État de l'art des technologies linguistiques pour la langue française

Gilles Adda, Annelies Braffort, Ioana Vasilescu, François Yvon, Jean-François Nominé

État de l'art des technologies linguistiques pour la langue française*

Gilles Adda Annelies Braffort Ioana Vasilescu

François Yvon

Avril 2022

Université Paris-Saclay, CNRS, LISN

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'études de l'état de l'art préparées dans le cadre du projet européen "European Language Equality", qui vise à établir une feuille de route pour parvenir à une "égalité linguistique" en Europe 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 présentent le contexte général de l'étude ; la première en présentant un bref état chiffré de la langue

*Ce document est la traduction en français d'un rapport préparé dans le cadre du projet européen *European Language Equality*, et qui correspond au livrable D-1.14. Cette traduction a été préparée par Jean-François Nominé (INIST-CNRS) et validée par les auteurs. La version officielle, en anglais, et les rapports équivalents pour les autres langues européennes, sont disponibles depuis sur le site du projet: https://european-language-equality.eu/deliverables/.

française et de la LSF à l'ère du numérique ; la seconde en introduisant très sommairement le domaine du traitement automatique des langues et les 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 (p. ex. : 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 aujourd'hui disponibles ainsi qu'à la disponibilité des ressources linguistiques associées. Une analyse comparative chiffrée de la situation des différentes langues européennes complète cette analyse. Dans la section 7, nous récapitulons 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'entraîner 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-exploités ; (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 aujourd'hui disponibles (e) soutenir les recherches interdiscipinaires sur les technologies linguistiques pour aller vers une compréhension profonde des langues, et accélérer leur diffusion dans tous les secteurs de la recherche et de l'industrie.

1 Introduction

La présente étude fait partie d'une série qui restitue les résultats d'une enquête sur le niveau de prise en charge technologique apportée aux langues de l'Europe. Elle s'adresse aux décideurs aux niveaux européen, national ou régional, aux communautés linguistiques, aux journalistes, etc. et se propose non seulement de brosser un état des lieux actualisé pour chacune des langues abordées par ces études mais aussi - et surtout - d'identifier les lacunes et les facteurs qui empêchent que la recherche et les technologies se développent davantage. L'identification de ces manques fondera une proposition complète inspirée par des données probantes et assortie des mesure nécessaires pour concrétiser l'égalité numérique entre langues en Europe d'ici à 2030.

À cette fin, plus de 40 partenaires scientifiques, experts dans une trentaine de langues européennes, ont suivi une procédure à la fois considérable et exhaustive de recueil de données qui a permis de dresser une cartographie détaillée, empirique et dynamique de la prise en charge technologique pour nos langues[2].

Ce rapport a été élaboré dans le cadre du projet pour l'égalité numérique des langues en Europe (ELE)[3].

2 La langue française et la langue des signes française à l'ère numérique

2.1 La langue française à l'ère numérique

Faits linguistiques sur le français

Le français est une langue romane occidentale à l'instar d'autres qui tirent leur origine du latin comme les dialectes italiens septentrionaux, l'espagnol et le portugais, avec lesquelles le français partage de nombreux traits linguistiques en raison de cette filiation commune et une longue histoire de contacts (connue sous le vocable d'"alliance aréale" des langues romanes) (de Carvalho, 2008; Walter, 2016; Pusch, 2011). Le français a hérité des traits des langues gauloises à partir des dialectes celtiques parlés par les groupes ethniques qui peuplaient par le passé le territoire conquis par les Romains. La cohabitation du latin vulgaire avec les dialectes de la Gaule a produit un nouvel idiome, la langue gallo-romaine, par la suite influencée par des dialectes germaniques en conséquence des invasions qui ont marqué la chute de l'Empire romain.

Comme toutes les langues romanes, le français a conservé l'alphabet latin. Du point de vue de ses traits typologiques (c'est-à-dire morphologiques, syntaxiques et en lien avec d'autres modèles synchroniques à orientation de contenu), le français est la langue romane qui s'est le plus écartée du latin et, en ce sens, passe pour la plus innovante de cette famille. Ainsi, elle possède le système vocalique le plus complexe avec ses 12 voyelles orales et 4 voyelles nasales ; la position de l'accent y est fixe, à la fin du mot phonologique. Concernant les traits morphologiques, la perte des terminaisons latines a entraîné une réorganisation spécifique et générale des systèmes nominal et verbal. Le français compte des caractéristiques communes avec les autres langues romanes : c'est,

par exemple, une langue de type nominatif-accusatif (SVO), cette distinction se codant par l'ordre des mots. Il s'agit aussi d'une langue à articles, qui partage avec d'autres idiomes romans le processus de grammaticalisation qui aboutit à l'apparition d'articles définis et indéfinis (Pusch, 2011; Smith, 2016).

Le français, langue de la "Francophonie"

Avec 128 millions de "locuteurs natifs et réels" dans le monde et une estimation de près de 300 millions de personnes pratiquantes du français entre autres langues à l'oral (Collectif, 2019), le français apparaît non seulement comme la 16e langue maternelle la plus utilisée dans le monde, mais aussi comme la 6e la plus parlée à cette échelle, après l'anglais, le chinois mandarin, l'espagnol, le hindi et le russe.[4]

En Europe, on estime que 129 millions d'humains parlent le français, ce qui le place au 3ème rang après l'anglais et l'allemand. Parmi les langues officielles, cette langue est la deuxième après l'anglais et on la trouve dans près de 30 pays dans le monde, plus particulièrement en Europe (65 millions de locuteurs en France, 7 millions en Belgique, 3 millions en Suisse et au Luxembourg), en Afrique, au Canada et en Haïti. Tous les pays où la langue française se parle constituent ce que l'on appelle "*La Francophonie*", avec l'"*Organisation Internationale de la Francophonie*" (OIF) qui assure la coordination des politiques multilatérales entre ses 88 États membres.

"*Le Conseil supérieur de la langue française*" (CSLF) est le nom attribué aux organismes nationaux de plusieurs pays francophones en charge de conseiller leurs gouvernements sur les questions liées à l'utilisation de la langue française. Une telle entité a existé en France et au Québec (pour l'application de la Charte de la langue française), et reste en activité en Belgique en tant qu'institution de la "*Fédération de Wallonie-Bruxelles*". En France et au Québec, les dossiers sur ces questions ont été transférés à leurs ministères respectifs, en particulier en France, à la "*Délégation générale à la Langue Française et aux Langues de France*" (DGLFLF) sous la tutelle du ministère de la Culture et de la Communication.[5] Celle-ci a pour mission d'élaborer les politiques relatives aux langues en relation avec tous les ministères, que ce soit pour la langue française et pour la variété des 80 langues parlées dans le pays. La DGLFLF a organisé les "*États-Généraux du Multilinguisme*" en 2008 et les "*États-Généraux du Multilinguisme en Outre-Mer*" en 2011 et en 2021. La France a toujours défendu la langue française avec vigueur sur la scène internationale, en tant que telle (celle-ci était avant le milieu du XXe siècle la langue prépondérante dans la diplomatie), ou dans le cadre du multilinguisme[6].

La constitution française stipule que la langue de la République française est le français. Les informations au consommateur et les publicités doivent être en français ou comporter une traduction française et tous les participants d'un débat scientifique sont en droit de s'exprimer dans cette langue. Les salariés doivent être libres de l'employer et ils doivent pouvoir utiliser des programmes de bureautique en français dans n'importe quelle entreprise. Tous les services audiovisuels qui émettent en France sont dans l'obligation d'utiliser le français. Les chaînes radios doivent respecter un quota de contenus dans cette langue, alors que les télévisions peuvent diffuser intégralement dans une langue étrangère. Seuls les sites web officiels sont dans l'obligation de publier en français

sur Internet. En même temps, la législation entend favoriser le plurilinguisme : quand une administration traduit des informations destinées au public, cette opération doit se faire dans deux langues étrangères, et la loi vise également l'enseignement de deux autres langues que le français dans les cursus scolaires.

L'*Académie Française* a été créée en 1635 pour être l'institution du pays prépondérante à traiter les questions portant sur la langue française, y compris la tenue d'un dictionnaire de référence. L'*Académie royale de langue et de littérature françaises* de Belgique, fondée en 1920, a également pour mission d'étudier et de promouvoir la langue française. Bien que leurs travaux n'aient pas d'impact sur l'usage du français dans le monde réel, ces institutions jouent un rôle actif dans la supervision des néologismes, dans le cadre de la "*Commission d'enrichissement de la langue française*"[7]. La *Fondation Alliance Française* représente une autre institution exerçant un rôle essentiel, puisque sa mission est de promouvoir la langue et la culture française hors de France, avec près de 800 représentations *Alliances Françaises* (contre 1000 en 2011) et 500 000 étudiants dans 132 pays partout dans le monde[8].

La Langue française sur Internet

Dans son édition de 2018, *The French Language in the World* (Pimienta, 2017; Collectif, 2019) faisait le constat que la langue française occupait la quatrième place sur Internet, derrière l'anglais, le chinois et l'espagnol, avec une confortable avance sur les langues suivantes : l'allemand, le portugais, le japonais, le russe, le hindi et l'arabe. Dans son édition de 2022, cet ouvrage établit que malgré son maintien à la quatrième place sur ce réseau :

- Elle se trouve côtoyée si ce n'est devancée par l'hindi qui fait montre d'un accroissement spectaculaire ;
- Son avance sur les autres langues situées plus bas dans le tableau (désormais le portugais, le russe, l'arabe, l'allemand, le japonais et le malais) s'est considérablement rétrécie, en raison de la conjonction des deux faits suivants : 1) les taux de connexion des francophones dans les pays industrialisés frisent la saturation (85% en moyenne), ce qui laisse peu de marge de progression, et 2) le fossé numérique dans les pays d'Afrique francophone est beaucoup plus lent à se combler que la croissance moyenne de la connectivité dans le monde.

À examiner les répartitions des documents par types et par sujets, l'avantage détenu par le français sur Internet se trouve du côté des livres, des MOOC et de la recherche (élargie à tous les sujets de la rubrique "science et technologie").

W3Techs[9] offre une image légèrement différente, avec une très forte montée de l'anglais depuis 2018 (de 51% à 63%) et un recul du français de 4,1 à 2,5% du nombre total de pages web (actuellement classé en 6è position et dépassé en particulier par l'espagnol).

Le français en tant que langue internationale

Le français est l'une des 24 langues officielles de l'UE et une des trois langues de travail de la Commission européenne, avec l'anglais et l'allemand. On a constaté une évolution défavorable dans l'utilisation du français comme langue de travail au sein des différentes institutions européennes. Bien plus qu'un déclin du français, cette tendance révèle un recul généralisé du multilinguisme (Lequesne, 2021).

Pour la rédaction de documents sources, sur les 69 000 documents produits par le Secrétariat général du Conseil en 2018, 1215 (2 %) l'ont été à l'origine en français. Par contre, 65 908 l'ont été au départ en anglais, soit 95 % du total de ces pièces. Les 3,1% restants représentent toutes les autres langues officielles de l'Union. On constate aussi la domination de l'anglais dans les documents sources publiés par la Commission. En 2019, 3,7% de ceux qu'elle avait envoyés en traduction avaient le français pour langue source, par rapport à 85,5% pour l'anglais. Vingt ans plus tôt, en 1999, 34% de ces documents étaient en français. Concernant les documents traités par les services des commissions du Parlement européen en 2019, seuls 11,7 % d'entre eux avaient le français comme langue source. Quant au SEAE (Service européen pour l'action extérieure), en 2019, seul 0,9% des documents que ses services ont fait traduire a été rédigé en français, contre 98,7% en anglais.

Le français fait aussi partie des langues de travail de l'OCDE (Organisation de coopération et de développement économiques), aux Nations unies (y compris à l'UNESCO, et l'OIT (Organisation internationale du travail), conjointement à l'anglais, l'espagnol, le russe, le chinois mandarin et l'arabe), il appartient aux trois langues utilisées lors des Jeux olympiques, avec l'anglais et la langue du pays organisateur, il figure avec l'anglais et l'allemand parmi les trois langues officielles de l'Office européen des brevets (OEB), et les quatre langues de travail de l'Union africaine, en compagnie de l'arabe, de l'anglais et du portugais.

Autre signe de la vitalité relative du français : le nombre de traductions dans le monde tel que mesuré par l'UNESCO. Le français est classé comme étant la deuxième langue source (loin derrière l'anglais) et la troisième langue cible après l'allemand et l'espagnol.[10]

Cela peut s'interpréter comme le fait que la production de biens intellectuels en français est importante et suscite l'intérêt des non-francophones, et qu'elle couvre déjà une quantité relativement grande des besoins des locuteurs francophones.

2.2 La langue des signes française à l'ère numérique

Les langues des signes - généralités

Les langues des signes (LS) sont des langues naturelles pratiquées au sein des communautés de Sourds. Si le terme "sourd" se rapporte à un statut auditif, le vocable "Sourd" employé avec un "S" majuscule désigne une identité culturelle propre aux personnes ayant une surdité qui détiennent une culture commune et qui possèdent usuellement une LS en partage qui est leur langue première.

Les LS sont des langues visuo-gestuelles grâce auxquelles une personne va s'exprimer à l'aide de nombreuses composantes corporelles (les mains et les bras, mais aussi à l'aide

d'expressions faciales, du regard, le buste, etc.) et son interlocuteur va percevoir son message par le canal visuel. Le système linguistique de la LS exploite ces canaux spécifiques : une quantité importante d'informations se trouve exprimée simultanément, s'organise dans l'espace, et l'iconicité y joue un rôle essentiel.

Les langues des signes ne sont pas universelles. Tout comme les langues parlées qui possèdent de multiples formes, des dialectes et des variations locales, il en va de même pour les LS bien qu'elles conservent certaines similitudes entre elles sur le plan grammatical. L'édition 2021 de l'Ethnologue[11] répertorie 150 LS, alors que l'Atlas of Sign Languages de SIGN HUB en répertorie plus de 200 et relève qu'il en existe d'autres qui n'ont pas encore été documentées ni identifiées.

Il existe également des langues des signes tactiles utilisées par les personnes avec surdicécité. Elles diffèrent significativement des LS visuelles en ce que des éléments comme l'expression faciale devront être remplacés par d'autres informations tactiles.

Une étude publiée en 2020 dans la revue Royal Society Open Science sur la diversité des LS et leurs processus évolutionnaires[12] montre que, dans l'échantillon étudié (qui ne prend pas en compte toutes les LS dans le monde), on est en présence de six grandes lignées européennes, avec trois groupes plus importants d'origines autrichienne, britannique et française, et trois groupes plus limités autour du russe, de l'espagnol et du suédois. Certaines LS actuellement utilisées semblent indépendantes de ces groupes, comme dans le cas norvégien. Les LS jouissent d'un statut qui diffère d'un pays à l'autre en Europe. Certains États ne reconnaissent pas leur LS, d'autres le font dans leur constitution, alors que d'autres encore en tout ou partie dans leur législation.

Comme toute autre langue parlée, les LS sont vulnérables et exposées au risque d'être mises en danger. Cela peut être le cas d'une LS utilisée par une petite communauté, qui peut même se voir abandonnée par ses utilisateurs qui adoptent une LS employée par une collectivité plus large. Même des LS reconnues nationalement courent ce risque en raison d'une augmentation de la pose précoce d'implants cochléaires, et l'encouragement de la part des médecins auprès des parents - qui sont généralement entendants - à favoriser la communication orale.

À ce jour, les LS ne possèdent pas de système d'écriture normalisé. Il en résulte qu'elles sont présentes sur Internet et les réseaux sociaux essentiellement sous forme de vidéos. Cette disponibilité sur ces supports varie grandement entre pays et, même dans le cas de ceux où elle est parmi les plus tangibles, celle-ci reste rare.

La Langue des Signes Française

La langue des signes française (LSF) est la langue signée pratiquée en France. Toutes les langues du monde sont sujettes à des variations sociolinguistiques, et la LSF ne fait pas exception : il n'existe pas une seule façon de signer, mais plusieurs qui dépendent des régions, villes, villages, écoles, des histoires et cultures familiales, et le lexique compte des variantes régionales significatives. En plus de leur parler local, la plupart des Sourds connaissent la LSF qui correspond peu ou prou au dialecte parisien, et s'adaptent sans difficulté à n'importe quel interlocuteur qui le pratique.

La LSF a été reconnue dans le cadre de la loi n°2005-102 du 11 février sur "l'égalité

des droits et des chances, la participation et la citoyenneté des personnes handicapées". Ainsi, l'article L. 312-9-1 énonce que : "La langue des signes française est reconnue comme une langue à part entière. Tout élève concerné doit pouvoir recevoir un enseignement de la langue des signes française. Le"*Conseil supérieur de l'éducation*" veille à favoriser son enseignement. Il est tenu régulièrement informé des conditions de son évaluation. [La LSF] peut être choisie comme épreuve optionnelle aux examens et concours, y compris ceux de la formation professionnelle. Sa diffusion dans l'administration est facilitée."

La LSF est pratiquée en France et dans la partie francophone de la Suisse. Les effectifs de signeurs en LSF (les personnes qui pratiquent la LSF), qui peuvent être des Sourds comme des entendants (par ex., les enfants, les membres de la famille et les proches de personnes Sourdes) ne sont pas connus avec précision. On cite le chiffre d'environ 169 000 personnes dans le monde, dont à peu près 100 000 en France en 2014.[13]

Dans l'éducation des jeunes Sourds, l'Article L. 112-3 du Code de l'éducation[14] pose le principe de la liberté de choix entre :

- une communication bilingue, en langue des signes française (LSF), et en langue française écrite,
- une communication en français écrit et oral, avec ou sans le recours à la langue française parlée complétée (LfPC)[15] ou à la langue des signes française (LSF).

Un centre d'enseignement, dans une aire géographique donnée, les ressources nécessaires pour soutenir les jeunes sourds scolarisés de l'école maternelle jusqu'au lycée, quel que soit leur projet linguistique. Pour un parcours bilingue, la LSF est la langue première de l'élève, elle est sa langue d'enseignement, mais aussi une langue enseignée. Les jeunes sourds scolarisés suivent des cours en LSF et apprennent le français progressivement, essentiellement par l'intermédiaire de l'écrit et à l'aide de la LSF. Tout au long de leur scolarité, les jeunes sont appelés à approfondir leur maîtrise de la LSF tout en intégrant par étapes les éléments de la culture Sourde. Dans la pratique, il semble que ce choix n'est pas simple parce qu'il n'existe pas toujours de structures adéquates dans toutes les régions de France.

La présence de la LSF dans les médias a augmenté ces dernières années, notamment depuis l'adoption de la loi de 2005. Cependant, les sites web entièrement bilingues restent extrêmement rares. Les deux principaux sont "*Media'Pi !*"[16], un média en ligne bilingue généraliste créé par des journalistes Sourds, et "*L'œil et la main*",[17] une émission de télévision bilingue de type documentaire développée et produite par une équipe mixte (Sourds et entendants).

3 Les technologies de la langue : qu'est-ce que c'est ?

La langue naturelle[18] est le moyen le plus courant et polyvalent pour les humains de transmettre des informations. Nous utilisons la langue, notre moyen de communication naturel, pour encoder, stocker, transmettre, partager et traiter des informations. Le traitement automatique des langues n'est pas une tâche triviale car celles-ci sont intrinsèquement complexes : les énoncés linguistiques sont souvent sujets à de multiples interprétations

(ambiguïté), et leur décodage nécessite des connaissances sur le contexte et le monde, et en parallèle il est possible d'user de représentations différentes pour dénoter le même sens (variation).

Le traitement informatique des langues humaines s'est établi en tant que domaine scientifique spécialisé connu sous les vocables de *linguistique informatique* (LI), de *traitement automatique des langues* (TAL), de *traitement automatique du langage parlé* (TALP) ou, plus généralement, de technologies de la langue (TL). Malgré des différences sur le plan des objectifs et de l'orientation, puisque la LI est davantage influencée par la linguistique, le TAL par l'informatique et le TALP par le traitement du signal, le terme TL est plus neutre. En fait, cet ensemble est largement multidisciplinaire par nature ; il associe la linguistique, l'informatique (et notamment l'intelligence artificielle (IA)), le traitement du signal, les mathématiques et la psychologie, entre autres. Dans la pratique, ces communautés travaillent en étroite collaboration, combinant des méthodes et des approches inspirées par chacune d'entre elles, pour constituer ensemble l'*IA axée sur le langage*.

On pourrait tenter la définition concise suivante : les technologies de la langue sont un domaine scientifique et technologique multidisciplinaire qui s'intéresse à l'étude et au développement de systèmes informatiques capables de traiter, d'analyser, de produire et de comprendre les langues humaines, qu'elles soient écrites, parlées ou corporalisées.

L'histoire des TL commence dans les années 1950 avec la description célèbre faite par Turing d'une machine intelligente (Turing, 1950), les premières tentatives de traduction automatique (Booth et Locke, 1955) ou de description formelle des structures grammaticales par des linguistes (Chomsky, 1957) et des informaticiens (Yngve, 1960). En France, les premiers travaux de recherche sont initiés à peu près à la même période (Cori et Léon, 2002). Elle a traversé par la suite une histoire faite de hauts et de bas (Wilks, 2005), les périodes d'espoirs (excessifs) alternant avec des phases de désillusions. Les années 1990 ont vu un changement méthodologique majeur, à une époque marquée par d'intenses efforts pour créer des ressources linguistiques à large couverture. Des corpus annotés, des thésaurus, etc. ont ainsi été étiquetés manuellement pour décrire des phénomènes linguistiques et pour être utilisés afin de parvenir à des règles lisibles par machine spécifiant la façon dont des éléments de langue peuvent être analysés et/ou générés automatiquement. Peu à peu, grâce à l'évolution de l'apprentissage automatique et ses progrès, les moteurs de règles ont été remplacés par des technologies basées sur des données et la mise au point de systèmes qui apprennent implicitement à partir d'exemples. Lors de la dernière décennie, dans les années 2010, l'apprentissage automatique a massivement investi l'utilisation de réseaux neuronaux multicouches capables de résoudre différents problèmes liés à la classification de textes, puis à celui de l'étiquetage séquentiel. La réussite de cette approche réside dans la capacité des réseaux neuronaux à apprendre des représentations vectorielles continues s'appliquant à des unités linguistiques (autrement dit, des plongements de mots ou de phrases) en utilisant des quantités considérables de données non étiquetées. À partir de là, il est possible d'entraîner des outils de traitement efficaces en se contentant d'utiliser un volume réduit de données étiquetées dans une phase de mise au point finale. Ces techniques ont suscité un intérêt considérable et continuent de progresser à un rythme soutenu, comme en témoigne le développement de représentations d'espaces continus contextualisés et multilingues.

Ces dernières années, la communauté des TL a vu l'émergence de techniques et d'outils d'apprentissage profond de plus en plus puissants, qui transforment la façon dont les tâches sont abordées dans ces domaines. On évolue progressivement d'une méthodologie où la mise en œuvre typique des solutions de TL reposait sur un enchaînement de multiples modules vers des architectures basées sur des réseaux neuronaux complexes intégralement entraînés à l'aide de très grandes quantités de données, aussi bien textuelles, audio que multimodales. Autre tendance massive : l'utilisation de modèles multilingues et multimodaux qui permettent un apprentissage par transfert entre tâches, langues et modalités. Si des succès dans ces domaines de l'IA ont été possibles, c'est grâce à la conjonction de quatre évolutions scientifiques différentes : 1) la maturation des algorithmes et des technologies des réseaux neuronaux profonds, 2) les masses de données (et pour les TL, des données multilingues diversifiées en grande quantité), 3) l'augmentation de la puissance de calcul haute-performance (HPC) grâce à l'utilisation massive de cartes GPU et 4) la mise en œuvre d'approches d'auto-apprentissage tout à la fois simples et efficaces.

Les TL tentent d'offrir des solutions principalement pour les domaines d'application suivants :

- L'**analyse de texte** qui vise à identifier et à étiqueter les informations linguistiques sous-jacentes à tout énoncé en langue naturelle. Cela concerne aussi bien la reconnaissance des limites de mots, de locutions, de phrases et de sections, l'identification des caractéristiques morphologiques des mots, celle des rôles syntaxiques et sémantiques, ainsi que l'acquisition des relations entre constituants du texte.

- Le **traitement de la parole** permet aux humains de communiquer avec des appareils électroniques par la voix. Parmi les principaux domaines de la technologie appliquées à la parole, citons la synthèse vocale, c'est-à-dire la génération de séquences parlées à partir d'un extrait de texte ; la transcription automatique d'énoncés oraux, autrement dit, la conversion d'un signal vocal en texte ; les systèmes de dialogue parlé, capables d'accomplir des tâches à partir d'une interaction orale, de reconnaître le locuteur et, plus généralement, le caractériser, et qui visent à déduire des informations à son sujet à partir d'enregistrements, comme des données démographiques élémentaires, son état émotionnel au moment de l'énonciation, etc.

- La **traduction automatique**, c'est-à-dire la traduction par ordinateur d'une langue naturelle dans une autre : celle-ci englobe la traduction d'énoncés écrits, parlés comme signés. L'interprétation automatique doit également intégrer des problèmes de traitement en temps réel et tenir compte du contexte extra-linguistique, comme les émotions ou les intentions du locuteur ;

- **L'extraction et la recherche d'informations**, dont la finalité est d'extraire des informations structurées à partir de documents non structurés, à retrouver des renseignements pertinents dans de vastes corpus d'une documentation hétérogène, comme Internet, et à fournir les documents ou des fragments textuels qui contiennent la réponse à la requête d'un utilisateur.

- **Génération automatique de textes (GAT)** La GAT est une tâche qui consiste à générer des textes par voie automatique. Le résumé automatique, la génération de paraphrases, la réécriture de textes, la simplification et la génération de questions sont autant d'exemples d'application de la GAT.

- L'**interaction homme-machine**, qui vise à développer des systèmes permettant à l'utilisateur de converser avec des ordinateurs en utilisant le langage naturel (texte, parole et signaux de communication non verbaux, tels que les gestes et les expressions faciales). Les agents conversationnels (mieux connus sous le nom de *chatbots*) constituent une application bien connue dans ce domaine, mais il faut également mentionner l'interaction homme-robot, l'interaction dans les environnements virtuels, l'interaction verbale dans les jeux, etc.

En outre, les TL répondent également aux besoins et aux objectifs de la linguistique empirique et appliquée, en ce qu'elles offrent aux linguistes et aux cogniticiens de nouveaux outils pour explorer les langues dans la réelle diversité de leurs usages, et plus généralement en permettant à des chercheurs de tous les domaines et disciplines relevant des humanités numériques à analyser les données non structurées qu'ils étudient.

Les TL font déjà partie intégrante de notre quotidien. À un niveau individuel, nous les utilisons peut-être sans même nous en rendre compte, quand nous vérifions et corrigeons nos erreurs de saisie dans les textes que nous produisons, quand nous faisons appel à des moteurs de recherche sur internet ou quand nous passons un appel téléphonique à notre banque pour effectuer une transaction. Il s'agit d'une composante importante, mais souvent invisible, des applications qui touchent à divers secteurs et domaines. N'en citons que quelques-uns : dans le domaine de *la santé*, les TL contribuent ainsi à la reconnaissance et à la classification automatiques de termes médicaux ou au diagnostic des troubles de la parole et de la cognition. Elle est de plus en plus intégrée dans des environnements et des *applications destinées aux apprenants*, p. ex. pour l'exploration de contenus pédagogiques, pour l'évaluation automatique de réponses en texte libre, pour offrir des boucles d'apprentissage aux apprenants et aux enseignants, pour l'évaluation de la prononciation dans une langue étrangère et bien plus encore. Dans le domaine du *droit,* les TL s'avèrent indispensables pour plusieurs tâches, que ce soient la recherche, la classification et la codification dans d'énormes bases de données juridiques ou encore la réponse à des questions dans ce domaine et la prédiction des décisions de justice.

Le large éventail d'applications des TL démontre non seulement que ces technologies comptent parmi les plus pertinentes pour la société, mais forment aussi l'un des domaines les plus importants de l'IA qui engendre un impact économique en croissance rapide.[19]

4 Les technologies de la langue et le français

4.1 Disponibilité des données et des outils linguistiques

La présente section se fonde sur l'analyse de plus de 1500 ressources, outils et modèles qui ont déjà été identifiés et documentés pour les langues françaises, ainsi que sur d'autres sources qui doivent encore être documentées par le consortium ELE.

Afin de procéder à un état des lieux en matière d'outils linguistiques, nous suivons pour l'essentiel la même organisation que dans (Mariani *et al.*, 2012) et étudions les mêmes groupes de technologies, au nombre de 15, auxquels s'ajoutent (a) des sections consacrées aux ressources génériques, lexicales et textuelles, qui peuvent servir à de nombreuses applications : dictionnaires, ressources terminologiques, corpus monolingues et modèles de langue ; (b) une section supplémentaire sur les TL pour la recherche en linguistique, théorique et appliquée. Ces thèmes sont abordés en détail plus loin. En ce qui concerne les méthodes les plus performantes dans le domaine du TAL (voir la section 3), il apparaît que l'écrasante majorité des outils et des applications les plus récents reposent presque exclusivement sur des technologies génériques d'apprentissage automatique, ce qui signifie que les ingrédients les plus importants pour construire des systèmes de traitement sont les données et, dans une moindre mesure, les ressources de calcul. C'est pour cela que les jeux de données seront donc abordés en même temps que les technologies qui s'y rapportent. Selon le ou les objectifs des applications qui sont présentées ci-dessous, il se peut toutefois que des contraintes opérationnelles spécifiques imposent de développer un logiciel dédié, p. ex. pour minimiser le coût d'entraînement, l'empreinte mémoire, pour augmenter la vitesse de traitement, ou adapter l'interaction avec l'utilisateur, etc. Il s'ensuit que les outils, modèles et algorithmes spécifiquement dédiés aux TL continuent à jouer un rôle important et qu'ils seront évoqués partout où cela sera nécessaire.

Dictionnaires monolingues et ressources terminologiques

Il existe un grand nombre de lexiques informatisés généralistes de grande taille pour le français, qui associent des lemmes, ou des formes de mots, à des descripteurs morpho-syntaxiques de base telles que la catégorie lexicale ("part-of-speech", ou "POS"), le genre, la classe de conjugaison (pour les verbes) ; selon la ressource considérée, il est possible d'y trouver d'autres informations telles que les propriétés morphologiques et syntaxiques détaillées, la prononciation, la fréquence, le niveau de difficulté. Parmi les lexiques disponibles, le Wiktionnaire[20] (le Wiktionary en français) est la ressource la plus à jour (et la plus importante), si ce n'est la plus étayée sur le plan linguistique. À l'ère des réseaux neuronaux, les représentations syntactico-sémantiques distribuées des mots du lexique constituent également une exigence élémentaire et de telles ressources pour le français le sont notamment par l'intermédiaire du projet FastText[21].

Diverses ressources lexicographiques historiques pour le français sont mises à disposition par l'ATILF[22], tandis que le site collaboratif "*Le Dictionnaire des Francophones*"[23] (DDF), qui donne accès à une compilation à grande échelle de définitions de dictionnaires en provenance de multiples sources, restitue peut-être un panorama plus diversifié et plus dynamique du lexique français. Le DDF ne retrace pas seulement le lexique générique de base, mais contient également des dictionnaires spécialisés et des terminologies. Un nombre considérable de ressources lexicales et terminologiques supplémentaires (vocabulaires spécialisés, terminologies monolingues ou multilingues, thésaurus, ontologies) de taille et de niveau de détail variables a également été rendu public - la banque de données terminologiques et linguistiques TERMIUM du gouvernement du Canada[24] et la base de données IATE avec les termes accumulés par les traducteurs de l'UE en sont des

exemples de grande taille[25].

Corpus monolingues, modèles de langage à grande échelle

Il n'existe pas de corpus national du français officiel, qui contiendrait un sous-ensemble représentatif de la langue, réparti selon les périodes, les genres et les domaines, comme cela peut exister pour d'autres langues. Des corpus importants (jusqu'à des milliards de tokens) de genres variés sont cependant accessibles et consultables, p. ex. via Frantext,[26] *Sketch Engine*,[27] ou sur le site web de la *Leipzig Corpora Collection*[28]. D'autres genres sont bien représentés, faciles à rechercher et à télécharger et comptent entre autres des' "écrits du Web" (comme Wikipédia, mais pas seulement), de la littérature (p. ex. via le projet Gutenberg[29]), des actualités (p. ex., le corpus NewsCrawl[30]), des écrits scientifiques, juridiques ou administratifs issus d'institutions nationales ou internationales ; certains d'entre eux sont mis à jour régulièrement.

Le projet CommonCrawl[31] agrège des données issues de l'exploration de la toile dont le volume est d'un ordre de grandeur qui dépasse largement ces ressources pour bon nombre de langues ; qui plus est, ce corpus est mis à jour régulièrement. L'utilisation de certaines parties du sous-ensemble français de CommonCrawl, éventuellement associée aux corpus davantage nettoyés que l'on vient d'évoquer, a permis d'entraîner des modèles de langue de type BERT à grande échelle. Ainsi, FlauBERT (Le *et al.*, 2020) est construit avec un corpus contenant environ 12 milliards d'occurrences de mots, CamemBERT (Martin *et al.*, 2020) utilise les 22 milliards de mots d'OSCAR, et ces chiffres continuent de croître, bien qu'à un rythme beaucoup plus lent que les corpus anglais correspondants. De gros modèles de langue de diverses moutures (adaptant au français les architectures BERT, ELMO, GPT,[32] ou mBART) ont désormais été rendus publics et sont disponibles pour des utilisations scientifiques comme commerciales, et d'autres devraient voir le jour (p. ex. grâce au projet BigScience[33]) ; ils ont permis de stimuler la progression dans des activités de pointe dans nombre de tâches en TAL. Il existe aussi un petit nombre de variantes spécialisées de ces ressources (p. ex., pour la modélisation des tweets). Comme pour d'autres langues, il a été possible de dériver des versions adaptées facilement (p. ex., pour les textes scientifiques et les brevets). Ces ressources offrent de nouvelles perspectives notamment pour la génération de textes en français[34].

Des bases de données de grande dimension d'enregistrements annotés (segmentés par phrases, locuteurs et intonations, transcrits) contenant des milliers d'heures de prises de son sont disponibles pour plusieurs genres (p. ex., actualités, lecture de livres, conférences). Il s'agit principalement d'échantillons du français standard (éduqué), et il est plus difficile de trouver de grands corpus pour d'autres conditions d'enregistrement (p. ex., conversations, émissions parlées spontanées, multipartites, téléphoniques, émotionnelles, bruitées, pathologiques) et d'autres variantes géographiques. La collecte d'enregistrements importants reste donc un enjeu urgent pour élargir l'applicabilité et l'efficacité du traitement de la parole en français, un objectif qui est abordé, p. ex., par le projet *Voice Lab*[35]. La situation est probablement encore plus difficile pour les données multimodales, qui posent encore plus de problèmes de confidentialité et de droits que les données vocales.

Outils linguistiques de base : tokenisation, étiquetage morphosyntaxique (POS tagging), analyse/génération de la morphologie

Voici un domaine où le terrain était déjà bien couvert en 2012 et qui a bénéficié de l'amélioration des outils d'apprentissage automatique. On dispose de tokeniseurs, lemmatiseurs et étiqueteurs morphosyntaxiques en licence libre de qualité industrielle pour le français, dans Spacy[36], Spark NLP[37] ou Stanza[38] et ceux-ci sont accessibles à l'aide de quelques lignes de code en Python. D'autres logiciels dédiés au traitement du français (comme MELT[39]) coexistent avec des suites de traitement multilingue développées au niveau national ou international en milieu académique ou industriel (NLTK[40], Freeling[41], GATE[42], etc). Il devrait être assez simple de créer d'autres outils de ce type, étant donné la disponibilité de composants génériques d'étiquetage de séquences entraînables et de grandes quantités de données annotées. Deux réserves s'imposent toutefois : (a) il n'existe pas de comparaison systématique récente des performances pour ces tâches ; (b) la plupart de ces outils sont conçus pour traiter le français "générique" et il en existe trop peu pour des variétés ou des genres plus spécifiques (p. ex., textes techniques, courriels, textos, transcriptions de chaînes parlées, contenu généré par l'utilisateur et autres, notamment pour les variétés qui évoluent très rapidement ; il en va de même pour les variétés moins représentées de français écrit ou parlé, p. ex., le français parlé hors de France, qui peut également comporter différents types de phénomènes d'alternance codique). Pour de tels matériels linguistiques, il est connu que la qualité du balisage morphosyntaxique et celle de l'analyse syntaxique diminuent rapidement (Plank *et al.*, 2014).

Une remarque complémentaire : dans de nombreuses architectures récentes en TAL, de tels outils ne sont même pas nécessaires une fois passées les étapes de segmentation de base des phrases et de tokenisation des mots, car le traitement y est effectué avec un processus de (sous-)tokenisation automatique utilisant des unités infra-lexicales calculées p. ex. par l'algorithme BPE (pour *Byte-Pair Encoding*) et entraîné de bout en bout, et en s'affranchissant du concept linguistique de "mot" et de ses propriétés de base (Sennrich *et al.*, 2016; Kudo et Richardson, 2018). C'est p. ex. le cas dans la bibliothèque de tokenisation HuggingFace[43]. De telles stratégies de prétraitement sont très efficaces, mais il leur manque les motivations linguistiques qui sont nécessaires pour les analyses plus profondes par d'autres composants de TL.

En ce qui concerne l'analyse morphologique profonde de mots ou de composés de création récente, la situation est beaucoup moins favorable car le seul outil significatif est DERIF (Namer, 2009), dédié au traitement de textes généraux plus qu'à celui de textes techniques. Grâce aux avancées du projet UniMorph[44], qui intègre le dictionnaire Lefff français (Sagot, 2010), les analyseurs morphologiques génériques sont plus faciles à entraîner, et des modèles ont commencé à apparaître également pour cette langue.

Analyse syntaxique profonde et de surface

En ce qui concerne l'analyse syntaxique, la situation est très similaire à celle de l'étiquetage morphosyntaxique, en raison de la disponibilité d'un vaste entrepôt ouvert multilingue hébergeant les corpus arborés (*treebank*) élaborés dans le cadre du projet "*Univer-*

sal Dependencies"[45], qui peuvent être facilement utilisés pour entraîner des analyseurs syntaxiques. Plusieurs corpus arborés pour le français (FTB, GTB, Sequoia, etc.) figurent dans cette collection et totalisent près de 1,2 million d'occurrences de mots issues de sources et de genres différents. L'apprentissage d'un analyseur en dépendances pour le français fournissant des informations syntaxiques de base est donc relativement simple et permet d'obtenir une précision proche de 90 % en termes de scores d'attachement non étiquetés (UAS), ce qui peut être suffisant pour de nombreuses applications. On trouvera des évaluations récentes dans (Zeman *et al.*, 2018). Ces types d'analyseurs syntaxiques ont été intégrés dans des outils de TAL génériques comme Spacy ou Stanza ; les développements de recherches basés sur les mêmes ressources sont relativement faciles à récupérer, à développer et à modifier. Il existe d'autres ressources intéressantes dans un objectif d'analyse syntaxique du français : mentionnons plusieurs dictionnaires de grande taille contenant des informations syntaxiques détaillées, ainsi que des corpus arborés et d'autres corpus plus récemment élaborés dans le cadre des projets français ANR Passage (de la Clergerie *et al.*, 2008) et PARSEME (Candito *et al.*, 2017) ; ; enfin, un ensemble important d'annotations a été collecté de manière participative dans le cadre d'un jeu sérieux, ZombiLingo (Fort *et al.*, 2014).

La prise en charge d'outils d'analyse syntaxique profonde qui permettraient d'obtenir une analyse fine est un peu moins développée.

Analyse sémantique au niveau de la phrase

L'analyse sémantique au niveau de la phrase pour le français peut faire appel à des réseaux sémantiques à grande échelle inspirés du projet anglais Wordnet, ou de leurs versions multilingues (p. ex., *Open Multilingual Wordnet*,[46] qui, pour le français, inclut le WOLF (Sagot et Fišer, 2008)). Ces ressources fournissent typiquement des éléments concernant les inventaires de sens des formes lexicales les plus courantes et permettent de désambiguïser l'utilisation de mots en contexte. Quelques corpus sémantiquement annotés de petite taille (p. ex., FrenchSemEval), constitués notamment dans le cadre des campagnes d'évaluation (*shared tasks*) organisées par la conférence SemEval, peuvent également être facilement utilisés pour entraîner ou évaluer des modules de désambiguïsation (*word sense disambiguation* ou WSD).

D'autres ressources sont précieuses pour l'analyse au niveau de la phrase : le FrameNet français Asfalda, dérivé du FrameNet original[47], ainsi que des corpus parallèles annotés de manière semi-automatique où les étiquettes de rôles sémantiques ont été projetées à partir du côté anglais du corpus. Malgré la disponibilité de ces ressources génériques, les outils et les modèles pour la désambiguïsation sémantique lexicale automatique[48] et l'étiquetage des rôles sémantiques sont moins avancés et intégrés que dans le cas de l'analyse syntaxique. Cela s'explique par le manque de corpus arborés sémantiquement annotées pour le français ; une autre raison possible serait que l'utilité des analyses profondes dans les applications en aval n'est peut-être plus aussi critique qu'elle l'était auparavant.

Une dernière tâche importante au niveau de la phrase concerne le calcul des relations d'implication textuelle (RTE ou NLI), un domaine dans lequel les méthodologies d'apprentissage automatique sont performantes ; en outre, les techniques de transfert

interlingue sont également pertinentes pour cette tâche, ce qui signifie que de grands ensembles de données en anglais peuvent être exploités pour développer des systèmes pour le français : de telles approches peuvent s'appuyer, p. ex., sur le corpus XNLI[49].

Analyse sémantique de documents

Pour l'analyse sémantique de documents, la résolution des coréférences est une tâche de base : des ressources annotées à grande échelle (p. ex. ANCOR-Centre et Democrat) pour l'apprentissage de tels systèmes existent à la fois pour les données parlées et écrites et sont utiles pour apprendre et évaluer des outils pour les genres de documents correspondants (Wilkens *et al.*, 2020). Le français fait également partie de l'ensemble des langues de base du projet CorefUd.[50] Malgré cela, le calcul des relations de coréférence n'est pas (encore) largement présent dans les outils standard de TAL, une exception étant l'outil d'analyse multilingue LIMA développé par le LIST du CEA[51].

Une analyse plus générale des structures discursives peut s'appuyer sur le corpus Annodis[52]ainsi que sur les outils dérivés et les repères produits pour les campagnes d'évaluation Disrpt (Zeldes *et al.*, 2021). Les traitements au niveau discursif jouent un rôle essentiel pour l'analyse des conversations et pour les systèmes de dialogue. Les dialogues réels (parlés ou écrits) sont ici assez rares, en raison des difficultés d'annotation que représentent ces types de ressources, qui concentrent nombre des pièges auxquels l'analyse textuelle à niveaux multiples (parole, émotions, étiquettes sémantiques, actes de dialogue, etc.) est confrontée, sans compter les problèmes de confidentialité liés à leur collecte. MEDIA (Bonneau-Maynard *et al.*, 2008), qui est le résultat d'un travail collectif important, a une portée limitée et est quelque peu obsolète ; les initiatives publiques visant à développer de nouvelles ressources et à enregistrer davantage d'interactions orales dans des domaines plus ouvert sont limitées (en termes de portée, de types d'annotation et de taille). Cela pourrait être considéré comme une faiblesse étant donné la floraison d'applications pour de tels systèmes (chatbots, agents conversationnels, assistants vocaux et autres) ; il est probable, cependant, que d'importantes bases de données privées commencent à exister pour de nombreuses langues, y compris le français.

Extraction d'information

Les systèmes de reconnaissance d'entités nommées (NER) sont largement présents sous forme de service de base à l'intérieur de bon nombre de suites logicielles d'analyse de texte (voir ci-dessus), où cependant la notion d'entité nommée se limite le plus souvent aux types et structures d'entités de base de style MUC (noms, lieux, organisations, dates, montants). Certaines d'entre elles comprennent également une forme d'extraction de relations. D'importants corpus avec des annotations de type NER pour les actualités et la presse françaises (écrites et parlées) existent (Galliano *et al.*, 2009; Sagot *et al.*, 2012; Dupont, 2019) et peuvent être utilisés pour entraîner des systèmes de NER efficaces (Ortiz Suárez *et al.*, 2020). Ces corpus sont toutefois soumis à une obsolescence rapide en raison de l'apparition de nouveaux noms et de nouvelles entités dans les contenus d'actualités. Les systèmes NER sont plus rares dans des domaines spécifiques (comme le

droit, la santé-médecine, les sciences) où les entités et relations peuvent être d'une nature différente (p. ex. : références légales, noms de jugements rendus ou noms de médicaments, symptômes et virus).

L'analyse d'opinions et de sentiments, ou la détection de discours haineux, sont généralement envisagées comme des tâches de classification de phrases qui peuvent être mises en œuvre avec peu, voire aucune, analyse linguistique ; leur développement repose sur la disponibilité de données annotées ouvertes et à grande échelle. Concernant le français, les données disponibles contiennent principalement des listes annotées de tweets et de critiques de produits avec une annotation en polarité (ironie et sarcasmes compris, voir (Karoui *et al.*, 2017) et la campagne DEFT 2017) ; des données d'étiquetage de la posture (*stance*) sont également disponibles pour le domaine politique[53]. En comparaison, il y a une pénurie de ressources publiques dédiées à d'autres types importants de tâches de classification de phrases telles que l'identification de contenus fallacieux ou haineux (voir cependant (Chiril *et al.*, 2020) pour la détection de discours sexistes ou les ressources distribuées par les producteurs de médias tels que l'AFP[54] ou "Le Monde" [55]). Il s'agit d'un domaine où le transfert interlinguingue à partir de l'anglais, lorsque de telles ressources existent, peut être envisagé - avec des résultats incertains, cependant, étant donné la dépendance vis à vis de la culture et de la langue de phénomènes subtils, mais importants, comme les propos humoristiques ou sarcastiques. Il faut aussi observer que pour ces applications, des acteurs du monde de l'entreprise possèdent et exploitent d'importantes bases de données multilingues (y compris avec des exemples français), soit pour un usage interne, soit pour une exploitation commerciale[56].

Recherche d'information et fouille de texte

Les technologies de recherche d'information et de fouille de textes sont déployées à grande échelle depuis un certain nombre d'années et font l'objet de moins de recherches qu'auparavant, l'accent étant progressivement mis sur des interactions plus complexes (p. ex., par le dialogue) avec des moteurs de recherche. Grâce à des années de campagnes d'évaluation telles que TREC[57], TRECVID[58], CLEF[59], et DEFT pour le français[60], de grandes collections appariant des requêtes avec des document existent pour une variété de types de données (textes, documents structurés, transcriptions de discours et de vidéos), de langues et de domaines. Des logiciels ouverts de recherche d'information robustes, efficaces et indépendants de la langue sont également largement disponibles.

Les outils de classification et de clustering sont d'autres exemples d'outils indépendants de la langue pour lesquels il est facile de trouver des ressources et des outils matures sur étagère. Là encore, les performances et le niveau de maturité des technologies existantes peuvent varier considérablement en fonction des types de données, des genres et des domaines. Les textes courts et bruités, les énoncés parlés continuent à poser des problèmes difficiles et justifient le développement de nouvelles ressources.

Trouver des ressources pour des tâches d'extraction de mots-clés et de terminologies est bien plus difficile, en particulier en comparaison de ce qui existe pour l'anglais pour lequel des bases de données comptant des centaines de milliers d'exemples ont été élaborées[61].

Génération automatique de textes

La génération automatique commence avec les outils d'assistance à la rédaction que sont les correcteurs orthographiques, grammaticaux et stylistiques, ainsi que les fonctions d'autocomplétion et de normalisation de textes (notamment pour des contenus générés par des utilisateurs). Des outils pour ces tâches sont largement disponibles et les logiciels les plus avancés existent dans le commerce et sont intégrés dans de nombreux éditeurs de texte, champs de saisie et boîtes de dialogue. Si les dictionnaires sont faciles à trouver, les corpus annotés comportant d'authentiques erreurs sont beaucoup plus rares ; il en va de même pour les corpus d'apprenants, qui sont collectés "en masse" par les institutions éducatives mais très rarement redistribués.

Le paysage de la génération automatique de textes a été profondément transformé par les développements de nouveaux modèles de langue de très grande taille que nous avons évoqués plus haut. Ces technologies sont porteuses de nouvelles applications utilisant la génération de textes contrôlés à partir de sources d'information structurées (statistiques, tableaux ou formules logiques, p. ex.) ou non-structurées (invites textuelles ou images). Au-delà des nombreuses considérations éthiques associées à l'entraînement et à l'utilisation de ces grands modèles de langue (Bender *et al.*, 2021), ces techniques sont également sujettes à des exploitations malveillantes telles que la production d'infox ou la génération de faux courriels, de sites web frauduleux, de fermes de liens, etc. L'évaluation des résultats des outils de génération automatique de textes requiert aussi des ressources de très grande taille qui n'existent pas aujourd'hui pour le français.

D'autres applications importantes de la génération de texte telles que la production automatique de résumé automatique et la traduction automatique sont abordées dans la suite.

Résumé automatique, réponse aux questions

Comme pour d'autres technologies de fouille de textes, la génération de résumé automatique est une tâche ancienne dont l'étude et l'évaluation peuvent s'appuyer sur un historique de campagnes d'évaluations portant principalement sur des textes en anglais, tout comme sur des sources multilingues. Bien que des corpus dédiés tels que PUCES ou RMP2[62] soient présents dans ce paysage depuis un certain temps, les applications de résumé de texte pour le français ne sont pas aussi développées que les autres applications de fouille de textes : alors que la dernière génération de systèmes était basée sur des techniques extractives, les progrès accomplis récemment dans les techniques statistiques ou neuronales de génération de texte, combinés à la disponibilité de jeux de données de données d'entraînement plus importants pour le résumé d'actualités (Scialom *et al.*, 2020) ou (Eddine *et al.*, 2020), pourraient changer cet état de fait et favoriser le développement de générateurs de résumés de texte abstractifs - au moins pour certains genres textuels bien couverts.

Les systèmes de réponse aux questions (*Question Answering*, ou QA) sont désormais une technologie mature, qui existe sous forme de service de base dans les agents conversationnels généralistes. Pour le français, ces systèmes peuvent bénéficier des ressources

développées au fil des années dans le cadre des campagnes d'évaluation TREC et CLEF déjà mentionnées. Les progrès des systèmes de réponse aux questions et des systèmes de compréhension automatique pour le français bénéficieront des récentes tentatives de mise au point de bases de données de grande taille pour la recherche sur les systèmes de question-réponse, comme PIAF (Keraron *et al.*, 2020) et FSquad (d'Hoffschmidt *et al.*, 2020), qui offrent à la communauté des dizaines de milliers de triplets question-passage-réponse, et permetteront d'aborder un plus large éventail de types et de difficultés de questions. Si ces ressources sont extrêmement utiles, nous constatons que, comme pour d'autres domaines technologiques, des ressources plus spécialisées, qui existent désormais pour d'autres langues, font encore défaut pour le français (p. ex., des bases de données de questions-réponses pour le domaine médical ou celui des renseignements pratiques).

Traduction automatique (TA)

Après avoir évolué vers des systèmes entièrement basés sur des corpus, et aujourd'hui entièrement neuronaux, l'existence de systèmes de TA pour le français dépend principalement de celle de corpus parallèles adéquats. En raison de son utilisation comme l'une des principales langues internationales, de telles ressources abondent pour le français, surtout lorsqu'elles sont associées à une traduction anglaise, dans un couple de langues pour lequel des centaines de millions de segments parallèles peuvent être exploités. C'est ainsi que des sources massives d'entraînement pour la traduction automatique à partir de données textuelles sont disponibles sur le site web d'Opus[63] (Tiedemann, 2012), sur celui de CommonCrawl (projet de l'UE), par le biais d'OpenCC de Facebook ou sur la page de ressources du "*Workshop for Machine Translation*" (WMT)[64]. Les données pour la traduction de la parole existent en quantités beaucoup plus faibles et pour des domaines très restreints (conférences, débats parlementaires) ; la série de campagnes d'évaluation IWSLT[65] diffuse des jeux de données d'entraînement et de test pour des applications de ce type. De nombreux moteurs de TA génériques de grande qualité sont également disponibles sur la toile pour la plupart des couples de langues bien dotées (qui varient selon les restrictions d'utilisation associées : e-translation, deepL, Google Traduction, Bing Translator, ou les moteurs de traduction neuronale de Systran et Reverso, de fabrication essentiellement française), y compris la traduction directe entre le français et la plupart des langues européennes, ainsi que l'arabe, le chinois ou le japonais. La situation par rapport aux autres langues est beaucoup moins favorable : si l'on peut puiser dans de grands corpus parallèles avec d'autres langues européennes, la situation pour les langues non européennes est plus contrastée, ce qui peut constituer une faiblesse pour la traduction du français vers le japonais, le chinois, le russe ou l'arabe, notamment dans les domaines spécialisés.

D'importantes bases de modèles de TA pré-entraînés pour le français ont été mis à disposition sur la plateforme HuggingFace[66] (principalement à l'initiative de l'Université d'Helsinki, mais aussi grâce à la politique d'ouverture pratiquée par des centres de recherche comme le FAIR de Meta, qui a rendu publics des modèles multilingues) ainsi que sur la place de marché de modèles de TA de Systran[67].

Il faut ajouter qu'il existe d'autres ressources pour la TA depuis et vers le français,

comme par exemple des dictionnaires bilingues de taille et de contenu variés, des bases terminologues multilingues pour de nombreux domaines ainsi que des jeux de test concernant divers aspects spécifiques de la TA (terminologie, coréférence, biais de genre, etc.). Dernières ressources utiles, et non des moindres, les outils d'alignement de phrases et de mots conservent un rôle important dans la préparation et l'analyse des systèmes de TA ; ils représentent également un atout précieux pour les travaux scientifiques en traductologie. Si les outils génériques sous licence logicielle ouverte sont relativement faciles à trouver, il n'existe qu'une poignée de corpus d'évaluation pour l'alignement, qui tous concernent la paire de langue français-anglais.

Systèmes de transcription de la parole

La prise en charge de la transcription de la parole peut être considérée comme satisfaisante, et les moteurs de reconnaissance vocale sont notamment un composant de base des assistants conversationnels (Siri, Alexa, Home, etc.) qui ont conquis un large public ; ils sont également disponibles pour les entreprises sous forme de services en nuage (IBM Watson, Google NLP Cloud services, Amazon Transcribe, etc.) Certains d'entre eux peuvent gérer plusieurs dialectes du français pour mieux répondre aux besoins des locuteurs belges, canadiens ou suisses. La qualité de la transcription automatique de la parole peut varier grandement en fonction des conditions d'enregistrement et des genres de parole. Il existe aussi des solutions commerciales de ces systèmes dédiées pour des cas d'usage particuliers comme la transcription de réunion, la reconnaissance vocale à bord de véhicules ou d'aéronefs, ou encore le sous-titrage de vidéos. Plusieurs PME françaises sont actives sur ces marchés (Vocapia Research, SNIPS/Sonos, ChapsVision, Linagora, etc.) avec également un dynamisme qui se traduit par l'arrivée de nouveaux acteurs (Amberscript, Zenidoc, Noota, etc.). Des données ouvertes appropriées pour l'entraînement des systèmes de transcription à grande échelle sont disponibles auprès de diverses sources (LibriSpeech, Mozilla Common voice), tandis que des modèles en logiciel ouvert pour transcrire le français sont également diffusés librement par le centre FAIR de Meta.

Les systèmes existants sont cependant fragiles et connus pour se dégrader rapidement lorsque le contexte de l'enregistrement implique du bruit ou de l'écho, ou lorsque le style et la voix du locuteur divergent des données d'entraînement : ces phénomènes peuvent être dus à des variations liées à l'âge, aux accents, aux types de données, mais aussi à des déficiences vocales temporaires ou permanentes. En tant que tel, le manque de robustesse par rapport à ces voix non-standard peut également être une cause d'exclusion de certains groupes d'utilisateurs. Le besoin de bases de données vocales plus variées qui seraient correctement représentatives de la diversité et des voix de la population générale reste donc un objectif légitime et important.

La parole est intrinsèquement multimodale et sa production implique un système moteur complexe qui mobilise plusieurs organes et structures physiologiques du corps humain (les articulateurs de la parole). Toute une branche de la recherche vise à développer des technologies vocales qui reposent non seulement sur le signal vocale mais aussi sur ces mouvements articulatoires (Schultz *et al.*, 2017). La reconnaissance vocale audio-visuelle en est un parfait exemple : une vidéo du visage du locuteur, traitée en même temps

que le signal de parole audio, permet d'améliorer la robustesse par rapport au bruit. D'autres applications sont la lecture labiale automatique (autrement dit, la reconnaissance exclusivement visuelle de la parole), la séparation vocale et la synthèse de visages parlants (voir infra). Alors que de très grandes bases de données audiovisuelles existent pour l'anglais, contenant des centaines d'heures, seules quelques-unes sont disponibles publiquement pour le français et sont d'un ordre de grandeur plus modeste (Petrovska-Delacrétaz *et al.*, 2008).

Les systèmes de reconnaissance vocale sont souvent assortis de modules pouvant effectuer des tâches associées de diverses natures comme la détection d'activité vocale, la détection de mots-clés, l'identification et la diarisation du locuteur. Pour ces tâches, il est en particulier possible d'exploiter les ressources développées au travers une série de campagnes d'évaluation (ESTER[68], ETAPE[69] et REPERE[70]).

Synthèse vocale

Jusqu'à récemment, la synthèse vocale, qui permet de générer de la parole à partir d'un texte (*text to speech*, TTS), mettait en jeu une chaîne de traitement composée de deux modules distincts : (i) le traitement du texte chargé d'extraire la séquence phonétique et la structure syntaxique de la phrase à synthétiser et (ii) un module de traitement du signal chargé de la synthèse du signal de parole. Cette dernière étape nécessite des corpus d'enregistrements bien conçus de données vocales acoustiquement propres, réalisées dans un environnement acoustique contrôlé. De telles ressources sont très coûteuses et rarement partagées entre les développeurs de technologies de synthèse vocale.

À l'ère de l'apprentissage profond, la synthèse vocale s'est écartée de ce schéma, pour aller ver un mode de traitement bout en bout complet où le signal vocal est généré directement à partir de la représentation orthographique (Shen *et al.*, 2018). Bien qu'il soit possible d'obtenir une qualité vocale raisonnable à partir de courts échantillons de la voix cible en association avec des transcriptions orthographiques, obtenir une synthèse vocale expressive de grande qualité nécessite toujours des échantillons de données propres richement annotés issus d'un seul locuteur : parmi les exemples de telles ressources pour le français on compte le corpus SynPaFlex[71] ainsi que des jeux de données produits à l'occasion du défi Blizzard[72].

Couplées au code source libre disponible pour l'entraînement de systèmes neuronaux avancés[73], ces bases de données peuvent être utilisées pour construire un système de synthèse vocale (quasiment) ex nihilo. Il n'en reste pas moins important de noter que la plupart des enregistrements disponibles correspondent à des productions de locuteurs éduqués utilisant un français conventionnel dans des conditions assez contrôlées, limitant ainsi l'étendue, la diversité et l'expressivité de la parole générée. À l'instar des systèmes de reconnaissance vocale, il est capital de poursuivre la collecte de différentes bases de données vocales, avec le recueil de parole spontanée, de parole en sitation d'interaction, de parole émotionnelle, etc.

Émotions dans la parole

La détection et la génération vocale d'émotions sont un domaine relativement récent. Celui-ci est associé au champ plus large des sciences affectives qui a émergé au début des années 2000. Les sciences affectives forment un domaine interdisciplinaire dont l'objectif déclaré est d'étudier les émotions et d'autres phénomènes affectifs (p. ex., les attitudes, les états cognitifs) grâce à la contribution de plusieurs disciplines. Dans ce cadre, l'informatique affective se concentre principalement sur la reconnaissance et la synthèse des expressions faciales et des inflexions de la voix (Picard, 2000). Ce domaine est connexe à plusieurs domaines des technologies de la langue. p. ex., la détection de motifs vocaux chez des locuteurs en fonction des états émotionnels présente un intérêt pour divers domaines des technologies de la langue comme la reconnaissance vocale, la caractérisation des locuteurs, les interactions homme-machine ; la synthèse vocale de voix émotionnelles est un autre domaine (voir les sections précédentes). Les applications qui impliquent la modélisation de l'état émotionnel des locuteurs concernent des domaines aussi variés que la santé, la sécurité, l'éducation, le divertissement ou les jeux sérieux.

Les technologies de la langue pour les études linguistiques

Comme pour la plupart des disciplines des sciences sociales et humaines (SHS), la révolution numérique ouvert de nouvelles voies pour l'analyse linguistique (Liberman, 2019). Les technologies ont aidé les linguistes de différents horizons non seulement à collecter, stocker, enrichir et échanger beaucoup plus facilement divers types de données, mais aussi à analyser différemment le matériau linguistique. Elles ont joué un rôle important dans le réexamen de questions de recherche classiques et a permis de poser de nouvelles hypothèses de recherche vérifiées dans de gros corpus ou du moins dans des données enrichies à l'aide d'une variété de méta-données. Elles ont également fait remonter à la surface la question de la reproductibilité, une préoccupation moins importante dans le domaine des sciences humaines mais qui devient d'actualité car les matériaux linguistiques disponibles sont désormais plus accessibles et plus vastes. Enfin, elles ont permis de créer de nouvelles collaborations avec d'autres domaines et ont favorisé l'interdisciplinarité. En matière de collaboration, les sciences informatiques sont naturellement privilégiées, mais de nouveaux ponts sont jetés avec bien d'autres disciplines comme la médecine, les neurosciences, la psychologie, la littérature, la sociologie, etc.

Pour certains domaines linguistiques, comme l'étude des langues en danger, nous assistons à un changement complet : grâce à de nouveaux outils et approches technologiques, il devient possible de documenter ces langues de manière plus fiable et plus complète, contribuant ainsi à leur revitalisation. La conséquence de ce changement méthodologique est que l'on peut s'attendre à ce que les générations futures de linguistes intègrent durablement dans leur paradigme scientifique des méthodes, des outils et des compétences adaptés à l'ère numérique et que ce mouvement vers les technologies se poursuive et s'amplifie.

En l'espèce, la France ne fait pas exception. Au niveau local, c'est-à-dire celui des universités ou des laboratoires, des initiatives spécifiques et/ou au niveau national, les

technologies ont un impact sur tous les domaines linguistiques, sous la forme de corpus, d'outils et de méthodes. Pour ce qui est des corpus, les variétés écrites et parlées de la langue française sont couvertes, les sources écrites historiques et des outils de textométrie associés étant les mieux représentés et plus visibles[74]. Des ressources comme Frantext[75], lancé il y a deux décennies, couvrent des corpus échantillonnés du 9ème au 21ème siècle avec un moteur de recherche qui permet d'effectuer des requêtes simples et complexes sur des formes, des lemmes ou des catégories grammaticales et d'afficher les résultats. On dispose également de ressources mixtes écrites et orales de grande taille, grâce à des projets tels que RHAPSODIE[76], ORFEO ("*Outils et Ressources sur le Français Ecrit et Oral*")[77], ESLO ("*Enquêtes socio-linguistiques d'Orléans*")[78]). En ce qui concerne les technologies nécessaires à l'analyse de ces données, le français est relativement bien pourvu grâce aux réalisations en TAL qui ont abouti à une variété d'outils d'analyse lexicale, morphologique, syntaxique et sémantique (p. ex. NooJ[79]). Un problème demeure quant à l'utilisabilité de ces outils, qui sont parfois difficiles à installer et à exploiter pour le non-expert.

Les TL sont également d'une grande utilité pour divers sous-domaines de la linguistique appliquée, notamment, comme mentionné dans la section 3, pour étudier des langues ou aider les apprenants - en particulier ceux qui ont des difficultés de lecture ou d'écriture. Si des ressources lexicales précieuses existent pour le français[80], l'étude des mesures de lisibilité, des techniques de simplification du texte, des outils de lecture automatique augmentée, de la génération et de la correction de questions, etc., est toujours entravée par la pénurie de bases de données publiques de grande taille.

Du côté de la langue parlée, les initiatives en termes de corpus ont été guidées par des besoins spécifiques, variables selons les domaines linguistiques[81] : les corpus CLAPI[82]sont destinés à l'analyse du discours ; PFC ("*Phonologie du Français Contemporain*")[83] a été initié par des phonéticiens et des phonologues intéressés par les accents, bien qu'*in fine* le corpus vise à répondre à des besoins dans d'autres domaines linguistiques. Pour ce qui est des applications aux données vocales, il existe divers outils : des aligneurs calculent une correspondance texte/signal vocal qui facilite l'analyse phonétique (EasyAlign[84], MAUS[85], etc.) tandis que des outils plus analytiques et des applications logicielles s'appuient sur PRAAT (Boersma et Weenink, 2009). Il résulte de la collaboration entre technologies de la langue et linguistique un partage avec la communauté des linguistes de corpus élaborés à des fins technologiques. Les corpus ESTER (Galliano *et al.*, 2009) et ETAPE (Gravier *et al.*, 2012) en sont des exemples, construits pour l'évaluation de campagnes françaises de reconnaissances vocales dont l'acquisition et l'exploration ont pu être exploitées avantageusement par des linguistes (phonéticiens, phonologues, sociolinguistes). Ces ressources s'avèrent particulièrement utiles car elles bénéficient d'une transcription manuelle et d'un alignement son/texte qui facilitent grandement la tâche d'analyse linguistique.

4.2 Projets, initiatives, acteurs

Recherche

La France compte environ 40 à 50 équipes de recherche actives sur la scène du TAL et quelques laboratoires en Belgique (Uni. Louvain-la-Neuve, Uni. Mons, Multitel lab), en Suisse (Uni. Genève, Uni. Lausanne, École Polytechnique Fédérale de Lausanne, l'IDIAP de Martigny), et au Canada (Montréal, Laval, Ottawa), contribuent aussi activement à la recherche sur les TL pour la langue française. Une initiative récente au Québec tente de mieux structurer la recherche en TAL sous l'égide du nouveau consortium CLIQ-ai[86]. Les équipes implantées en France sont réparties sur l'ensemble du territoire, et dans la plupart des cas affiliées à des universités, avec une éventuelle affiliation mixte avec le CNRS, ainsi que (dans de plus rares cas) avec l'Inria[87]. En revanche, la recherche en TAL est presque entièrement absente du périmètre des grandes écoles d'ingénieurs françaises (*Grandes Ecoles*), à l'exception notable de Télécom Paris à Palaiseau, et l'Eurecom à Sophia Antipolis. Les centres les plus importants sont établis à Aix-Marseille, Avignon, Besançon, Grenoble, Le Mans, Nancy, Nantes, Paris, Orsay, Rennes, Sophia-Antipolis, Strasbourg et Toulouse. Historiquement, les laboratoires accueillant la recherche en TAL ont mis principalement l'accent sur l'informatique ou le traitement du signal, d'une part, soit sur la linguistique et les sciences du langage, d'autre part, avec très peu de chevauchement ou de centres véritablement multidisciplinaires. Cette activité est structurée au niveau national par le CNRS, qui porte deux réseaux nationaux de recherche sur les technologies et ressources langagières[88], avec des thèmes et des objectifs complémentaires. Outre le CNRS et l'Inria, deux autres institutions nationales de recherche mènent une activité importante dans le domaine du TAL : le CEA-LIST à Palaiseau et l'INRAe à Jouy-en-Josas.

La recherche multidisciplinaire sur les TL est accompagnée pour l'essentiel par trois sociétés savantes dédiées (l'ATALA[89], principalement autour de l'étude de la langue écrite, l'AFCP [90], celle de la langue parlée, et l'ARIA[91], spécialisée dans les travaux sur la recherche d'information), ainsi que par l'AFIA[92], plus généraliste. Conjointement ou séparément, elles organisent des manifestations scientifiques annuelles ou bisannuelles : les "*Journées d'Études sur la parole*" (JEP, depuis 1970), la conférence "*Traitement automatique des langues naturelles*" (TALN, depuis 1994), et la "*Conférence en Recherche d'Information et Applications*" (CORIA, à partir de 2004), qui continuent d'attirer une population diversifiée de chercheurs issus de tous les sous-domaines des TL. Les travaux de recherche sont diffusés dans des revues comme "*Traitement automatique des langues*" (TAL) et "*Corpus*", ainsi que dans "*Discours*" ou encore la "*Revue Française de linguistique appliquée*".

Dans tous ces milieux, l'élaboration, la diffusion et la documentation de ressources et d'outils linguistiques sont désormais une pratique bien développée : dans une analyse sur 15 années de comptes-rendus dans les actes de la conférence LREC, Mariani *et al.* (2014) mentionnent que plus de 10% des articles sont co-écrits par une équipe française ; les résultats de ces travaux concrétisés dans des corpus, des modèles et des outils sont également bien présents dans les bases de données de l'ELG. Toutes ces activités font

aujourd'hui l'objet d'une reconnaissance plus équitable dans les évaluations académiques.

Les entreprises et industries de la langue

En raison de son rôle en tant que langue internationale, de la taille relativement importante et du développement avancé des marchés francophones, le français est relativement bien couvert par les prestataires internationaux de services en TL et dans le secteur des langues. Ainsi, le français et l'anglais ont formé parmi les premiers couples de langues à être traduits sur internet et des versions françaises de Siri, Amazon Echo et Google Home existent depuis de nombreuses années. Cela signifie que le développement des technologies portant sur le français dépasse les frontières de la France ou dans des autres pays francophones.

Pour disposer d'une vision d'ensemble, on peut regrouper les acteurs industriels présents dans le paysage français en trois grandes catégories. La première compte les fournisseurs internationaux de technologies actifs dans le développement et la commercialisation de solutions basées sur l'Intelligence Artificielle, avec des technologies linguistiques tombant dans le périmètre de leurs filiales ou succursales françaises comme Apple, Fujitsu, Huawei, IBM, Google/DeepMind, META/FAIR, Microsoft, NaverLabs, Samsung, Sony ; Orange, Thales et Dassault-System font également partie de ce groupe. La politique de diffusion en licence logicielle ouverte de la part de certains de ces acteurs a permis de proposer des outils et des ressources multilingues de grande envergure qui sont également utiles pour la langue française, comme l'illustre la collection de ressources `fairseq`[93] du centre META/FAIR.

Deuxième famille d'acteurs, dont la part est plus difficile à évaluer : celles de grandes entreprises extérieures du secteur informatique qui développent ou intègrent des technologies de la langue pour leurs besoins et produits internes. Des entreprises de divers secteurs industriels comme Airbus, l'AFP, EDF, Engie, Renault, Sanofi, la SNCF, la Société Générale, BNP Paribas, ainsi que de nombreuses entreprises de plus petite taille du secteur des services et des médias font montre de leur intérêt pour les TL, même si leur contribution nette actuelle au paysage des ressources reste comparativement modeste.

Le troisième groupe, de loin le plus important, rassemble les PME élaborant des services et des logiciels dédiés. Ce groupe s'est développé très rapidement au cours des dernières années et réunit un éventail d'acteurs "historiques" tels que Systran et Reverso (TA), Druide et Synapse (correction orthographique), Synomia, Sinequa ou Pertimm (moteurs de recherche), qu'accompagne une variété de très petites entreprises et de start-ups qui sont apparues avec les progrès des technologies en IA. La plupart des domaines applicatifs, allant de la recherche d'information (Qwant, Exalead-DS) à la génération de texte (Syllabs), de la transcription vocale (Vocapia Research) à la reconnaissance optique de caractères et au traitement de documents (Jouve, A2IA), de l'analyse de texte (Expert Systems) à la synthèse vocale (Acapela group, Voxygen) sont pris en charge, avec une recrudescence récente d'acteurs développant des agents conversationnels, notamment pour la relation client (Davi, Hellomybot.io, Julie Desk, Konverso, Kwalys, Linagora, ViaDialog, Vivoka, Zaion, etc). Grâce à des années de projets en collaboration avec des équipes universitaires, ces acteurs ont contribué au développement d'une myriade de ressources

et d'outils, dont certains sont accessibles au public ou distribués sous licence logicielle libre.

Instruments et plateformes pour le partage des ressources

Les activités de recherche en linguistique peuvent s'appuyer sur Huma-Num[94], une infrastructure nationale de recherche dédiée aux sciences humaines et sociales et aux humanités numériques, mise en œuvre par le ministère de l'Enseignement supérieur et de la Recherche et soutenue par le Centre National de la Recherche Scientifique, Aix-Marseille Université et le Campus Condorcet, grâce à laquelle les linguistes peuvent accéder à la fois à des outils et des données. Huma-Num affiche pour objectif de développer, mettre en œuvre et préserver les programmes de recherche - leurs données et leurs outils - sur le long terme dans un contexte de science ouverte et de partage des données. À cela viennent s'ajouter des initiatives comme ORTOLANG[95], un "EquipEx" (pour "Equipement d'Excellence") sélectionné dans le cadre du "*Programme Investissements d'Avenir*" (PIA) et dont l'objectif est de fournir des services linguistiques spécialisés, en complément d'Huma-Num. Dans le domaine de l'acquisition, du traitement et du partage de données, des projets comme le consortium CORLI (Corpus, Languages, Interactions) au sein de la structure Huma-Num contribuent à diffuser des corpus, des outils ainsi que des méthodes de travail et d'exploration de ces corpus. Les services proposés recouvrent une aide financière et méthodologique pour enrichir/finaliser des corpus, des formations consacrées aux nouvelles méthodologies d'investigation de différents corpus linguistiques, ou encore des ateliers et des écoles d'été pour les étudiants désireux d'améliorer leurs compétences en matière de traitement de corpus numériques.

Projets nationaux dans le domaine des TL

On peut noter l'absence d'appels récurrents dédiés aux TL en France ces dernières années. Quelques appels ont été lancés au cours des dix années écoulées, parmi lesquels on peut noter les appels à projets "*Langues et Numérique*" pilotés par la Délégation générale à la langue française et aux langues de France (DGLFLF) avec le soutien du Secrétaire d'État chargé du numérique et de l'innovation (2017 et 2018) avec un financement total d'environ 500k€ pour chaque appel[96]. Dans le cadre du Programme National de Recherche en Intelligence Artificielle (PNRIA)[97] lancé en 2018, doté d'un budget de 1,5 milliard d'euros sur fonds publics, un réseau de 4 instituts opérationnels Interdisciplinaires d'Intelligence Artificielle (3IA) a été lancé, où la langue est en partie présente ; sur les quelque 190 chaires de recherche et d'enseignement associées à ces 3IA ou financées par des appels ultérieurs, 10 à 15 portent sur l'informatique linguistique et les technologies de la langue. Des programmes liés à l'IA ont également été lancés en Belgique, notamment en Wallonie francophone sous la bannière "DigitalWallonia4.ai", et plus récemment, par le biais des *trusted AI Labs* (TRAIL[98]).

La plupart des projets concernant les TL sont financés par l'ANR (Agence Nationale de la Recherche) dans le cadre de l'appel à projets générique qui a lieu chaque année. Si certains axes de recherche, parmi la cinquantaine existante, mentionnent explicitement

le traitement du langage, il faut noter qu'aucun n'est spécifiquement dédié aux TL. Il en résulte une grande disparité dans le nombre de projets et dans le montant des sommes consacrées aux TL (tous instruments financés par l'ANR confondus) : 9 projets pour un montant de 2,64 M€ en 2021, 6 projets pour un montant de 2,10 M€ en 2020 (en partie en raison d'une réaffectation à des thématiques spécifiques de la COVID), 21 projets pour un montant de 7,90 M€ en 2019, 11 projets pour 3,98 M€ en 2018. Cette grande variabilité (du simple au quadruple) d'une année à l'autre, n'est pas favorable à la planification, à l'embauche régulière de doctorants et de post-doctorants. Les spécialistes de linguistique appliquée peuvent également obtenir un financement de l'ANR par le biais d'appels à la fois généraux et collaboratifs.

Il existe également des financements régionaux tels que ceux fournis par les LabEx ("*laboratoires d'excellence*") et RNMSH[99], parfois dans le cadre d'appels interdisciplinaires à orientation technologique et linguistique comme l'initiative développée par la MSH de Paris-Saclay et l'institut DATAIA[100].

Une autre possibilité pour des projets applicatifs de grande envergure s'est présentée sous la forme de programmes nationaux dédiés financés par le Ministère de l'Industrie ou le Ministère de la Défense par l'intermédiaire de Bpifranbce [101] : ces dernières années, des projets comme ROSETTA (sur le sous-titrage et la génération en langue des signes), Linto (un assistant intelligent en licence logicielle libre), et le *Voice Lab* (collecte de données vocales) ont bénéficié de financements importants. Cependant, l'ensemble des montants distribués par ces canaux sont difficiles à consolider.

Enfin, les recherches en TL, comme dans d'autres sous-domaines de l'IA, ont largement bénéficié du déploiement de la plateforme Jean Zay[102] (à partir de 2019), une infrastructure ouverte de calcul haute performance hébergeant des milliers de cartes GPU modernes. Sans celles-ci, plusieurs des projets à grande échelle précités n'auraient pas été possibles.

5 Les technologies de la langue et la langue des signes française

Les langues des signes (LS) sont des langues sous-dotées en ressources : très peu d'ouvrages de référence, une connaissance partielle de la grammaire, lexiques ou corpus limités, très peu de technologies sur ces langues, très peu de ressources en général. Par ailleurs, les travaux de recherche en traitement automatique sont beaucoup plus récents que pour les langues parlées ou écrites, et bien que la recherche dans ce domaine soit active tant en matière de reconnaissance, de génération que de traduction, il n'existe pas encore d'outils utilisables, à l'exception de quelques rares produits mais qui ne sont *a priori* pas complètement automatiques.

5.1 Disponibilité des données et des outils linguistiques

Un document récent[103], qui constitue un livrable du projet européen EASIER[104], recense les ressources linguistiques qui peuvent être utilisées pour le traitement des LS et indique

dans quelles conditions le public y a accès. Plus précisément, il répertorie :

- les corpus dans les langues des signes européennes de taille "substantielle" (pour les LS) qui peuvent être utilisés comme données d'entraînement de haute qualité pour la traduction automatique,
- les tâches de collecte de données utilisées dans plus d'un de ces corpus linguistiques,
- les ressources lexicales des langues des signes européennes.

Il existe deux types de ressources : les corpus linguistiques et les données télédiffusées. Les données télédiffusées sont disponibles en quantité relativement importante mais comporte généralement des sous-titres synchronisés avec la parole, tandis que les corpus linguistiques offrent des données de qualité élevée grâce à une transcription riche et à une annotation linguistique mais sont rares et de taille réduite. Une autre différence importante tient au fait que la première est une interprétation en direct, et donc une langue des signes soumise aux contraintes de temporalité et de la structure du discours oral, alors que la seconde est une langue des signes produite durectement par des signeurs Sourds dont c'est la langue première.

Comme il n'existe pas encore jusqu'ici de technologie qui permette d'étiqueter ou d'annoter automatiquement les données en LS avec la qualité exigée pour une annotation linguistique, les créateurs de corpus ont dû assumer la difficulté de devoir annoter manuellement ces données. Comme nous l'avons déjà indiqué, les langues des signes ne disposent pas d'un système d'écriture normalisé, ni même d'un système graphique de transcription (équivalent à l'alphabet phonétique international), et à ce jour, des conventions d'annotation différentes sont utilisées sur l'ensemble des différents corpus.

Pour la langue des signes française (LSF), les principaux dictionnaires monolingues et ressources terminologiques sont les suivants :

- Ocelles,[105] un site collaboratif entièrement bilingue français-LSF qui rassemble signes, définitions, informations sur des projets et organismes à titre de ressource pédagogique. Pour chaque concept, au moins une définition et souvent des exemples associés dans différents domaines de connaissance sont proposés. Les utilisateurs peuvent téléverser des informations (p. ex. textes, images, vidéos, présentation) qui sont examinés par des experts sur la forme et le contenu avant d'être mis en ligne.
- Sign'Maths[106], un glossaire consacré aux mathématiques. Le groupe Sign'Maths mène ses recherches en organisant des ateliers mensuels en LSF, réunissant des professeurs de mathématiques et de LSF, de l'enseignement primaire, secondaire et supérieur, mais aussi des étudiants Sourds et des interprètes. Ils examinent divers concepts mathématiques, domaine par domaine, jusqu'à ce qu'une unité lexicale (un signe) spécifique soit établie, répondant à la fois aux contraintes linguistiques et aux critères mathématiques.
- Elix[107], un dictionnaire bilingue LSF/Français fonctionnant sous forme de moteur de recherche. La recherche peut s'effectuer par mots-clés français, les résultats

listent les signes correspondants et leur définition en LSF. Elix peut être utilisé comme plateforme web en ligne et comme application. À ce jour, il contient plus de 21 000 définitions françaises traduites en LSF et plus de 15 300 signes.

- le Lexique Dicta Sign [108], un lexique multilingue pour les langues des signes britannique, grecque, allemande et française, l'anglais, le grec, l'allemand et le français. Environ 1 000 concepts sont fournis pour chacune des LS du projet. La liste commune des concepts choisis pour le lexique correspond à l'usage quotidien ou est spécifiquement liée au domaine des voyages en Europe.

À ce jour, les trois principaux corpus de LSF dont on dispose sont :

- CREAGEST, un corpus constitué d'une part de Langue des Signes Française (LSF) d'adultes et d'enfants et d'autre part de gestes naturels. Il se décompose en trois sous-corpus : un premier d'enfants portant sur l'acquisition, un autre de dialogues entre adultes Sourds et un troisième de gestes naturels. Pour l'acquisition des données, 65 enfants Sourds et 17 adultes sourds ont été enregistrés par quatre enquêteurs Sourds. Pour la base de données de dialogues, 51 entretiens ont été menés par quatre enquêteurs Sourds. Pour les données gestuelles, celles-ci sont constituées d'enregistrements recueillis auprès de cinq binômes entendants-entendants, Sourds-Sourds et Sourds-entendants. Au total, plus de 500 heures produites par plus de 250 signeurs ont été enregistrées. À ce jour, seule une petite partie du corpus (1 heure), disponible sur le site web d'Ortolang[109], a été annotée.

- Dicta-Sign-LSF-v2, une version étendue du sous-corpus LSF du corpus créé au cours du projet européen Dicta-Sign, fournissant les données primaires (vidéos), les données d'élicitation et d'annotation avec un guide d'annotation associé, ainsi que des données prétraitées sur les signeurs comprenant les expressions faciales et de la partie supérieure du corps et des estimations de la forme des mains. Il contient neuf sessions de dialogue avec 18 signeurs en LSF sur le thème des voyages en Europe. Les données ont été annotées de manière détaillée et un réseau d'apprentissage convolutif récurrent a été entraîné sur celles-ci, en s'appuyant sur une modélisation compacte et généralisable des signeurs pour obtenir une base de référence pour la reconnaissance des signes lexicaux et des structures non-lexicales. Il contient 11 heures de vidéos annotées et traduites (en français).[110]

- Mediapi-Skel, un corpus de LSF sous forme de squelette 2D en LSF avec sous-titres en français. Le corpus se compose de 368 vidéos sous-titrées produites par Média'Pi[111], un média en ligne qui produit des contenus bilingues en LSF et en français écrit. Il contient environ 27 heures de LSF et 17 000 tokens provenant de sous-titres.[112]

5.2 Projets, initiatives, acteurs

En France, il n'existe pas de programme national, d'infrastructures ou de fournisseurs de TL en lien avec des technologies portant sur la LSF ou les langues des signes. Certains

projets de recherche sont financés par des agences comme l'ANR ou la DGLFLF. Depuis une période plus récente, des projets collaboratifs public/privé sont financés par Bpifrance. Sur les 5 dernières années et actuellement, les principaux projets qui impliquent des technologies sur les langues des signes sont les suivants :

- ROSETTA[113], un projet public/privé français qui a étudié les solutions d'accès aux contenus audiovisuels, qui inclut une étude exploratoire liée à la traduction automatique de sous-titres en LSF affichés par des avatars signeurs (ou signeurs virtuels).
- *Serveur Gestuel*, un projet public/privé français qui vise à créer un serveur gestuel, c'est-à-dire l'équivalent d'un serveur vocal mais en LSF, intégrant donc des technologies de reconnaissance et de génération.

6 Comparaison inter-langues

Cette section a été rédigée par les porteurs du projet européen ELE et n'a pas fait l'objet d'une traduction vers le français.

7 Résumé et conclusions

7.1 Observations générales

La recherche et le développement dans les TL et en IA pour la langue française sont bien avancés et une large gamme d'applications en TL à destination du grand public sont disponibles, et pénètrent également de plus en plus dans l'univers de l'entreprise et de l'industrie. Il s'agit notamment des systèmes de dialogue et d'agents virtuels, de technologies embarquées comme les claviers intelligents, la reconnaissance automatique de la parole, l'analyse et la compréhension du langage, les systèmes de questions-réponses, la synthèse vocale ou encore, pour d'autres usages, la traduction automatique, qui sont tous destinés au consommateur francophone. Il convient de noter que ces applications recourent souvent à des composants en licence logicielle ouverte, ce qui réduit le coût du développement des nouveaux services et logiciels.

Des ressources de moyennes à grandes tailles ont été progressivement développées pour la langue française et sont facilement accessibles pour toutes ces tâches et applications importantes des technologies de TAL. Nombre de ces ressources ont été collectées de manière opportuniste dans le cadre de projets de recherche ou en imitant des ressources similaires développées pour l'anglais, grâce à un intérêt accru pour le TAL multilingue, ou par le biais de tentatives coordonnées de collecte de corpus d'entraînement et de test parallèles ou comparables pour de nombreuses tâches (p. ex., dans le projet Universal Dependency). Bien que les bases de données utilisables pour étudier des problèmes algorithmiques et de calcul, ou pour évaluer la performance de certains outils de traitement de la langue pour le français ne manquent pas, elles sont loin de couvrir toutes les variétés linguistiques, tous les genres textuels, registres ou domaines. Cela reste un problème à la

fois pour les travaux de recherche en linguistique et pour le développement d'applications industrielles en TAL pour le français.

Avec l'essor des technologies à base d'IA, de nouvelles possibilités et de nouveaux acteurs industriels sont récemment apparus, élargissant ainsi l'éventail des services, domaines et types de données pris en charge. La recherche académique sur les TL a moins directement bénéficié de ce développement rapide des investissements dans les applications d'IA. Il en résulte que l'on peut considérer que la langue française est raisonnablement bien dotée en technologies et peut se comparer plus ou moins favorablement à cet égard à d'autres langues européennes comme l'allemand, l'espagnol et l'italien. Cependant, l'écart sur le plan de la profondeur, de la couverture et de la qualité des outils existants, observé il y a dix ans avec l'anglais, continue à se creuser, amplifiant les inégalités linguistiques entre anglophones et non-anglophones, ces derniers bénéficiant de meilleurs services et applications en TL.

Comme nous l'avons abordé dans la section 3, une avancée récente dans les TL s'est matérialisée dans le fantastique succès qu'a représenté l'arrivée en force des méthodologies basées sur l'apprentissage automatique, méthodologies qui s'appliquent (pratiquement) aussi bien à l'anglais qu'au français (et à de nombreuses autres langues). Ces méthodes s'appuient sur (a) des sources massives de données linguistiques, qu'elles soient textuelles, audio ou multimodales, qui peuvent servir à l'auto-entraînement de représentations lexicales (multilingues) généralistes de grande échelle ; (b) des corpus d'entraînement et d'évaluation bien conçus et nettoyés destinés à un éventail d'applications le plus large possible. En outre, ces réussites ont pu entretenir l'illusion que (c) la nécessité de développer des outils spécifiques à la langue n'importait pas.

En ce qui concerne le point (a), notre analyse a montré que les ressources disponibles pour le français étaient d'un ordre de grandeur plus réduit et moins diversifié que pour l'anglais. Ayant accumulé et annoté de gigantesques répertoires de données linguistiques (y compris une partie substantielle du français), et ayant développé des outils de traitement ML optimisés et des infrastructures informatiques à grande échelle, certains acteurs internationaux (p. ex. aux Etats-Unis ou en Chine) sont maintenant en mesure de développer des modèles, puis des outils et des services qui sont hors de portée de la capacité de développement des acteurs locaux qui n'utiliseraient que des ressources publiques. Grâce à l'ouverture de certains acteurs privés, des modèles et des échantillons de ces ensembles de données ont été mis à la disposition de la recherche et annotés (p. ex. pour l'analyse de sentiment ou la détection des contenus haineux) : ces ressources sont toutefois fragiles et leur caractère ouvert peut changer de manière inopinée.

En ce qui concerne le point (b), notre étude a montré que les ressources françaises disponibles ne couvrent pas, il s'en faut de beaucoup, le spectre complet des applications, des domaines, des genres et des modalités. Comme indiqué ci-dessus, dans les situations où les données d'apprentissage ne sont pas disponibles, l'apprentissage par transfert (à partir de l'anglais) ou l'utilisation de la traduction automatique était parfois une alternative viable, qui s'accompagne souvent d'une perte de performance ou d'effets secondaires indésirables (biais). Cette perte est elle-même rarement signalée ou documentée, sauf lorsqu'elle a des implications sociales ou éthiques visibles, comme lorsqu'un moteur de reconnaissance vocale fait plus d'erreurs pour les voix avec accent que pour les voix

parlant un français standard. Cela ne signifie pas que ces différences de performances n'existent pas, mais qu'elles ne sont généralement pas mesurées dans le cadre de campagnes d'évaluation ouvertes, en raison de l'absence de jeux de tests appropriées. En ce qui concerne les ressources linguistiques, la conception de nouveaux jeux d'essai pour les TL est soumise aux choix des organisateurs de l'évaluation qui, dans un paysage de plus en plus multilingue, peuvent ne pas considérer le français comme une langue intéressante au motif qu'il s'agit déjà d'une langue très bien dotée, qui partage de nombreuses similitudes linguistiques avec l'anglais, etc.

Ceci nous amène au point (c) : en raison de cette proximité avec l'anglais, de nombreuses ressources existantes pour l'anglais (écrit) peuvent être utiles pour améliorer le traitement du français. C'est à la fois une bonne chose, mais entretient aussi l'illusion que les deux langues peuvent être traitées de manière analogue, malgré d'importantes différences linguistiques et culturelles qui doivent parfois être examinées avec soin (p. ex. pour des tâches comme la détection d'émotions ou de sarcasmes).

Dernière observation : les ressources et outils disponibles identifiés dans notre étude[114] sont éparpillés sur de multiples plateformes et souffrent de l'absence d'un inventaire centralisé, associé à des méta-données détaillées, malgré de nombreuses tentatives pour offrir un tel service, par le biais d'institutions telles que CLARIN, de projets tels que FlareNet, ou d'initiatives louables comme LRE Map[115], ISLRN[116] ou le site *European Language Grid* (ELG) récemment lancé. L'association ELRA (*European Language Resource Association*), installée en France, a été un acteur majeur dans ces démarches et a largement contribué aux inventaires de ressources existants. Observons que, ces dernières années, ces initiatives se sont vues bousculées par le développement d'acteurs industriels et par le mouvement de la science ouverte, qui ont conduit à une floraison d'initiatives visant des objectifs similaires (p. ex. les bibliothèques de données de Huggingface, de `tensorflow` ou du site Paperwithcode[117]).

Plusieurs des observations formulées plus haut s'appliquent également à la LSF qu'il faut, par ailleurs, considérer comme une langue faiblement dotée et pour laquelle la production de ressources et le développement d'outils restent très rares et dispersés.

7.2 Quelques recommandations

7.2.1 Concernant les ressources et la collecte de données

De nombreux corpus français existant dans le domaine ouvert sont le résultat d'initiatives non coordonnées et couvrent par conséquent les besoins d'applications de domaines spécifiques, avec leurs annotations, leurs méta-données et leur finalité propres. Cet état de fait se traduit par (a) un manque de visibilité des outils et des données qui ne sont connus que de communautés restreintes, bien que la situation se soit grandement améliorée grâce aux infrastructures européennes et nationales susmentionnées, (b) un manque d'initiatives interdisciplinaires soutenues au niveau national et au-delà.

Au vu de la dispersion et de l'hétérogénéité des jeux de données existants, une première recommandation serait d'institutionnaliser des politiques plus claires et des incitations à la déclaration et à l'archivage des ressources linguistiques pour le français lorsqu'elles

sont produites par des projets de recherche financés par des fonds publics, comme cela est déjà fait pour les publications scientifiques sur la plateforme HAL. Disposer d'un point d'entrée bien identifié pour les ressources linguistiques, les modèles et les outils pour le français, et leur documentation associée, serait d'une grande importance pour de nombreuses disciplines scientifiques, notamment dans le domaine des sciences humaines et sociales (SHS), ainsi que pour de nombreuses entreprises industrielles.

La taille pose également souvent un problème, compte tenu des besoins croissants des logiciels d'apprentissage automatique avides de données. Une réponse possible consisterait à ouvrir les grandes bases de données produites par l'administration et d'autres institutions publiques (p. ex. dans les domaines de la santé, de la culture, des médias, de la justice ou de l'éducation) qui restent enfouis et difficiles d'accès - parfois pour des raisons légitimes (problèmes de confidentialité, droits d'auteur imprécis). Des incitations et des politiques publiques devraient être mises en place pour poursuivre et amplifier les actions entreprises dans le cadre du programme européen CEF/ELRC[118], avec le développement de dépôts publics avec des règles d'accès claires analogues à celles qui existent ou celles en vigueur pour les données de santé et pour les articles scientifiques accessibles à travers l'initiative ISTEX[119]. Là où les données sont absentes actuellement et où des lacunes ont été identifiées, il est essentiel de continuer à soutenir le développement et l'annotation de nouvelles ressources ouvertes à grande échelle. C'est en particulier le cas pour la langue des signes française, qui pourrait grandement bénéficier d'initiatives similaire au DGS-Korpus allemand[120].

Les applications qui font appel aux données des réseaux sociaux nécessitent des actions spécifiques, car elles sont souvent associées à des questions juridiques délicates (liées aux droits de la propriété intellectuelle et industrielle ou à ceux relatifs aux informations personnelles) qui limitent leur diffusion et leur exploitation ultérieure. La recherche sur la langue française peut être ici trop dépendante de la politique de données actuelle des détenteurs de contenus qui entrave le développement de recherches sur la prospection d'opinions, la détection d'infox et des discours de haine, la vérification des faits, sur les biais et les questions éthiques, pour n'en citer que quelques-unes. Deux problématiques sont en jeu : (a) sécuriser l'accès aux données sensibles à des fins de recherche ; (b) faciliter la diffusion des bases de données et des modèles produits publiquement.

Ce manque de coordination observé laisse finalement entrevoir la nécessité de définir une feuille de route stratégique pour identifier, construire, assurer la conservation, annoter et sécuriser les ressources pour les variétés ou les domaines critiques pour la recherche, l'industrie ou l'administration dans chaque pays francophone, à partir d'une analyse précise des lacunes dans les jeux de données existants (certaines ont été évoquées plus haut). Cette feuille de route devrait également identifier les scénarios dans lesquels les ressources pourraient être transférées à partir de leur équivalent anglais, et veiller à ce que les technologies de traduction de haute qualité nécessaires soient largement disponibles en tant que ressource publique.

Notons que le même argument vaut pour d'autres langues pour lesquelles il pourrait être approprié de transférer des ressources *à partir du français* – c'est par exemple le cas de nombreuses langues à peu dotées en ressources linguistiques, comme les langues régionales, ou les langues qui ont historiquement coexisté avec le français dans diverses

régions du monde. La collecte de ressources et le développement de la TA pour traduire ces langues depuis et vers le français sont donc susceptibles d'entraîner des retombées importantes pour les deux parties.

Enfin, avec un passage des données accru à l'échelle, il sera également essentiel de continuer à soutenir les infrastructures informatiques publiques, avec des règles d'accès facilitées, pour la recherche et les entreprises (startups, PME).

7.2.2 Concernant l'évaluation

Les TL étant intégrées dans un nombre croissant d'applications utilisées couramment par le grand public, dans un nombre de tâches de plus en plus variées, il devient d'autant plus important d'évaluer de manière ouverte et de faire connaître les performances des systèmes existants dans des conditions réelles, de diagnostiquer leurs biais potentiels et de mieux documenter leurs défauts et leurs effets néfastes potentiels. Des actions visant à garantir que des campagnes d'évaluation ciblant spécifiquement le français pour un nombre suffisamment important d'applications et de domaines soient organisées régulièrement devraient être entreprises, autant de fois que possible en coordination avec les campagnes d'évaluation internationales afin d'accroître leur visibilité et la participation. Étant donné que des systèmes de base "génériques" et des protocoles d'évaluation existent pour de nombreuses tâches, on estime que de telles évaluations pourraient être effectuées à un coût réduit, nécessitant principalement la création de nouvelles données d'essai pour des cas d'utilisation réalistes.

Une source d'inspiration pourrait être les évaluations systématiques des TL pour le français entreprises sous l'égide de l'Aupelf/AUF ("*AgenceUniversitaire de la Francophonie*", Association des Universités Francophones) (dans les années 90) ou dans le cadre du programme TechnoLangue (au début des années 2000)[121]. Ces actions ont réussi à consolider les procédures d'évaluation et à favoriser la création de données d'évaluation annotées ; quinze ans plus tard, le besoin de systèmes d'étalonnage qui pourraient aider à mieux analyser et diagnostiquer les biais et les limites des TL actuelles est toujours aussi pressant.

7.2.3 Autres priorités en recherche

Comme indiqué précédemment, le développement de TL dédiées au traitement du français et de la langue des signes française nécessite encore de stimuler le développement de la recherche fondamentale et de nouvelles ressources. Outre les thèmes déjà évoqués, on peut notamment citer : (a) les algorithmes et technologies d'analyse profonde des langues (y compris les corpus arborés de grande échelle contenant des informations sémantiques et annotées au niveau du discours pour une multiplicité de genres, de domaines et de tâches), dans le but de parvenir à une compréhension profonde des langues pour certaines applications représentatives ; un exemple pourrait être le développement d'agents collaboratifs, capables d'interactions sociales par le biais du langage, mais également dotés de capacités d'apprentissage, de raisonnement et de résolution de problèmes ; (b) des ressources multimodales pour l'étude de l'émergence non supervisée des compétences

langagières et de leur développement par le biais d'interactions et d'ancrage situationnel ; (c) des ressources et des outils pour des recherches dans le traitement informatique du langage pathologique. En grande partie, ces recherches sont pluridisciplinaires par essence, et devraient être menées avec les communautés de recherche concernées.

Remerciements

Les auteurs souhaitent remercier I. Aldabe, F. Béchet, B. Daille, T. François, T. Hueber, J. Mariani, P. Müller, B. Sagot, V. Vandeghinste pour leurs commentaires et suggestions sur la première version de ce rapport. Ils expriment également leur gratitude envers l'aide que leur ont prodigué M. Wang-Castejon et L. Khennou qui ont participé au recueil de certaines des sources de données et de la documentation utilisées dans cette enquête. Ils remercient enfin chaleureusement V. Arranz (ELRA) pour sa contribution à l'inventaire des ressources linguistiques.

Bibliographie

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 et Aitor Soroa : Report on the state of the art in language technology and language-centric AI. Rapport technique, European Language Equality, Deliverable D1.2, 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/10/ELE_Deliverable_D1_2.pdf. [cité page 42]

Itziar Aldabe, Georg Rehm, German Rigau et Andy Way : Report on existing strategic documents and projects in LT/AI. Rapport technique, European Language Equality, Deliverable D3.1, 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf. [cité page 42]

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major et 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. URL https://doi.org/10.1145/3442188.3445922. [cité page 18]

Paul Boersma et David Weenink : Praat : doing phonetics by computer (version 5.1.13), 2009. URL http://www.praat.org. [cité page 23]

Hélène Bonneau-Maynard, Alexandre Denis, Frédéric Béchet, Laurence Devillers, Fabrice Lefèvre, Matthieu Quignard, Sophie Rosset et Jeanne Villaneau :

Media : évaluation de la compréhension dans les systèmes de dialogue. *In* Stéphane Chaudiron et Khalid Choukri, éditeurs : *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. [cité page 16]

Andrew D. Booth et William N. Locke, éditeurs. *Machine translation of languages : fourteen essays*. Press of the Massachusetts Institute of Technology, Cambridge, Massachusetts, 1955. [cité page 9]

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Yannick Parmentier, Caroline Pasquer et Jean-Yves Antoine : Annotation d'expressions polylexicales verbales en français. *In* Jean-Yves Antoine Iris Eshkol, éditeur : *24e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Actes de TALN, volume 2 : articles courts, pages 1–9, Orléans, France, juin 2017. URL https://hal.archives-ouvertes.fr/hal-01537880. [cité page 15]

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi et 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. [cité page 17]

Noam Chomsky : *Syntactic structures*. The Hague : Mouton, 1957. [cité page 9]

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. [cité page 4], [cité page 5]

Marcel Cori et Jacqueline Léon : La constitution du TAL. *Revue TAL*, 43(3):21–55, 2002. URL https://halshs.archives-ouvertes.fr/halshs-00158854. [cité page 9]

Joaquim Brandão de Carvalho : Western romance in lenition and fortition. *In* Joaquim Brandão de Carvalho, Tobias Scheer et Philippe Ségéral, éditeurs : *Lenition and Fortition*. De Gruyter Mouton, 2008. ISBN 9783110211443. URL https://doi.org/10.1515/9783110211443. [cité page 3]

Eric Villemonte de la Clergerie, Olivier Hamon, Djamel Mostefa, Christelle Ayache, Patrick Paroubek et Anne Vilnat : PASSAGE : from French parser evaluation to large sized treebank. *In* Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis et Daniel Tapias, éditeurs : *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/. [cité page 15]

Martin d'Hoffschmidt, Maxime Vidal, Wacim Belblidia et Tom Brendlé : Fquad : French question answering dataset. *CoRR*, abs/2002.06071, 2020. URL https://arxiv.org/abs/2002.06071. [cité page 19]

Yoann Dupont : Un corpus libre, évolutif et versionné en entités nommées du français. TALN 2019 - Traitement Automatique des Langues Naturelles, juillet 2019. URL https://hal.archives-ouvertes.fr/hal-02448590. Poster. [cité page 16]

Moussa Kamal Eddine, Antoine J.-P. Tixier et Michalis Vazirgiannis : BARThez : a skilled pretrained French sequence-to-sequence model. *arXiv preprint arXiv :2010.12321*, 2020. [cité page 18]

Karën Fort, Bruno Guillaume et 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. URL https://doi.org/10.1145/2594776.2594777. [cité page 15]

Sylvain Galliano, Guillaume Gravier et Laura Chaubard : The Ester II evaluation campaign for the rich transcription of French radio broadcasts. *In Proceedings of InterSpeech*, 2009. [cité page 16], [cité page 23]

Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel et 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. [cité page 23]

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco et 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. [cité page 17]

Rachel Keraron, Guillaume Lancrenon, Mathilde Bras, Frédéric Allary, Gilles Moyse, Thomas Scialom, Edmundo-Pavel Soriano-Morales et 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, mai 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.673. [cité page 19]

Taku Kudo et 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, novembre 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-2012. [cité page 14]

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier et 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, mai 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.302. [cité page 13]

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. [cité page 6]

Mark Liberman : Corpus phonetics. *Annual Review of Linguistics*, 5(1):91–107, 2019. [cité page 22]

Joseph Mariani, Patrick Paroubek, Gil Francopoulo et 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 et Stelios Piperidis, éditeurs : *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. [cité page 24]

Joseph Mariani, Patrick Paroubek, Gil Francopoulo, Aurélien Max, François Yvon et 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. [cité page 12]

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah et Benoît Sagot : CamemBERT : a Tasty French Language Model. *In ACL 2020 - 58th Annual Meeting of the Association for Computational Linguistics*, Seattle / Virtual, United States, juillet 2020. URL https://hal.inria.fr/hal-02889805. [cité page 13]

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. [cité page 14]

Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary et 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, mai 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.569. [cité page 16]

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 et 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, septembre 2008. URL https://hal.archives-ouvertes.fr/hal-00765334. [cité page 21]

Rosalind W Picard : *Affective computing*. 2000. [cité page 22]

Daniel Pimienta : étude sur la présence de la langue française dans le cyberspace. Rapport technique 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. [cité page 5]

Barbara Plank, Anders Johannsen et 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. [cité page 14]

Karl Pusch : The Romance languages : Typology. *In* Bernd Kortmann et Johan van der Auwera, éditeurs : *The Languages and Linguistics of Europe, A Comprehensive Guide*, volume 1 de *The World of Linguistics*, pages 69–96. Berlin/New York : Mouton de Gruyter, 2011. [cité page 3], [cité page 4]

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. [cité page 14]

Benoît Sagot et Darja Fišer : Building a free French wordnet from multilingual resources. *In OntoLex*, Marrakech, Morocco, mai 2008. URL https://hal.inria.fr/inria-00614708. [cité page 15]

Benoît Sagot, Marion Richard et Rosa Stern : Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées. *In* Georges Antoniadis, Hervé Blanchon et Gilles Sérasset, éditeurs : *Traitement Automatique des Langues Naturelles (TALN)*, volume 2 - TALN de *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, Grenoble, France, juin 2012. URL https://hal.inria.fr/hal-00703108. [cité page 16]

Tanja Schultz, Michael Wand, Thomas Hueber, Dean J. Krusienski, Christian Herff et Jonathan S. Brumberg : Biosignal-based spoken communication : A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2257–2271, 2017. [cité page 20]

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski et 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, novembre 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.647. [cité page 18]

Rico Sennrich, Barry Haddow et 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, août 2016. URL https://www.aclweb.org/anthology/P16-1162. [cité page 14]

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. [cité page 21]

John Charles Smith : French and northern Gallo-Romance. *In* Adam Ledgeway et Martin Maiden, éditeurs : *The Oxford Guide to the Romance Languages*. Oxford University Press, 2016. [cité page 4]

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 et Stelios Piperidis, éditeurs : *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. [cité page 19]

Alan M. Turing : Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950. ISSN 0026-4423. URL https://doi.org/10.1093/mind/LIX.236.433. [cité page 9]

Henriette Walter : *Le français dans tous les sens*. Points - Le goût des mots. Seuil, 2016. [cité page 3]

Rodrigo Wilkens, Bruno Oberle, Frédéric Landragin et Amalia Todirascu : French coreference for spoken and written language. *In Proceedings of the 12th Language Resources and Evaluation Conference*, pages 80–89, Marseille, France, mai 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.10. [cité page 16]

Yorick Wilks : The history of natural language processing and machine translation. *Encyclopedia of Language and Linguistics*, 2005. [cité page 9]

Victor H Yngve : A model and an hypothesis for language structure. *Proceedings of the American philosophical society*, 104(5):444–466, 1960. [cité page 9]

Amir Zeldes, Yang Janet Liu, Mikel Iruskieta, Philippe Muller, Chloé Braud et 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, novembre 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.disrpt-1.1. [cité page 16]

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre et 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. [cité page 15]

Notes

1. european-language-grid.eu
2. Les résultats de cette procédure de collecte d'informations seront intégrés dans la plateforme de la grille européenne des langues (*European Language Grid*, http://european-language-grid.eu) afin de pouvoir les étudier à l'aide de mécanismes avancés de recherche et de navigation, et de réaliser des visualisations comparatives entre langues.
3. http://european-language-equality.eu S'appuyant sur un large consortium composé de 52 partenaires couvrant tous les pays européens, la recherche et l'industrie et toutes les initiatives paneuropéennes majeures, le projet ELE détaille un ordre du jour stratégique de recherche, d'innovation et d'application ainsi qu'une feuille de route pour réaliser une égalité numérique totale des langues en Europe d'ici 2030.
4. Les estimations varient légèrement selon les sources et l'approche la plus prudente serait de ranger le français dans un groupe de langues qui compterait entre 250 et 300 millions de locuteurs, avec le russe, l'arabe, l'ourdou et le portugais https://www.ethnologue.com/.
5. https://www.culture.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France
6. http://www.efnil.org/documents/declarations/dublin-declaration-1
7. Connue anciennement sous la désignation de "*Commission Générale de Terminologie et Néologie*'. Des listes de termes sont publiées sur le site web de FranceTerme : http://www.culture.fr/franceterme.
8. https://www.fondation-alliancefr.org/
9. https://w3techs.com/technologies/history_overview/content_language/ms/y
10. La traduction à l'UNESCO : http://databases.unesco.org/xtrans/stat/xTransStat.html.
11. https://www.ethnologue.com/subgroups/sign-language
12. https://royalsocietypublishing.org/doi/10.1098/rsos.191100
13. https://fr.wikipedia.org/wiki/Langue_des_signes_française

14. https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000019911145

15. Il s'agit de la version française du "Cued Speech","un mode de communication visuel qui utilise des configurations et des placements de main associés à des mouvements labiaux et de la parole pour permettre de différencier entre eux les phonèmes de la langue orale" (Notre traduction, source : https://en.wikipedia.org/wiki/Cued_speech)

16. https://media-pi.fr/

17. https://www.france.tv/france-5/l-oeil-et-la-main/

18. Cette section est la traduction d'un résumé adapté du livrable ELE *D1.2 Report on the state of the art in Language Technology and Language-centric AI* (Agerri *et al.*, 2021) et des sections 1 and 2 du livrable ELE *D3.1 Report on existing strategic documents and projects in LT/AI* (Aldabe *et al.*, 2021).

19. Dans un rapport récent de 2021, le marché mondial des TL était déjà évalué à 9,2 milliards de dollars en 2019 et on estime son taux de croissance annuelles à 18,4 % de 2020 à 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). Un autre rapport de 2021 estime qu'en pleine crise de la COVID-19, le marché mondial du TAL s'élevait à 13 milliards d'USD en 2020 et devrait atteindre 25,7 milliards d'USD d'ici 2027, avec un taux de croissance annuelle de 10,3 % (https://www.researchandmarkets.com/reports/3502818/natural-language-processing-nlp-global-market).

20. https://fr.wiktionary.org/wiki/Catégorie:français

21. http://fasttext.cc

22. https://www.atilf.fr/ressources/tlfi/

23. https://www.dictionnairedesfrancophones.org/

24. http://https://www.btb.termiumplus.gc.ca/

25. https://www.eurotermbank.com

26. https://www.frantext.fr/

27. https://www.sketchengine.eu/corpora-and-languages/french-text-corpora/

28. https://corpora.uni-leipzig.de/en?corpusId=fra_mixed_2012

29. https://www.gutenberg.org/

30. http://data.statmt.org/news-crawl/README

31. https://commoncrawl.org

32. p. ex. Pagnol : https://pagnol.lighton.ai

33. https://bigscience.huggingface.co/

34. https://cedille.ai/

35. http://www.levoicelab.org/

36. https://spacy.org

37. https://nlp.johnsnowlabs.com/

38. https://stanfordnlp.github.io/stanza/

39. https://gitlab.inria.fr/almanach/alTextProcessing/melt
40. https://www.nltk.org/
41. https://nlp.lsi.upc.edu/freeling/
42. https://gate.ac.uk/
43. https://github.com/huggingface/tokenizers
44. https://unimorph.github.io/
45. https://universaldependencies.org/
46. http://compling.hss.ntu.edu.sg/omw/summx.html
47. https://framenet.icsi.berkeley.edu/fndrupal/
48. Comme Babelfyhttp://babelfy.org/)
49. https://github.com/facebookresearch/XNLI
50. https://ufal.mff.cuni.cz/corefud
51. https://github.com/aymara/lima/
52. http://redac.univ-tlse2.fr/corpus/annodis/
53. https://github.com/ZurichNLP/xstance
54. P. ex. dans le projet européen InVid https://www.invid-project.eu/invid-datasets/ et WeVerify.
55. https://s1.lemde.fr/mmpub/data/decodex/hoax/hoax_debunks.json
56. p. ex. https://storyzy.com/?lang=frpour la détection d'infox.
57. https://trec.nist.gov/
58. https://trecvid.nist.gov/
59. http://www.clef-initiative.eu/
60. https://deft.lisn.upsaclay.fr
61. https://github.com/boudinfl/ake-datasets
62. http://juanmanuel.torres.free.fr/corpus/rpm2/doc_resumes_fr.html
63. https://www.opus.eu
64. Voir http://statmt.org/WMT21 pour l'occurrence la plus récente.
65. https://iwslt.org/
66. https://hugginface.co
67. https://www.systransoft.com/marketplace-catalog/?lang=fr
68. http://www.islrn.org/resources/055-636-352-982-9/
69. http://islrn.org/resources/425-777-374-455-4/

70. http://www.elra.info/en/projects/archived-projects/repere/

71. http://synpaflex.irisa.fr/

72. https://zenodo.org/record/4580406

73. p. ex. https://github.com/NVIDIA/tacotron2

74. V. p. ex. https://txm.gitpages.huma-num.fr/textometrie/

75. https://www.frantext.fr/

76. https://rhapsodie.modyco.fr/

77. https://repository.ortolang.fr/api/content/cefc-orfeo/11/documentation/site-orfeo/index.html

78. http://eslo.huma-num.fr/index.php

79. https://www.nooj-association.org/index.html

80. p. ex. https://cental.uclouvain.be/cefrlex/

81. http://icar.cnrs.fr/ressources-base-donnees/

82. http://icar.cnrs.fr/ressources-base-donnees/

83. https://www.projet-pfc.net/

84. http://latlcui.unige.ch/phonetique/easyalign.php

85. http://clarin.phonetik.uni-muenchen.de/BASWebServices/interface

86. https://fr.cliq-ai.quebec/

87. Le CNRS (*Centre National de la Recherche Scientifique*) et l'Inria (*Institut National de Recherche en informatique et en Automatique*) sont des établissements publics nationaux de recherche, employant des chercheurs et des ingénieurs à plein temps.

88. Le GdR ("*Groupement de Recherche*"). TAL (https://gdr-tal.ls2n.fr/) et LIFT (https://gdr-lift.loria.fr/).

89. *Association pour le Traitement Automatique des Langues.*

90. *Association Française pour la Communication Parlée.*

91. *Association Francophone de Recherche d'Information et Applications.*

92. *Association Française d'Intelligence Artificielle.*

93. https://github.com/pytorch/fairseq/

94. https://www.huma-num.fr/

95. https://www.ortolang.fr

96. Le ministère de la Culture lance depuis dix ans un appel à projets sur "l'action culturelle et la langue française" pour la maîtrise du français, doté d'un budget d'un million d'euros par an, mais dont seule une part marginale concerne le développement des TL.

97. https://www.intelligence-artificielle.gouv.fr/fr/thematiques/programme-national-de-recherche-intelligence-artificielle-pnria

98. https://trail.ac/trail4wallonia/

99. https://www.msh-reseau.fr/

100. http://msh-paris-saclay.fr/appel-a-projets-excellence-2021-msh-paris-saclay-institut-dataia-23-04-2021/

101. Bpifranbce est une banque publique d'investissement pour le financement et le développement des entreprises. Voir : https://www.bpifrance.fr/.

102. http://www.idris.fr/eng/jean-zay/jean-zay-presentation-eng.html

103. 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

104. https://www.project-easier.eu/

105. https://ocelles.inshea.fr/

106. https://signmaths.univ-tlse3.fr/

107. https://dico.elix-lsf.fr/

108. https://www.sign-lang.uni-hamburg.de/dicta-sign/portal/

109. https://www.ortolang.fr/market/corpora/ortolang-000926

110. https://www.ortolang.fr/market/corpora/dicta-sign-lsf-v2/

111. https://media-pi.fr/

112. https://www.ortolang.fr/market/corpora/mediapi-skel

113. https://rosettaccess.fr/index.php/home-page-english/

114. Il est également probable que de nombreuses ressources précieuses sont passées inaperçues, ou ont tout simplement été perdues.

115. http://www.elra.info/en/catalogues/lre-map/

116. http://www.islrn.org/

117. https://paperswithcode.com/

118. https://elrc-share.eu/

119. https://www.inist.fr/services/analyser/istex-textes-corpus/

120. https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/welcome.html

121. http://www.technolangue.net/