Task	Task 1.3 Language Technology Support of Europe's Languages in 2020/2021
Authors	Gilles Adda, Annelies Braffort, Ioana Vasilescu, François Yvon
Reviewers	Kepa Sarasola, Khalid Choukri
Editors	Maria Giagkou, Stelios Piperidis, Georg Rehm, Jane Dunne
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences)	IPIPAN	PL
42	Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science)	FCULisbon	PT
43	Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy)	ICIA	RO
44	University of Cyprus, French and European Studies	UCY	CY
45	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences)	JULS	SK
46	Institut Jožef Stefan (Jozef Stefan Institute)	JSI	SI
47	Centro Nacional de Supercomputación (Barcelona Supercomputing Center)	BSC	ES
48	Kungliga Tekniska högskolan (Royal Institute of Technology)	KTH	SE
49	Universität Zürich (University of Zurich)	UZH	CH
50	University of Sheffield	USFD	UK
51	Universidad de Vigo (University of Vigo)	UVIGO	ES
52	Bangor University	BNGR	UK

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
ANR	French National Research Agency
ASR	Automatic Speech Recognition
CEF	Connecting Europe Facility
CL	Computational Linguistics
CLARIN	Common Language Resources and Technology Infrastructure
CNRS	Centre National de la Recherche Scientifique
CORLI	Corpus, Languages, Interactions
DDF	Le Dictionnaire des Francophones
DGLFLF	General Delegation for the French language and languages of France
DLE	Digital Language Equality
EEAS	European External Action Service
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELG	European Language Grid (EU project, 2019-2022)
ELRA	European Language Resource Association
ELRC	European Language Resource Coordination
EPO	European Patent Office
EU	European Union
GA	Grant Agreement
GPU	Graphics Processing Unit
HPC	High-Performance Computing
ILO	International Labour Organization
LM	Language Model
LSF	French Sign Language
LR	Language Resources/Resources
LREC	Language Resource and Evaluation Conference
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MOOC	Massive Open Online Courses
MT	Machine Translation
NER	Named Entity Recognition

NLG	Natural Language Processing
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
POS	Part-of-Speech
QA	Question Answering
SL	Sign Language
SLP	Spoken Language Processing
SME	Small and Medium-sized Enterprise
SSH	Social Sciences and the Humanities
SVO	Subject Verb Object
TTS	Text-to-Speech
UAS	Unlabeled Attachment Scores
UNESCO	United Nations Educational, Scientific and Cultural Organization
WSD	Word Sense Disambiguation

Abstract

This report presents a survey of the current state of technologies for the automatic processing of the French language as well as French Sign Language (FSL). Similar reports have been prepared independently for all languages of the European Union. It is based on a thorough analysis of existing tools and resources for French, and also provides an accurate presentation of the domain and of its main stakeholders. This report is organized in three main parts: first, two background sections that document notably the presence of French on the internet, as well as defining in broad terms the domain of language technologies. The core of the report is made of the following two sections, describing respectively the state of play for French and French sign language. The last two sections first summarize the main findings of a quantitative analysis performed within the ELE project; then spell out some general conclusions and formulate a series of recommendations, the implementation of which could improve the current technological support for French and FSL.

Résumé long

Ce rapport présente un panorama du domaine des technologies linguistiques pour le français, *en se focalisant principalement sur la question des ressources et outils* aujourd'hui disponibles pour le traitement automatique de la langue française et pour le traitement automatique de la langue des signes française (LSF). Il fait partie d'un ensemble d'états des lieux préparés dans le cadre du projet Européen "European Language Equality", qui vise à établir une feuille de route pour parvenir à une "égalité linguistique" en dans l'Union Européenne à l'horizon 2030, en s'appuyant sur des recensements précis des ressources, modèles et outils existants pour les langues européennes. Un inventaire de l'ensemble des ressources identifiées est consultable en ligne sur le site du projet "European Language Grid".[1] Décrire l'état des technologies pour la langue française dépasse largement la question des ressources et ce rapport présente également les principaux acteurs impliqués dans le développement de méthodes, données, outils et services pour le traitement de données linguistiques : équipes et laboratoires de recherche académiques, laboratoires privés, PME et startups, ainsi que les agences de financement, en essayant de couvrir aussi largement que possible l'ensemble du paysage européen (essentiellement en Belgique, France et Suisse), ainsi qu'au Canada.

Ce rapport est structuré en trois parties principales : les sections 2 et 3 introduisent le contexte général de l'étude : la première en présentant un bref état chiffré de la langue française et de la LSF à l'ère du numérique ; la seconde en introduisant très sommairement au domaine du traitement automatique des langues et aux principales technologies associées à ce domaine. Un constat majeur est la place centrale des systèmes de traitement basés sur l'apprentissage automatique profond, qui exploitent des grands corpus pour en extraire des modèles numériques utiles pour de multiples tâches. Les sections 4 et 5 concentrent l'essentiel des analyses concernant l'état des technologies respectivement pour la langue française et pour la langue des signes française. Une quinzaine de grandes familles de technologies (par exemple : la recherche d'information, la traduction automatique, la transcription de parole, etc) y sont passées en revue et analysées relativement à l'état de maturité et aux performances des systèmes de traitement automatique de 2022, ainsi qu'à la disponibilité des ressources linguistiques associées. Une analyse comparative chiffrée de la situation des différentes langues européennes complète cette analyse.[2] Dans la section 7, nous récapitulons

[1] https://european-language-grid.eu

[2] Cette analyse a été produite par les partenaires du projet ELE indépendamment du reste du rapport, et insérée à l'identique par les éditeurs dans chacun des états de l'art produit par le projet. Elle ne reflète pas nécessairement l'opinion des auteur-e-s du présent rapport.

un certain nombre d'observations réalisées tout au long de cette étude et formulons un ensemble de recommandations qui permettraient de progresser vers une plus grande égalité linguistique en Europe.

Pour résumer ces conclusions dans leurs grandes lignes, cette étude démontre la bonne vitalité du domaine des technologies linguistiques pour la langue française, qui peut aujourd'hui s'appuyer sur un solide réseau d'équipes de recherche académiques, ainsi que sur un nombre croissant d'acteurs privés offrant des services variés sur toute la gamme des technologies, depuis la synthèse vocale jusqu'à la détection d'infox, ou encore l'extraction d'informations spécialisées. L'ensemble de l'éco-système peut s'appuyer sur un large ensemble de ressources linguistiques (données, modèles, outils) accumulées au fil des années, qui permettent d'entrainer et d'évaluer des systèmes de traitement pour de nombreuses applications. Pour autant, l'écart avec le niveau des ressources disponibles pour l'anglais ne cesse de s'accentuer, aussi bien pour ce qui concerne la variété linguistique, la diversité des domaines et applications, que pour ce qui concerne la taille des données accessibles. Une conséquence majeure est la moindre qualité générale des outils de traitement automatique aujourd'hui disponibles pour la langue française par comparaison aux outils qui existent pour les locuteurs de l'anglais.

Pour faire évoluer cet état de fait, plusieurs propositions sont formulées qui visent (a) consolider l'effort de production, d'archivage et de diffusion des sources de données existantes ; (b) ouvrir plus largement l'accès à des grandes sources de données linguistiques qui sont aujourd'hui sous-exploitées ; (c) mieux coordonner, pour certains grands domaines critiques, l'effort de développement de nouvelles ressources qui sont aujourd'hui manquantes pour le français et la LSF ; (d) relancer l'effort d'évaluation des technologies linguistiques, en proposant de nouveaux jeux de données pour des tâches réalistes, qui permettront de diagnostiquer précisément les biais et limitations des systèmes de traitement actuels (e) soutenir les recherches interdiscipinaires sur les technologies linguistiques pour aller vers une compréhension profonde des langues et accélérer leur diffusion dans tous secteurs de la recherche et de l'industrie.

1 Introduction

This study is part of a series that reports on the results of an investigation of the level of support the European languages receive through technology. It is addressed to decision makers at the European and national/regional levels, language communities, journalists, etc. This series of reports seeks to not only delineate the current state of affairs for each of the European languages covered, but to additionally – and most importantly – to identify the gaps and factors that hinder further development of research and technology. Identifying such weaknesses will lay the grounds for a comprehensive, evidence-based, proposal of required measures for achieving Digital Language Equality in Europe by 2030.

More than 40 research partners with expertise about language technologies in more than 30 European languages have conducted an enormous and exhaustive data collection that provided a detailed, empirical and dynamic map of technology support for our languages.[3]

The report has been developed by the European Language Equality (ELE) project.[4] With a large and all-encompassing consortium consisting of 52 partners covering all European countries, research and industry and all major pan-European initiatives, the ELE project develops a strategic research, innovation and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030.

[3] The results of this data collection procedure have been integrated into the European Language Grid so that they can be discovered, browsed and further investigated by means of comparative visualisations across languages.

[4] http://european-language-equality.eu

2 The French Language and the French Sign Language in the Digital Age

2.1 The French Language in the digital age

Linguistic facts about French

French is genetically a Western Romance language, together with other languages whose origin is Latin including Northern Italian dialects, Spanish and Portuguese, with whom French shares many linguistic features due to the common origin and to a long contact history (known as the "areal alliance" of Romance languages) (de Carvalho, 2008; Walter, 2016; Pusch, 2011). French inherited Gaulish features, from the Celtic dialects spoken by the ethnic groups that previously populated the territory conquered by the Romans. The cohabitation of Vulgar Latin with the Gaulish dialects resulted in a new idiom, the Gallo-Romanic language, later on influenced by Germanic dialects as a consequence of the invasions that marked the fall of the Roman Empire.

Like all modern Romance languages, French has retained the Latin alphabet. In terms of typological features (ie. morphological, syntactic and other content-oriented synchronic patterns), French is the Romance language that has diverged most from Latin, and, in that sense, is considered the most innovative Romance language. For instance, French has the most complex vowel system with 12 oral and four nasal vowels and the accent position is fixed at the end of the phonological word. In terms of morphological features, the loss of Latin endings led to a specific and overall reorganisation of nominal and verbal systems. French shares linguistic features with the other Romance languages: for instance, French is a nominative-accusative language (SVO), the distinction being coded by word order. It is also an article language, sharing with the other Romance idioms the process of grammaticalisation that lead to the emergence of definite and indefinite articles (Pusch, 2011; Smith, 2016).

French, the language of "Francophonie"

With 128 million "native and real speakers" worldwide and an estimate of close to 300 million persons speaking French 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.[5]

In Europe, it is estimated that 129 million people speak French making it the 3rd most spoken second language, after English and German. It is ranked second after English as an official language in close to 30 countries around the world, most notably in Europe (France, 65 million speakers, Belgium, 7 million speakers, Switzerland, 3 million speakers, and Luxembourg), Africa, Canada and Haiti. All French-speaking countries together constitute "*La Francophonie*", with the "*Organisation Internationale de la Francophonie*" (OIF) coordinating multilateral policies between 88 associated states and entities.

"*Le Conseil supérieur de la langue française*" (CSLF) is the name given to the national bodies in several French-speaking countries that are responsible for advising their governments on issues related to the use of the French language. Such a body has existed in France and Québec (for the application of the Charter of the French Language), and is still active in Belgium as an institution of the "*Fédération de Wallonie-Bruxelles*". In France and Québec, matters related to these issues have been transferred to their respective ministries; in particular in France to the "*Délégation générale à la Langue Française et aux Langues de France*"

[5] Estimates vary slightly depending on sources: a more cautious statement would rank French in a set of languages with about 250m to 300m speakers, with Russian, Arabic, Urdu and Portuguese https://www.ethnologue.com.

(General Delegation for the French Language and the Languages of France, DGLFLF) under the aegis of the French Ministry of Culture and Communication.[6] Its mission is to elaborate the policies regarding languages in relationship with all ministries, both for the French language and for the various 80 languages spoken in France. DGLFLF organized the "*États-Généraux du Multilinguisme*" in 2008 and the "*États-Généraux du Multilinguisme en Outre-Mer*" in 2011 and 202 which have focused on the regional and minority languages spoken in France. France has always strongly defended the French language on the international scene, either as such (it was prior to the mid 20th century the pre-eminent language of diplomacy), or in the framework of Multilingualism.[7]

The French constitution states that the language of the French Republic is French. By law,[8] information to the consumer and advertising should be in French or have a French translation, and all participants in a scientific debate in France have the right to express themselves in French. Employees should be free to use French and should have access in French to office systems in any company. All audio-visual services that broadcast in France are required to use the French language. Radios should include a quota of French content, while TV may fully broadcast in a foreign language. Only official Web sites are required to use French on the Internet. At the same time, the legislation aims at promoting plurilingualism: where an administration translates information intended to the public, it should be done at least in two foreign languages, and the law also aims at two languages other than French in education.

The *Académie Française* has been established in 1635 as the pre-eminent French body to address matters related to the French language, including the maintenance of a reference dictionary. The Belgium *Académie royale de langue et de littérature françaises*, established in 1920, also aims to study and promote the French language. Although their work does not really impact the usage of French in the real word, such institutions play an active role in the control of neologisms, within the "*Commission d'enrichissement de la langue française*".[9] Another key institution for the French language is the "*Fondation Alliance Française*", whose mission is to promote the French language and culture outside France, with close to 800 (to be compared to 1 000 in 2011) "*Alliances Françaises*" representations and 500,000 students in 132 countries all over the world.[10]

The French language on the Internet

In the 2018 edition of *The French Language in the World* (Pimienta, 2017; Collectif, 2019), it was established that the French language occupied the fourth place on the Internet, behind English, Chinese and Spanish, with a comfortable lead over the following languages: German, Portuguese, Japanese, Russian, Hindi and Arabic. In the 2022 edition of *The French Language in the World* (Pimienta, 2022), it is established that although French remains in fourth place on the Internet:

- It is now accompanied, and perhaps already surpassed,[11] by Hindi which is showing a spectacular rise;
- its lead over the following languages (now Portuguese, Russian, Arabic, German, Japanese, and Malay) has narrowed considerably, as a result of the combination of the following two facts: 1) Internet connection rates of Francophones in industrialized countries are

[6] https://www.culture.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France
[7] http://www.efnil.org/documents/declarations/dublin-declaration-1
[8] https://www.legifrance.gouv.fr/loda/id/LEGITEXT000005616341/
[9] Formerly known as the "*Commission Générale de Terminologie et Néologie*'. Term lists are published on the FranceTerme web site: http://www.culture.fr/franceterme.
[10] https://www.fondation-alliancefr.org
[11] India's official Internet connectivity data appears to be grossly underestimated. Moreover, if languages that are variants of Hindi were counted with Hindi, then it would be ahead of French.

close to saturation (85% on average), leaving little room for growth, and 2) the digital divide in Francophone African countries is much slower to close than the average growth of connectivity worldwide.

Looking at the distributions of documents types and themes, the strengths of French on the Internet are found in books, MOOCs and research (extended to all the themes of the "science and technology" section).

W3Tech[12] gives a slighlty different picture, with a very strong rise of English since 2018 (from 51% to 63%) and a fall of French from 4.1 to 2.5% of the total web pages (currently in 6th position, surpassed by Spanish in particular).

French as an international language

French is one of the 24 official languages of the EU and one of the three working languages of the European Commission, with English and German. There has been an unfavorable trend in the use of French as a working language within the various European institutions; more than a decline in the use of French, this trend reveals a general decline of multilingualism (Lequesne, 2021).

As documented in this report, for the drafting of source documents, of the 69,000 documents produced by the General Secretariat of the Council in 2018, 1215 (2 %) were in originally in French. On the other hand, 65,908 documents were written originally in English, amounting to 95% of the total number of documents. The remaining 3.1% represent all the other official languages of the EU. The dominance of English is also visible in the source documents published by the Commission. In 2019, 3.7% of the Commission documents sent for translation had French as their source language, compared to 85.5% for English. Twenty years earlier, in 1999, 34% of documents were in French. For the source documents produced by the Parliament's committee services, in 2019, only 11.7% of the documents had French as their source language. At the European External Action Service (EEAS), in 2019, only 0.9% of the documents sent for translation to the EEAS services were written in French, compared to 98.7% in English.

French is also a working language at the OECD (Organisation for Economic Co-operation and Development), at the United Nations (including UNESCO and ILO (International Labour Organization), together with English, Spanish, Russian, Mandarin Chinese and Arabic), one of the three languages of the Olympic Games, together with English and the language of the organizing country, one of the three official languages, with English and German, at the European Patent Office (EPO), and one of the four working languages of the African Union, together with Arabic, English and Portuguese.

Another sign of French vitality is given by the number of translations worldwide as measured by the UNESCO: French is ranked 2nd as a source language (far behind English), and 3rd as a target language, after German and Spanish.[13] This can be interpreted as the fact that the production of intellectual assets in French is important and of interest for non-francophones, and that it already covers a relatively large amount of the needs of francophone speakers.

2.2 The French Sign Language in the digital age

General facts about Sign Languages

Sign Languages (SLs) are natural languages practised within Deaf communities. While the word "deaf" refers to the hearing statut, "Deaf" with a capital D indicates a cultural identity

[12] https://w3techs.com/technologies/history_overview/content_language/ms/y
[13] Translation at UNESCO: http://databases.unesco.org/xtrans/stat/xTransStat.html.

for deaf people who share a common culture and who usually have a shared SL that is their first language.

SLs are visual-gestural languages: a person expresses himself in SL using many bodily components (hands and arms, but also facial expressions, gaze, chest, etc.) and his interlocutor perceives the message through the visual channel. The linguistic system of SL exploits these specific channels: a lot of information is expressed simultaneously and is organised in space, and iconicity plays a central role.

Sign languages are not universal. Just as a spoken language has many forms, dialects and local variations, so does SL, although there are also some similarities among them at the grammatical level. The 2021 edition of Ethnologue[14] lists 150 SLs, while the SIGN-HUB Atlas of Sign Languages[15] lists over 200 and notes that there are more which have not been documented or discovered yet.

There are also tactile sign languages used by people with deafblindness. There is a significant degree of difference with visual SL as elements such as facial expression will have to be replaced by additional manual information.

A study published in 2020 by the Royal Society Open Science on the diversity of SLs and their evolutionary processes[16] shows that, in the sample studied (which does not take into account all the world SLs), there are six main European lineages, with three larger groups of Austrian, British and French origin, and three smaller groups centred on Russian, Spanish and Swedish. Some SLs in current use appear to be independent of these groups, such as Norwegian. The status of SLs in Europe differs from country to country. Some countries do not recognise their SLs, others recognise them in their constitutions, and others provide for full or partial recognition by law.

As with any spoken language, SLs are also vulnerable to becoming endangered. For example, a SL used by a small community may be endangered and even abandoned as users shift to a SL used by a larger community. Even nationally recognised SLs can be endangered due to the increase in early implantation of cochlear implants, and the encouragement by doctors to parents – who are generally hearing – to favour oral communication.

To date, SLs do not have a standard writing system. As a result, the presence of SLs on the Internet and the social media is mainly in the form of videos. The presence of SLs on these media greatly varies from country to country and, even for those countries where it is most present, it remains rare.

French Sign Language

The French Sign Language (LSF) is the signed language used in France. All the languages of the world are subject to sociolinguistic variation, and LSF is no different: there is not just one, but several ways of signing, depending on regions, towns, villages, schools, family histories and cultures, and there are significant regional variations in the lexicon. In addition to their local language, most Deaf people know the LSF that corresponds more or less to the Parisian dialect, and adapt without difficulty to any interlocutor who uses it.

LSF was recognised under the Law No. 2005-102 of 11 February 2005 “for equal rights and opportunities, participation and citizenship of people with disabilities”. Quoting art. L. 312-9-1:[17] “The French Sign Language is recognised as a language in its own right. Every pupil concerned must be able to receive instruction in French Sign Language. The *Conseil Supérieur de l’Éducation* ensures that its teaching is encouraged. It is regularly kept informed of the conditions of its evaluation. LSF may be chosen as an optional test in examinations

[14] https://www.ethnologue.com/subgroups/sign-language
[15] https://www.sign-hub.eu/sign-language/
[16] https://royalsocietypublishing.org/doi/10.1098/rsos.191100
[17] https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006524761/

and competitions, including in the context of professional training. Its dissemination in the administration is facilitated."

LSF is practised in France and in the French-speaking part of Switzerland. The number of LSF signers (persons who practise LSF), who may be deaf as well as hearing (e. g. children of deaf parents, family or relatives of deaf people), is not precisely known. It is said to be practised by about 169,000 people worldwide, including about 100,000 in France in 2014.[18]

In the education of young deaf children, Article L. 112-3 of the Education Code[19] sets out the principle of freedom of choice between:

- bilingual communication: French sign language (LSF) and written French language,
- communication in written and oral French: with or without the support of French completed spoken language (LfPC)[20] or French sign language (LSF).

A teaching centre for young deaf people (*Pôle d'enseignement des jeunes sourds* - PEJS) brings together, in a given geographical area, the resources needed to support deaf pupil from kindergarten to high school, regardless of their linguistic project. For the bilingual pathway, LSF is the pupils' first language: it is the language of instruction but also a taught language. Deaf pupils follow the lessons in LSF and learn French progressively, essentially via the written language and thanks to LSF. Throughout their schooling, pupils will deepen their mastery of LSF while gradually integrating elements of the Deaf culture. In practice, it seems that this choice is not simple because there are not always adequate structures for all regions of France.

The presence of LSF in the media has increased in recent years, particularly since the adoption of the 2005 law. However, fully bilingual websites are still extremely rare. The two main ones are "*Média'Pi!*",[21] a generalist bilingual online media created by Deaf journalists, and "*L'œil et la main*",[22] a bilingual television program of the documentary type developed and produced by a mixed team (deaf and hearing people).

3 What is Language Technology?

Natural language[23] is the most common and versatile way for humans to convey information. We use language, our natural means of communication, to encode, store, transmit, share and process information. Automatic processing language is a non-trivial, intrinsically complex task, as language is subject to multiple interpretations (ambiguity), and its decoding requires knowledge about the context and the world, while in tandem language can elegantly use different representations to denote the same meaning (variation).

The computational processing of human languages has been established as a specialized scientific field known as *Computational Linguistics* (CL), *Natural Language Processing* (NLP), *Spoken Language Processing* (SLP) or, more generally, Language Technology (LT). While there are differences in focus and orientation, since CL is more informed by linguistics, NLP by computer science and SLP by signal processing, LT is a more neutral term. In fact, LT is largely multidisciplinary in nature; it combines linguistics, computer science (and notably

[18] https://fr.wikipedia.org/wiki/Langue_des_signes_française
[19] https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000019911145/
[20] This is the French version of "Cued Speech", "a visual mode of communication that uses hand shapes and placements in combination with the mouth movements and speech to make the phonemes of spoken language look different from each other." (source: https://en.wikipedia.org/wiki/Cued_speech
[21] https://media-pi.fr
[22] https://www.france.tv/france-5/l-oeil-et-la-main/
[23] This section has been provided by the editors. It is an adapted summary of Agerri et al. (2021) and of sections 1 and 2 of Aldabe et al. (2021).

AI), signal processing, mathematics and psychology, among others. In practice, these communities work closely together, combining methods and approaches inspired by all, together making up *language-centric AI*.

A concise definition may just be the following: Language Technology is the multidisciplinary scientific and technological field that is concerned with studying and developing systems capable of processing, analysing, producing and understanding human languages, whether they are written, spoken or embodied.

LT's history started in the 1950s with Turing´s renowned description of a intelligent machine (Turing, 1950), initial attempts at automatic translation (Booth and Locke, 1955) or at the formal description of grammatical structures by linguists (Chomsky, 1957) and computer scientists (Yngve, 1960). In France, early research efforts were initiated around the same period (Cori and Léon, 2002). LT then enjoyed a bumpy history (Wilks, 2005), alternating periods of (excessive) hopes with times of disillusions. A major methodological shift occurred in the 1990s, a period signalled by intense efforts to create wide-coverage linguistic resources, such as annotated corpora, thesauri, etc. which were manually labelled for various linguistic phenomena and used to elicit machine readable rules which dictated how language can be automatically analysed and/or produced. Gradually, with the evolution and advances in machine learning, rule-based systems have been displaced by data-based technologies and the development of systems that learn implicitly from examples. In the recent decade of 2010s, machine learning massively embraced to the use of multilayer neural networks able to solve various text classifications, then sequential labelling, problems. The success of this approach lies in the ability of neural networks to learn continuous vector representations for linguistic units (i. e. word or sentence embeddings) using vast amounts of unlabelled data; based on these, effective processing tools can be trained using only a small amount of labelled data in a fine-tuning stage. Such techniques have attracted a tremendous interest and are still progressing at a fast pace, illustrated by the development of contextualized and multilingual continuous space representations.

In recent years, the LT community has been witnessing the emergence of ever-more powerful deep learning techniques and tools that are transforming the way in which LT tasks are approached. The field is gradually moving from a methodology in which a pipeline of multiple modules was the typical way to implement LT solutions, to architectures based on complex neural networks trained end-to-end with vast amounts of data, be it text, audio or multimodal. Another massive trend is the use of multilingual and multimodal models that allow transfer learning between tasks, languages and modalities. The success in these areas of AI has been possible because of the conjunction of four different research trends: 1) maturation of deep neural network algorithms and technology, 2) large amounts of data (and for LTs, large and diverse multilingual data), 3) increase in high performance computing (HPC) power in the form of GPUs, and 4) application of simple but effective self-learning approaches.

LT is trying to provide solutions for the following main application areas:

- **Text Analysis** which aims at identifying and labelling the linguistic information underlying any text in natural language. This includes the recognition of word, phrase, sentence and section boundaries, recognition of morphological features of words, of syntactic and semantic roles as well as capturing the relations that link text constituents together.
- **Speech processing** allows humans to communicate with electronic devices through voice. Some of the main areas in Speech Technology are Text to Speech Synthesis, i. e. the generation of speech given a piece of text, Automatic Speech Transcription, i. e. the conversion of speech signal into text, Spoken Dialog Systems, capable to fullfil tasks based on a spoken interaction, Speaker Recognition and more generally Speaker Char-

acterisation, aimed at deriving information about a speaker from recordings such as basic demographic information, current emotional state, etc.

- **Machine Translation**, i. e. the automatic translation from one natural language into another: this encompasses the machine translation of written, spoken, and signed languages utterances; Machine Interpretation, mediating the communication between speakers of different languages, also needs to integrate real time processing issues, and to take into account extra-linguistic context such as the speaker's emotions or intents;
- **Information Extraction and Information Retrieval** which aim at extracting structured information from unstructured documents, finding appropriate pieces of information in large collections of unstructured material, such as the internet, and providing the documents or text snippets that include the answer to a user's query. This latter task is known as Question Answering (QA).
- **Natural Language Generation (NLG)**. NLG is the task of automatically generating texts. Summarisation, i. e. the generation of a summary, the generation of paraphrases, text re-writing, simplification and generation of questions are some example applications of NLG.
- **Human-Computer Interaction**, which aims at developing systems that allow the user to converse with computers using natural language (text, speech and non-verbal communication signals, such as gestures and facial expressions). A very popular application within this area are conversational agents (better known as chatbots), but one must also mention human-robot interaction, interaction in virtual environments, verbal interaction in games, etc.

In addition, LTs are also serving the needs and purposes of empirical and applied linguistics, providing linguists and cognitive scientists with new tools to explore languages in the actual diversity of their uses, more generally helping scholars througout all the fields and disciplines of Digital Humanities to analyze their unstructured data.

LT is already fused in our everyday lives. As individual users, we may be using it without even realizing it, when we check our texts for spelling errors, when we use internet search engines or when we call our bank to perform a transaction. It is an important, but often invisible, ingredient of applications that cut across various sectors and domains. To name just very few, in the *health* domain, LT contributes for instance to the automatic recognition and classification of medical terms or to the diagnosis of speech and cognitive disorders. It is more and more integrated in *educational* settings and applications, for instance for educational content mining, for the automatic assessment of free text answers, for providing feedback to learners and teachers, for the evaluation of pronunciation in a foreign language and much more. In the *law/legal* domain, LT proves an indispensable component for several tasks, from search, classification and codification of huge legal databases to legal question answering and prediction of court decisions.

The wide scope of LT applications evidences not only that LT is one of the most relevant technologies for society, but also one of the most important AI areas with a fast growing economic impact.[24]

[24] In a recent report from 2021, the global LT market was already valued at USD 9.2 billion in 2019 and is anticipated to grow at an annual rate of 18.4% from 2020 to 2028 (https://www.globenewswire.com/news-release/2021/03/22/2196622/0/en/Global-Natural-Language-Processing-Market-to-Grow-at-a-CAGR-of-18-4-from-2020-to-2028.html. Another report from 2021 estimates that amid the COVID-19 crisis, the global market for NLP was at USD 13 billion in the year 2020 and is projected to reach USD 25.7 billion by 2027, growing at an annual rate of 10.3% (https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market).

4 Language Technologies for French

4.1 Language Data and Tools

This section is based on analysis of the 1500+ resources, tools, and models that were already identified and documented for the French languages, as well as other sources that are yet to be documented by the ELE consortium.

To review the state of play in terms of language tools, we mostly follow the same organisation as in (Mariani et al., 2012) and consider the same 15 groups of technologies, with the addition of (a) dedicated sectionss for generic resources, lexical and textual, that can serve for many applications: dictionaries, terminological resources, monolingual corpora and language models; (b) a supplementary section on LTs for formal and applied linguistic research. These themes are detailed below. With respect to the recent advances of the field (see Section 3), it appears that the overwhelming majority of state-of-the-art tools and applications rest almost exclusively on generic machine learning technologies, meaning that the most important ingredients for system building are data and, to a lesser extent, computing resources. Datasets will thus be discussed together with the related technologies. Depending on the target application(s), specific operational constraints might require dedicated software development, for instance to minimize the training cost, memory footprint, increase the output speed, adapt the interaction with the user, etc. This implies that tools, models and algorithms dedicated to LTs continue to play an important role and will be mentionned wherever needed.

Monolingual dictionaries and terminological resources

Large-scale, general purpose computational lexica for French are widely available,[25] associating lemmas or word forms to basic morpho-syntactic information such as part-of-speech (POS), grammatical gender, conjugation class (for verbs); depending on the resource, other information such as detailed morphological and syntactic properties, pronunciation, frequency, difficulty level, might also be available. Among these, the French Wiktionary[26] is the most up-to-date (and the largest) resource, if not the most linguistically informed. In the Neural Net era, syntactico-semantic distributed representations are also a basic requirement and such resources for French are notably distributed through the FastText project.[27]

Historical lexicographic resources for French are made available by the ATILF,[28] while the collaborative website *Le Dictionnaire des Francophones*[29] (DDF), which gives access to a large scale compilation of dictionary definitions from multiple sources, perhaps reflects a more diverse and dynamic view of the French lexicon. DDF not only documents the basic general lexicon, but also includes specialized dictionaries and term lists. A considerable number of additional lexical and terminological resources (specialized vocabularies, monolingual or multilingual lists of terms, thesaurus, ontologies) of variable size and level of details have also been made available - large size examples are the Government of Canada's terminology and linguistic data bank TERMIUM[30] and the IATE list of terms accumulated by the translators of the EU.[31]

[25] For instance, dozens of French lexica are listed in the ELRA catalogue.
[26] https://fr.wiktionary.org/wiki/Cat%C3%A9gorie:fran%C3%A7ais
[27] http://fasttext.cc
[28] https://www.atilf.fr/ressources/tlfi/
[29] https://www.dictionnairedesfrancophones.org
[30] https://www.btb.termiumplus.gc.ca
[31] https://www.eurotermbank.com

Monolingual corpora, large-scale language models

There is no official French National Corpus, that would contain a representative subset of the language, balanced accros periods, genres and domains, as may exist for other languages. Searchable sizable corpora (up to billions of tokens) of mixed genres are however accessible and searchable e. g. on via Frantext,[32] the Sketch Engine,[33] or on the Leipzig Corpora Collection website.[34] Other well represented genres that are easy to search and download include web texts (Wikipedia, but not only), literature (eg via the Gutenberg project[35]), news (e. g. the NewsCrawl corpus[36]), science, and legal / administrative written texts from national or international institutions; some of these are updated on a regular basis.

The CommonCrawl project[37] aggregates Web crawled data that is orders or magnitude larger than these resources for many languages; furthermore this corpus is being updated on a regular basis. By using parts of the French subset of CommonCrawl, possibly conjoined with the more curated corpora alluded to above has enabled to train large-scale BERT-style Language Models (LMs) – FlauBERT (Le et al., 2020) is built with a corpus containing about 12B running words, CamemBERT (Martin et al., 2020) uses the 22B words OSCAR, and these numbers continue to grow, albeit at a much slower pace than the corresponding English corpora. Large LMs for French of various guises (BERT-style, ELMO-style, GPT-style,[38] mBART-style) are now published and available for research and commercial uses, and more are to come (e. g. thanks to the BigScience initiative); they have enabled to boost the state-of-the art in multiple NLP tasks. A small number of specialized variants (e. g. for modeling tweets) are also available. As for other languages, adapted versions could easily be derived (e. g. for scientific texts and patents). These resources offer new perspectives notably for text generation in French.[39]

Large scale databases of annotated (segmented in sentences, speakers and turns, transcribed) recordings, containing thousands of hours of recordings are available for several genres (e. g. news, read books, talks). This mostly concerns standard French and large corpora for other recording conditions (e. g. conversational, spontaneous, multi-party, telephone, emotional, noisy, pathological speech) and/or other geographical variants are more difficult to find. The collection of large sets of recordings thus remains a pressing issue to widen the applicability and effectiveness of speech processing in French, an objective that is addressed e. g. by Mozilla's Common Voice[40] or the Voice Lab project.[41] The situation is probably even more difficult for multi-modal data, which pose even more privacy and right issues than spoken data.

Basic language pack: tokenisation, POS tagging , morphology analysis/generation

This is an area where the ground was already well covered in 2012 and has benefited from the improvement of machine learning tools. Open source industrial strength tokenizers, lemmatizers and POS taggers for French are available for instance in Spacy,[42] Spark NLP[43]

[32] https://www.frantext.fr
[33] https://www.sketchengine.eu/corpora-and-languages/french-text-corpora/
[34] https://corpora.uni-leipzig.de/en?corpusId=fra_mixed_2012
[35] https://www.gutenberg.org/
[36] http://data.statmt.org/news-crawl/README
[37] https://commoncrawl.org
[38] e. g. Pagnol: https://pagnol.lighton.ai
[39] https://cedille.ai
[40] https://commonvoice.mozilla.org/fr
[41] http://www.levoicelab.org
[42] https://spacy.org
[43] https://nlp.johnsnowlabs.com

or Stanza[44] and can be accessed with a couple of lines of Python code. Several other software packages dedicated to the processing of French (such as MELT[45]) coexist with multilingual processing suites developed nationally or internationally in the academia or the industry (NLTK,[46] Freeling,[47] GATE,[48] etc.). Building more of such tools should be quite straightforward given the availability of generic, trainable sequence labelling tools and of large amounts of annotated data. Two caveats are however in order: (a) no recent systematic performance comparison exist for these tasks; (b) most of these tools are meant to process "generic" French and too little exists for more specific varieties or genres (e. g. technical texts, e-mails, text messages, speech transcripts, user generated content and the like, notably for those varieties that evolve very quickly; the same holds for less represented varieties of written or spoken French, e. g. spoken outside of France, which may also contain various kinds of code-switching phenomena). For such linguistic material, the quality of POS-tagging and parsing is known to decrease quickly (Plank et al., 2014).

A side note: in many recent NLP pipelines, such tools are not even needed beyond basic sentence segmentation and word tokenisation, as processing is performed with automatic (sub)tokenisation process using e. g. Byte Pair Encoding (BPE) units which are trained end-to-end and do not require the notion of a linguistic word and of its basic properties (Sennrich et al., 2016; Kudo and Richardson, 2018). This is for instance the case in the HuggingFace tokenisation library.[49] Such preprocessing strategies are very effective, but lack the linguistic motivations that is needed for the analysis of LT components.

With respect to deep morphological analysis of newly coined words or compounds, the situation is much less favourable as the only tool of significance is DERIF (Namer, 2009), dedicated to the processing of general texts more than technical texts. Thanks to the progress of the UniMorph project,[50] which embedds the French Lefff (Sagot, 2010), generic morphological analyzers are easier to train, and models have started to appear also for French.

Parsing, deep and shallow

The situation with respect to parsing is very similar to that of POS tagging, owing to the availability of a large multilingual open repository for treebanks developped in the "Universal Dependencies" project,[51] which can be readily used to train parsers. Several French treeBanks (FTB, GTB, Sequoia, etc.) are included in this collection, with a grand total of close to 1.2M running words from varying sources and genres. Training a dependency parser for French with basic syntactic information is thus relatively straightforward and yields accuracy in the low 90s in terms of Unlabeled Attachment Scores (UAS), which may be good enough for many applications. Recent evaluation results are in (Zeman et al., 2018). Such parsers have been integrated in generic NLP tools such as again Spacy or Stanza; academic developments based on the same resources are relatively easy to access, develop and modify. Other useful resources for parsing French include several large-scale dictionaries with detailed syntactic information, as well as treebanks and corpora developed in the French projects EASY (Technolangue programme), ANR/Passage (de la Clergerie et al., 2008) and ANR/PARSEME (Candito et al., 2017); finally, a sizeable set of crowd-sourced annotations have been collected through the ZombiLingo serious game (Fort et al., 2014).

[44] https://stanfordnlp.github.io/stanza/
[45] https://gitlab.inria.fr/almanach/alTextProcessing/melt
[46] https://www.nltk.org
[47] https://nlp.lsi.upc.edu/freeling/
[48] https://gate.ac.uk
[49] https://github.com/huggingface/tokenizers
[50] https://unimorph.github.io
[51] https://universaldependencies.org

Support for deep parsing tools that would deliver a fine grained analysis is somewhat less developed.

Sentence level semantic analysis

Sentence level semantic analysis for French can make use of large-scale semantic networks inspired by the English Wordnet project, or their multilingual versions (e. g. Open Multilingual Wordnet,[52] which includes the French WOLF (Sagot and Fišer, 2008)). Such resources typically provide information regarding the sense inventories of the most common word forms and help to disambiguate word use in context. A few small scale semantically annotated corpora (e. g. French SemEval), developed notably in the context of SemEval shared tasks, can also be readily used to train or evaluate word sense disambiguation (WSD) modules.

Other valuable resources for sentence level analysis are the Asfalda French FrameNet, a derivation of the original FrameNet,[53] as well as semi-automatically annotated parallel corpora where semantic role labels have been projected from the English side of the corpus. In spite of the availability of these generic resources, tools and models for automatic WSD[54] and semantic role labeling are less advanced and integrated than syntactic analysis tools. One reason is a lack of semantically annotated treebank for French; another reason may be that the usefulness of deep analyses in downstream applications might not be as critical as it used to be.

A last important sentence-level task is the computation of textual entailment relationships (RTE or NLI) is also an area where machine learning methodologies perform well; additionally, cross-lingual transfer techniques are also relevant for this task, meaning that large English datasets can be leveraged to develop systems for French: such approaches can rely, for instance, on the XNLI dataset.[55]

Discourse-level semantic analysis

A basic task for discourse-level analysis is coreference resolution: large scale annotated resources (e. g. ANCOR-Centre and Democrat) for learning such systems exist both for spoken and written texts and are useful to train and evaluate tools for the corresponding textual genres (Wilkens et al., 2020). French is also part of the core set of languages for the CorefUd Project.[56] Despite this, the computation of coreference relationships is not (yet) widely available in standard NLP tools, one exception being the LIMA multilingual analysis tool developed by CEA LIST.[57]

More general discourse structure analysis can rely on the Annodis corpus[58] and derived tools and benchmarks produced for the Disrpt shared tasks (Zeldes et al., 2021). Discourse level processing is also critical for the analysis of conversations and for dialogue systems. Actual (spoken or written) dialogues are here quite sparse, owing to the difficulty of annotating such resource, which compound many of the pitfalls of textual analysis at multiple levels (speech, emotions, semantic labels, dialog acts, etc.), notwithstanding privacy issues related to their collection. MEDIA (Bonneau-Maynard et al., 2008), which was the result of a significant collaborative effort, is narrow in scope and somewhat outdated; public initiatives to develop new resources and record more open-domain spoken interactions are limited (in

[52] http://compling.hss.ntu.edu.sg/omw/summx.html
[53] https://framenet.icsi.berkeley.edu/fndrupal/
[54] Such as Babelfy (http://babelfy.org).
[55] https://github.com/facebookresearch/XNLI
[56] https://ufal.mff.cuni.cz/corefud
[57] https://github.com/aymara/lima/
[58] http://redac.univ-tlse2.fr/corpus/annodis/

scope, annotation types and size). This might be seen as a weakness given the blooming of applications for such systems (chatbots, dialogue agents, vocal assistants and the like); it is likely, though, that large private databases have been collected for many languages, including French.

Information extraction

Named Entity Recognition (NER) systems are widely available as basic service in multiple text analysis suites (see above), where the notion of a name entity is however mostly limited to the basic MUC-style entity types and structures (names, places, organisations, dates, amounts). Some of these also include some form of relation extraction. Existing large corpora with NER annotations for French News (both written and spoken) are available (Galliano et al., 2009; Sagot et al., 2012; Dupont, 2019) can be used to trained effective NER systems (Ortiz Suárez et al., 2020); a concern is their quick obsolescence as new names and entities keep appearing in the news. For specific subdomains (e. g. legal, health, science), where the important entities and relations may be of a different nature (e. g. legal references, cases names or drugs, symptoms and virus) public NER systems are more scarce.

Opinion and sentiment mining, or hate speech detection are mostly conceived as pure sentence classification tasks that can be implemented with little, if any, linguistic analysis; their development rests on the availability of open, large scale annotated data. For French, available data mostly contain annotated lists of tweets and product reviews with polarity annotation (including irony and sarcasm (Karoui et al., 2017) and the DEFT 2017 campaign); stance label data is also available for the politic domain.[59] In comparison, there is a shortage of public resources dedicated to other important types of sentence classification tasks such as identifying fake, or hateful content (see however (Chiril et al., 2020) for sexism detection or resources distributed by media producers such as AFP[60] or "Le Monde"[61]). This is an area where cross-lingual transfer from English, where such resources exist, may be considered – with uncertain results, though, given the cultural and language dependency of subtle, yet important, phenonemas such as humorous and sarcastic language. Also note that for such applications, large scale multilingual databases (including French examples) are detained and exploited by industrial stakeholders, either for internal use or for commercial exploitation.[62]

Information retrieval and text mining

Information retrieval and text mining technologies have been deployed at scale for a number of years and are less actively researched than they used to, the focus progressively shifting towards more challenging interactions (e. g. dialogue). Thanks to years of evaluation campaigns such as TREC,[63] TRECVID,[64] CLEF,[65] and AMARYLLIS and DEFT for French,[66] large collections of query-document pairs exist for a variety of data types (texts, structured documents, speech and videos transcripts), languages and domains. Robust, effective, language-independent open-source IR tools are also widely available.

Classification and clustering tools are other instances of mostly language independent tools for which resources and mature tools on-the-shelf technology is easy to find. Here again,

[59] https://github.com/ZurichNLP/xstance
[60] E. g. in the EU project InVid https://www.invid-project.eu/invid-datasets/ and WeVerify
[61] https://s1.lemde.fr/mmpub/data/decodex/hoax/hoax_debunks.json
[62] e. g. https://storyzy.com/?lang=fr for Fake News detection.
[63] https://trec.nist.gov
[64] https://trecvid.nist.gov
[65] http://www.clef-initiative.eu
[66] https://deft.lisn.upsaclay.fr

the performance and maturity level of existing technology may greatly vary depending on the data types, genres and domain, with short, noisy texts or speech utterances still posing difficult challenges, and warranting the development of new resources.

French resources for the keyword or terminology extraction tasks are much more difficult to find, especially in comparison to what exists for the English language where datasets with hundreds of thousands examples have been developed.[67]

Natural Language Generation

Natural language generation starts with writing aids providing services such as spell, grammar and style checking, as well as autocompletion and text normalisation features (notably for User Generated contents). Tools for this are widely available and the most advanced softwares are commercially available and embedded in many text editors and text entry boxes. If dictionaries are easily found, real world corpora annotated with genuine errors are much more scarce; the same holds for learner corpora, whih are collected “en masse” by educational institutions but very rarely redistributed.

The technical landscape of text generation has been deeply impacted by the new developments of very large language models alluded to above. These techniques are enabler for new applications using controlled text generation from structured (e. g. statistics, tables, or logical formulas) or unstructured (text prompt or images) signal. Notwithstanding the many ethical considerations associated to the training and use of these large language models (Bender et al., 2021), these techniques are also prone to malicious exploitations such as the production of fake news / e-mails generation, fraudulent web sites, link farms, etc. Evaluating the output of NLG tools also requires large-scale resources, which currently do not exist for French.

Other important applications of text generation: automatic summarisation and machine translation are discussed below.

Text summarisation, Question Answering

Text Summarisation has long been studied and evaluated as other text mining technologies, with a history of shared task mostly addressing English texts as well as multilingual sources. Even though dedicated corpora such as PUCES or RMP2[68] have been around for a while, text summarisation applications for French are not as developed as other text mining applications: while the past generation of system was based on extractive techniques, recent progresses in statistical / neural text generation techniques, coupled with the availabililty of a larger training dataset for summarizing news (Scialom et al., 2020) or (Eddine et al., 2020), may change this state of play and foster the development of abstractive text summarizers - at least for some well covered textual genres.

Question-Answering (QA) is now a mature technology, available as a basic service in generic conversational agents. QA for French can benefit from the resources developed over the years in the framework of EQUER, TREC and CLEF evaluation campaigns mentioned above. Progresses for QA and reading comprehension in French, so as to address a wider range of question types and difficulty will benefit from recent attempts at designing large scale datasets for QA research such as PIAF (Keraron et al., 2020) and FSquad (d’Hoffschmidt et al., 2020), which are making available tens of thousands of question-passage-answer triplets to the community. If these resources are extremely useful, we note that like for other technological areas, more specialized resources, which are now available for other languages, are still lacking for French (e. g. QA databases for the medical or the “how-to” domains).

[67] https://github.com/boudinfl/ake-datasets
[68] http://juanmanuel.torres.free.fr/corpus/rpm2/doc_resumes_fr.html

Machine Translation

Having moved to fully corpus-based, nowadays fully-neural, the availability of MT systems for French mostly depends on the availability of appropriate parallel corpus. Owing to its use as one major international language, such resources exist for French, especially when paired with an English translation, a pair for each hundreds of millions of parallel segments can be exploited. Massive sources of translation training on text data are for instance available on the Opus website[69] (Tiedemann, 2012), in the CommonCrawl (EU project) website, through Facebook's OpenCC or on the resource page of the "Workshop for Machine Translation" (WMT).[70] Data for speech translation exists in much smaller amounts, and for very restricted domains (talks, parliamentary debates); the IWSLT series of benchmarks[71] distributes training and test datasets for such applications. Many high quality, generic MT engines are also available on the web for most well-resourced language pairs (with various usage restrictions: e-translation, deepL, Google Translate, Bing Translator, or the mostly French-made Systran NMT and Reverso Translation), including the direct translation between French and most European languages, as well as Arabic, Chinese or Japanese. The situation with respect to other languages is much less favourable: if large parallel corpus with other European languages can easily be collected, the situation for non-European languages is more contrasted, which may come as a weakness for translating French into Japanese, Chinese, Russian, or Arabic, especially in specialized domains.

Large sets of pretrained MT models for French have been made available on the HuggingFace platform[72] (mostly thanks to University of Helsinki's efforts, but also thanks to the open source policy of research centers such as Meta's FAIR, which has released open-source multilingual models) as well as on Systran's marketplace.[73]

Additional resources for MT from and into French include bilingual dictionaries of various sizes and content, multilingual term lists for many domains as well as evaluation benchmarks for specific aspects of MT (terminology, coreference, gender bias, etc.). Last but not least, sentence and word alignment tools still play an important role for the preparation and analysis of MT systems; they also constitute valuable assets for translation studies. While generic open-source tools are relatively easy to find, only a handful of evaluation corpora exist, all of them matching French with English.

Speech transcription

The support for speech transcription can be considered as satisfactory, and speech recognition engine are notably a component of conversational assistants (Siri, Alexa, Home, etc.) that have reached the general public and are also widely available as cloud-based services for businesses (IBM Watson, Google NLP Cloud services, Amazon Transcribe, etc.). Some of them can handle several dialects of French to better serve the needs of the Belgium, Canadian or Swiss speakers. ASR quality varies greatly depending on the recording conditions and speech type, and dedicated commercial ASR solutions for specific usecases such as meeting transcription, speech recognition in cars or planes, automating subtitling of videos, are also available. Several small to medium size French companies are active on these markets (Vocapia Research, SNIPS/Sonos, ChapsVision, Linagora etc.) with a lively scene of new players (Amberscript, Zenidoc, Noota, etc). Appropriate open data for training ASR systems at scale is available from various sources (LibriSpeech, Mozilla Common voice).

[69] https://www.opus.eu
[70] see http://statmt.org/WMT21 for the most recent occurrence.
[71] https://iwslt.org
[72] https://hugginface.co
[73] https://www.systransoft.com/marketplace-catalog/?lang=fr

Existing ASR systems are however brittle and known to quickly degrade when the recoding conditions include noise or echo, or when the speaker style and voice diverge from the training data: this may be due to variations in age, accents, data types, but may also be caused by temporary or permanent voice impairements. As such, the lack of robustness with respect to these impredicted voices may also be a cause of exclusion. The need for more varied speech databases that would correctly represent the diversity and voices of the general population thereby remains a legitimate and important goal.

Speech is intrinsically multimodal and its production is a complex motor system involving the movement of several organs and physiological structures of the human body (the speech articulators). A whole line of research aims at developing speech technologies based not only on the sound of the voice but also on these articulatory movements (Schultz et al., 2017). A typical example is audio-visual speech recognition where a video of the speaker's face, processed together with the audio speech signal, improves robustness to noise. Other applications are automatic lip reading (i. e. a visual-only speech recognition system), speech separation and talking face synthesis (see below). While very large audiovisual speech database exist for English, containing hundreds of hours, only a few of them are available for French and are order of magnitude smaller (Petrovska-Delacrétaz et al., 2008).

ASR is often accompanied with modules for performing various related tasks such as speech activity detection, speaker identification and diarisation. For such tasks, one can rely notably on resources developed in the course of a series of evaluation campaigns (ESTER,[74] ETAPE,[75] REPERE[76]).

Speech generation

Until recently, a typical pipeline for text-to-speech (TTS) synthesis was composed of two distinct modules: (i) text processing in charge of extracting the phonetic sequence and the syntactic structure of the sentence to be synthesized and (ii) a signal processing module in charge of the audio synthesis. The latter step requires well designed and curated recordings of clean speach made in a controlled acoustic environnement, a very costly resource that is rareply shared between TTS developers. In the deep learning era, TTS has switched to a full end-to-end mode where the speech signal is generated directly from the orthographic representation (Shen et al., 2018). While a reasonable speech quality can be obtained from short samples of the target voice, associated with orthographic transcripts, high quality, expressive TTS still requires richtly annotated samples of clean data from one single speakers: examples of such resources for French include the SynPaFlex-Corpus[77] as well as datasets produced for the Blizzard challenge.[78]

Coupled with the open-source code available for training advanced neural systems,[79] these databases can be used to build a text-to-speech system (almost) from scratch. It is still important to note that most available recordings correspond to productions of educated speakers of standard French in quite controled conditions, thereby limitating the scope, diversity, and expressiveness of the generated speech. Like for ASR, it is therefore crucial to continue the collection of more diverse speech databases, such as spontaneous speech, speech in interaction, emotional speech, etc.

[74] http://www.islrn.org/resources/055-636-352-982-9/
[75] http://islrn.org/resources/425-777-374-455-4/
[76] http://www.elra.info/en/projects/archived-projects/repere/
[77] http://synpaflex.irisa.fr
[78] https://zenodo.org/record/4580406
[79] e. g. https://github.com/NVIDIA/tacotron2

Emotions in speech

Emotion detection and generation in speech is a relatively recent domain. It is associated to the broader field of affective sciences that emerged in the early 2000. Affective sciences is an interdisciplinary field whose stated goal is to study emotions and other affective phenomena (e. g. attitudes, cognitive states) through the contribution of several humanistic and digital disciplines. Within it, affective computing focuses primarily in the recognition and synthesis of facial expression and of voice inflections (Picard, 2000). It thus relates to several language technology areas. For instance, the detection of speaker's voice patterns as function of emotional states is of interest for various language technology domains such as speech recognition, speaker characterisation, human-computer interactions; speech synthesis of emotional voices is another field (see preceding sections). Applications that involve the modeling of the emotional state of speakers concern fields as varied as health, security, education, entertainment or serious games.

Language technologies for linguistic studies

As for most of Social Sciences and Humanities (SSH) domains, the digital revolution has created new avenues for language analysis (Liberman, 2019). Technology helped linguists from different backgrounds not only to much more easily collect, store, enrich and exchange various types of data, but also to analyse differently the linguistic material. Technology played a significant role in reconsidering classical research questions and paved the way to new research hypotheses verified in large-scale corpora or at least in data enriched with various meta-data. It also raised the issue of reproducibility, a concern shared with all experimental sciences. Finally, it helped building new collaborations with other domains and fostered interdisciplinarity. With respect to the collaboration, computer sciences are naturally privileged, however new bridges are built with various others disciplines such as medicine, neuroscience, psychology, literature, sociology etc.

This methodology is also happening in France. Through local, that is at universities or laboratories levels, specific initiatives and/or at national level, technology impacts all linguistic domains, under the form of corpora, tools and methods. In terms of corpora, both written and spoken varieties of the French language are covered although historically written sources and related textometry tools are better represented and more visible.[80] Resources such as Frantext[81] initiated two decades ago covers sampled corpus from the 9th to the 21st century and a research tool that allows users to perform simple and complex searches on forms, lemmas or grammatical categories and to display the results. Mixed written and large scale spoken resources are also available, thanks to projects such as RHAPSODIE,[82] ORFEO ("*Outils et Ressources sur le Français Ecrit et Oral*"),[83] ESLO ("*Enquêtes socio-linguistiques d'Orléans*"[84]). With respect to the technology needed to analyze such data, French is relatively well equipped thanks to the NLP accomplishments resulting in a variety of tools for lexical, morphology, syntactic and semantic analysis (e. g. NooJ[85]): a remaining issue concerns the usability of these tools, which are sometimes difficult to install and exploit for the non-experts. LTs are also of great value for various subfields of applied linguistics, notably, as mentioned in Section 3, to study or assist language learners - especially those with reading or writing difficulties. If valuable lexical resources exist for French,[86] the study of readabil-

[80] See e. g. https://txm.gitpages.huma-num.fr/textometrie/
[81] https://www.frantext.fr
[82] https://rhapsodie.modyco.fr
[83] https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/index.html
[84] http://eslo.huma-num.fr/index.php
[85] https://www.nooj-association.org
[86] e. g. https://cental.uclouvain.be/cefrlex/

ity measures, text simplification techniques, automatic augmented reading tools, question generation and correction etc., is still hindered by the lack of large scale public datasets.

On the spoken side of the language, initiatives in terms of corpora have been driven by specific needs, dependent of linguistic domains: CLAPI[87] corpora are aimed for discourse analysis; PFC ("*Phonologie du Français Contemporain*")[88] was initiated by phoneticians and phonologists interested in accents, although in fine the corpus is could fulfill needs in other linguistic domains. In terms of applications for spoken data, various tools are available: aligners compute a text/speech signal correspondence that facilitates phonetic analysis (EasyAlign,[89] MAUS,[90] etc.) while more analytic tools and software applications are based on PRAAT (Boersma and Weenink, 2009). A consequence of the collaboration between LTs and linguistics is the sharing of corpora built for technological purposes with the linguistic community. Examples are the ESTER (Galliano et al., 2009) and ETAPE (Gravier et al., 2012) corpora, built for evaluation campaigns for speech recognition that were retrieved and fruitfully explored and exploited by linguists (phoneticians, phonologists, sociolinguists). Such resources are particularly useful as they benefit from manual transcription and sound/text alignment greatly facilitating the linguistic analytic task.

4.2 Projects, Initiatives, Stakeholders

Academic research

France counts approximately 40-50 research teams that are active on the NLP scene and a handful of laboratories in Belgium (Uni. Louvain-la-Neuve, Uni. Mons, Multitel lab), in Switzerland (Uni. Geneva, Uni. Lausanne, EPFL, Martigny's IDIAP), and in Canada (Montréal, Laval, Ottawa), are also actively contributing to the research on LTs for the French language. A recent initiative in Québec attempts to better structure research on NLP under the umbrella of the newly launched CLIQ-ai.[91] French teams are distributed over the entire territory, and are in most cases affiliated to universities, with a possible joint affiliation to the CNRS, and as well as (in rarer cases) to Inria.[92] By contrast, research in NLP is almost entirely absent from the scope of the French engineering schools (*Grandes Ecoles*), with the notable exception of Telecom Paris in Palaiseau. The largest centres are established in Aix-Marseille, Avignon, Besançon, Grenoble, Le Mans, Nancy, Nantes, Paris, Orsay, Rennes, Sophia-Antipolis, Strasbourg and Toulouse. Historically, laboratories hosting NLP research have been either predominantly focusing on Computer Science / Signal Processing on the one hand or Linguistic and Language Studies, on the other hand, with very little overlap or truly multidisciplinary centers. This activity is structured at the national level by CNRS, which supports two national research networks on language technologies and resources,[93] with complementary themes and objectives. In addition to CNRS and Inria, two other national research institutions host a significant activity in NLP: CEA-LIST in Palaiseau; INRAe in Jouy-en-Josas.

Multidisciplinary research on LTs is predominantly supported by three dedicated scientific associations (ATALA,[94] predominantly for the study of the written language, AFCP,[95]

[87] http://icar.cnrs.fr/ressources-base-donnees/
[88] https://www.projet-pfc.net
[89] http://latlcui.unige.ch/phonetique/easyalign.php
[90] http://clarin.phonetik.uni-muenchen.de/BASWebServices/interface
[91] https://fr.cliq-ai.quebec
[92] CNRS (*Centre National de la Recherche Scientifique*) and Inria (*Institut National de Recherche en informatique et en Automatique*) are national public reseach institutions, employing full-time researchers and engineers.
[93] The GdR ("*Groupement de Recherche*") TAL (https://gdr-tal.ls2n.fr) and LIFT (https://gdr-lift.loria.fr).
[94] "*Association pour le Traitement Automatique des Langues*".
[95] "*Association Française pour la Communication Parlée*".

mostly for the study of the spoken language, and ARIA,[96] which supports research on information retrieval), as well as the more generalist AFIA.[97] Conjointly or separately, they organize yearly or bi-yearly scientific events: "*Journées d'Études sur la parole*" (JEP, as of 1970), "*Traitement automatique des langues naturelles*" (TALN, as of 1994), and "*Conférence en Recherche d'Information et Applications*" 'CORIA, as of 2004), which continue to attract a diverse set of scholars from the all the subdomains of LTs. Research on LTs is also disseminated in journals such as "*Traitement automatique des langues*" (TAL) and "*Corpus*", as well as in "*Discours*" or "*Revue Française de linguistique appliquée*".

In all these circles, the development, dissemination and documentation of linguistic resources and tools are nowadays a well-developed practice: analyzing 15 years of LREC (Language Resource and Evaluation Conference) proceedings, Mariani et al. (2014) mention that more than 10% of the papers are co-authored by a French group; the tangible outcome of this work consisting of corpora, models and tools, is also well reflected in ELG's databases. All these activities are nowadays getting a more fair recognition in academic evaluations.

Businesses and industry

Owing to its role as an international language and the comparative large size and advanced development of the French speaking markets, French is relatively well covered by international LT services and language providers: for instance, French-English has been one of the earliest translation pair on the internet, and French versions of Siri, Amazon Echo and Google Home have been available for several years. This means that the development of technologies for French far exceeds developments happening in France or other French-speaking countries.

For a bird's eye view, industrial actors present on the French scene can be grouped into three main categories. The first contains international technology provider involved in the development and commercialisation of AI-based solutions, with language technologies in the scope of their French offices: Apple, Fujitsu, Huawei, IBM, Google/DeepMind, META/FAIR, Microsoft, NaverLabs, Samsung, Sony, as well as Orange, Thales and Dassault-System belong to this group. The open-source policy of some of these actors has resulted in the release of some large-scale multilingual tools and resources that are also of use for the French language, as illustrated by FAIR's fairseq collection of resources.[98]

A second family of players, whose share is harder to evaluate, contains large companies outside the IT sector that are developing or integrating LT technologies for their internal needs and products. Companies from various industrial sectors such as Airbus, AFP, Cap Gemini, EDF, Engie, Renault, Sanofi, SNCF, Société Générale as many smaller businesses from the service sector have shown an interest for LTs, even though their current net contribution to the resource landscape remains comparatively small.

The third group, by far the largest, gathers SMEs developing dedicated services and softwares. This group has been developing very quickly over the last years and includes a mix of "historical" actors such as Systran and Reverso (MT), Druide and Synapse (spell checking), Synomia, Sinequa or Pertimm (search engines) and a variety of very small companies and startups which have emerged along with the development of IA technologies. Most application domains, from search (Qwant, Exalead-DS) to text generation (Syllabs), from speech transcription (Vocapia Reseach) to optical character recognition and document processing (Jouve, A2IA), from text analytics (Expert Systems) to speech synthesis (Acapela group, Voxygen), are addressed. Recent years have witnessed a notable upsurge of players developing conversational agents for the customer management relationship (Davi, Hellomybot.io, Julie

[96] "*Association Francophone de Recherche d'Information et Applications*"
[97] "*Association Française d'Intelligence Artificielle*"
[98] https://github.com/pytorch/fairseq/

Desk, Konverso, Kwalys, Linagora, ViaDialog, Vivoka, Zaion, etc.). Through years of joint collaborative projects with academic teams, these actors have contributed to the development of a myriad of resources and tools, some of them publicly accessible or distributed with an open-source licence.

Instruments and platforms for resources sharing

The linguistic research can rely on Huma-Num,[99] a national research infrastructure dedicated to the humanities, social sciences, and digital humanities, implemented by the Ministry of Higher Education and Research and supported by the National Center for Scientific Research, Aix-Marseille University and the Condorcet Campus, through which social scientists can access both tools and data. Huma-Num has as stated goal to develop, implement and preserve research programs – their data and tools – over the long term in a context of open science and data sharing. It is supplemented by initiatives such as ORTOLANG,[100] an "EquipEx" (for "*Équipement d'Excellence*") validated within the framework of "*Programme Investissements d'Avenir*" (PIA: Programme of Investments for the Future), whose aim is to provide specialized linguistic services, complementary to Huma-Num. In terms of data acquisition, processing and sharing, initiatives such as the CORLI (Corpus, Languages, Interactions) consortium within the Huma-Num structure helps to disseminate corpora, tools and methods of work and exploration of these corpora. The proposed services provide financial and methodological help to enrich/finalize corpora, training dedicated to new methodologies in various linguistic corpora investigation, or workshops and summer schools for students interested in improving their digital corpus processing skills.

National projects in the area of LTs

A first observation is the absence of recurrent calls for project proposal specifically targeting LTs in France in the recent years. A few calls have been launched in the last ten years, among which we can note the calls for projects "*Langues et Numérique*" ("Languages and the Computer") led by the General Delegation for the French language and languages of France (DGLFLF) with the support of the Secretary of State for Digital Affairs and Innovation (2017 and 2018) with a total funding of about 500,000 euros for each call.[101] In the framework of National Artificial Intelligence Research Programme (PNRIA)[102] launched in 2018, with a public budget of 1.5 billion euros, a network of 4 operational Interdisciplinary Artificial Intelligence (3IA) has been launched, where language is partly present; out of the approximately 190 Chairs in Research and Teaching associated to these 3IA or funded by subsequent calls, 10-15 address language processing and language technologies. AI related programs have also been launched in Belgium, notably in the French speaking Wallonia under the umbrella of the "DigitalWallonia4.ai", and more recently, through the Trusted AI Labs (TRAIL[103]).

In France, most projects concerning LTs are funded by the ANR (the French National Research Agency) in the framework of the generic call for projects, which takes place every year. If some research areas, among the fifty or so existing ones, explicitly mention language processing, it should be noted that none is specifically dedicated to LT. The result is a great disparity in the number of projects and in the amount of money dedicated to LT each year

[99] https://www.huma-num.fr
[100] https://www.ortolang.fr
[101] The Ministry of Culture has been issuing a call for projects for the past ten years on "Cultural action and the French language" for the proficiency in French, with a budget of 1 million Euros per year, but only a marginal part of it concerns the development of LTs.
[102] https://www.intelligence-artificielle.gouv.fr/fr/thematiques/programme-national-de-recherche-intelligence-artificielle-pnria
[103] https://trail.ac/trail4wallonia/

(all instruments financed by the ANR): 9 projects for a sum of 2.64 million euros in 2021, 6 projects for a sum of 2.10 million euros in 2020 (partly due to a reallocation to specific COVID topics), 21 projects for a sum of 7.9 million euros in 2019, 11 projects for 3.98 million euros in 2018. This great variability (from simple to quadruple) from one year to another, is not favorable to planning, to the regular hiring of PhD students and post-doctoral fellows. Applied linguists can also get support from ANR both through general and collaborative calls. Some regional fundings are also available such as labex ("*laboratoires d'excellence*") and RNMSH calls,[104] sometimes under interdiscplinary technology and linguistic oriented calls such as the initiative developed by Paris-Saclay MSH and the DATAIA institute.[105]

Another opportunity for large scale application projects has existed through dedicated national programmes funded by the Ministry of Industry or the Ministry of Defense through BPIFrance.[106]: in the recent years, projects such as ROSETTA (on subtitling and sign language generation), Linto (an open source smart assistant), and le Voice Lab (speech data collection) have benefited from significant fundings. However, the overall amount of money distributed via such channels is difficult to consolidate.

Finally, research on LTs, like for other subdomains of AI, has greatly benefited from the development of the Jean Zay platform[107] (as of 2019), an open high-performance computing infrastructure hosting thousands of modern GPU units. Without Jean Zay, several large scale projects mentioned above would not have been possible.

5 Language Technology for French Sign Language

Sign Languages (SLs) are under-resource Languages: very few reference books, partial knowledge of grammar, limited lexicons or corpora, very few SL technologies, very few resources in general. Moreover, research in automatic processing is much more recent than that for spoken or written languages, and although research in this area is active in both recognition, generation and machine translation, there is no usable tools yet, except for a few rare products but which are *a priori* not completely automatic.

5.1 Language Data and Tools

A recent document,[108] which forms a deliverable of the European project EASIER,[109] lists linguistic resources that can be used for SL processing and the extent to which they are publicly available. More specifically, it lists:

- linguistic corpora of European SLs of "substantial" (for SLs) size that can be used as high-quality training data for automatic translation,
- data collection tasks used in more than one of these linguistic corpora,
- lexical resources of European sign languages.

There is two kinds of resources: linguistic corpora and broadcast data. While broadcast data are available in comparatively large quantity, linguistic corpora offer high quality data through rich transcription and linguistic annotation. Another important difference is that

[104] https://www.msh-reseau.fr
[105] http://msh-paris-saclay.fr/appel-a-projets-excellence-2021-msh-paris-saclay-institut-dataia-23-04-2021/
[106] https://www.bpifrance.fr Bpifrance is a public investment bank for the financing and development of companies.
[107] http://www.idris.fr/eng/jean-zay/jean-zay-presentation-eng.html
[108] https://www.project-easier.eu/wp-content/uploads/sites/67/2021/08/EASIER-D6.1-Overview-of-Datasets-for-the-Sign-Languages-of-Europe.pdf
[109] https://www.project-easier.eu

the former is interpreting, and therefore a sign language that is subject to the constraints of temporality and structure of oral discourse, whereas the latter is a sign language produced by Deaf signers for whom it is their first language.

As the technology to automatically tag or annotate SL data in the quality required for linguistic annotation does not exist yet, corpus creators have met the challenge of having to manually annotate the data. As said before, SLs do not have a standard writing system, or even a graphical system for transcription (equivalent to the international phonetic alphabet), and to date, different conventions for the annotation are used throughout the different corpora.

For French Sign Language (LSF), the main monolingual dictionaries and terminological resources are the following:

- Ocelles,[110] a collaborative website entirely bilingual in French and LSF which collects signs, definitions, information on projects and organisations as a teaching resource. For each concept at least one definition and its associated descriptors in various knowledge fields are proposed. Users can upload information (e. g. texts, pictures, videos, presentation) which is examined by experts on form and content before being released online.
- Sign'Maths,[111] a glossary dedicated to mathematics. The Sign'Maths group conducts its research by organising monthly workshops in LSF, bringing together teachers of Maths and LSF, from primary, secondary and higher education, but also Deaf students and interpreters. They discuss various mathematical concepts, field by field, until a specific lexical unit (or sign) is established that meets both linguistic constraints and mathematical criteria.
- Elix,[112] a dictionary for LSF/French working like a search engine. French keywords can be searched, hits show associated signs and their definition in LSF. Elix can be used as an online web platform and as an application. To date, it contains over 21,000 French definitions translated into LSF and over 15,300 signs.
- Dicta Sign Lexicon,[113] a multilingual lexicon for BSL, GSL, DGS, LSF, English, Greek, German and French. Approximately 1,000 concepts are provided for each of the project SLs. The shared list of concepts chosen for the lexicon is of everyday use or specifically related to the field of travels in Europe.

To date, the three main LSF corpora are the following:

- CREAGEST, a corpus of adult and child French Sign Language (LSF) and of natural gestures. It consists of three sub corpora: a child acquisition dataset, a dataset of dialogues between Deaf adults and a dataset of natural gestures. For the acquisition data 65 Deaf children and 17 Deaf adults were recorded by four Deaf investigators. For the dialogue dataset 51 interviews were conducted by four Deaf investigators. For the gestural dataset pairs of five hearing-hearing, five Deaf-Deaf and Deaf-hearing individuals were recorded. In total more than 500 hours of over 250 signers have been recorded. To date, only a small part of the corpus has been annotated (1 hour) and is available on the Ortolang website.[114]
- Dicta-Sign-LSF-v2, an extended version of the LSF sub-corpus of the corpus created during the Dicta-Sign European project, providing primary data (videos), elicitation data,

[110] https://ocelles.inshea.fr
[111] https://signmaths.univ-tlse3.fr
[112] https://dico.elix-lsf.fr
[113] https://www.sign-lang.uni-hamburg.de/dicta-sign/portal/
[114] https://www.ortolang.fr/market/corpora/ortolang-000926

annotation data and a related annotation guide, as well as preprocessed signer data including facial pose, upper body pose and hand shape estimates. It contains nine dialogue sessions with 18 signers of LSF covering the topic of travel in Europe. The data was annotated in more detail and a convolutional-recurrent learning network was trained on the data, drawing on a compact and generalisable modeling of the signers to provide a baseline for the recognition of lexical signs and non-lexical structures. It contains 11 hours of annotated and translated (into French) videos.[115]

- Mediapi-Skel, a 2D-skeleton video corpus of LSF with French subtitles. The corpus consists of 368 subtitled videos produced by Média'Pi,[116] a media company producing bilingual content with LSF and written French. It contains abour 27 hours of LSF and 17,000 tokens from subtitles.[117]

5.2 Projects, Initiatives, Stakeholders

In France, there is no national program, infrastructures or LT providers related to LSF or SL technology. Some research projects are funded by agencies such as the ANR or the DGLFLF. More recently, collaborative public/private projects are funded by BPIFrance. Over the past 5 years and currently, the main projects involving SL technology are the following:

- ROSETTA,[118] a French public/private project that studied access solutions for audiovisual content, including an exploratory study related to automatic translation of subtitles in LSF displayed through virtual signing avatars.
- *Serveur Gestuel*, a French public/private project that aims to create a gestural server, i. e. the equivalent of a voice server but in LSF, thus including recognition and generation technologies.

6 Cross-Language Comparison

The LT field[119] as a whole has evidenced remarkable progress during the last years. The advent of deep learning and neural networks over the past decade together with the considerable increase in the number and quality of resources for many languages have yielded results unforeseeable before. However, is this remarkable progress equally evidenced across all languages? To compare the level of technology support across languages, we considered more than 11,500 language technology tools and resources in the catalogue of the European Language Grid platform (as of January 2022).

6.1 Dimensions and Types of Resources

The comparative evaluation was performed on various dimensions:

- The current state of technology support, as indicated by the availability of tools and services[120] broadly categorised into a number of core LT application areas:

[115] https://www.ortolang.fr/market/corpora/dicta-sign-lsf-v2/
[116] https://media-pi.fr
[117] https://www.ortolang.fr/market/corpora/mediapi-skel
[118] https://rosettaccess.fr/index.php/home-page-english/
[119] This section has been provided by the editors.
[120] Tools tagged as "language independent" without mentioning any specific language are not taken into account. Such tools can certainly be applied to a number of languages, either as readily applicable or following fine-tuning, adaptation, training on language-specific data etc., yet their exact language coverage or readiness is difficult to ascertain.

  - Text processing (e. g. part-of-speech tagging, syntactic parsing)
  - Information extraction and retrieval (e. g. search and information mining)
  - Translation technologies (e. g. machine translation, computer-aided translation)
  - Natural language generation (e. g. text summarisation, simplification)
  - Speech processing (e. g. speech synthesis, speech recognition)
  - Image/video processing (e. g. facial expression recognition)
  - Human-computer interaction (e. g. tools for conversational systems)
- The potential for short- and mid-term development of LT, insofar as this potential can be approximated by the current availability of resources that can be used as training or evaluation data. The availability of data was investigated with regard to a small number of basic types of resources:
  - Text corpora
  - Parallel corpora
  - Multimodal corpora (incl. speech, image, video)
  - Models
  - Lexical resources (incl. dictionaries, wordnets, ontologies etc.)

6.2 Levels of Technology Support

We measured the relative technology support for 87 national, regional and minority European languages with regard to each of the dimensions mentioned above based on their respective coverage in the ELG catalogue. For the types of resources and application areas, the respective percentage of resources that support a specific language over the total number of resources of the same type was calculated, as well as their average. Subsequently each language was assigned to one band per resource type and per application area and to an overall band, on a four-point scale, inspired by the scale used in the META-NET White Paper Series, as follows:

1. **Weak or no support**: the language is present (as content, input or output language) in <3% of the ELG resources of the same type
2. **Fragmentary support**: the language is present in $\geq$3% and $<$10% of the ELG resources of the same type
3. **Moderate support**: the language is present in $\geq$10% and $<$30% of the ELG resources of the same type
4. **Good support**: the language is present in $\geq$30% of the ELG resources of the same type[121]

The overall level of support for a language was calculated based on the average coverage in all dimensions investigated.

[121] The thresholds for defining the four bands were informed by an exploratory k-means 4-cluster analysis based on all data per application and resource type, in order to investigate the boundaries of naturally occurring clusters in the data. The boundaries of the clusters (i. e. 3%, 10% and 30%) were then used to define the bands per application area and resource type.

6.3 European Language Grid as Ground Truth

At the time of writing (January 2022), the ELG catalogue comprises more than 11,500 metadata records, encompassing both data and tools/services, covering almost all European languages – both official and regional/minority ones. The ELG platform harvests several major LR/LT repositories[122] and, on top of that, more than 6,000 additional language resources and tools were identified and documented by language informants in the ELE consortium. These records contain multiple levels of metadata granularity as part of their descriptions.

It should be noted that due to the evolving nature of this extensive catalogue and differing approaches taken in documenting records, certain levels of metadata captured are not yet at the level of consistency required to carry out a reliable cross-lingual comparison at a granular level. For example, information captured on corpora size, annotation type, licensing type, size unit type, and so on, still varies across records for many languages, while numerous gaps exist for others. As the ELG catalogue is continuously growing, the comprehensiveness, accuracy and level of detail of the records will naturally improve over time. Moreover, the Digital Language Equality (DLE) metric will allow for dynamic analyses and calculations of digital readiness, based on the much finer granularity of ELG records as they mature.[123]

For the purposes of high-level comparison in this report, the results presented here are based on relative counts of entries in the ELG for the varying types of data resources and tools/services for each language. As such, the positioning of each language into a specific level of technology support is subject to change and it reflects a snapshot of the available resources on January 2022.

That said, we consider the current status of the ELG repository and the higher level findings below adequately representative with regard to the current existence of LT resources for Europe's languages.

6.4 Results and Findings

As discussed above, our analysis takes into account a number of dimensions for data and tools/services. Table 1 reports the detailed results per language per dimension investigated and the classification of each language into an overall level of support.

The best supported language is, as expected, English, the only language that is classified in the *good support* group. French, German and Spanish form a group of languages with *moderate support*. Although they are similar to English in some dimensions (e. g. German in terms of available speech technologies and Spanish in terms of available models), overall they have not yet reached the coverage that English has according to the ELG platform. All other official EU languages are clustered in the *fragmentary support* group, with the exception of Irish and Maltese, which have only *weak or no support*. From the remaining languages, (co-)official at national or regional level in at least one European country and other minority and lesser spoken languages,[124] Norwegian and Catalan belong to the group of languages with *fragmentary support*. Basque, Galician, Icelandic and Welsh are borderline cases; while they are grouped in the *fragmentary support* level, they barely pass the threshold from the lowest level. All

[122] At the time of writing, ELG harvests ELRC-SHARE, LINDAT/CLARIAH-CZ, CLARIN.SI, CLARIN-PL and HuggingFace.

[123] Interactive comparison visualisations of the technology support of Europe's languages will be possible on the ELG website using a dedicated dashboard, which dynamically analyses the resources available in the ELG repository, from the middle of 2022 onwards.

[124] In addition to the languages listed in Table 1, ELE also investigated Alsatian, Aragonese, Arberesh, Aromanian, Asturian, Breton, Cimbrian, Continental Southern Italian (Neapolitan), Cornish, Eastern Frisian, Emilian, FrancoProvencal (Arpitan), Friulian, Gallo, Griko, Inari Sami, Karelian, Kashubian, Ladin, Latgalian, Ligurian, Lombard, Lower Sorbian, Lule Sami, Mocheno, Northern Frisian, Northern Sami, Picard, Piedmontese, Pite Sami, Romagnol, Romany, Rusyn, Sardinian, Scottish Gaelic, Sicilian, Skolt Sami, Southern Sami, Tatar, Tornedalian Finnish, Venetian, Võro, Walser, Yiddish.

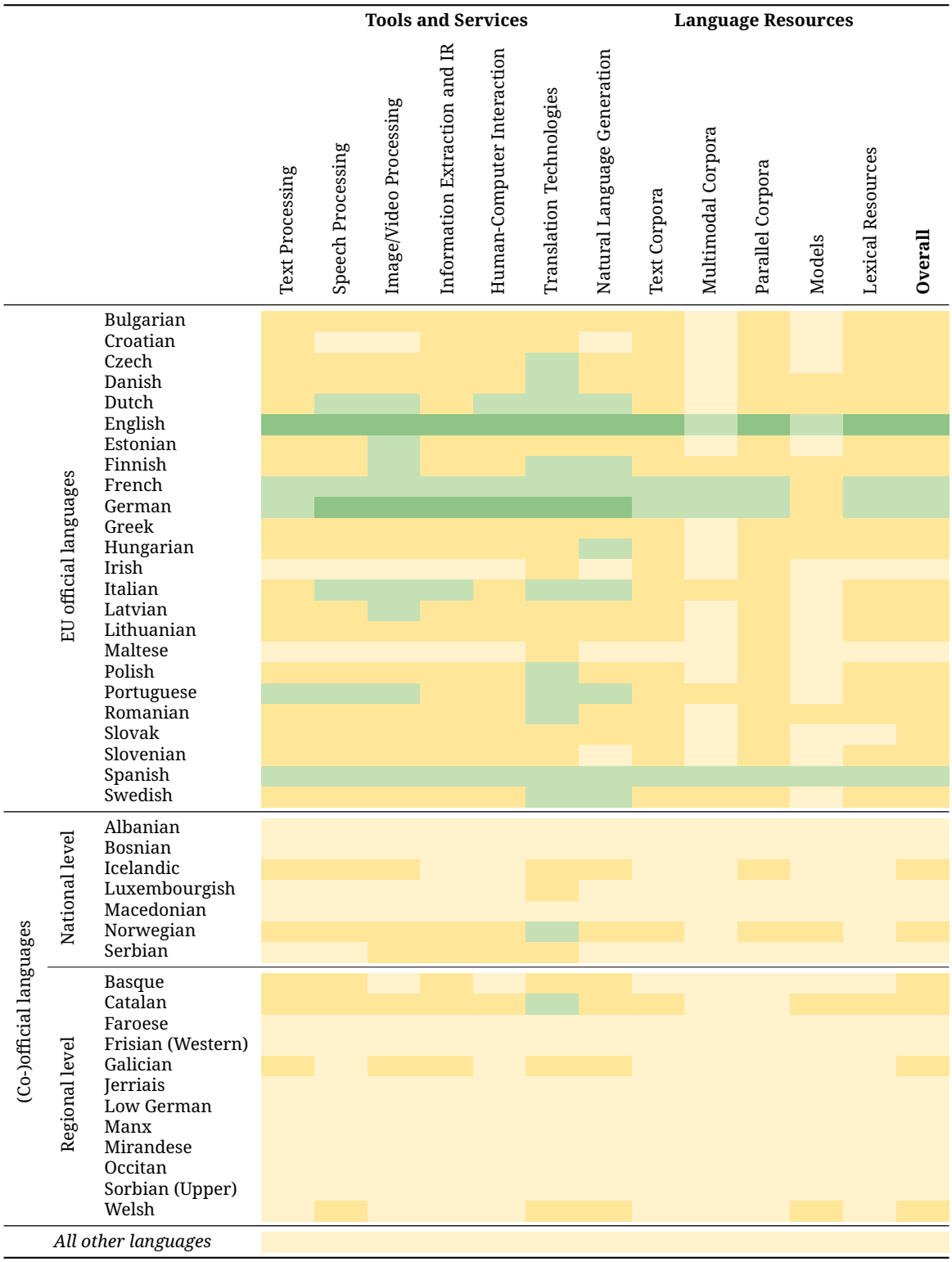

Table 1: State of technology support, in 2022, for selected European languages with regard to core Language Technology areas and data types as well as overall level of support (light yellow: weak/no support; yellow: fragmentary support; light green: moderate support; green: good support)

other languages are supported by technology either weakly or not at all. Figure 1 visualises our findings.

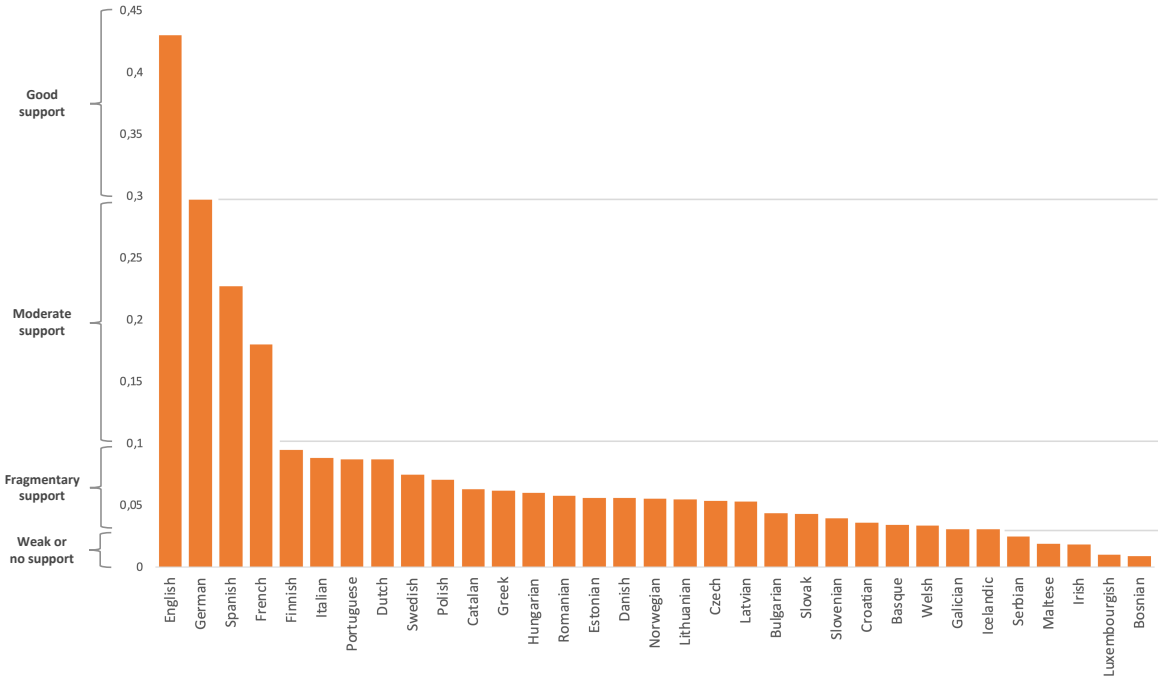

Figure 1: Overall state of technology support for selected European languages (2022)

While a fifth level, *excellent support*, could have been foreseen in addition to the four levels described in Section 6.2, we decided not to consider this level for the grouping of languages. Currently no natural language is optimally supported by technology, i. e. the goal of *Deep Natural Language Understanding* has not been reached yet for any language, not even for English, the best supported language according to our analysis. While recently there have been many breakthroughs in AI, Computer Vision, ML and LT, we are still far from the grand challenge of highly accurate deep language understanding, which is able to seamlessly integrate modalities, situational and linguistic context, general knowledge, meaning, reasoning, emotion, irony, sarcasm, humour, culture, explain itself at request, and be done as required on the fly and at scale. A language can only be considered as excellently supported by technology if and when this goal of Deep Natural language Understanding has been reached.

The results of the present comparative evaluation reflect, in terms of distribution and imbalance, the results of the META-NET White Paper Series (Rehm and Uszkoreit, 2012). The complexities of the analyses clearly differ across 2012 and 2022 studies, and as such, a direct comparison between the two studies can therefore not be made. However, we can instead compare the relative level of progress made for each language in the meantime. It is undebatable that the technology requirements for a language to be considered digitally supported today have changed significantly (e. g. the prevalent use of virtual assistants, chat bots, improved text analytics capabilities, etc.). Yet also the imbalance in distribution across languages still exists.

The results of this analysis are only informative of the relative positioning of languages, but not of the progress achieved within a specific language. The LT field as a whole has significantly progressed in the last ten years and remarkable progress has been achieved for specific languages in terms of quantity, quality and coverage of tools and language resources. Yet, the abysmal distance between the best supported languages and the minimally

supported ones is still evidenced in 2022. It is exactly this distance that needs to be ideally eliminated, if not at least reduced, in order to move towards Digital Language Equality and avert the risks of digital extinction.

7 Summary and Conclusions

7.1 General Observations

Research and development of LTs/AIs for the French language is well-advanced with a large range of ubiquitous LT applications targeting the general public, and more and more, also penetrating businesses and industries. This notably concerns dialog systems and virtual agents, embedded technologies such as auto-completing keyboards, automatic speech recognition, language analysis and understanding, question answering, speech synthesis, or, for other uses, machine translation, all of which exist for the French speaking consumer. It is noteworthy that these applications often rely on open-source components, which lower the cost of developing new services and software.

Accordingly, medium to large-scale resources for the French language have progressively been developed and are easily accessible for all these important tasks and applications of NLP technologies. Many of these resources have been collected opportunistically in research projects or by imitating similar resources for English, thanks to an increased focused on multilingual NLP, or through coordinated attempts to collect parallel or comparable training and test corpora for many tasks (e. g. in the Universal Dependency project). While there is no shortage of usable datasets for studying the algorithmic and computational challenges, or to evaluate the performance of some language processing tools for French, the lack of diversity with respect to language varieties, textual genres, register and domains remains an issue both for linguistic studies and for the development of industrial NLP applications for French.

With the booming of AI-based technologies, new opportunities and new industrial actors have recently emerged, thereby widening the range of addressed services, domains and data types. Academic research on LTs has less directly benefited from this rapid development of investment in AI applications. As a result, the French language can be considered to be reasonably well-resourced in terms of technologies and is more or less comparable in this respect to other European languages such as German, Spanish and Italian. However, the gap in depth, coverage and quality of existing tools that was observed 10 years ago with English has continued to widen, amplifying the linguistic inequalities between English and non-English speakers, as they get better services and LT applications.

As discussed in Section 3, a recent development of LTs has been the amazing successes of brute-force machine learning based methodologies, which apply (almost) equally well to English and French (and to many other languages). These methods rely on (a) massive sources of linguistic data, be they textual, audio, or multimodal that can serve to self-train (multilingual) general purpose lexical representations at scale; (b) well-designed and curated training and evaluation corpora for the largest possible range of applications. Additionally, these successes may have given the illusion that (c) the need to develop language-specific tools was unimportant.

Regarding point (a), our survey has showed that the available resources for French were, in order of magnitude, smaller and less diverse than what was available for English. Having accumulated and annotated gigantic repertoires of linguistic data (including a substantial portion of French), and having developed optimized ML processing tools and large-scale computing infrastructure, some international players (e. g. in the US or in China) are now in a position to develop models, then tools and services that are out of reach of what can be developed by local actors using public resources. Thanks to the open-source policy of some

private actors, models and samples of these datasets have been made available to research and have been further annotated (e. g. for sentiment analysis or hateful content detection): these resources however are fragile and their status may change without notice.

Regarding point (b), our survey has showed that available French resources fail, by a large margin, to cover the full spectrum of applications, domains, genres, and modalities. As noted above, in situations where training data is not available, transfer learning (from English) or the use of machine translation was sometimes a viable alternative, that is often accompanied with a loss in performance or with undesirable side effects (biases). This loss is hardly reported or documented, except when it has visible social or ethical implications, as when a speech recognition engines make more errors for female voices than for male voices. This does not mean that these performance differences do not exist, but that they are not usually measured in open evaluation campaigns, due to the lack of appropriate testing sets. As for language resources, the design of new test sets for LTs is subject to choices from the evaluation organizers who, in an increasingly multilingual landscape, may not see French as an interesting testing language on the grounds that it is already a high-resource language, which shares many linguistic similarities with English, etc.

This brings us to point (c): owing to the linguistic similarity to English, many of the existing resources for (written) English may be useful to improve the processing of French. This is both a good thing, but also entertains the illusion that both languages can be preprocessed analogously, despite important linguistic differences, as well as cultural differences that sometimes need to be carefully looked at (e. g. for tasks such as emotion, or sarcasm detection).

A last observation is that the available resources and tools identified in this survey[125] are scattered among multiple platforms and lack a centralized inventory, associated with detailed meta-data, in spite of many attempts at providing such a service, through institutions such as CLARIN, projects such as FlareNet, or laudable initiatives such as LRE Map,[126] ISLRN[127] or the newly launched European Language Grid. The European Language Resource Association (ELRA), installed in France, has been a major actor in these ventures and has largely contributed to existing resource inventories. Note that in the recent years, such initiatives have been challenged by the development of industrial actors and the open science movement, which have both fostered more attempts to achieve similar goals (e. g. Huggingface's or tensorflow's Dataset libraries or Paperwithcode[128]).

Several of the observations described above also apply to LSF, which, moreover, must be considered a poorly endowed language for which the production of resources and the development of tools is still very rare and scattered.

7.2 Recommendations

Regarding resources and data collection

Many French corpora that exist in the open domain are the result of uncoordinated initiatives and consequently partially cover the needs of specific domain applications, with their specific annotations, meta-data and goals. This state of affairs results in (a) a lack of visibility of tools and data that are only known to restricted communities, although the situation has greatly improved thanks to above-mentioned European and national infra-structures, (b) a lack of interdisciplinary initiatives supported at national level and going beyond.

In the view of the dispersion and heterogeneity of existing datasets, a first recommendation would be to institutionalise clearer policies and incentives for the declaration and

[125] It is also likely that many valuable resources have gone unnoticed, or have simply been lost.
[126] http://www.elra.info/en/catalogues/lre-map/
[127] http://www.islrn.org
[128] https://paperswithcode.com

archival of language resources for French when they are produced by publicly funded research projects, as is already done for scientific publications on the HAL platform. Having a well-identified entry point for linguistic resources, models and tools for French, and their associated documentation, would be of great significance for many scientific disciplines, notably in the Social Sciences and Humanities (SSH) area, as well as for many industrial companies.

Size is also often an issue, with the growing needs of data-eager machine learning softwares. A possible answer would be to open the large datasets produced by the administration and other public institutions (e. g. in the domains of health, culture, media, justice or education) which remain buried and hard to access – sometimes for legitimate reasons (privacy issues, unclear copyright). Incentives and public policies to support and amplify the actions undertaken in the European CEF/ELRC programme,[129] with the development of public repositories with clear access rules analogous to what exists or for health data and for scientific papers through the ISTEX initiative.[130] Where data is currently lacking and where gaps have been identified, it is essential to continue to support the development and annotation of novel large scale open-resources. This is for instance the case for French Sign Language, which could greatly benefit from an initiative similar to the German DGS-Korpus.[131]

Applications that involve social network data require specific actions, as they are often associated with delicate legal issues (related to proprietary rights or personal information) that limit their dissemination and further exploitation. Research on the French language may be here overly dependent on the current data policy of content holders, and hinders the development of studies on opinion mining, fake news and hate speech detection, fact checking, on biases and ethical issues, to name a few. Two issues are at stake (a) to secure the access to sensible data for research purposes; (b) to facilitate the dissemination of publicly produced databases and models.

This observed lack of coordination finally hints at the need to define a strategic roadmap for identifying, building, curating, annotating and securing resources for 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 scenarios where resources could be transferred from their English equivalent, and make sure that the necessary high-quality translation technologies are widely available as a public commodity at scale.

Note that the same argument holds for other languages for which it might be proper to transfer resources *from French* – this is for instance the case of many low-resource languages, such as regional languages, or languages that have historically co-existed with French in various areas of the world. Collecting resources and developing MT to translate these languages from and into French is thus likely to result in large pay-offs for both sides.

Finally, as data continues to increase in scale, it is also essential to continue to support public computing infrastructures, with generous access rules, for the research and industries (startups, SMEs).

Regarding evaluation

With LTs being embedded in a growing number of applications that are routinely used by the general public, in an increasing varied number of tasks, there is a growing need to openly evaluate and publicize the performance of existing systems in real world conditions, and to diagnose their potential biases and better document their potential defects and harmful impacts. Actions to ensure that evaluation campaigns that specifically target French for a suf-

[129] https://elrc-share.eu
[130] https://www.inist.fr/services/analyser/istex-textes-corpus/
[131] https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html

ficiently large number of applications and domains are organized on a regular basis should be undertaken, whenever possible in coordination with international evaluation campaigns so as to increase their visibility and participation. Given that baseline, "generic" systems and evaluation protocols exist for many tasks, it is felt that running such evaluations could be performed at a reduced cost, requiring mostly the creation of novel test data for realistic use cases.

A potential source of inspiration might be the systematic evaluations of LTs for French undertaken under the aegis of Aupelf/AUF ("*Agence Universitaire de la Francophonie*", the Association of Francophone Universities[132]) in the 90s or within the TechnoLangue programme in the early 2000s.[133] These actions have been successful in consolidating evaluation procedures and fostering the creation of annotated evaluation data; fifteen years later, the need for benchmarks that could help better analyse and diagnose the biases and limits of current LTs is no less pressing.

Other items on the research agenda

As previously noted, the development of LTs dedicated to the processing of French and French Sign language still requires stimulation for the development of fundamental research and of new resources. In addition to the themes already evoked above, one may notably cite: (a) deep language analysis algorithms and technologies (including large scale treebanks with semantic and discourse-level information for a multiplicity of genres, domains and tasks), with the objective to achieve deep language understanding for some representative applications; one example could be the development of collaborative agents, capable of social interactions through language but also equipped with learning, reasoning, and problem-solving abilities; (b) multimodal resources for the study of the unsupervised emergence of language and language development though interactions and grounding; (c) resources and tools for the study of pathological language processing. Much of this research is pluridisciplinary in essence, and should be conducted with the relevant research communities.

Acknowledgements

The authors wish to thank I. Aldabe, F. Béchet, B. Daille, K. Choukri, T. François, T. Hueber, J. Mariani, P. Müller, B. Sagot, V. Vandeghinste for comments and suggestions on an early draft of this report. They also acknowledge the help of M. Wang-Castejon and L. Khennou who have participated to the collection of some of the data sources and documentation used in this survey. They finally warmly thank V. Arranz (ELRA) for her contribution to the inventory of linguistic resources.

References

Rodrigo Agerri, Eneko Agirre, Itziar Aldabe, Nora Aranberri, Jose Maria Arriola, Aitziber Atutxa, Gorka Azkune, Arantza Casillas, Ainara Estarrona, Aritz Farwell, Iakes Goenaga, Josu Goikoetxea, Koldo Gojenola, Inma Hernaez, Mikel Iruskieta, Gorka Labaka, Oier Lopez de Lacalle, Eva Navas, Maite Oronoz, Arantxa Otegi, Alicia Pérez, Olatz Perez de Viñaspre, German Rigau, Jon Sanchez, Ibon Saratxaga, and Aitor Soroa. European Language Equality D1.2: Report on the state of the art in Language Technology and Language-centric AI, September 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf.

[132] https://www.auf.org
[133] http://www.technolangue.net

Itziar Aldabe, Georg Rehm, German Rigau, and Andy Way. European Language Equality D3.1: Report on existing strategic documents and projects in LT/AI, November 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Paul Boersma and David Weenink. Praat: doing phonetics by computer (version 5.1.13), 2009. URL http://www.praat.org.

Hélène Bonneau-Maynard, Alexandre Denis, Frédéric Béchet, Laurence Devillers, Fabrice Lefèvre, Matthieu Quignard, Sophie Rosset, and Jeanne Villaneau. Media : évaluation de la compréhension dans les systèmes de dialogue. In Stéphane Chaudiron and Khalid Choukri, editors, *L'évaluation des technologies de traitement de la langue, les campagnes Technolangue*, Cognition et traitement de l'information, pages 209–232. Hermès, Lavoisier, 2008. URL https://hal.archives-ouvertes.fr/hal-00337343.

Andrew D. Booth and William N. Locke, editors. *Machine translation of languages: fourteen essays*. Press of the Massachusetts Institute of Technology, Cambridge, Massachusetts, 1955.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Yannick Parmentier, Caroline Pasquer, and Jean-Yves Antoine. Annotation d'expressions polylexicales verbales en français. In Jean-Yves Antoine Iris Eshkol, editor, *24e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Actes de TALN, volume 2 : articles courts, pages 1–9, Orléans, France, June 2017. URL https://hal.archives-ouvertes.fr/hal-01537880.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. An annotated corpus for sexism detection in French tweets. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1397–1403, 2020.

Noam Chomsky. *Syntactic structures*. The Hague: Mouton, 1957.

Collectif. *La langue française dans le monde*. Gallimard - Organisation internationale de la Francophonie, 2019. URL http://observatoire.francophonie.org/wp-content/uploads/2021/04/LFDM-20Edition-2019-La-langue-fran%C3%A7aise-dans-le-monde.pdf.

Marcel Cori and Jacqueline Léon. La constitution du TAL. *Revue TAL*, 43(3):21–55, 2002. URL https://halshs.archives-ouvertes.fr/halshs-00158854.

Joaquim Brandão de Carvalho. Western romance in lenition and fortition. In Joaquim Brandão de Carvalho, Tobias Scheer, and Philippe Ségéral, editors, *Lenition and Fortition*. De Gruyter Mouton, 2008. ISBN 9783110211443. doi: doi:10.1515/9783110211443. URL https://doi.org/10.1515/9783110211443.

Eric Villemonte de la Clergerie, Olivier Hamon, Djamel Mostefa, Christelle Ayache, Patrick Paroubek, and Anne Vilnat. PASSAGE: from French parser evaluation to large sized treebank. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). ISBN 2-9517408-4-0. http://www.lrec-conf.org/proceedings/lrec2008/.

Martin d'Hoffschmidt, Maxime Vidal, Wacim Belblidia, and Tom Brendlé. Fquad: French question answering dataset. *CoRR*, abs/2002.06071, 2020. URL https://arxiv.org/abs/2002.06071.

Yoann Dupont. Un corpus libre, évolutif et versionné en entités nommées du français. TALN 2019 - Traitement Automatique des Langues Naturelles, July 2019. URL https://hal.archives-ouvertes.fr/hal-02448590. Poster.

Moussa Kamal Eddine, Antoine J.-P. Tixier, and Michalis Vazirgiannis. BARThez: a skilled pretrained French sequence-to-sequence model. *arXiv preprint arXiv:2010.12321*, 2020.

Karën Fort, Bruno Guillaume, and Hadrien Chastant. Creating Zombilingo, a game with a purpose for dependency syntax annotation. In *Proceedings of the First International Workshop on Gamification for Information Retrieval*, GamifIR '14, page 2–6, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450328920. doi: 10.1145/2594776.2594777. URL https://doi.org/10.1145/2594776.2594777.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. The Ester II evaluation campaign for the rich transcription of French radio broadcasts. In *Proceedings of InterSpeech*, 2009.

Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel, and Olivier Galibert. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *LREC - Eighth international conference on Language Resources and Evaluation*, page na, Turkey, 2012. URL https://hal.archives-ouvertes.fr/hal-00712591.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *15th European Chapter of the Association for Computational Linguistics (EACL 2017)*, volume 1 - long pap, pages 262–272, Valencia, ES, 2017. Association for Computational Linguistics (ACL). URL https://oatao.univ-toulouse.fr/18921/. Thanks to Association for Computational Linguistics (ACL). This papers appears in volume 1 of Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics ISBN 978-1-945626-34-0 The definitive version is available at: http://eacl2017.org/images/site/Proceeding/book_long.pdf.

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales, and Jacopo Staiano. Project PIAF: Building a native French question-answering dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5481–5490, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.673.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://www.aclweb.org/anthology/D18-2012.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.302.

Christian Lequesne. Diversité linguistique et langue française en Europe. report of the task force *Diversité linguistique et langue française dans les institutions européennes*, October 2021. Presented to Clément Beaune, Secretary of State for European Affairs, and Jean-Baptiste Lemoyne, Secretary of State for Tourism, for the French Presidency of the European Union, Secretary of State for Tourism, French citizens living abroad and the French-speaking world.

Mark Liberman. Corpus phonetics. *Annual Review of Linguistics*, 5(1):91–107, 2019. doi: doi:10.1146/annurev-linguistics-011516-033830.

Joseph Mariani, Patrick Paroubek, Gil Francopoulo, Aurélien Max, François Yvon, and Pierre Zweigenbaum. *The French language in the digital age / La Langue française à l'ère du numérique*. Springer Verlag, Berlin, 2012. URL https://link.springer.com/book/10.1007/978-3-642-30761-4.

Joseph Mariani, Patrick Paroubek, Gil Francopoulo, and Olivier Hamon. Rediscovering 15 years of discoveries in language resources and evaluation: The LREC anthology analysis. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph

Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a Tasty French Language Model. In *ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*, Seattle / Virtual, United States, July 2020. doi: 10.18653/v1/2020.acl-main.645. URL https://hal.inria.fr/hal-02889805.

Fiammetta Namer. *Morphologie, Lexique et Traitement Automatique des Langues: l'analyseur DériF*. TIC et sciences cognitives. Hermès-Lavoisier, 2009. URL https://hal.archives-ouvertes.fr/hal-00413337. ISBN 978-2-7462-2363-9.

Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary, and Benoît Sagot. Establishing a new state-of-the-art for French named entity recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4631–4638, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.569.

Dijana Petrovska-Delacrétaz, Sylvie Lelandais, Joseph Colineau, Liming Chen, Bernadette Dorizzi, Emine Krichen, Mohamed Anouar Mellakh, Anis Chaari, Souhila Guerfi, Moshen Ardabilian, Johan D'Hose, and Boulbaba Ben Amor. The IV2 Multimodal Biometric Database (Including Iris, 2D, 3D, Stereoscopic and Talking Face Data) and the IV2-2007 Evaluation Campaign. In *2nd IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS 2008)*, page (elec. proc), Crystal City, Washington DC, United States, September 2008. doi: 10.1109/BTAS.2008.4699323. URL https://hal.archives-ouvertes.fr/hal-00765334.

Rosalind W Picard. *Affective computing*. 2000.

Daniel Pimienta. étude sur la présence de la langue française dans le cyberspace. Technical Report Rapport final #2, MAAYA - Réseau mondial pour la diversité linguistique, 2017. URL http://observatoire.francophonie.org/wp-content/uploads/2019/04/2018-Place-francais-sur-Internet-Pimienta-MAAYA.pdf.

Daniel Pimienta. La place du français sur internet. In *La langue française dans le monde 2022*. OIF/Gallimard, 2022. to be published.

Barbara Plank, Anders Johannsen, and Anders Søgaard. Importance weighting and unsupervised domain adaptation of POS taggers: a negative result. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–973, Doha, Qatar, October 2014. Association for Computational Linguistics.

Karl Pusch. The Romance languages: Typology. In Bernd Kortmann and Johan van der Auwera, editors, *The Languages and Linguistics of Europe, A Comprehensive Guide*, volume 1 of *The World of Linguistics*, pages 69–96. Berlin/New York: Mouton de Gruyter, 2011.

Georg Rehm and Hans Uszkoreit, editors. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages, Heidelberg etc., 2012. Springer.

Benoît Sagot. The Lefff, a freely available, accurate and large-coverage lexicon for French. In *Proceedings of the 7th Language Resource and Evaluation Conference*, La Valette, Malte, 2010.

Benoît Sagot and Darja Fišer. Building a free French wordnet from multilingual resources. In *OntoLex*, Marrakech, Morocco, May 2008. URL https://hal.inria.fr/inria-00614708.

Benoît Sagot, Marion Richard, and Rosa Stern. Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées. In Georges Antoniadis, Hervé Blanchon, and Gilles Sérasset, editors, *Traitement Automatique des Langues Naturelles (TALN)*, volume 2 - TALN of *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, Grenoble, France, June 2012. URL https://hal.inria.fr/hal-00703108.

Tanja Schultz, Michael Wand, Thomas Hueber, Dean J. Krusienski, Christian Herff, and Jonathan S. Brumberg. Biosignal-based spoken communication: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2257–2271, 2017. doi: 10.1109/TASLP.2017.2752365.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.647. URL https://aclanthology.org/2020.emnlp-main.647.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. doi: 10.18653/v1/P16-1162. URL https://www.aclweb.org/anthology/P16-1162.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *Proc. of ICASSP*, pages 4779–4783, 2018.

John Charles Smith. French and northern Gallo-Romance. In Adam Ledgeway and Martin Maiden, editors, *The Oxford Guide to the Romance Languages*. Oxford University Press, 2016.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433.

Henriette Walter. *Le français dans tous les sens*. Points - Le goût des mots. Seuil, 2016.

Rodrigo Wilkens, Bruno Oberle, Frédéric Landragin, and Amalia Todirascu. French coreference for spoken and written language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 80–89, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.10.

Yorick Wilks. The history of natural language processing and machine translation. *Encyclopedia of Language and Linguistics*, 2005.

Victor H Yngve. A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466, 1960.

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud, and Sonia Badene. The DISRPT 2021 shared task on elementary discourse unit segmentation, connective detection, and relation classification. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 1–12, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.disrpt-1.1.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/K18-2001.