UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

Dipartimento di Studi Linguistici e Letterari

Corso di Laurea Magistrale in

Lingue Moderne per la Comunicazione e la Cooperazione

Tesi di Laurea

Traduire l'innovation : les défis de la terminologie liée à l'intelligence artificielle

Relatrice
Prof. Federica Vezzani

Laureanda
Giulia Sartor
n° matr. 2036519 / LMLCC

Anno Accademico 2023 / 2024

À Giulia

Où que tu sois

Vis ce jour comme s'il était le tien

Pour toi

Pour toutes les femmes victimes de féminicide

Pour toutes

Remerciements

É arrivato il momento che tanto aspettavo e sola ora posso dire di essere riuscita a raggiungere un traguardo che all'inizio sembrava lontano in termini di competenze, costanza e capacità. Sento quindi il bisogno di ringraziare tutti coloro che mi hanno accompagnata in questo lungo cammino.

Vorrei anzitutto ringraziare la mia relatrice, la Professoressa Federica Vezzani, che si è subito proposta di seguirmi in questo percorso. La ringrazio per la sua disponibilità, attenzione e per i suoi preziosi consigli che mi hanno aiutata molto, sia durante l'anno accademico, sia durante la stesura della tesi.

Un altro ringraziamento va fatto al mio correlatore, il Professor Giorgio Maria Di Nunzio, che si è occupato della parte inerente alla materia di sua competenza: l'informatica.

Vorrei poi ringraziare mia mamma Gianna e mio papà Franco, che in questi anni d'università mi sono stati vicini ogni giorno. Non importava quale situazione ci affliggesse, loro sono sempre stati pronti a supportarmi. Ringrazio anche mia zia Graziella che sia da vicino che da lontano mi ha sempre offerto il suo appoggio e il suo amore incondizionato. Vi amo con tutto il cuore.

Un caloroso ringraziamento va alla mia madrina Monika, a sua mamma Giovanna, a sua sorella Antonella e al mio padrino Daniele, che mi hanno vista crescere, aiutandomi ad affrontare il mondo con forza e vitalità. Ringrazio particolarmente Monika e Daniele che mi hanno fatto da guide turistiche per la prima volta in giro per le città europee e per il loro costante amore e supporto. Ringrazio anche il fratello di Monika e Antonella, Nicola che, sebbene ci abbia abbandonato, rimane sempre nel mio cuore.

Ringrazio la mia roccia, Giorgia, che dal 1998 non mi ha mai abbandonata. Per me sei come una sorella, da te e con te ho imparato tante cose. Mai dimenticherò i pomeriggi passati insieme e l'allegria che mi invadeva (e che mi invade) stando con te. Ringrazio anche tutta la famiglia di Giorgia: il fratello Leonardo, la mamma Marta e il papà Paolo.

Ringrazio poi le mie amiche e colleghe di lavoro Sabrina, Beatrice, Desirée, Hasnaa e Sara che mi tengono compagnia durante i servizi e che tanto mi sopportano. Ringrazio anche il mio titolare Yuri e la compagna Ozana, che mi hanno sempre permesso di studiare quando ne avevo bisogno rimanendo sempre comprensivi nei miei confronti.

Ringrazio le mie amiche Noemi, Clara e Raiane e i miei amici Antonio e Cristian. Grazie per i bei momenti passati insieme e per quelli che verranno.

Un ringraziamento speciale va alla mia super psicoterapeuta Luisa che da circa un anno mi sta aiutando a combattere le mie ansie. La ringrazio per aver sempre compreso ogni singola cosa di me, istruendomi con empatia e dedizione.

Un altro importante ringraziamento va alla mia migliore amica Erica. Mai avrei pensato che la nostra amicizia, cominciata nei banchi dell'università, potesse trasformarsi in qualcosa di così forte, superando anche la distanza che ci separa. Grazie di tutto, ti voglio un mondo di bene. Spero di vivere tante altre avventure insieme a te.

Ultimo, ma non meno importante, ringrazio l'amore della mia vita, Simone. Per me sei stato come una ventata di aria fresca e lo sei ancora. Ogni giorno mi insegni cose nuove e ogni giorno ringrazio l'universo per avermi portata da te. Sei tutto ciò che ho sempre desiderato, sei il compagno perfetto e la persona più buona e gentile che io conosca. Ti amo immensamente e non vedo l'ora di affrontare insieme cosa la vita ci riserva per il futuro. Ringrazio anche Michela, mamma di Simone, il papà Christian, il fratello Alex e la sua dolce metà Irene. Vi ringrazio per avermi accolta con le braccia aperte nella vostra famiglia. Grazie per tutte le cene e i pranzi e grazie per il costante aiuto e supporto che ogni giorno mi date.

Per concludere, ringrazio me stessa per non aver mai mollato e per aver sempre cercato di vedere oltre le difficoltà, nonostante i grandi ostacoli da affrontare. Mi ringrazio per essere sempre riuscita a rialzarmi dopo ogni caduta, arrivando fino a questo indimenticabile giorno.

Résumé

La terminologie de l'intelligence artificielle et du traitement automatique du langage (TAL) est confrontée à des défis constants. Les termes qui caractérisent ces domaines sont nés et naissent dans des contextes anglo-américains. Cela comporte la présence de nombreux anglicismes dans la langue française. Cependant, les réalités francophones s'engagent pour arrêter l'introduction de mots étrangers dans le lexique français, en rédigeant des milliers de fiches terminologiques des termes relatifs à l'intelligence artificielle afin d'assurer le développement et la diffusion du vocabulaire français de l'intelligence artificielle. Ce mémoire veut analyser le processus utilisé pour étudier les termes de l'intelligence artificielle et du TAL, en présentant un projet terminographique multilingue qui comprend l'étude du rôle du corpus, des fiches terminologiques et leur examen qualitatif.

Mots-clés : intelligence artificielle, traitement automatique du langage, terminologie, projet terminographique, corpus, fiche terminologique.

Abstract

The terminology of artificial intelligence and natural language processing (NLP) faces lots of challenges. The terms that characterize these fields of study are created in Anglo-American contexts. As a result, there are many anglicisms in the French language. However, French-speaking countries are committed to stop the introduction of foreign words into the French lexicon, by writing thousands of terminology records of terms relating to artificial intelligence in order to ensure the development and diffusion of the French vocabulary of artificial intelligence. The aim of this dissertation is to analyse the process used to study artificial intelligence and NLP terms, by presenting a multilingual terminographical work that includes the study of the role of the corpus, terminology records and their qualitative evaluation.

Keywords: artificial intelligence, natural language processing, terminology, terminographical work, corpus, terminological record.

Table des matières

Liste des tableaux et des figures 13

Liste des abréviations 15

Introduction 19

CHAPITRE 1 25
Aperçu historique de l'intelligence artificielle et coup d'œil sur le traitement du langage naturel (TAL) 25

1.1 L'évolution de l'homme, de ses origines à nos jours 25
1.1.1 Les inventions les plus remarquables de l'homme 26

1.2 L'intelligence artificielle et ses fondements au cours de l'histoire 30
1.2.1 Les algorithmes, la base où l'IA s'appuie 30
1.2.2 La définition d'intelligence artificielle 31
1.2.3 La naissance de l'intelligence artificielle, les premiers travaux et les systèmes experts 32
1.2.4 L'intelligence artificielle fait un bond en avant grâce à l'apprentissage automatique 36
1.2.5 L'intelligence artificielle générale veut égaliser l'être humain 37

1.3 L'intelligence artificielle et les multiples disciplines qui l'utilisent 38
1.3.1 Des numéros de l'IA 38
1.3.2 L'aide de l'IA en différents champs d'études 40
1.3.3 Les problèmes liés à l'IA 43

1.4 Le traitement du langage naturel (TAL) 45
1.4.1 Le TAL et ses liens avec la langue 47
1.4.2 L'emploi du TAL et ses problèmes 49

CHAPITRE 2 53
La terminologie : entre intelligence artificielle et traitement automatique du langage 53

2.1 Introduction à la terminologie et à la terminographie 53
2.1.1 Un peu d'histoire de la terminologie 54

2.2 La terminologie générale et les autres approches alternatives 55
2.2.1 La centralité du social dans la socioterminologie 56
2.2.2 La dimension du texte dans la terminologie textuelle 57
2.2.3 Le rôle de la culture dans la terminologie culturelle 58
2.2.4 La différence entre « concept classique » et « concept sociocognitif » 59
2.2.5 Les multiples faces du terme dans la théorie communicative de la terminologie 61

2.3 La terminographie et ses pratiques 61
2.3.1 Les sept étapes du travail terminographique 62
2.3.2 Description des données terminologiques 62

2.4 La terminotique, un lien être informatique et terminologie 63
2.4.1 Les changements apportés par l'informatique 64

2.5 La terminologie à l'aide des autres disciplines 66
2.5.1 Communication spécialisée, traduction, documentation et terminologie 66
2.5.2 Informatique, ingénierie des connaissances et terminologie 69

2.6 La terminologie de l'intelligence artificielle et du TAL 71
2.6.1 Introduction à la *lingua franca* 71
2.6.2 Comment la France fait face aux anglicismes 72
2.6.3 Les termes du domaine de l'intelligence artificielle et du TAL 74

CHAPITRE 3 77
Présentation théorique et analyse de la méthodologie pour la construction d'un corpus 77

3.1 Exposé général du corpus 77
3.1.1 Définitions et caractéristiques 77
3.1.2 L'origine du terme « corpus » 80
3.1.3 Le texte spécialisé à la base du corpus 81

3.2 Le corpus spécialisé et les autres types de corpus 82
3.2.1 Les corpus multilingues 84
3.2.2 Le corpus annoté : annotation et étiquetage 86

3.3 La linguistique de corpus 91
3.3.1 L'histoire de la linguistique de corpus 91
3.3.2 Linguistique de corpus et informatique 92
3.3.3 Emploi actuel de la linguistique de corpus 93

3.4 Construire un corpus 94
3.4.1 Choix des textes 94
3.4.2 La fiabilité des sources 96
3.4.3 L'analyse des systèmes de gestion de corpus 98
3.4.4 Les difficultés 102

CHAPITRE 4 105
Les termes, leur extraction et la compilation des fiches terminologiques 105

4.1 Les termes et leurs caractéristiques 105
4.1.1 Terme simple et terme complexe 105
4.1.2 Formation des termes 106
4.1.3 Termes non prédicatifs et termes prédicatifs 114
4.1.4 Identifier les termes 116

4.1.5 Le terme dans le texte spécialisé 117

4.2 L'extraction de termes 119
4.2.1 Les bases d'un extracteur de termes 120
4.2.2 Comparer des corpus pour extraire les termes 121
4.2.3 Les autres techniques pour l'extraction de termes 122
4.2.4 Le logiciel TermoStat 125
4.2.5 Les problèmes les plus courants 128

4.3 Le stockage de termes 130
4.3.1 Les bases de données et les documents structurés 130
4.3.2 Structures et modèles des données 132

4.4 FAIRterm et les fiches terminologiques 134
4.4.1 La fiche terminologique 135
4.4.2 La compilation de fiches terminologiques avec FAIRterm 137

CHAPITRE 5 143
Analyse qualitative du projet terminographique et considérations finales 143

5.1 Analyse des corpus du projet 143
5.1.1 Le corpus en français 143
5.1.2 Le corpus en italien 151
5.1.3 Le corpus en anglais 159
5.1.4 Analyse conclusive du corpus 165

5.2 Analyse de l'extraction de termes 167
5.2.1 L'exigence d'un nettoyage manuel 168

5.3 Analyse de la compilation de fiches terminologiques 171
5.3.1 Quelques cas concrets 171

5.4 Considérations finales sur le projet 174
5.4.1 Les outils employés 174
5.4.2 Évaluation de la méthodologie 176

Conclusion 179

Annexes 185

Bibliographie 201

Sitographie 205

Résumé en italien 219

Liste des tableaux et des figures

Figure 1 – Exemple d'un corpus aligné (cf. L'Homme 2020, 149) 85
Figure 2 – La première fenêtre du logiciel AntConc 98
Figure 3 – La première fenêtre du logiciel Sketch Engine 100
Figure 4 – La première fenêtre du logiciel TermoStat 101
Figure 5 – Exemple d'une fiche terminologique monolingue 136
Figure 6 – Exemple d'une fiche terminologique bilingue 137
Figure 7 – L'ajout d'un terme sur FAIRterm 137
Figure 8 – Fiche terminologique du terme « base de connaissance » 138

Tableau 1 – Les étiquettes et leur description 89
Tableau 2 – Exemple pratique d'une phrase étiquetée par TreeTagger 90
Tableau 3 – Exemples de préfixes en français scientifique 109
Tableau 4 – Exemples de suffixes de la langue française 110
Tableau 5 – Première partie de la liste des candidats-termes du logiciel TermoStat pour le projet *YourTerm TECH* 127

Liste des abréviations

AAAS	American Association for the Advancement of Science
ACM	Association for Computing Machinery
AGT	Anzani Trading Group
AGT AI	Anzani Trading Group Artificial Intellegence
AI	Artificial intelligence
ALPAC	Automatic Language Processing Advisory Committee
AMIA	American Medical Informatics Association
ANSI	American National Standards Institute
ARPANET	Advanced Research Projects Agency Network
B2B	Commerce entreprise à entreprise
BNC	British National Corpus
CAT	Computer-aided translation
ChatGPT	Chat Generative Pre-trained Transformer
CNR	Consiglio Nazionale delle Ricerche
CNRS	Centre National de la Recherche Scientifique
CRIAD	Centro di Ricerca per l'Informatica Applicata alla Didattica e all'Educazione
CR-ROM	Compact disc – read-only memory
CSTN	Commissions spécialisées de terminologie et de néologie
CSV	Valeurs séparées par une virgule
CT	Candidats termes
DAPRA	Defense Advanced Research Projects Agency
DENDRAL	DENDRitic ALgorithm
DevOps	Développement et opérations
DGLFLF	Délégation générale à la langue française et aux langues de France
EOSC	European Open Science Cloud
FTP	File Transfer Protocol
HTTP	Protocole de transfert hypertexte
IA	Intelligence artificielle
IATE	InterActive Terminology for Europe

IBM	International Business Machines Corporation
ICMCCE	International Conference on Mechanical, Control and Computer Engineering
IGSG	Istituto di Informatica Giuridica e Sistemi Giudiziari
Inria	Institut national de recherche en sciences et technologies du numérique
IoT	Internet des objets
ISIR	Institut des Systèmes Intelligents et de Robotique
ISO	Organisation internationale de normalisation
ITTIG	Istituto di Teoria e Tecniche dell'Informazione Giuridica
JAMIA	Journal of the American Medical Informatics Association
LDC	Linguistic Data Consortium
LISP	List processing
MIT	Massachusetts Institute of Technology
NAE	National Academy of Engineering
NAM	National Academy of Medicine
NAS	National Academy of Sciences
NLP	Natural Language Processing
OCI	Oracle Cloud Infrastructure
ONU	Organisation des Nations unies
PDF	Portable Document Format
PNAS	Proceedings of the National Academy of Sciences
REF	Revue d'économie financière
RES	Research for Enterprise Systems
RIOS	Rete Italiana Open Source
SGBD	Système de gestion de base de données
SGML	Standard Generalized Markup Language
SIDA	Syndrome d'immunodéficience acquise
SNFC	Société nationale des chemins de fer français
SOTT	Standardisation-oriented terminology theory
TAL	Traitement automatique du langage
TALN	Traitement automatique du langage naturel

TAO	Traduction Assistée par Ordinateur
TBX	TermBase Exchange
TCP/IP	Transmission Control Protocol/Internet Protocol
TermCoord	Terminology Coordination Unit
TF-IDF	Term frequency-inverse document frequency
TSV	Tab separated values
UBS	Union des banques suisses
UE	Union européen
UNICEF	Fonds des Nations unies pour l'enfance
URSS	Union des républiques socialistes soviétiques
UTF-8	Universal Character Set Transformation Format - 8 bits
WEBIST	International Conference on Web Information Systems and Technologies
XLSX	Microsoft Excel Spreadsheet
XML	eXtensible Markup Language

Introduction

Le présent mémoire vise à fournir une analyse approfondie de la terminologie de l'intelligence artificielle et du traitement automatique du langage (TAL), en proposant une série exhaustive de termes spécialisés des deux domaines cités, ainsi que la présentation générale et historique des deux domaines, la présentation de la terminologie en tant que discipline et l'analyse de la mise en œuvre d'un projet de terminographie.

Les raisons pour lesquelles nous avons décidé de traiter cette thématique sont multiples. En premier lieu, le cours de traduction française spécialisée à la deuxième année du master nous a permis d'entrer en contact pour la première fois avec la terminologie et sa complexité, en développant un majeur intéresse pour la discipline. Un des aspects les plus intéressants était la compilation de fiches terminologiques, car nous avons réalisé à quel point le travail des terminographes peut être long, complexe et articulé. En deuxième lieu, la proposition de traiter la terminologie de l'intelligence artificielle et du traitement automatique du langage nous a semblé être un défi qui pourrait être transformé en un projet démontrant l'importance d'avoir à disposition des données terminologiques complètes, surtout pour ce qui concerne les domaines spécialisés et l'échange d'informations entre les experts et les non-experts. En troisième lieu, l'intelligence artificielle et le traitement automatique du langage sont deux disciplines pas à la portée de tous, relativement difficiles à comprendre, il est donc nécessaire de s'en occuper pour garantir leur vulgarisation et connaissance. Aujourd'hui, l'intelligence artificielle est devenue partie intégrante de nos vies : nous l'utilisons lorsque nous sommes au téléphone, par exemple, nous faisons une question à un assistant virtuel comme Siri ; elle a été insérée dans les usines et les entreprises et les travailleurs l'utilisent comme une aide pour effectuer de nombreuses tâches ; et elle est exploitée également afin d'entraîner les machines à traiter et interpréter le langage humain. C'est donc quelque chose que doit être étudiée et connue, car elle sera de plus en plus présente et importante dans la vie de tous les jours. En dernier lieu, en nous occupant de ces deux domaines spécialisés, nous avions la possibilité de participer au projet *Terminology without borders* ou Terminologie sans frontières, en français. L'objectif principal du projet est d'améliorer la communication dans plusieurs domaines en adaptant la terminologie aux besoins des citoyens. Ce projet est géré par l'Unité de coordination de la terminologie, également

appelée *TermCoord*, qui soutient l'objectif de la Direction générale de la Traduction du Parlement européen de communiquer avec les citoyens en utilisant un langage clair. Le projet est ainsi inséré dans le cadre de la collaboration avec les Universités de Terminologie et il donne l'opportunité de collaborer avec des agences de l'UE et de l'ONU et avec d'autres organisations internationales. Les macro-domaines scientifiques dont le projet s'occupe sont les suivants : la médecine (MED), l'environnement (ENVI), la culture (CULT), le droit de la femme (FEM), la justice et le droit (JURI), les affaires maritimes et la pêche (MARE), l'instruction (EDU), la gastronomie et l'industrie alimentaire (FOOD), la technologie (TECH) et l'économie et la finance (FIN). Nous avons adhéré au projet *YourTerm TECH* (Knowledge Centre on Interpretation, s. d.). Pour ce projet, nous devrons trouver 150 termes concernant les deux domaines d'étude mentionnés qui seront analysés grâce aux fiches terminologiques.

Le mémoire est structuré dans cinq chapitres que nous introduisons ci-dessous. Le premier chapitre vise à bien présenter le domaine de l'intelligence artificielle et le domaine du traitement automatique du langage. Nous commençons donc par une petite analyse des inventions de l'homme jusqu'à arriver à la technologie. Nous continuons par exposer le domaine de l'intelligence artificielle et du traitement automatique du langage. De toute évidence, il est nécessaire de fournir une description complète des domaines afin de mieux comprendre comment réaliser le projet terminographique. Nous allons ainsi présenter les passages historiques les plus importants pour la naissance et le développement de l'intelligence artificielle. Puis, nous nous intéresserons à ses applications principales, par exemple, les mathématiques, la bio-informatique et la robotique et aussi à ses applications moins connues, par exemple, la psychologie et les sciences humaines et sociales. Il faudra également que nous présentions les problèmes liés à l'intelligence artificielle, car elle ne manque pas de créer des difficultés. Pour terminer ce chapitre, nous présenterons le traitement automatique du langage qui fait partie du domaine général de l'intelligence artificielle et s'occupe d'entraîner les ordinateurs afin de traiter de manière automatique le langage humain, et il mérite donc un examen approfondi.

Le deuxième chapitre est consacré à la terminologie et à la terminographie. Dans cette partie, les deux disciplines seront décrites de manière exhaustive. Elles représentent en fait le cœur de ce projet, car elles nous fournissent les lignes directrices pour sa

réalisation. Nous allons les analyser, en se concentrant sur leur histoire, les théories principales qui les caractérisent, les travaux concrets, les données terminologiques et les changements apportés par la technologie. De plus, nous allons décrire comment la terminologie est au service d'autres domaines qui l'appliquent, par exemple, la traduction spécialisée, l'ingénierie des connaissances et la communication. À la fin du chapitre, nous ferons un bond vers la terminologie de l'intelligence artificielle et du traitement automatique du langage pour aller au cœur du projet. Les questions liées à la langue française et aux termes spécialisés seront discutées.

Le troisième chapitre va se concentrer sur le corpus, qui est une partie fondamentale pour la réalisation du projet terminographique. Tout d'abord, nous allons présenter le corpus en soulignant sa définition et ses caractéristiques les plus importantes. Nous continuons le chapitre en analysant le corpus spécialisé, qui est le corpus que nous intéresse le plus, et également les autres typologies, telles que le corpus multilingue et le corpus annoté. La partie suivante du chapitre trois sera dédiée à une discipline dont la base est précisément le corpus : la linguistique de corpus. Nous analyserons donc son histoire, ses liens avec la technologie et l'informatique et son emploi actuel. Pour terminer, nous passerons à la partie pratique : nous allons expliquer comment construire un corpus. Les points principaux que nous allons présenter sont : le choix des textes, la fiabilité des sources où les textes ou les informations proviennent, l'analyse des systèmes de gestion de corpus et les difficultés que les terminographes peuvent rencontrer pendant la construction.

Le quatrième chapitre se penche plus en détail sur les termes, sur l'extraction de termes et sur la compilation des fiches terminologiques, qui sera une partie essentielle pour le projet. Nous présenterons donc les termes pour ce qu'ils sont, c'est-à-dire que nous allons analyser la différence entre terme simple et terme complexe, le processus de formation et la différence entre les termes prédicatifs et non. En ajoute, nous examinerons la méthode pour identifier les termes et l'environnement linguistique des termes, c'est-à-dire le texte spécialisé. Puis, ce sera le tour de la partie consacrée à l'extraction terminologique. Nous discuterons donc les extracteurs automatiques de termes, comment ils fonctionnent, les méthodes les plus courantes pour extraire les termes, le logiciel que nous utiliserons pour travailler au projet et les problèmes qui peuvent se présenter pendant l'extraction. Ensuite, il est nécessaire d'analyser les techniques de stockage de termes et

leurs caractéristiques, telles que les bases de données et les banques de terminologies. Pour conclure ce chapitre, nous allons présenter le logiciel que nous sert pour la compilation des fiches terminologiques du projet *YourTerm TECH*. Pour garantir un examen complet et exhaustif, nous parlerons des fiches terminologiques en général et puis comment doit être faite leur compilation.

Le dernier chapitre, le numéro cinq, est le chapitre conclusif qui résume l'âme du projet. Ce chapitre est consacré à l'analyse qualitative de l'entier projet. Tout d'abord, nous allons examiner la construction des trois corpus. Pour chaque corpus, nous expliquerons pourquoi les documents qui les composent ont été choisis. Nous discuterons donc de la fiabilité des sources, en soulignant le type de source et ses principales caractéristiques. Par exemple, des sites web ou d'articles de revues scientifiques seront analysés. En concluant l'analyse de corpus, nous indiquerons quel critère nous avons utilisé pour choisir tous les documents du corpus et quels problèmes nous avons rencontrés pendant la recherche. Ensuite, nous présenterons un examen de l'extraction de termes : nous reprendrons brièvement les principaux types de fonctionnement de l'extraction et nous expliquerons pourquoi nous avons dû effectuer un nettoyage manuel après avoir pris vision des résultats fournis par l'extracteur automatique. Nous allons donc traiter la compilation concrète des fiches terminologiques. Pour donner des exemples réels, nous allons inclure quelques cas de compilation parmi les plus intéressants. Pour terminer ce dernier chapitre, nous présenterons nos considérations finales en soulignant l'importance des outils employés et de la méthodologie adoptée.

La rédaction de ce mémoire et la réalisation du projet *YourTerm TECH* se basent sur l'emploi fondamental de deux outils : TermoStat et FAIRterm. Le premier sera employé pour l'extraction de termes et le second est l'outil grâce auquel nous compilerons les fiches terminologiques. Ces deux ressources informatiques seront présentées en détail, en expliquant leur fonctionnement, leur importance et leur efficience.

Pour ce qui concerne la méthodologie relative à la rédaction du mémoire, elle se base principalement sur l'emploi des sources fiables qui traitent les thématiques que nous s'intéressent, à savoir l'intelligence artificielle, le traitement automatique du langage, la terminologie, la linguistique de corpus et les méthodes d'extraction de termes. Nous nous sommes servis de livres, des articles de revues scientifiques, d'essais, de mémoires et de thèses, de diverses recherches scientifiques sur la terminologie et sur l'informatique, et

encore sur de nombreux sites web qui font autorité. Nous avons également consulté plusieurs dictionnaires en ligne.

Le projet et le mémoire ont deux principaux objectifs. Le premier objectif que nous voulons atteindre est de démontrer comment la terminologie est essentielle pour transmettre les concepts et les informations relatifs à un domaine spécialisé. Les termes occupent une place prépondérante en ce qui concerne l'étude, la compréhension et la vulgarisation de tous domaines spécialisés et ils doivent être bien analysés, précis et clairs afin de garantir de bonnes ressources pour les chercheurs de toutes disciplines et aussi pour les non-experts. Le deuxième objectif de ce mémoire rappelle les objectifs du projet *Terminology without borders*, c'est-à-dire assurer l'amélioration de la communication dans les domaines spécialisés en adaptant la terminologie aux besoins de tous. Dans notre cas, nous devrons s'intéresser aux domaines de l'intelligence artificielle et du traitement automatique du langage pour permettre à tous d'utiliser les données terminologiques qui les concernent.

CHAPITRE 1

Aperçu historique de l'intelligence artificielle et coup d'œil sur le traitement du langage naturel (TAL)

1.1 L'évolution de l'homme, de ses origines à nos jours

Ce premier chapitre veut expliquer de manière détaillée le domaine concernant l'intelligence artificielle, en soulignant son histoire, le travail de l'homme, ses applications et ses difficultés, ainsi qu'un coup d'œil dans le monde du traitement automatique du langage naturel.

L'évolution est le « processus continu de transformation, passage progressif d'un état à un autre. » (Centre National de Ressources Textuelles et Lexicales, s. d.). Pour analyser et comprendre l'évolution, il faut travailler sur la comparaison des états dans le temps, mais également entre les organismes et leurs adaptations. Par conséquent, nous pouvons apprendre que l'évolution est la somme de plusieurs transitions et elles peuvent être majeures ou mineures et liées à l'apparition (ou à la disparition) d'espèces entières ou de traits particuliers. L'une des plus extraordinaires évolutions dans la Création est notre progrès, notre métamorphose : c'est l'évolution humaine, qui a été étalée sur un ensemble des transitions qui se sont vérifiées sur une période de 5 millions d'années (cf. Foley/Martin/Mirazón 2016, 2). Depuis lors, l'homme a conçu et produit des merveilles pour s'adapter au monde dont il est l'hôte.

Il existe plusieurs chercheurs qui ont étudié l'évolution humaine, par exemple, Charles Darwin, un paléontologue anglais, qui a travaillé sur l'évolution des espèces vivantes en révolutionnant la biologie. Puis, Kenneth Page Oakley était un anthropologue physique, paléontologue et géologue anglais. Il a écrit de nombreuses publications qui ont développé le domaine de l'évolution humaine. L'une de ces publications les plus reconnues est *Man the Tool-Maker*. Ce livre est une étude fondamentale qui a révolutionné la compréhension de l'évolution humaine. Il a été publié pour la première fois en 1949 et il a été réédité plusieurs fois, pour un total de six éditions distinctes en 1976 (Wikipédia, 2023). Dans son livre, Kenneth Oakley (1961, 1) a défini l'homme

comme un *social animal*, c'est-à-dire un « animal social ». L'homme diffère des autres mammifères, car la culture est partie intégrante de sa vie ; il est capable de communiquer ses idées avec ses pairs et il sait comment construire des outils, qui seront nécessaires pour vivre dans son environnement. Au contraire, les autres mammifères ont eu une évolution très différente : ils se sont adaptés au milieu et aux certaines situations, par conséquent ils ont développé des équipements physiques. Un exemple est le cheval, qui a développé des dents et des sabots pour manger de la végétation et pour vivre sur des plaines herbeuses. L'homme est devenu la créature la plus adaptable au monde, grâce à l'invention et à la production des équipements qui ne sont pas physiques et qui peuvent être modifiés en fonction des circonstances (cf. Oakley 1961, 1).

1.1.1 Les inventions les plus remarquables de l'homme

Depuis la nuit des temps, l'homme vit en contact étroit avec la technologie et l'innovation. Le terme technologie signifie « Science des techniques, étude systématique des procédés, des méthodes, des instruments ou des outils propres à un ou plusieurs domaine(s) technique(s), art(s) ou métier(s). » (Centre National de Ressources Textuelles et Lexicales, s. d.). La technologie fait partie de nos jours et elle influence tous les domaines de la vie humaine (cf. Spataro/Furholt 2020, 13). Elle joue un rôle fondamental dans le développement et l'application d'outils techniques. Son principal objectif est la résolution de problèmes pratiques et, pour le faire, l'homme, avec l'aide de la technologie, se base sur des connaissances scientifiques, par exemple les mathématiques et l'informatique (Treccani. Vocabolario online, s. d.).

Tout comme la technologie, l'innovation est un élément essentiel du progrès humain. Joseph Schumpeter était un économiste qui naît en 1883, il était professeur à Harvard dans les années 1930 et ses études les plus notables ont par objet l'innovation dans le domaine de l'économie. Schumpeter parle de « destruction créatrice », car les grandes phases de croissance et de crise sont liées à l'alternance de technologies nouvelles et anciennes. Toutes les technologies obsolètes traversent une crise qui a été provoquée par une innovation de rupture qui peut donner lieu à une cascade des autres innovations qui contribuent au progrès et à la croissance économique.

La technologie et l'innovation sont en constante évolution. L'homme a commencé à créer les produits de l'innovation technique avant la découverte du feu. L'une des plus importantes et utiles innovations de l'homme est la poterie. La technologie céramique et

les poteries sont nées pendant le Néolithique, une période qui fait partie de la Préhistoire, où l'homme commence à cultiver les champs, à élever le bétail et à construire les premiers villages sédimentaires à l'aide de l'organisation sociale des groupes humains. À l'origine, la production de poteries consistait en leur réalisation sans l'emploi du feu. Dans cette phase, les poteries ne sont pas cuites, elles étaient séchées au soleil. Plus tard, nous assistons au développement de la technologie céramique : les céramiques sont cuites par le feu. Pour ce qui concerne le but de cette invention, la poterie a été considérée comme une technologie de « prestige » dans les sociétés de chasseurs-cueilleurs : elle a été utilisée pour conserver des aliments spéciaux. Comme indiqué précédemment, les nouvelles inventions stimulent d'autres inventions qui peuvent appartenir également à de différents domaines. Dans le cas de la poterie, elle a permis d'innover par ailleurs les techniques de transformation des aliments et les techniques de leur conservation, par exemple, le trempage et la fermentation. Pour conclure, l'innovation déterminée par la poterie présente aussi des explications économiques. Les céramiques étaient utilisées selon les dynamiques de l'offre et de la demande : la céramique a été adoptée, car les autres types de récipients ne résistaient pas à la croissance de la demande causée par les naissants modes de préparation et de conservation des aliments (cf. Spataro/Furholt 2020, 139).

L'histoire de l'homme est donc influencée par l'innovation et les inventions naissent grâce à l'homme lui-même qui utilise la technologie et son génie pour créer des produits et pour concevoir des idées de plus en plus innovantes. L'histoire peut être divisée en plusieurs périodes, chacune caractérisée par des inventions spécifiques. Nous avons déjà discuté de la naissance de la poterie pendant le Néolithique, toutefois elle est seulement le début. Nous pourrions écrire des centaines de pages sur les innovations et les technologies que l'homme a conçues (cf. Curley 2010, 13).

Au cours de la Renaissance, les inventions les plus remarquables sont la presse à imprimer et les caractères mobiles. L'art de l'imprimerie était déjà connu en Orient, surtout en Chine et au Japon. Cependant, c'est Johannes Gutenberg qui a créé en Allemagne le premier atelier d'impression sur vaste échelle. Par conséquent, l'imprimerie s'est développée et elle est devenue un instrument de communication utilisé aux niveaux politique, social, religieux et scientifique, ainsi qu'un moyen pratique pour la diffusion de nouvelles et d'informations (cf. Curley 2010, 40).

Une autre période très riche en innovation est la révolution industrielle, une étape de l'histoire qui s'étend de la fin du 18e siècle jusqu'au début du 19e siècle. Cette révolution, qui naît en Angleterre, jette les bases pour la mécanisation, c'est-à-dire que nous assistons à l'invention et à la mise en œuvre de machines capables d'automatiser les processus de fabrication, de production et de distribution. Cette avant-garde a permis d'apporter des améliorations et des changements pas seulement dans la naissante industrie, mais aussi dans l'agriculture et dans le commerce. Grâce à l'industrialisation et à l'automatisation, les usines pouvaient s'organiser selon la dynamique de la production en série des biens, qui pouvaient être vendus à des prix réduits, en permettant leur achat aux gens qui appartenaient à des classes sociales inférieures (cf. Curley 2010, 13).

En bref, l'homme a modelé son existence en créant des outils, des produits, des machines et il a élaboré ses idées et ses idéologies. L'innovation et la technologie nous ont permis de survivre sur la Terre et elles sont essentielles pour la communication parmi tous. Les inventions que nous avons rapportées auparavant sont simplement des exemples : l'homme a également inventé les outils les plus élémentaires pour vivre, comme la roue, l'aqueduc, le calendrier, le papier ; il a créé un réseau pour les transports au niveau global qui est devenu de plus en plus vaste et qui comprend les voitures, les trains, les bateaux et les avions. Il a expérimenté les opportunités lui offert par la science et il a travaillé sur le corps humain pour le comprendre et le sauvegarder : la médecine a fait des pas de géant grâce aux inventions comme la vaccine contre la variole, l'anesthésie, l'insuline, la pilule contraceptive et bien plus encore.

Enfin, le 21e siècle est caractérisé par la « révolution numérique » à laquelle nous associons de nouvelles technologies comme le Web, le *streaming* ou les médias sociaux. Comme la machine à vapeur ou le chemin de fer, les technologies numériques ont provoqué (et elles provoquent encore) des changements capitaux sur notre société. Elles touchent tous les champs de la vie humaine et c'est pourquoi elles sont connues comme « technologies à usage général »[1]. Ce processus de digitalisation a renversé également les relations sociales, les marchés, les rapports humains et la politique et ses règles (cf. Quarta/Smorto 2020, 1–2). Dans cette révolution, l'internet a joué un rôle fondamental et il est l'une des inventions les plus extraordinaires. Internet est défini comme :

[1] De l'anglais *general purpose technologies*.

> Ensemble de réseaux mondiaux interconnectés qui permet à des ordinateurs[2] et à des serveurs de communiquer efficacement au moyen d'un protocole de communication commun (IP). Ses principaux services sont le Web, le FTP (*File Transfer Protocol*), la messagerie et les groupes de discussion (Institut national de la statistique et des études économiques, 2020).

L'ancêtre de l'internet s'appelle ARPANET, qui est l'acronyme de Advanced Research Projects Agency Network, le premier réseau à transfert de paquets de données conçu aux États-Unis par la DAPRA (Defense Advanced Research Projects Agency) en 1969 (cf. Curley 2010, 95). À ce moment-là, la guerre froide[3] battait son plein et l'armée des États-Unis voulait créer un réseau d'ordinateurs sans un cœur central. L'objectif du département de la défense nord-américaine était d'entraver une éventuelle attaque informatique par les autres pays (en particulier par l'Union soviétique) qui pourrait compromettre le réseau en entier ou son noyau, mais seulement une partie. Nous pouvons donc constater que l'internet est né pour accomplir des objectifs de nature militaire dans un contexte de tensions militaires et politiques. Toutefois, dans les années 1970, cette formidable structure de communication devient opérative également dans d'autres domaines. L'internet se développe rapidement dès lors qu'un langage commun à tous les ordinateurs est mis en œuvre : c'est le TCP/IP, de l'anglais *Transmission Control Protocol/Internet Protocol*. Ces protocoles sont utilisés par les réseaux existants pour dialoguer : nous assistons à la naissance d'un réseau global qui est appelé « réseau de réseaux ». Après l'invention d'internet, le moment est venu également pour le Web ou, en anglais, *World Wide Web*. L'ingénieur anglais Tim Berners-Lee est le père du Web. Avant sa création, les ordinateurs pouvaient s'envoyer des informations, s'ils étaient reliés au même réseau, mais grâce au Web il est possible de créer un système hypertexte public qui fonctionne sur internet. Nous ne devons pas confondre l'internet avec le Web, il faut donc éclaircir que ses termes ne sont pas synonymes : l'internet est l'infrastructure qui permet la transmission des données et il accueille de nombreux services, par exemple,

[2] Au cours de la lecture de ce mémoire, vous trouverez le terme « ordinateur », mais également les termes « calculateur » et « machine ».

[3] « La guerre froide […] est le nom donné à la période de fortes tensions géopolitiques durant la seconde moitié du XXe siècle entre, d'une part, les États-Unis et leurs alliés constitutifs du bloc de l'Ouest et, d'autre part, l'Union des républiques socialistes soviétiques (URSS) et ses États satellites formant le bloc de l'Est. La guerre froide s'installe progressivement à partir de la fin de la Seconde Guerre mondiale, dans les années 1945 à 1947, et dure jusqu'à la chute des régimes communistes en Europe en 1989, rapidement suivie de la dislocation de l'URSS en décembre 1991 » (Wikipédia, 2024).

la messagerie électronique et les groupes de discussion ; alors que, grâce au Web, les différents types de contenu Web peuvent être transformés sous forme de pages Web, qui pourront être regroupées à leur tour dans les sites Web. Tous ces contenus Web sont connectés par les hyperliens[4] et, à l'aide de protocoles standardisés, ils sont devenus accessibles à tous (cf. Quarta/Smorto 2020, 12–13).

L'homme a donc réussi non seulement à s'adapter, mais il a réussi à créer des choses qui semblent sortir d'un film de science-fiction. Ce sentiment augmente si nous pensons à l'intelligence artificielle (IA).

1.2 L'intelligence artificielle et ses fondements au cours de l'histoire

1.2.1 Les algorithmes, la base où l'IA s'appuie

Pour parler de l'intelligence artificielle, nous devons d'abord définir le terme « intelligence » : ce mot-là signifie « Fonction mentale d'organisation du réel en pensées chez l'être humain, en actes chez l'être humain et l'animal. » (Centre National de Ressources Textuelles et Lexicales, s. d.). L'être humain est donc capable de mettre en place ses idées et les utiliser pour résoudre des problèmes. La résolution des différents problèmes a lieu grâce aux algorithmes, c'est-à-dire une séquence finie d'instructions, ou règles claires et univoques afin d'accomplir des opérations, dans une période déterminée, dans le but d'arriver à la solution du problème. Le terme « algorithme » remonte au 9^{e} siècle : Muḥammad ibn Mūsā al-Khwārizmī était un mathématicien Persan qui a théorisé le concept d'algorithme pour la première fois. Par conséquent, le terme « algorithme » n'est pas utilisé exclusivement pour le domaine de l'informatique, comme la plupart des personnes pensent. Pour mieux comprendre le sens de ce terme, il existe un exemple très simple pour nous aider. Nous avons dit qu'un algorithme correspond à une série d'instructions ; nous pouvons donc dire que, pour faire un café, nous avons besoin d'un algorithme : il faut dévisser la cafetière ; puis il faut remplir la partie inférieure de la cafetière avec de l'eau fraîche ; après il est essentiel d'ajouter le café moulu dans le filtre métallique jusqu'au bord et quand la cafetière est prête, il faut la mettre sur le feu. Voilà un algorithme qui n'appartient pas au monde informatique, mais à la vie quotidienne. Au contraire, quand nous parlons des algorithmes en informatique, nous parlons d'une suite finie d'instructions claires que l'homme donne aux ordinateurs pour la résolution des

[4] L'hyperlien est un « lien associé à un élément d'un document hypertexte, qui pointe vers un autre élément textuel ou multimédia » (Dictionnaire en ligne Larousse, s. d.).

problèmes. L'exécution de ces règles logiques par les calculateurs s'appelle codage informatique, ou, en anglais, *coding*. L'homme utilise donc un langage de programmation (codage) pour communiquer avec la machine qui traduira les instructions formant les algorithmes afin d'exécuter les commandements destinés à elle (cf. Quarta/Smorto 2020, 14–15).

Nous devons mentionner une différence décisive entre les hommes et les machines quand nous parlons des algorithmes et aussi d'intelligence. L'homme est capable de résoudre des problèmes sans le besoin de recevoir des instructions claires, univoques et précises : nous avons l'habileté de comprendre même les règles les plus vagues, car nous tirons des enseignements de l'expérience et nous faisons nos évaluations selon les différentes circonstances qui se présentent. Cependant, la machine ne peut pas exécuter d'instructions ambiguës : il est impératif de donner aux calculateurs des informations les plus transparentes et univoques possibles (cf. Quarta/Smorto 2020, 14–15).

1.2.2 La définition d'intelligence artificielle

Maintenant, nous avons tous les éléments pour commencer à définir l'intelligence artificielle et ses applications les plus capitales et pour tracer son histoire. Tout d'abord, nous expliquons ce que « intelligence artificielle » signifie : c'est la « Recherche de moyens susceptibles de doter les systèmes informatiques de capacités intellectuelles comparables à celles des êtres humains. » (Centre National de Ressources Textuelles et Lexicales, s. d.). Pour utiliser des mots plus simples à comprendre, nous pouvons dire que l'intelligence artificielle est une sorte d'ensemble de plusieurs technologies qui est entrainée à penser (à peu près) et à agir comme l'homme le ferait. Selon le Conseil de l'Europe, l'intelligence artificielle est :

> […] en réalité une discipline jeune d'une soixante d'années, qui réunit des sciences, théories et techniques (notamment logique mathématique, statistique, probabilités, neurobiologie computationnelle et informatique) et dont le but est de parvenir à faire imiter par une machine les capacités cognitives d'un être humain (Conseil de l'Europe, s. d.).

1.2.3 La naissance de l'intelligence artificielle, les premiers travaux et les systèmes experts

Pendant l'été 1956, John McCarthy et Marvin Minsky ont organisé des rencontres dans le Dartmouth College, qui est une université privée située dans la ville de Hanover, dans l'État du New Hampshire, aux États-Unis (cf. Konieczny/Parde 2020, 7). John McCarthy (1927-2011) était un informaticien et mathématicien américain qui est considéré comme l'un des pères de l'intelligence artificielle avec son collègue Marvin Minsky. McCarthy défendait la logique mathématique en relation avec l'intelligence artificielle, il a contribué directement à la robotique et il est l'inventeur du langage de programmation LISP (*list processing*), qui a favorisé le développement de l'informatique symbolique (Treccani. Vocabolario online, s. d.). Marvin Minsky (1927-2016) était un informaticien américain, il s'occupait de sciences cognitives et il était professeur au MIT (Massachusetts Institute of Technology). Comme déjà anticipé, il est l'un des fondateurs de l'intelligence artificielle, il a notamment travaillé sur la représentation des connaissances et sur l'apprentissage automatique (Treccani. Vocabolario online, s. d.). Initialement, d'autres chercheurs qui participaient aux rencontres à Hanover n'ont pas donné leur appui aux deux pionniers, en voyant dans leurs idées seulement un traitement complexe de l'information. Cependant, McCarthy et Minsky ont reçu le support d'autres chercheurs, par exemple, Claude Shannon (1916-2001), un mathématicien américain qui est considéré comme le père de la théorie de l'information[5] et Nathaniel Rochester (1919-2001), un informaticien américain qui a conçu le premier ordinateur commercial (cf. Konieczny/Parde 2020, 7).

De toute évidence, l'intelligence artificielle naît grâce aux multiples contributions d'autres érudits, spécialisés pas seulement en informatique, mais aussi en divers domaines qui se sont révélés vitaux pour jeter les bases de l'intelligence artificielle. Warren McCulloch (1898-1969) était un chercheur américain ; il a étudié de nombreuses disciplines, parmi elles, nous trouvons la théologie, la philosophie, la psychologie et la physique mathématique, et il a également atteint un doctorat en médecine. Pour ce qui concerne les expérimentations sur l'intelligence artificielle, McCulloch a travaillé sur

[5] « La théorie de l'information […] est une théorie probabiliste permettant de quantifier le contenu moyen en information d'un ensemble de messages, dont le codage informatique satisfait une distribution statistique précise » (Techno-Science.net, s. d.).

l'informatique et sur les fonctions neuronales (University of Illinois Board of Trustees, 2014). Walter Pitts (1923-1969) était un autodidacte américain qui a approfondi le fonctionnement des neurones en relation avec l'informatique, les réseaux neuronaux dynamiques et l'apprentissage (Washington University in St. Louis, s. d.). McCulloch et Pitts ont travaillé ensemble sur la base de la neurophysiologie : ils ont mis au point un modèle mathématique pour la compréhension des relations neuronales et ils ont contribué aux principes fondamentaux de la cybernétique (University of Illinois Board of Trustees, 2014). Les deux pionniers ont publié l'article *Un calculateur logique des idées immanentes dans l'activité nerveuse* (en anglais *A Logical Calculus of the Ideas Immanent in Nervous Activity*). Cet article présente la possibilité d'un système de neurones artificiels d'exécuter des fonctions logiques essentielles. Conformément aux idées de McCulloch et Pitts, un calculateur pouvait apprendre comme l'homme le fait, c'est-à-dire avec l'expérience, qui permet à la machine de reconnaître les tentatives et les erreurs faites, en renforçant ou affaiblissant les connexions entre les neurones. Pour publier cet article, les deux collègues avaient consulté les études faites par des biologistes dans les années 1940 : ils ont constaté que les signaux entre les neurones de notre cerveau étaient la source de l'apprentissage et de l'intelligence. En se conformant donc à ces biologistes, qui se sont occupés de cette recherche pour la première fois, la théorie qui continue à être validée à ce jour confirme l'importance de la fréquence des communications et des tentatives : les deux sont fondamentales pour fortifier les connexions entre les neurones (cf. Signorelli 2017). John von Neumann (1903-1957) était un mathématicien et physicien américano-hongrois qui a travaillé sur la théorie quantique et la nouvelle mécanique statistique (Treccani. Vocabolario online, s. d.). Alan Turing (1912-1954) était un mathématicien et cryptologue britannique qui est reconnu comme l'un des pionniers de l'intelligence artificielle (Treccani. Vocabolario online, s. d.). Turing voulait résoudre le problème de la décidabilité posé en avant par le mathématicien David Hilbert et il a élaboré la machine de Turing en 1936 : c'est un modèle de machine abstrait qui doit vérifier la validité d'une proposition énoncée d'un système logique. Cette machine est « un outil indispensable […] largement utilisé en théorie de la calculabilité, en théorie de la complexité ou en théorie de l'approximation. » (JP Zanotti, 2023). John von Neumann et Alan Turing ont travaillé sur la technologie qui deviendra le squelette de l'IA. Les calculateurs du 19^{e} siècle étaient basés sur une logique décimale, c'est-à-dire

que les calculateurs utilisaient dix symboles, de 0 à 9. Grâce aux Neumann et Turing, les ordinateurs changent leur logique décimale et ils commencent à utiliser la logique binaire qui adopte seulement deux symboles : 0 et 1. De cette manière, les deux chercheurs ont élaboré l'architecture des calculateurs modernes (Conseil de l'Europe, s. d.).

Comme nous pouvons donc affirmer grâce aux derniers concepts, les années 1950 ont été caractérisées par des idées innovantes concernant la possibilité de créer de « machines pensantes » (cf. Konieczny/Parde 2020, 7). Les scientifiques, qui étaient d'accord avec McCarthy et Minsky, soutenaient la théorie que les machines peuvent reproduire les connaissances de l'homme et l'intelligence humaine. Pour créer une telle machine, les chercheurs étaient convaincus que les calculateurs devaient recevoir des instructions très claires, des descriptions sans ambiguïtés pour exécuter l'algorithme demandé. À la suite des réunions qui se sont tenues à Hanover, les premiers travaux se consacraient à instruire attentivement les machines. Le résultat de ces expérimentations est les systèmes experts, en anglais *expert systems*. Un système expert est un « Ensemble de logiciels dont les capacités de résolution de problèmes nouveaux dans un domaine donné sont assimilables à celle d'un expert humain spécialiste de ce domaine. » (Dictionnaire en ligne Larousse, s. d.). Pour mieux comprendre, nous pouvons faire un exemple très simple : un système expert est capable d'automatiser certains processus qui étaient réservés exclusivement à l'homme, comme le dosage de médicaments (cf. Quarta/Smorto 2020, 15–16).

Les expérimentations et les études faites après les rencontres organisées par les deux chercheurs John McCarthy et Marvin Minsky sont seulement une sorte d'épreuve dans une période où l'intelligence artificielle était encore un embryon. Les programmes informatiques des années 1950 créés pour le développement de l'intelligence artificielle étaient certainement nouveaux et très recherchés, mais ils n'étaient pas totalement différents de ceux qui existaient déjà (cf. Quarta/Smorto 2020, 15–16). En fait, les systèmes experts des machines du temps ont reçu des règles selon la logique du « si…alors… »[6]. Cette règle s'appelle instruction conditionnelle avec *if* et elle « […] permet d'évaluer une condition logique, et selon le résultat, d'exécuter ou non une instruction. » (Université du Québec en Outaouais, s. d.). Ces règles donc sont des règles

[6] En anglais, cette logique s'appelle *if then*.

préméditées, car elles disent aux machines « si *x* se produit, alors exécuter *y* ». Avec ce type d'instruction, la machine est très lointaine d'apprendre des nouvelles fonctions, en présentant ainsi d'importantes limites. Dans cette première phase d'expérimentation et de développement de l'intelligence artificielle, les chercheurs ont rencontré plusieurs difficultés, par exemple, les coûts très élevés et le problème concernant la fiabilité des décisions prises par les machines. Nous pouvons dire que les travaux initiaux, qui ont intéressé cette partie de l'informatique, ont donné de bons résultats, mais aussi ils ont affronté de déplaisantes défaites (cf. Quarta/Smorto 2020, 16).

Il existe plusieurs exemples de programmes informatiques qui sont très voisins à être une intelligence artificielle. Le système ELIZA est l'un des premiers *chatbots* de l'histoire. La paternité de ce programme revient à Joseph Weizenbaum (1923-2008), un informaticien allemand. Le nom que Weizenbaum a donné à son programme est vraiment curieux : ELIZA prend son nom en référence à Eliza Doolittle de la pièce *Pygmalion* où elle joue le rôle d'une fleuriste avec un accent faubourien et qui reçoit l'aide d'un professeur de phonétique, Henry Higgins, pour parler comme une duchesse. La dynamique du programme ELIZA est la même : ELIZA répond à son interlocuteur de manière qu'il/elle pense de parler avec un autre humain, dans ce cas-là avec un psychothérapeute (cf. Tarnoff 2023). Weizenbaum a créé ELIZA en 1965 pour étudier le langage naturel et pour analyser la communication entre l'homme et les machines. Pour fonctionner de manière performante, ELIZA a besoin d'identifier les mots clés et le contexte de base ; sur la base de règles de décomposition, ELIZA transforme des phrases faites et doit posséder une bonne capacité d'édition et de révision (cf. Weizenbaum 1966, 36). Voici un exemple d'une conversation en anglais entre le *chatbot* (a) et un humain (b) :

b: Men are all alike.

a: IN WHAT WAY.

b: They're always bugging us about something or other.

a: CAN YOU THINK OF A SPECIFIC EXAMPLE.

b: Well, my boyfriend made me come here.

a: YOUR BOYFRIEND MADE YOU COME HERE.

b: He says I'm depressed much of the time.

a: I AM SORRY TO HEAR YOU ARE DEPRESSED (cf. Tarnoff 2023).

Un exemple remarquable de système expert est DENDRAL (1965) : c'est le premier système expert et il est l'acronyme de *DENDRitic ALgorithm*. Ses créateurs sont Edward Feigenbaum et Joshua Lederberg. DENDRAL avait une tâche précise : ce système expert devait cartographier la structure des molécules dans le but d'aider les chimistes pour identifier la présence de molécules organiques inconnues. Pour reconnaître les structures moléculaires, DENDRAL utilise une série de règles et des données/informations (nous avons vu qu'il s'agit d'un algorithme) pour définir les structures possibles, les comparer et déterminer les bonnes structures. Comme nous avons déjà remarqué auparavant, les systèmes experts sont capables d'automatiser la prise de décision et la résolution de problèmes à la place de l'être humain : c'est le but du programme DENDRAL, qui donne un support aux chimistes (Klondike, 2021).

1.2.4 L'intelligence artificielle fait un bond en avant grâce à l'apprentissage automatique

Les années qui suivent les premières expérimentations sur l'IA sont toujours caractérisées par le scepticisme. Le fait de donner aux machines une sorte de « intelligence » semblable à celle de l'être humain était considéré comme quelque chose d'inimaginable. Les années 1970 et 1980 sont appelées « grand hiver de l'intelligence artificielle » de l'anglais *AI winter* et elles apportent des petites innovations et de nouvelles études, mais les résultats n'étaient pas si productifs comme les objectifs atteints dans les années 1950. Il faut dire que l'intelligence artificielle entre dans son âge d'or avec le nouveau millénaire où elle se développe hors de proportions (cf. Quarta/Smorto 2020, 16).

Avant les années 2000, nous assistons à la création de différents systèmes experts dont nous avons déjà expliqué la fonction et présenté deux exemples. Les chercheurs, qui s'approchaient du nouveau millénaire, ont travaillé abondamment à la conception de nouvelles techniques. Les modernes études sur l'intelligence artificielle se sont concentrées sur l'abandon de la logique « si…alors… », autrement connue par la définition anglaise *rule based*, c'est-à-dire la déjà mentionné logique – très traditionnelle – qui s'appuie sur une séquence finie d'instructions et de données univoques, qui ont caractérisé les calculateurs dans les années que nous avons vu jusqu'à maintenant. Dans cette période l'apprentissage automatique entre en jeu, en anglais nous utiliserons le terme *machine learning*, qui est devenu un terme international utilisé de manière uniforme.

Grâce à l'apprentissage automatique, les ordinateurs ne doivent plus recevoir d'instructions univoques et finies. Maintenant ils peuvent prendre des décisions sans qu'ils soient programmés par l'homme : ils apprennent donc de l'expérience. Il existe un exemple très facile pour comprendre les nouvelles fonctionnalités d'un calculateur qui suit la logique du *machine learning* : le jeu d'échecs. Une machine traditionnelle, qui est caractérisée par la logique « si…alors… », effectuera une série de coups programmés. Ces coups sont téléchargés dans sa mémoire et la machine peut exécuter tous ces coups, mais elle n'en peut pas générer de nouveaux. Au contraire, les ordinateurs, qui sont caractérisés par la logique du *machine learning*, peuvent utiliser les mêmes coups de différents matchs d'échecs, mais ils ont la possibilité de jouer avec eux-mêmes afin de découvrir de nouveaux coups qui n'ont pas été téléchargés dans leurs mémoires. Par conséquent, nous pouvons dire que les technologies du *machine learning* peuvent être comparées aux capacités de l'être humain, car elles tirent des enseignements de manière constante de l'expérience et selon cette dernière elles basent ses comportements, les changent et peuvent arriver à trouver des solutions inédites (l'homme ferait le même processus) (cf. Quarta/Smorto 2020, 16–17).

L'intelligence artificielle est donc passée de l'utilisation de la logique très traditionnelle du « si…alors… », à la logique de l'apprentissage automatique, qui a permis aux chercheurs d'amplifier ce domaine.

1.2.5 L'intelligence artificielle générale veut égaliser l'être humain

Nous avons exploré les différentes phases de la naissance et du développement de l'intelligence artificielle : la première phase qui était caractérisée par les travaux réalisés grâce aux contributions de plusieurs chercheurs et par les systèmes experts et la seconde phase qui était marquée par la technologie du *machine learning*.

Actuellement, l'IA est en train de traverser sa troisième phase de développement. Jusqu'à présent, l'intelligence artificielle avait le but de s'occuper de domaines spécifiques. Cet emploi de l'IA représente une limite pour sa croissance et les chercheurs ont décidé d'identifier son spécifique trait comme « intelligence artificielle faible »[7], précisément pour désigner qu'il agit d'une intelligence artificielle performante et capable de surpasser l'être humain dans certaines circonstances, mais en même temps elle se remet à zéro chaque fois que nous apportons des changements aux tâches données. Un

[7] En anglais, nous utiliserons l'expression *narrow AI*.

exemple est de demander à la machine de jouer au jeu de dames avec ses connaissances du jeu d'échecs : c'est impossible ! L'être humain possède deux attributs que les machines ne possèdent pas (ou elles les possèdent, mais dans une mesure très limitée) : ils sont la flexibilité et la polyvalence. L'homme peut donc utiliser ses connaissances et se baser sur son expérience même quand le domaine change. Les chercheurs veulent faire de grands progrès pour ce qui concerne l'intelligence artificielle et une idée sera de mettre en œuvre des machines qui disposent des mêmes compétences de l'homme (cf. Quarta/Smorto 2020, 18).

Nous avons vu qu'il existe une expression spécifique pour identifier une IA à peu près « limitée » et il existe également une expression pour distinguer une autre typologie d'IA : c'est « intelligence artificielle générale »[8]. Cette « super » intelligence artificielle sera capable de raisonner de manière transversale et elle pourra innover et concevoir des idées complètement nouvelles, sans le besoin d'une série d'instructions détaillées à la base et donc fournies par l'homme. L'adjectif utilisé par les chercheurs est justement « générale », car il s'agit d'une IA capable d'opérer n'importe quel domaine. En plus, dans cette troisième phase, les recherches sur l'intelligence artificielle ont l'objectif de lui donner la capacité de communiquer avec l'homme et lui expliquer les motivations qui l'ont poussée à prendre telles décisions et solutions (cf. Quarta/Smorto 2020, 18).

Pour conclure, l'intelligence artificielle devra favoriser l'automation, supporter le travail de l'homme et dialoguer avec lui. L'analyse des différentes étapes du progrès concernant l'intelligence artificielle et de ses éléments a nous permis de comprendre les multiples phases de son évolution ; les nombreux chercheurs qui ont pris en charge sa naissance et ses innovations techniques au cours des années ; et les attentes pour le futur.

1.3 L'intelligence artificielle et les multiples disciplines qui l'utilisent

1.3.1 Des numéros de l'IA

Aujourd'hui l'intelligence artificielle fait partie de nos vies et elle va les remplir de plus en plus. Conformément aux recherches du cabinet d'études de marché Next Move Strategy Consulting, le marché de l'intelligence artificielle augmentera considérablement. En 2021, le marché mondial de l'IA a atteint 95,60 milliards de dollars américains de chiffre d'affaires. Cependant, d'ici à 2030, son chiffre d'affaires va atteindre près de

[8] En anglais, nous utiliserons le terme *general AI*.

1847,58 milliards de dollars américains, en enregistrant un taux de croissance économique de 32,9 % (Next Move Strategy Consulting, 2023). Même la multinationale International Business Machines Corporation, ou IBM[9], a commissionné une recherche sur l'IA pour analyser son impact au niveau global. Cette recherche s'intitule *IBM Global AI Adoption Index 2022*. Selon IBM, en 2022, 35 % des entreprises utilisent l'intelligence artificielle dans leurs activités et 42 % des entreprises sont en train d'explorer le monde de l'IA pour l'utiliser. L'adoption de l'IA et ses technologies dans les processus de production des entreprises sont donc en croissance constante. L'exploitation de l'intelligence artificielle est également considérée à propos de la main-d'œuvre : 30 % des professionnels de l'informatique affirment que les employés gagnent du temps grâce à l'IA, en termes de nouveaux logiciels et d'automatisation, en limitant les tâches répétitives. En plus, l'intelligence artificielle peut aider l'environnement en favorisant le développement durable : 66 % des entreprises appliquent (ou appliqueront dans le futur) l'IA pour atteindre leurs objectifs *green* (cf. IBM 2022, 3).

En suivant, nous pouvons observer des taux concernant l'adoption de l'IA dans différents pays. Le pays qui utilise dans une plus large mesure l'intelligence artificielle est la Chine (58 %), suivie par l'Inde (57 %). Les États-Unis utilisent l'IA seulement pour 25 %, contrairement à ce que la plupart de personnes pensent. Au-delà des États-Unis, les pays qui ont un pourcentage d'utilisation de l'intelligence artificielle très basse sont : la Corée du Sud (22 %), l'Australie (24 %), le Royaume-Uni (26 %), le Canada (29 %) et l'Amérique latine (29 %). Les pays plus voisins à nous ont des taux plus hauts, par exemple, l'Espagne et la France (31 %), l'Allemand (34 %) et l'Italie (42 %). Au niveau global, le pourcentage compte 34 % d'utilisation de l'IA (cf. IBM 2022, 4).

Pour ce qui concerne notre contexte géopolitique, selon le Parlement européen « La croissance et la richesse de l'Europe sont étroitement liées à la manière dont elle utilisera les données et les technologies connectées » (Parlement Européen, 2023). En Europe (et possiblement dans le monde entier), les opportunités liées à l'emploi de l'IA offrent un terrain fertile pour les citoyens et pour les entreprises. L'intelligence artificielle peut améliorer beaucoup de services qui sont fondamentaux pour le bien-être social, par exemple, la santé, l'éducation et les transports, mais aussi elle peut favoriser un procédé

[9] « International Business Machines Corporation, connue sous le sigle IBM, est une entreprise multinationale américaine présente dans les domaines du matériel informatique, du logiciel et des services informatiques. La société est née le 16 juin 1911 […] » (Wikipédia, 2024).

industriel de plus en plus efficace et, en général, une économie progressivement circulaire et verte. Grâce aux avantages obtenus par l'IA, les consommateurs sont plus protégés et également les services publics, par exemple, l'énergie et la gestion de déchets, amélioreront en termes de réduction des coûts et d'offre de nouvelles opportunités. Ces prévisions sont confirmées par des estimations : le Parlement européen a envisagé d'ici 2035 une notable augmentation de la productivité du travail associé à l'IA (de 11 % à 37 %) et il a calculé que l'exploitation de l'intelligence artificielle pourrait réduire les émissions mondiales de gaz à effet de serre d'ici 2030 : de 1,5 % à 4 % (Parlement Européen, 2023).

L'intelligence artificielle et les autres innovations qu'elle va produire seront des partenaires cruciaux pour aider l'homme à faire face dans un monde où l'environnement est en danger et où les guerres sont à l'ordre du jour.

1.3.2 L'aide de l'IA en différents champs d'études

L'intelligence artificielle est utilisée dans plusieurs domaines. Il existe des domaines qui semblent plus enclins à se servir de l'IA, par exemple, la robotique, les mathématiques et la bio-informatique, mais il existe également d'autres disciplines, qui auraient une tendance inférieure à recourir à l'IA, par exemple, la psychologie et les sciences humaines et sociales.

Si nous pensons à une discipline qui est strictement liée au domaine de l'IA, nous pensons très probablement à la robotique : c'est un « Ensemble des techniques permettant la conception, la réalisation de machines automatiques, de robots ; […], utilisation de ces machines dans un domaine ou un contexte donné. » (Centre National de Ressources Textuelles et Lexicales, s. d.). Un robot est « […] un agent physique réalisant des tâches dans l'environnement dans lequel il évolue. » (cf. Matignon 2011). Il peut fonctionner par suite de plusieurs composants : les actionneurs, les capteurs, les moyens de communication et les moyens de calcul. Grâce à sa structure science-fictive, un robot est capable de percevoir lui-même et l'environnement entourant et il peut accomplir une série de tâches en exploitant de l'autonomie totale ou partielle et de la robustesse. La difficulté la plus éclatante au sujet du fonctionnement de robots est la variabilité : si l'environnement et les tâches sont variables, l'autonomie du robot rencontrera des problèmes relatifs à l'interprétation et à la conduite de ses actions (cf. Konieczny/Parde 2020, 47–48). L'histoire de la robotique remonte à l'antiquité, car l'homme a toujours

essayé de construire des machines qui pourraient travailler à sa place. Léonard de Vinci aurait construit au 16e siècle l'*Automa cavaliere*, en français nous le traduisons le Chevalier mécanique, qui aurait été éventuellement le premier androïde capable de coordonner ses mouvements. Plus tard, au 18e siècle, l'inventeur français Jacques de Vaucanson créera un canard automate qui pouvait boire, se nourrir, caqueter et digérer les aliments (cf. Guillot 2003). En approchant nos jours, le premier robot mobile autonome a été inventé par Grey Walter en 1948 : il s'agissait d'une tortue qui était capable d'aller vers des sources de lumière. Malgré les premières créations de robots (pas programmables), leur mise en place a été possible seulement grâce à la création des transistors[10] et circuits intégrés pendant les années 1950. En plus, pour ce qui concerne le développement initial de la robotique, les industries ont joué un rôle décisif pour l'exploitation des robots dans leurs usines, par exemple, la société General Motors, constructeur automobile américain, qui a créé et utilisé pour la première fois un robot industriel, qui s'appelait UNIMATE. Vers les années 1970, les premiers robots mobiles sont nés et nous rappelons le robot *Shakey*, qui possédait les fonctions de perception, planification et exécution (cf. Matignon 2011). Pendant les premières expérimentations sur la robotique, l'intelligence artificielle n'était pas si présente qu'aujourd'hui, mais la liaison entre ces deux disciplines est presque « naturelle ». L'intelligence artificielle est utilisée pour les études et les projets concernant la robotique, car elle s'occupe de problèmes liés à la perception, la décision et l'action des robots. L'IA représente un point tournant pour la création de robots intelligents, qui auront l'opportunité d'améliorer l'interprétation et la modélisation de l'environnement, la planification et l'exécution de ses mouvements et de tâches à accomplir, la communication et le dialogue avec l'homme et l'organisation de ses fonctions sensori-motrices et cognitives (cf. Konieczny/Parde 2020, 48–49).

Un autre domaine d'étude qui tire profit de l'intelligence artificielle est la neuroscience, c'est-à-dire la science qui étudie le système nerveux. Notre cerveau est un sujet très complexe à examiner pour son anatomie et ses fonctions et il nécessite une étude globale et interdisciplinaire, car il faut intégrer des disciplines comme la chimie, la biologie et la psychologie, pour n'en citer quelques-unes. Nous avons déjà vu que les

[10] « Composant électronique constitué de matériaux semi-conducteurs utilisé pour redresser, amplifier, interrompre des oscillations électriques à la place d'un tube électronique. » (Centre National de Ressources Textuelles et Lexicales, 2012–).

pionniers de l'intelligence artificielle pendant les années 1950 voulaient créer des « machines pensantes » et les neurosciences les ont aidés. Par exemple, les neurosciences ont aidé les chercheurs quant à l'apprentissage, en introduisant les premiers modèles de neurones artificiels et, en même temps, l'intelligence artificielle s'est occupée de la création des outils pour traiter les données et les connaissances que les neurosciences génèrent et acquièrent. Ces deux sciences collaborent également sur le traitement de l'information et sur la compréhension de processus de résolution de problèmes. Nous devons donc reconnaître que l'intelligence artificielle n'a pas seulement contribué à la croissance des neurosciences, mais elle a même reçu un bagage d'éléments fondamentaux pour son développement. Nous pouvons ainsi parler d'un échange réciproque et riche entre les deux disciplines (cf. Konieczny/Parde 2020, 60–63).

Au cours des dernières années, l'intelligence artificielle a influencé et changé l'interaction humain-machine. L'homme et la machine ont désormais un lien solide et indissoluble, qui ne cesse pas de se renforcer. Les principaux problèmes relatifs à cette relation sont la qualité, la performance et l'expérience de l'homme avec la machine en général. L'interaction humain-machine doit être multimodale et la machine doit prendre en compte plusieurs signaux émis par l'humain. Aujourd'hui, les chercheurs travaillent sur des machines qui aspirent à avoir les mêmes capacités d'un être humain. Elles devraient bien comprendre les intentions et les émotions de leurs interlocuteurs humains, elles doivent prendre les décisions les plus appropriées et elles ont besoin de s'adapter à leur individualité et culture (cf. Konieczny/Parde 2020, 50). En plus, la relation entre la machine pilotée par l'intelligence artificielle et l'homme devrait être caractérisée par la coopération, qui ne peut pas exister sans une confiance et une adaptation réciproques. Pour franchir cette ligne d'arrivée, l'intelligence artificielle doit collaborer avec la psychologie. Actuellement, les chercheurs de ces deux champs d'études ont trouvé trois défis concernant cette coopération insolite. Le premier défi fait référence à la complémentarité : une machine intelligente, qui va substituer l'humain, doit être capable de fournir une performance meilleure, car elle peut découvrir les limites et les erreurs de l'homme, par exemple, la négligence des informations d'importance. Le second défi est l'explicabilité : en donnant des conclusions, un agent artificiel devrait être en mesure de fournir une justification. La présence d'une justification est particulièrement grave lorsque nous sommes confrontés à des erreurs. Si nous pouvons comprendre et remonter

au pourquoi, nous maintenons notre confiance, mais, d'une manière opposée, nous la perdrons. Le dernier défi est l'acceptabilité éthique, qui concerne la prise des décisions par les agents artificiels qui comporte des conséquences pour l'humain. Afin de confier l'intelligence artificielle, l'homme devrait accepter les principes éthiques des machines et, pour le faire, la psychologie devra travailler ensemble aux chercheurs pour fournir aux citoyens une intelligence artificielle efficace et transparente, qui tient en compte l'éthique (cf. Konieczny/Parde 2020, 63–64).

Pour conclure cette partie dédiée aux applications de l'intelligence artificielle, nous parlerons de son emploi dans quelques sciences humaines et sociales. Dans les années 1990, l'IA est utilisée en économie : l'intelligence artificielle cherche à programmer des agents qui peuvent modéliser le comportement humain, en utilisant, par exemple, la théorie de la décision et la théorie du choix social ; et elle peut également questionner les théories économiques sur la base de méthodes de simulation et d'expérimentation. Les résultats obtenus par l'intelligence artificielle sont visibles aussi en matière de droit. En ce cas-là, l'apprentissage automatique se rend très utile pour l'analyse des nombreux verdicts juridiques et aider la jurisprudence à les choisir. Ce processus se base sur d'algorithmes qui disposent d'une quantité énorme des données. En même temps, les décisions prises par les algorithmes de l'IA peuvent être injustes pour certains citoyens, une éventualité qui nous ramène au problème de l'éthique. L'intelligence artificielle est également l'objet des études du droit : en fait, le législateur s'occupe de régler le domaine de l'intelligence artificielle, et, en particulier, ce qui concerne les robots, les drones et les véhicules autonomes (cf. Konieczny/Parde 2020, 65).

1.3.3 Les problèmes liés à l'IA

Dans les pages précédentes, nous avons pu apprendre que l'intelligence artificielle est aujourd'hui une partie intégrante de nos activités et elle représente l'aide dont l'homme a besoin pour travailler de manière efficace et efficiente. L'intelligence artificielle est devenue essentielle pour certaines disciplines qui sont relativement les plus ardues à étudier et nous les avons vu, par exemple, les neurosciences et la robotique, mais nous pouvons mentionner aussi l'ingénierie aérospatiale, la physique, la biotechnologie et la bio-informatique. Également dans le domaine médical, l'intelligence artificielle apporte des améliorations : elle peut optimiser les soins des patients et personnaliser les traitements selon les variables exigences ; elle permettra de détecter plus facilement et

plus rapidement les maladies et les diagnostics ; et elle pourra fournir aux étudiants de médecine une éducation et une formation professionnelle meilleure. En tout cas, l'intelligence artificielle entre également dans nos maisons et dans nos vies et activités quotidiennes. Grâce à l'IA, nous pouvons chercher des informations sur internet ; nous pouvons nettoyer la maison avec des robots aspirateurs dotés d'une intelligence artificielle ; et nous pouvons imaginer des véhicules qui seront de plus en plus autonomes (cf. Konieczny/Parde 2020, 70–71).

Nous avons donc la possibilité de disposer de « petits oracles informatiques » (cf. Konieczny/Parde 2020, 71). Ces oracles informatiques nous servent pour les demander conseils, pour les déléguer certaines tâches, en nous permettant d'avoir un processus de décision plus rapide et d'économiser notre temps afin de donner la juste attention aux tâches plus importantes. Cette opportunité qui se présente à l'homme n'est pas toujours le choix le plus fiable à suivre, car il existe des problèmes de confiance, de morale et d'éthique. Les décisions les plus critiques ne devront jamais être prises par l'intelligence artificielle : la prise de décision est à l'homme lui-même, car la machine doit seulement donner son aide et ne pas avoir de pouvoir décisionnaire. Cette éventualité est autrement grave lorsque nous parlons des décisions qui comportent des conséquences pour les citoyens. En fait, l'intelligence artificielle, s'elle tombe dans les mains de sujets irresponsables et motivés par des intentions hostiles, peut être utilisée à de mauvaises fins. L'exemple le plus évident est la sécurité. Nous ne pouvons pas permettre à l'intelligence artificielle d'utiliser librement les systèmes d'armes ou accéder à toutes nos données. Par conséquent, les chercheurs qui s'en occupent travaillent pour encadrer et limiter l'IA au moment où elle peut devenir un problème (cf. Konieczny/Parde 2020, 70).

Il existe la conviction que l'intelligence artificielle pourra dépasser l'homme. Selon diverses opinions, l'IA n'est pas toujours une bonne affaire et, dans les films de science-fiction, nous avons vu plusieurs fois que les robots et l'intelligence artificielle ont pris le pouvoir du monde. L'éventualité que la situation nous échappe à tout contrôle est évidemment une possibilité lointaine, mais elle ne peut pas être exclue et les chercheurs et les législateurs doivent en tenir compte. En fait, l'homme est partiellement responsable pour les conquêtes et les œuvres de l'intelligence artificielle. Les machines peuvent vanter de nombreuses facultés, par exemple, le raisonnement, la mémorisation, l'apprentissage et la perception, seulement sur la base des mérites de l'homme. En plus, toutes les données

que nous avons fournies aux machines consentent un processus d'apprentissage qui pourra ne plus nécessiter de l'intervention de l'être humain. Sans la médiation humaine et avec une autonomie considérable, les machines pourraient devenir imprévisibles et dangereuses. Dans ce cadre spécifique, la question éthique devient centrale : il est donc indispensable de s'intéresser « […] aux principes qui régissent les comportements individuels et aux conséquences sociales et morales du développement des sciences et de leurs applications pratiques. » (cf. Konieczny/Parde 2020, 72–73).

1.4 Le traitement du langage naturel (TAL)

« Le traitement automatique du langage naturel (TAL ou TALN) est une discipline de l'informatique et des sciences du langage […] » (cf. Konieczny/Parde 2020, 27). En anglais, nous utiliserons le terme *Natural Language Processing* ou son sigle NLP. L'histoire du TALN commence en 1954 avec la création du premier traducteur automatique. Dans cette période-là, la guerre froide caractérisait l'ordre géopolitique du monde et l'URSS se développait dans de nombreux domaines tels que l'astronautique et donc la course à l'espace. Par conséquent, les Américains souhaitaient se mettre au courant avec les publications techniques des Soviétiques. Le problème le plus ardu était la langue : le russe. Tous les militaires américains ne connaissaient pas la langue russe et pour éviter de la faire apprendre, ils ont opté pour la traduction automatique. Cette première expérience était encore rudimentaire : le vocabulaire comptait seulement 250 mots et la grammaire 6 règles, mais elle va être le premier travail de développement parmi tant d'autres dans ce domaine d'étude (cf. Yvon 2007, 1). Nous avons déjà déclaré que les premiers travaux sur l'intelligence artificielle et sur le TAL étaient caractérisés par le scepticisme et le niveau d'optimisme relevé parmi certains chercheurs était excessif. Depuis 1954, les techniques de traduction consistent en la seule et simple traduction des textes mot à mot et les chercheurs se sont concentrés sur la création et la manipulation de dictionnaires électroniques. Cependant pour arriver à une traduction acceptable nous avons besoin de nombreuses connaissances contextuelles et encyclopédiques. Il existe un exemple qui explique parfaitement ce problème : la phrase anglaise « *The spirit is willing but the flesh is weak* »[11] a été traduite en russe, puis traduite encore en anglais. La traduction finale (en anglais) était : « *The vodka is strong but the meat is rotten* »[12], il est

[11] En français : « l'esprit est fort mais la chair est faible ».
[12] En français : « la vodka est forte mais la viande est pourrie ».

donc nécessaire d'avoir les justes connaissances pour bien traduire la phrase et comprendre que dans ce cas-là le mot *« spirit »* indique l'esprit, non de l'alcool (cf. Yvon 2007, 2).

En 1956, nous assistons à la naissance de l'intelligence artificielle. Les chercheurs qui y ont travaillé voulaient créer des machines intelligentes capables de simuler le langage humain et en général l'intelligence de l'homme. Pour ce qui concerne l'analyse du traitement du langage et son appareillage cognitif, un pas en avant est fait grâce à Noam Chomsky, un linguiste américain. Chomsky a étudié la syntaxe des langues naturelles et, en étudiant le langage, il a formulé des hypothèses curieuses sur la cognition en 1957. Le langage est une faculté à la fois universelle, car tous les êtres humains le développent spontanément et elle est également spécifique à l'espèce humaine, car le langage humain et son système de communication n'ont pas d'égal en termes de richesse et complexité. Les progressives études sur le traitement du langage et les conquêtes concernant l'intelligence artificielle ont permis le développement de divers systèmes pour un traitement simple du langage et sa compréhension automatique, sur la base de mots-clés. L'exemple le plus extraordinaire est le système ELIZA dont nous avons déjà discuté en précédence : c'est un système qui simule une conversation entre un psychiatre et un patient. Bien qu'il s'agisse de systèmes innovants pour l'époque, les contextes communicatifs de ces systèmes sont fortement restreints : ils utilisent quelques formes grammaticales pour le traitement des phrases, mais ils n'utilisent pas la syntaxe tous azimuts, et surtout ils n'utilisent pas la sémantique ou la pragmatique. Au fil des années, nous assistons à des changements : la syntaxe est reléguée au second plan et la sémantique devient centrale. Ce n'est pas un hasard que, dans les années 1970, le contexte et les connaissances contextuelles et encyclopédiques acquièrent un rôle fondamental dans la compréhension d'un texte et, par conséquent, la recherche concernant le TAL est élargie à des unités de textes plus grandes, par exemple, des dialogues et des récits. Bien entendu, les études sur la syntaxe ne se sont pas arrêtées là et les modèles syntaxiques sont devenus plus performants et raffinés. En plus, dans les années 1980, les chercheurs travaillaient également sur la morphologie et la phonologie. Aujourd'hui, le traitement du langage naturel intéresse le grand public et il a de nombreuses applications telles que la traduction automatique. La recherche sur ce champ de l'intelligence artificielle a donc pour ambition la création des mécanismes automatiques pour l'apprentissage de connaissances

syntaxiques, grammaticales, sémantiques et pragmatiques, afin d'accomplir des progrès qui seront essentiels pour améliorer l'expérience d'utilisation du TAL (cf. Yvon 2007, 2–3).

1.4.1 Le TAL et ses liens avec la langue

Le traitement du langage naturel est fondamental pour entraîner les ordinateurs afin de traiter de manière automatique le langage humain, en particulier sa forme écrite, mais aussi sa forme orale. Une machine dotée d'un programme informatique spécifique, qui est capable de comprendre notre langage naturel et qui reçoit un certain stimulus linguistique, est une machine qui peut adopter des comportements que l'être humain adoptera en recevant les mêmes stimuli (cf. Dahl 2010, 66). Cette branche de l'intelligence artificielle s'applique dans la modélisation et l'automatisation des processus cognitifs langagiers et tente d'élaborer des programmes qui peuvent émuler la communication humaine. Pour le faire, la machine a besoin des algorithmes capables de comprendre le langage humain ; des données langagières avec d'éventuelles descriptions linguistiques ; et des outils et des architectures logicielles spécifiques. Il est donc essentiel de définir et analyser la langue : c'est un « Système de signes vocaux et/ou graphiques, conventionnels, utilisé par un groupe d'individus pour l'expression du mental et la communication. » (Centre National de Ressources Textuelles et Lexicales, s. d.). Nous pouvons donc dire que la langue est formée par différents moyens d'expression et ses unités peuvent être définies seulement grâce à leurs relations (cf. Konieczny/Parde 2020, 27–28). Le traitement du langage naturel comprend plusieurs niveaux afin de comprendre un énoncé ou un texte : ce processus permet d'analyser toutes les dimensions de la langue et ses relations. À partir du moment où la machine doit examiner un énoncé, la technologie du TAL commence son travail. Les principales étapes pour une bonne compréhension sont : la segmentation de l'énoncé (ou du texte) en mots ; l'identification des composants lexicaux (traitement lexical) ; l'identification des constituants de l'énoncé et ses relations (traitement syntaxique) ; la représentation du sens et l'association des concepts dans un monde de référence (traitement sémantique) ; et l'identification de la fonction de l'énoncé dans le contexte dans lequel il a été produit (traitement pragmatique). La segmentation est aussi appelée « traitement de bas niveau » et elle est facilitée grâce à la présence de séparateurs explicites, par exemple, les espaces, les virgules, les points et bien d'autres. Pour ce qui concerne les documents électroniques, le

problème se pose lorsque nous parlons de leur format. Depuis quelques années, les documents informatiques sont disponibles dans des formats qui contiennent d'informations utiles telles que la fin de paragraphe et les changements de fontes. Alors qu'auparavant, les formats de ces documents ne fournissaient pas de telles informations sur la structure du texte. Le niveau lexical permet d'analyser les mots et donc de reconnaître les caractéristiques lexicales de plusieurs unités linguistiques. Cette étape comprend l'aide de la morphologie qui est caractérisée par les processus de suffixation, c'est-à-dire l'adjonction d'un affixe, suffixe ou préfixe ; de dérivation, c'est-à-dire la création de nouvelles formes à partir de formes existantes ; et de composition, qui permet de créer des mots composés. Les technologies du TAL ont donc besoin des lexiques électroniques et des analyseurs morphologiques qui peuvent garantir une bonne couverture d'une langue. Le niveau syntaxique comprend la grammaire et il s'occupe de la validité de certaines suites de mots, qui doivent former des phrases acceptables. L'analyse syntaxique présente des obstacles pour la machine. Tout d'abord, les textes ne sont pas toujours corrects, ils peuvent contenir des fautes d'orthographe et la syntaxe doit éliminer les séquences grammaticalement invalides. Puis, les énoncés ont une hiérarchie spécifique et ils ne sont pas seulement formés par les mots, mais il existe également des unités plus hautes : les syntagmes. Dans ce cas-là, le traitement syntaxique a le but d'identifier les différents constituants de l'énoncé, leurs relations et leurs fonctions syntaxiques. Enfin, il se présente aussi sous la possibilité de construire des paraphrases d'un même énoncé, par exemple, l'utilisation du passif. Maintenant, il n'existe pas un analyseur de syntaxe complet pour aucune des langues naturelles, cependant nombreux lemmatiseurs sont capables de désambiguïser un énoncé au niveau morpho-syntaxique, en identifiant aussi la structure des constituants et les relations syntaxiques entre eux. L'étape sémantique s'occupe du sens des énoncés et de respecter les contraintes qui joignent une expression linguistique avec son sens. Le sens dépend du contexte et il peut être représenté par un concept désigné (expression référentielle) ou par un prédicat (expression prédicative). L'un des problèmes plus grands est l'ambigüité et les objets désignés ou les concepts doivent être identifiés de manière univoque. Nous devrons donc disposer d'un contexte concret. La pragmatique s'occupe des attitudes du locuteur, par exemple, la vérité, la désirabilité et la probabilité. La pragmatique est liée au niveau argumentatif du langage et elle comprend aussi les opérations logiques que nous faisons

quand nous devons reconnaître les attitudes de nos interlocuteurs. Pour mieux comprendre le concept de pragmatique, nous utiliserons un exemple très classique : « Il fait plutôt froid ici ». En le disant, le locuteur espère que son interlocuteur ira fermer la fenêtre ou la porte, car il existe une relation causale entre le fait qu'il fait froid dehors et le fait qu'il fait froid dedans. Pour l'interlocuteur c'est logique qu'il ferme la fenêtre ou la porte. Le niveau pragmatique est moins compliqué que le niveau sémantique, bien que les outils qui lui correspondent soient encore sous-développés. Identifier l'intention argumentative du locuteur ou de l'auteur est important pour différentes applications, par exemple, la gestion de dialogue, le résumé de texte et la traduction automatique (cf. Yvon 2007, 6–19).

1.4.2 L'emploi du TAL et ses problèmes

L'emploi du TAL est maintenant varié et désormais quotidien. Nous pouvons faire une distinction entre les différentes typologies d'application du TAL : la lecture de documents, la production de documents et les interfaces homme-machine. L'application du TALN est devenue très utile pour le traitement de nos ressources en langage naturel. L'exemple le plus significatif est la traduction automatique. Nous avons vu que l'histoire du TAL commence grâce à la création du premier traducteur automatique et aujourd'hui la traduction reste une discipline dans laquelle il faut investir. En fait, nous pouvons profiter de nombreuses entreprises qui ont créé des systèmes de traduction automatique, par exemple, Systran et Logos, et de certains moteurs de recherche comme Google qui proposent la traduction automatique des pages web. Puis le TAL est fondamental pour l'indexation automatique de documents électroniques ; la recherche de documents ; la lecture automatisée de documents (pour les stocker ou pour obtenir des résumés) ; et l'analyse d'un corpus concernant des thématiques précises. Dans le domaine de la production de textes, le TAL est également d'une importance décisive. Nous distinguons par exemple les claviers « auto-correcteurs » et la reconnaissance optique de caractères[13], offerts par exemple par des logiciels comme Recognita ou Omnipage. En plus, le TAL s'applique aux correcteurs d'orthographe ou de syntaxe, qui sont déjà disponibles dans les systèmes de traitement de texte ; il est utilisé dans le domaine de la rédaction, en

[13] « OCR ou reconnaissance optique de caractères est également appelé reconnaissance de texte ou extraction de texte. Les techniques OCR […] vous permettent d'extraire du texte imprimé ou manuscrit à partir d'images, comme des affiches, des plaques de rue ou des étiquettes de produits, ainsi qu'à partir de documents comme des articles, des rapports, des formulaires et des factures » (Microsoft, 2023).

favorisant de « bonnes » pratiques rédactionnelles ; dans l'apprentissage assisté par ordinateur ; et pour la génération automatique de documents à partir de spécifications formelles. Le TAL est aussi profitable pour les interfaces naturelles : les technologies du Tal sont appliquées à l'interrogation par l'être humain de moteurs de recherche sur la toile et aux interfaces vocales, qui exploitent les modules de reconnaissance et synthèse de parole, de génération et gestion de dialogue (cf. Yvon 2007, 19–22).

Le TAL peut être un bon allié pour l'homme, mais les difficultés ne manquent pas. La communication humaine est particulièrement difficile à comprendre pour les machines, car elle est caractérisée par la présence de références contextuelles et empiriques du monde. Ces éléments sont considérés comme acquis par l'homme, mais une machine a besoin d'être entraînée et elle nécessite une série de données claires et explicites pour travailler. Nous avons déjà compris que les machines doivent recevoir des instructions univoques et précises afin d'exécuter un algorithme, mais, pour ce qui concerne le domaine du TAL, le langage naturel est riche en ambiguïté. C'est absolument normal pour l'être humain et il/elle ne nécessite pas d'analyser plusieurs alternatives afin de donner un *feedback*. Au contraire, une machine pourrait considérer plus d'une alternative. Pour mieux comprendre ce concept, nous présentons un exemple très simple. Considérons la suivante requête : « Quel est le prix d'un meuble à quatre tiroirs ? »[14]. En ce cas-là, l'être humain ne voit qu'une seule solution, c'est-à-dire que l'interlocuteur veut savoir le prix d'un meuble à quatre tiroirs : rien n'est plus simple pour nous. En revanche, si nous analysons la « pensée » de la machine, nous nous apercevons qu'elle va examiner deux possibilités : elle doit décider si le syntagme prépositionnel « à quatre tiroirs » modifie le nom « prix » ou le nom « meuble ». Si, d'une part, pour un être humain, il est frappant que le nom « prix » ne peut jamais être modifié par le syntagme « à quatre tiroirs », car un prix ne peut pas posséder de tiroirs, d'autre part, la machine examine les deux alternatives qui seront toutes deux probables, si nous ne la programmons pas à comprendre le sens de l'absurde (cf. Dahl 2010, 68–69).

En plus, le langage naturel présente différents niveaux : la prosodie, qui étudie la poésie ; la phonologie, qui s'occupe des sons ; la morphologie, qui étudie les morphèmes[15] ; la syntaxe, qui explore les règles de combinaison des mots pour former

[14] Le document original est écrit en anglais. Cette phrase a été traduite par l'autrice de ce mémoire de l'anglais : « *Which is the price of a cabinet with four drawers?* ».
[15] Un morphème est une sous-unité syntaxique d'un mot.

des phrases ; la sémantique qui s'occupe du sens ; et la pragmatique qui étudie la langue par rapport à la connaissance du monde. Nous n'avons aucun problème à reconnaître et saisir simultanément ces différents niveaux, tandis qu'il est laborieux de les transmettre précisément à un ordinateur. Aujourd'hui les ordinateurs ont développé surtout la reconnaissance vocale, cependant il n'existe pas une machine qui peut substituer l'humain ou qui peut égaler les capacités humaines. En parlant de pragmatique, une autre difficulté du traitement du langage naturel concerne le besoin de partager avec les machines une connaissance pragmatique du monde, qui est une connaissance implicite chez l'humain. Nous pouvons présenter un autre exemple pour expliquer ce concept grâce à la phrase affirmative « J'ai été pris en train de griller un feu rouge et le cochon m'a infligé une amende »[16]. Nous n'avons pas de problèmes à entendre qu'il s'agit d'un agent de police et pas d'un vrai animal, au contraire une machine rencontrera des difficultés à interpréter la même phrase à l'égal de l'homme (cf. Dahl 2010, 69–71).

Un autre problème que la machine peut rencontrer est la présence d'informations implicites. Une conversation quotidienne entre deux êtres humains est accompagnée d'un contexte d'interaction qui favorise la désambiguïsation d'un dialogue. L'être humain est capable de participer au processus communicatif naturel, car il/elle dispose d'une connaissance d'arrière-plan qui permet de comprendre même les informations implicites. Les machines n'ont pas cette possibilité : sans des connaissances, comme la connaissance du monde (ou du domaine), la connaissance générale et la connaissance sur le contexte de l'énonciation, les machines doivent faire face à des problèmes de compréhension épineux (cf. Yvon 2007, 5).

Dans ce premier chapitre, nous avons donc présenté le tableau général historique et actuel de l'intelligence artificielle, en exposant l'importance du génie de l'homme qui nous a conduits dans une ère où la technologie est un facteur majeur dans notre progrès. En plus, nous avons défini et appris la discipline du traitement automatique du langage, qui voit l'union multidisciplinaire de l'intelligence artificielle, de la linguistique et de l'informatique.

[16] Le document original est écrit en anglais. Cette phrase a été traduite par l'autrice de ce mémoire de l'anglais : « *I was caught running a red light and the pig fined me for it* ».

CHAPITRE 2

La terminologie : entre intelligence artificielle et traitement automatique du langage

2.1 Introduction à la terminologie et à la terminographie

Dans ce chapitre, nous allons explorer le monde de la terminologie dans son ensemble et la terminologie liée à l'intelligence artificielle. Nous découvrirons les bases sur lesquelles la terminologie et la terminographie s'appuient, en analysant ses méthodes théoriques et sa mise en œuvre. En plus, nous allons présenter le domaine de l'intelligence artificielle et du traitement automatique du langage (TAL) dans une perspective terminologique.

Tous les domaines d'études se fondent sur une base théorique et sur une base pratique : les mathématiques, par exemple, sont une discipline qui est caractérisée par une partie de théorie qui explique ses lois, axiomes et théorèmes, et une partie concernant la pratique qui s'occupe de leurs applications. La terminologie et la terminographie de la même manière que les mathématiques sont deux faces d'une même pièce : la terminologie s'intéresse à la théorie et au cadre conceptuel de l'étude des termes ; alors que, la terminographie s'occupe de plusieurs activités telles que la collecte des termes, leur compilation et leur gestion (cf. L'Homme 2020, 17). Comme nous pouvons le déduire, ces deux mots « terminologie » et « terminographie » dérivent du nom « terme ». Un terme est une désignation linguistique d'un concept (ISO 1087 : 2019)[1] relevant d'un domaine de spécialité. Dans la plupart des cas, un domaine correspond à une activité socioprofessionnelle. Pour faire un exemple, nous pouvons regrouper les termes : « barre de tâches », « caractères spéciaux », « cliquer » et « curseur » qui sont des termes du domaine de l'informatique (cf. L'Homme 2020, 24).

[1] « L'ISO (Organisation internationale de normalisation) est une fédération mondiale d'organismes nationaux de normalisation (comités membres de l'ISO). L'élaboration des Normes internationales est en général confiée aux comités techniques de l'ISO. Chaque comité membre intéressé par une étude a le droit de faire partie du comité technique créé à cet effet. Les organisations internationales, gouvernementales et non gouvernementales, en liaison avec l'ISO participent également aux travaux » (ISO, 2019).

2.1.1 Un peu d'histoire de la terminologie

Avant de commencer à examiner plus en détail les caractéristiques et les applications de la terminologie et de la terminographie, nous étudierons un petit morceau de leur histoire. Dès le 18^{e} siècle, les chercheurs avaient averti le besoin de donner une dénomination aux concepts scientifiques et pour le faire ils avaient la nécessité d'avoir une série de règles à suivre. Parmi ces chercheurs, citons Claude Louis Berthollet et Antoine-Laurent de Lavoisier qui se sont intéressés à la chimie, et Carl von Linné qui s'occupait de botanique et zoologie. Au fur et à mesure que les technologies se développent, il était nécessaire de nommer les nouveaux concepts et de discuter à propos des termes utilisés. Pendant longtemps, ni les linguistes ni les spécialistes des sciences sociales se sont intéressés particulièrement à la terminologie et seulement quand les scientifiques et les techniciens l'ont pris en considération, nous arrivons à un significatif développement de la discipline (cf. Cabré 1998, 1–2).

Les bases jetées pour parler de terminologie trouvent leurs origines aussi dans le contexte social. À l'époque, la société était encore rurale, les gens vivaient à la campagne et la plupart d'entre eux étaient analphabètes et n'allaient pas à l'école. Cependant, grâce au développement de l'industrie et de l'économie au 20^{e} siècle, la société a changé et l'éducation est devenue de plus en plus importante. Il est évident que de grands changements culturels ont également eu lieu pendant ces années, en identifiant une nouvelle civilisation basée sur la recherche de biens matériels, sur l'individualisme et la compétition et sur le pouvoir et le succès. Nous pouvons donc identifier deux évidents changements culturels : le premier est la technologisation de la société et le second est la valeur de l'information. Par conséquent, les effets majeurs de ces changements se sont concrétisés sur la langue et la communication interpersonnelle, en démontrant ainsi la nécessité pour l'homme de développer de nouveaux produits linguistiques, de nouvelles professions liées à la langue et de nouveaux modes d'organisation de la communication (cf. Cabré 1998, 3).

La terminologie a donc été influencée par les changements sociaux et par les nouveaux besoins linguistiques. Nous pouvons considérer, par exemple, l'incroyable développement de la science et de la technologie qui a conduit à la naissance de nouveaux concepts et champs conceptuels, qui nécessitent évidemment de nouvelles dénominations. Le progrès de la technologie a également créé de nouveaux moyens de

communication et de nouveaux vocabulaires, pour lesquels des mises à jour constantes sont nécessaires. Un autre facteur de changement est le multilinguisme, qui affecte le transfert de connaissances et les nouveaux espaces d'échanges, en soulignant le besoin de normaliser les systèmes et les unités de base du transfert. Ensuite, il y a l'augmentation de la quantité d'informations, qui a créé une masse énorme de données, dont il en résulte l'exigence de stockage et de récupération, ainsi que de systèmes standardisés pour le transfert automatique du contenu. La terminologie développe même grâce au développement de la communication de masse, car la vulgarisation est devenue un point de rencontre entre les experts et le grand public et les termes spécifiques deviennent donc partie intégrante de la culture populaire caractérisée presque entièrement par un lexique général. Le dernier facteur qui va affecter la terminologie est l'intervention de l'État, car il faut dire que les innovations scientifiques et technologiques sont les filles des puissances économiques mondiales, ce qui entraîne un transfert de connaissances caractérisé par un degré élevé d'emprunts dans les autres pays (cf. Cabré 1998, 4).

2.2 La terminologie générale et les autres approches alternatives

Comme nous l'avons déjà prévu, la terminologie s'occupe de la partie théorique de cette discipline et elle offre des modèles théoriques qui servent pour formuler ses principes généraux. L'objectif de la terminologie est de normaliser ou standardiser la communication dans les domaines de spécialité. Dans le 20e siècle, le problème le plus grand était la normalisation de la communication entre les experts pour éviter les ambiguïtés. Il était donc nécessaire de créer des lignes directrices pour travailler dans la terminologie en général (cf. L'Homme 2020, 29).

Nombreux chercheurs ont travaillé sur la terminologie et ses lois, mais le premier chercheur qui a développé les principes de la théorie classique de la terminologie est Eugen Wüster. Il était un ingénieur autrichien qui en 1930 a contribué à l'optique conceptuelle de la terminologie : selon cette optique chaque domaine spécialisé est caractérisé par une série de termes qui sont organisés en fonction des connaissances du domaine lui-même (cf. L'Homme 2020, 26). Tout d'abord, nous avons les termes : ils forment le vocabulaire d'une discipline et ils sont des désignations qui font référence à un concept à travers la langue (pour simplifier, nous dirons qu'un terme désigne un concept). Désigner signifie « pointer » vers quelque chose, par exemple, vers un objet, et, pour chaque objet du monde réel, l'être humain a des images mentales à l'attribuer. Puis

nous avons le concept, qui est une « unité de connaissance créée par une combinaison unique de caractéristiques » (ISO 1087 : 2019). Ainsi, grâce au concept, tous les domaines sont marqués par des structures conceptuelles ou systèmes conceptuels, qui sont formés en raison de leurs caractéristiques et par des « opérations de classement » (cf. L'Homme 2020, 27). En langage clair, l'objectif de cette optique conceptuelle est d'identifier les caractéristiques d'un concept et puis le terme qui désigne le concept. Cette approche est la plus utilisée par les êtres humains, car il s'agit d'un comportement qui nous sert pour nommer les objets du monde réel : cette démarche est appelée démarche onomasiologique.

Comme indiqué ci-dessus, la terminologie s'intéresse à la normalisation des termes et à l'élimination des ambiguïtés dans la communication. Un de rôle de cette discipline est donc d'intervenir sur des phénomènes linguistiques, qui sont la synonymie et la polysémie, pour favoriser la biunivocité, c'est-à-dire que « […] à une forme correspond un seul concept et un concept est exprimé par une seule forme. » (cf. L'Homme 2020, 29). Au contraire, la synonymie est un phénomène linguistique qui est caractérisé par l'utilisation de plusieurs entités linguistiques pour désigner un seul concept ou référent[2]. La polysémie, cependant, indique un phénomène linguistique impliquant la présence d'une seule entité linguistique pour désigner plusieurs concepts[3]. La biunivocité comporte la sélection d'un identificateur unique pour étiqueter et représenter un concept et, s'il existe plusieurs formes linguistiques pour désigner le même concept, une seule de ces formes sera choisie (cf. L'Homme 2020, 29).

2.2.1 La centralité du social dans la socioterminologie

La théorie générale de la terminologie élaborée par Eugen Wüster n'est pas la seule théorie existante : dans les dernières années, plusieurs chercheurs ont étudié différentes et nouvelles approches, en comblant les lacunes de la théorie classique (cf. L'Homme 2020, 34). Les évolutions doctrinales qui ont caractérisées la terminologie intéressent quatre principales disciplines : la sociolinguistique théorique, la sociolinguistique de terrain, la linguistique générale, la linguistique de corpus. Ces sources ont permis d'apporter diverses révisions de la terminologie, par exemple, le

[2] Par exemple, les deux noms *maison* et *habitation* sont synonymes ; les deux adjectifs *beau* et *joli* sont synonymes. Ces deux paires de synonymes désigne un seul concept.
[3] Par exemple, le mot *souris* est une seule entité linguistique qui désigne plusieurs concepts : la souris d'ordinateur, l'animal et la viande d'agneau.

développement de la politique linguistique, la reprise du statut du terme et du lien entre termes et référents et les nouvelles méthodes de la gestion informatisée des écrits (cf. Gaudin 2005, 80–81).

Un premier exemple d'une approche terminologique alternative est la socioterminologie proposée par François Gaudin et qui s'occupe de la circulation des termes en synchronie[4] et en diachronie[5], mais aussi de l'analyse et de la modélisation des significations et des conceptualisations. Dans ce cas-là, les termes sont examinés par rapport à leurs usages sociaux (cf. Gaudin 2005, 80–81). Le besoin d'analyser la dimension sociale de la terminologie naît sous la nécessité d'un accès démocratique aux savoirs contemporains afin de favoriser le respect de toutes identités culturelles. Nous savons que la *lingua franca* dans la communication scientifique est l'anglais, mais elle privilégie une approche conceptuelle qui veut rationaliser les contacts translinguistiques et qui prévoit une hégémonie linguistique et culturelle particulière. Pour cette raison, il était nécessaire d'expérimenter de nouvelles méthodes : c'est le cas du modèle glottopolitique, qui a l'objectif d'éliminer les oppositions qui existent entre langue, discours et parole, en raison du fait que la société les influence toutes (cf. Gaudin 2005, 84). Selon cette nouvelle nuance de l'étude de la terminologie et selon la nouvelle doctrine de Gaudin, les termes ne sont plus des étiquettes de concepts, et il adopte une vision dynamique qui prévoit leur analyse au sein des échanges langagiers dont ils émergent. En plus, les termes doivent être considérés en relation avec les types d'interactions qui les concernent et donc nous pouvons dire qu'il n'existe pas un terme juste en soi, car les termes seront considérés appropriés sur la base des interactions définies (cf. Gaudin 2005, 85–86). Gaudin soutient que l'étude de termes doit également réfléchir sur les tensions concernant les communautés langagières et des idéologies linguistiques (cf. Gaudin 2005, 85).

2.2.2 La dimension du texte dans la terminologie textuelle

La terminologie textuelle est une autre alternative à la terminologie classique de Wüster. Cette démarche proposée par Didier Bourigault et Monique Slodzian examine les termes en tenant compte du texte dans lequel ils sont contenus et utilise des outils de

[4] « État de langue considéré à un point donné du temps en fonction de sa structure propre et sans référence à l'évolution qui a pu amener à cet état » (Dictionnaire en ligne Larousse, s. d.).
[5] « Caractère des faits linguistiques considérés du point de vue de leur évolution dans le temps ; succession de synchronies constituant l'histoire de telle ou telle langue » (Dictionnaire en ligne Larousse, s. d.).

la linguistique de corpus[6] afin de mener les tâches terminologiques. Si selon la vision wüstérienne, nous devons d'abord classifier et désigner le concept et puis trouver le terme, dans cette vision textuelle nous devons identifier le terme pour après définir le concept (cf. L'Homme 2020, 34). La terminologie textuelle a émergé quand même la linguistique de corpus se développait (cf. Condamines 2005, 36). Le corpus est un « ensemble de textes établi selon un principe de documentation exhaustive, un critère thématique ou exemplaire en vue de leur étude linguistique. » (Centre National de Ressources Textuelles et Lexicales, s. d.) et il collecte des termes sur un sujet spécifique, pour maîtriser sa terminologie et sa phraséologie[7]. Dans cette approche, le texte est considéré comme une production langagière effective, mais, selon Wüster, l'utilisation des textes pour constituer de terminologies ne pouvait pas être prise en compte, car les usages qui émergent dans les textes ne représentent pas la base de la création de terminologies, qui sont également menacées par le discours (cf. Condamines 2005, 42). Cependant, pour Bourigault et Slodzian, la pratique terminologique se révèle tout autre, car les textes sont fondamentaux pour leur création et pour les terminologues, qui ne peuvent s'appuyer sur leurs seules intuitions linguistiques dans des domaines où ils n'ont pas de compétence (cf. Condamines 2005, 42).

2.2.3 Le rôle de la culture dans la terminologie culturelle

Une autre approche à la terminologie est liée à la culture : c'est la terminologie culturelle qui a été suggérée par Marcel Diki-Kidiri (cf. L'Homme 2020, 35). Cette nouvelle démarche présente des similitudes avec la socioterminologie, car elles sont caractérisées par une orientation sociale. Pour formuler les principes de la terminologie culturelle, Diki-Kidiri a analysé la manière dont les communautés culturelles appréhendent des concepts et comment elles les représentent, en soulignant les notions de reconceptualisation et d'adaptation de l'expression. Il a basé ses études sur les communautés africaines et en particulier sur le domaine des ravageurs de coton et de l'utilisation des pesticides. Grâce à son travail, nous savons que les concepts et leur représentation influencent le choix des termes et il nous montre que la modalité dont les concepts sont appréhendés parmi plusieurs communautés est différente, même s'il s'agit

[6] « Branche de la linguistique qui se propose d'extraire d'un corpus les connaissances linguistiques essentielles à l'enseignement des langues. Cette discipline a proposé un nouveau regard sur le langage en plaçant le sens dans le discours et non dans la pensée du locuteur. » (Linternaute, 2021).
[7] Nous analyserons plus en détail la question des corpus dans le chapitre suivant.

d'un seul champ de connaissances. Diki-Kidiri a choisi de se concentrer sur les langues africaines, car un de son objectif est l'affirmation identitaire qu'une communauté donnée peut acquérir ; ainsi, il est devenu l'un des premiers linguistes qui ont travaillé pour la « défense » des langues en Afrique (cf. Diagne 2022, 2–3). Cette approche culturelle promeut la conceptualisation comme un ensemble d'étapes : elle part d'un concept premier qui englobe l'idée essentielle d'un objet réel ou non ; puis, elle présente un plan de réalisation ; et pour conclure, les objets sont insérés dans les classes d'objets qui concrétisent le concept. La reconceptualisation, qui est au cœur de la terminologie culturelle, est « […] décrite comme le procédé par lequel les locuteurs reprennent à leur compte un concept conformément à leurs modes de pensée et d'action » (cf. Diagne 2022, 4). La reconceptualisation peut être également définie comme un processus de déconstruction et reconstruction. Cette dernière n'implique pas l'annulation du travail déjà fait, mais elle sert pour déterminer tous les éléments constitutifs de l'objet et sa structure. De cette façon, il sera possible de relever les éléments fondamentaux à considérer pour concevoir des concepts et désignations analogues vers la langue cible. Pour ce qui concerne la partie pratique, Diki-Kidiri a élaboré une méthode qui permet d'individuer le cadre d'émergence sociale du travail terminologique ; de comprendre un concept par rapport à la langue source ; de reconceptualiser le concept dans la culture cible ; et enfin d'implémenter les termes normalisés (cf. Diagne 2022, 4–5). En bref, la terminologie culturelle a l'objectif de retrouver les éléments qui correspondent le mieux aux réalités et concepts qui sont représentatifs d'une communauté locutrice afin de l'outiller terminologiquement (cf. Diagne 2022, 13).

2.2.4 La différence entre « concept classique » et « concept sociocognitif »

L'approche sociocognitive proposée par Rita Temmerman au début des années 2000 représente un autre possible remplacement de la théorie générale de la terminologie de Wüster (cf. L'Homme 2020, 34). Temmerman et Wüster ont présenté des idées très différentes. Tout d'abord, selon Wüster et la théorie classique de la terminologie, les concepts sont des entités fixes et clairement délimitées[8], tandis que selon Temmerman et la terminologie sociocognitive, les concepts représentent des unités de compréhension qui

[8] « Firstly, SOTT starts from concepts which are believed to be clearly delineated […] » (cf. Temmerman 2000, 453). L'abréviation SOTT signifie *standardisation-oriented terminology theory*, la dénomination anglaise pour indiquer la théorie classique de la terminologie.

ont le plus souvent une structure prototypique[9]. Puis, la théorie générale de la terminologie déclare que les concepts possèdent une place dans une structure conceptuelle, alors que l'approche sociocognitive propose l'idée selon laquelle une unité de compréhension est caractérisée par une structure intra-catégorielle et inter-catégorielle et fonctionne dans des modèles cognitifs. Wüster et l'école de Vienne ont attribué au terme à un et un seul concept et les termes et les concepts sont étudiés de manière synchrone, mais pour Temmerman la synonymie et la polysémie sont considérées des traits fonctionnels dans la progression de la compréhension et les unités de compréhension sont étudiées dans leur évolution constante. Pour conclure les différences entre les principes de ces deux approches, il faut expliquer comme le concept de la définition est articulé : selon l'école de Vienne, un concept doit être défini par le biais d'une définition intensionnelle[10] et/ou une définition extensionnelle[11] ; tandis que dans l'approche sociocognitive, la définition de l'unité de compréhension dépend du niveau et du type de spécialisation de l'émetteur et du récepteur et les informations essentielles ou moins essentielles peuvent varier. Pour analyser les unités de compréhension, Temmerman a proposé différentes méthodes d'analyse : elle fournit une analyse intensionnelle et extensionnelle du degré de prototypicité ; une analyse pour dessiner des représentations visuelles de définitions qui peuvent être liées à des modules d'information de différents types et des analyses du degré de pertinence de ces modules d'information ; et des analyses de l'évolution historique (cf. Temmerman 2000, 453–455). La théorie sociocognitive de la terminologie aide donc les terminographes à concevoir des modèles pour décrire les unités de compréhension appartenant à une discipline particulière et elle permet de construire une base de données terminologiques de meilleure qualité (cf. Temmerman 2000, 459).

[9] « Prototypique est employé pour désigner ce qui est le plus caractéristique d'une classe d'objets. Prototypique désigne ce qui appartient à un prototype, qui représente l'idéal de quelque chose » (Linternaute, 2021).
[10] « […] an intensional definition specifies all the properties that the objects have in common » (cf. Park 2016).
[11] « An extensional definition specifies all the objects that a term can be correctly applied to » (cf. Park 2016).

2.2.5 Les multiples faces du terme dans la théorie communicative de la terminologie

La dernière approche alternative qui sera présentée est la théorie communicative de la terminologie proposée par Maria Teresa Cabré. Cette approche ne se concentre pas seulement sur les aspects cognitifs et linguistiques de la terminologie, mais aussi sur la dimension communicative. Selon Cabré, le terme est une sorte de polyèdre et il est donc comparé à une forme géométrique ayant trois dimensions (cf. L'Homme 2020, 35). La théorie communicative de la terminologie présente les termes comme des moyens linguistiques qui ont le but d'exprimer la connaissance spécialisée et ils peuvent être analysés sous trois angles (qui correspondent à la forme du polyèdre) : cognitif, linguistique et communicatif. Conformément à cette vision, les unités terminologiques ont un statut privilégié puisqu'elles sont des éléments qui entrent dans la structure conceptuelle d'un domaine et ils sont également des unités lexicales. Étant donné que nous sommes en présence des unités lexicales, les termes peuvent coïncider avec des unités de langue générale, en appartenant aux parties du discours, par exemple, le verbe et le nom, mais ils peuvent également coïncider avec des structures plus complexes, qui sont les syntagmes nominaux, adjectivaux, verbaux et adverbiaux (cf. L'Homme 2005, 1116). Dans ce point de vue, nous pouvons apprendre que la terminologie est un champ multidimensionnel et interdisciplinaire, et Cabré voulait créer un modèle théorique qui était capable « […] de répondre aux besoins multiples relatifs au transfert des savoirs spécialisés » (cf. Vušović 2014, 86). Selon Cabré, le traitement multidimensionnel attribué aux termes se complète grâce à la troisième dimension : la dimension communicative est relative à la situation, qui intervient sur l'activation du sens spécialisé (cf. Vušović 2014, 86).

2.3 La terminographie et ses pratiques

La terminographie utilise les modèles théoriques de la terminologie et son objectif principal est de « […] décrire des termes dans les dictionnaires spécialisés ou les banques de terminologie. » Nous avons dit que la terminologie et la terminographie s'occupent des termes relativement aux domaines de spécialités. Chacun de ces domaines est caractérisés par des dictionnaires spécialisés qui contiennent tous ses termes spécialisés. Par exemple, des domaines de spécialités sont le droit, la médecine, l'informatique et ils ont leurs propres dictionnaires et termes. Ces dictionnaires peuvent être en format papier

ou en format électronique ; ils peuvent être écrits dans une seule langue ou ils peuvent être bilingues et aussi multilingues. Le format électronique de dictionnaires spécialisés présents des avantages, par exemple, l'absence d'une limite relative à la quantité et à la variété de données qui peuvent être insérées (cf. L'Homme 2020, 50). Les personnes impliquées dans la terminographie et qui y travaillent sont appelées les terminographes : c'est un métier relativement récent dont le but est principalement la collecte et l'organisation des données terminologiques. Le travail du terminographe devient essentiel également pour les travaux des autres professionnels, par exemple, les traducteurs, les rédacteurs spécialisés ou les étudiants (cf. L'Homme 2020, 23–24).

2.3.1 Les sept étapes du travail terminographique

La terminographie se concrétise en sept tâches, qui ensemble déterminent une recherche que nous appellerons recherche thématique, car le travail terminographique finira avec une collecte des termes d'un même domaine. La première étape du travail terminographique est la mise en forme d'un corpus qui est constitué par une série de textes spécialisés pour chercher les termes. Quand le/la terminographe aura à disposition son corpus, il/elle pourra passer au repérage des termes. Ensuite le/la terminographe s'occupera de la collecte de données sur les termes trouvés ; ces données sont une première collection des renseignements. Une fois les données et les renseignements seront collectés, le travail terminographique procédera avec leur analyse et leur synthèse. Pour conclure, il y a l'encodage, l'organisation et la gestion des données terminologiques : l'encodage consiste dans l'insertion des données dans un dictionnaire spécialisé ou une banque de terminologie ; l'organisation sert pour ordonner les données selon divers paramètres, par exemple, en ordre alphabétique ou thématique ; et la gestion a l'objectif d'ajouter ou de corriger les données. Dans le cas où le/la terminographe doit travailler sur plusieurs langues, ces sept tâches sont exécutées pour chacune langue (cf. L'Homme 2020, 52–53).

2.3.2 Description des données terminologiques

Comme nous pouvons l'imaginer, les données terminologiques correspondent à une multitude de renseignements relatifs à une série de termes et le point de départ pour commencer la collecte des termes sont les textes spécialisés qui divulguent les connaissances d'un domaine spécifique. Lorsque le/la terminographe a fini la recherche de termes dans ces textes et il/elle créera une première liste des données. Ce catalogue

initial de termes ne sera jamais égal au catalogue final, car le travail terminographique comprendra des changements et des révisions, comme nous avons vu. Il faut préciser que les terminographes peuvent être spécialisés ou non spécialisés : dans ce cas-là, le/la terminographe non spécialisé(e), qui ne maîtrise pas un précis domaine, a la possibilité de consulter un expert qui pourra l'aider à organiser les données terminologiques et à éclairer les aspects opaques (cf. L'Homme 2020, 42).

À partir du moment où le/la terminographe aura collecté les termes et les données terminologiques, il/elle procédera à la compilation d'un dictionnaire spécialisé ou il/elle aura l'occasion d'enrichir une banque terminologique existante. Les renseignements qui forment les données terminologiques peuvent être synthétisés dans une liste très simple : nous avons tout d'abord l'entrée, qui sera le terme choisi et qui comprendra également les formes lui associés comme les synonymes, les variantes et les sigles ou les abréviations ; après nous trouvons l'information grammaticale, qui indique le genre, la partie du discours ou, en présence de verbes, s'ils sont transitifs ou intransitifs ; puis les marques d'usages sont mentionnées, c'est-à-dire les informations liées à l'emploi du terme, par exemple, l'aire géographique ou le niveau socioprofessionnel où le terme est utilisé; l'une d'information la plus importante est le domaine ou le sous-domaine d'emploi du terme choisi ; ensuite il y a la définition qui explique le sens du terme ; cette liste des données comprend aussi le contexte qui aide la compréhension du terme grâce à une phrase où le terme est utilisé ; enfin, nous trouvons les cooccurrents qui sont des autres termes qui se bien combinent avec le terme objet. Ces éléments sont les principaux, mais ils existent aussi d'autres comme les notes, les illustrations, la prononciation, les familles de mots et les liens entre les termes. Pour compléter ce travail, il sera approprié d'ajouter aussi les données bibliographiques où les sources documentaires seront indiquées. Pour une correcte gestion de toutes ces données, le/la terminographe les réunit sur des fiches de terminologie ou fiches terminologiques (cf. L'Homme 2020, 42–49).

2.4 La terminotique, un lien être informatique et terminologie

Comme nous l'avons déjà prévu, l'informatique et en général la technologie sont des parties devenues fondamentales dans nos vies de tous les jours et elles sont utilisées même dans le domaine de la terminologie. L'informatique et la terminographie ont commencé à travailler ensemble dans les années 1960, lorsque les terminographes avaient de plus en plus de problèmes à diffuser et à gérer de milliers de termes sur papier. Grâce

à l'évolution de la technologie, aujourd'hui la plupart des textes ont un format électronique et les terminographes ont à leur disposition de nombreux instruments informatiques qui leur permettent de systématiser le travail terminographique. C'est ainsi qu'est née la terminotique, qui lie la terminologie à l'informatique (cf. L'Homme 2020, 52–53).

2.4.1 Les changements apportés par l'informatique

Pour approfondir cette thématique, nous analyserons aussi le travail fait par Maria Teresa Cabré, que nous avons déjà rencontré dans ce chapitre. Selon Cabré (1998, 161), les relations entre l'informatique et la terminologie contribuent à leur croissance mutuelle : d'une partie l'informatique influence les activités terminologiques et leur méthodologie, et de l'autre la terminologie donne sa contribution à la recherche en linguistique informatique[12]. Cabré affirme que « La terminologie, un élément essentiel de la communication spécialisée, est donc de plus en plus importante en tant que moyen de transfert de la pensée et de la technologie. » (cf. Cabré 1998, 162).

La croissance continue de l'informatique a contribué à un changement radical de la méthodologie terminologique. Aujourd'hui les terminographes peuvent utiliser les corpus en format électronique et les bases des données informatisées, et le traitement de grandes quantités d'informations et l'analyse de texte assistée par ordinateur ont modifié les bases de la compilation terminologique et aussi le degré d'intervention de l'homme sur le travail terminologique. Alors qu'auparavant, les terminographes devaient s'adresser à un expert pour rechercher un terme qui n'était pas présent dans un dictionnaire, maintenant il leur suffit d'accéder à des bases de données terminologiques ou à des bases de données textuelles spécialisées et de les consulter, en changeant leur manière de travailler. Le/la terminographe pourra chercher un corpus de textes représentatif d'un spécifique domaine qui peut être comparé à d'autres corpus toujours relatifs au même domaine. Tout cela grâce au travail des machines. En même temps, l'énorme quantité de données à disposition des terminologues donne l'opportunité d'obtenir beaucoup d'informations sur les termes qu'ils/elles étudient. Tous ces avantages ont permis aux terminographes de prendre leurs décisions de manière plus solide et efficace, et ils contribuent à un travail

[12] « La linguistique informatique est un champ interdisciplinaire basé sur une modélisation symbolique (à base de règles) ou statistique du langage naturel établie dans une perspective informatique. » En anglais, *computational linguistics* (DataFranca.org, 2024).

terminologique plus flexible. Une autre innovation qui intéresse la méthodologie terminologique est l'accès à des bases de données d'images et de les utiliser. Lorsque le/la terminographe est confronté à des situations particulières pour lesquelles il/elle doit décrire un concept qui fait référence à un objet du monde réel, l'image est le moyen le plus simple et clair à comprendre. Les images donc peuvent être utilisées pour intégrer les descriptions des termes analysés par les terminographes (cf. Cabré 1998, 162–164).

Les tâches de la terminotique sont les mêmes de la terminographie traditionnelle, mais elles présentent quelques nouveautés qui ont été apportées par l'informatique. Les corpus utilisés pour la recherche de termes seront en format électronique qui est exploitable par les logiciels, en permettant de rassembler les textes plus facilement. Le repérage des termes devient plus rapide grâce à la présence et à l'utilisation des extracteurs de termes qui rédigent une liste de candidats-termes et le/la terminographe devra ensuite les vérifier et les sélectionner, afin d'écarter les termes qui ne seront pas plausibles. D'autre part, en ce qui concerne la collecte de données, le concordancier vient en aide : c'est un logiciel et « […] un outil de référence très utile aux linguistes qui permet de faire la recherche dans un corpus d'un mot accompagné de son contexte, que ce soit pour attester son usage ou l'étudier. » (Techno-Science.net, s. d.). Ils existent également des systèmes de gestion de bases de données pour ce qui concerne l'encodage des données terminologiques et aussi leur organisation, qui sont par exemple les logiciels de terminologie ou les banques de terminologie. En plus, les terminographes ont à leur disposition un langage de structuration de documents très avantageux pour organiser les données terminologiques : l'*eXtensible Markup Language* ou XML, qui permet de définir et de stocker les données[13] afin de les partager. Pour conclure, pendant la phase de la gestion des données terminologiques sur support informatique, le/la terminographe pourra utiliser des fonctions de recherche et d'indexation pour corriger, ajouter ou supprimer les renseignements à modifier (cf. L'Homme 2020, 52–55).

[13] Dans notre cas, il s'agit de métadonnées, qui sont des données qui servent à définir des autres données.

2.5 La terminologie à l'aide des autres disciplines

La terminologie et la terminographie représentent un domaine d'étude qui s'est avéré bénéfique pour de nombreuses autres disciplines, par exemple, la communication, la traduction, la rédaction, la documentation, mais encore l'informatique, les sciences cognitives dont l'intelligence artificielle fait partie et l'ingénierie des connaissances.

2.5.1 Communication spécialisée, traduction, documentation et terminologie

La terminologie est fondamentale pour la communication en générale et en particulier pour la communication spécialisée. Quand nous parlons de communication, il y a une importante différence à expliquer : pour communiquer, nous pouvons utiliser la langue générale ou la langue spécialisée ou spéciale. La langue générale est utilisée pour la vie de tous les jours, entre pairs ; tandis que la langue spéciale est utilisée pour l'étude d'un domaine spécialisé, pour la communication spécialisée, entre les experts, mais aussi entre les experts et le grand public, quand la vulgarisation[14] est nécessaire. Les langues spécialisées, par exemple le langage médical, sont riches en terminologie, en les différenciant de la langue générale. Par conséquent, la communication spécialisée se distingue de la communication générale soit par le type de textes produits, soit par l'emploi des termes très spécifiques. L'utilisation d'une terminologie standardisée est avantageuse pour rendre la communication entre les experts plus efficace. Les textes généraux et les textes spécialisés sont différents dans le contenu et la structure : les textes généraux sont caractérisés par l'expression, la variété et l'originalité ; tandis que les textes spécialisés ou textes scientifiques préfèrent la concision, la précision et l'adéquation. Ces trois particularités du texte scientifique assurent un haut niveau communicatif, car elles réduisent les possibilités de distorsion de l'information et permettent au texte d'être précis en raison de la nature des sujets scientifiques et techniques expliqués. En plus, quand nous parlons de communication spécialisée, les textes scientifiques doivent répondre aux besoins de communication d'un domaine spécialisé et ils doivent s'adapter à la situation de communication et aux caractéristiques des interlocuteurs et à leur niveau de connaissance. Dans ce précis contexte, le rôle de la terminologie est central, car elle contribue à la bonne production des textes scientifiques et à la standardisation : les termes

[14] La vulgarisation ou vulgarisation scientifique a l'objectif « de faire partager à un large public les nouvelles découvertes scientifiques et de favoriser chez ce dernier l'acquisition d'une certaine culture scientifique » (Université de Sherbrooke, s. d.).

utilisés pour désigner un concept spécialisé sont presque toujours concis ; l'utilisation d'un terme au lieu d'une explication d'un concept ou d'une paraphrase contribue à une majeure précision ; et la création et l'utilisation d'une terminologie normalisée favorisent un environnement communicatif où les experts peuvent se référer à un domaine spécialisé qu'ils partagent (cf. Cabré 1998, 45–47).

Une autre discipline qui est liée à la terminologie est la traduction, qui est le processus que nous utilisons pour communiquer avec des locuteurs qui parlent une langue différente. Avant, nous avons parlé de communication spécialisée, dans ce cas-là nous parlons de traduction technique ou traduction spécialisée, c'est-à-dire la traduction concernant des textes propres à un art ou à une science. Le processus de traduction spécialisée comprend deux étapes spécifiques : le décodage ou la démarche terminologique et le transcodage ou la démarche traductologique. Le décodage est la compréhension des informations spécialisées contenus dans le texte source, l'identification des termes techniques et la collecte des données terminologiques. Le transcodage est l'étape où le traducteur devra transférer les informations spécialisées dans la culture cible : c'est l'identification des candidats termes, la collecte des données terminologiques et la pratique propre de la traduction. Les traducteurs spécialisés sont des personnalités professionnelles interdisciplinaires, qui doivent connaître les termes, exprimer le même contenu que le texte source, et utiliser les formes qu'un lecteur natif de la langue cible utiliserait. La terminologie joue donc un rôle capital dans le travail des traducteurs, car ils peuvent avoir le besoin de consulter les vocabulaires bilingues ou multilingues ou ils travaillent comme les terminologues, en cherchant eux-mêmes les équivalents aux termes qui ne figurent pas dans les vocabulaires disponibles ou dans les banques de données spécialisées. La terminologie et ses produits doivent également proposer aux traducteurs d'autres informations sur les termes, par exemple le contexte qui est fondamental pour comprendre la terminologie et pour choisir la bonne traduction (cf. Cabré 1998, 47–48).

La terminologie est devenue importante également pour la documentation. Il s'agit d'un domaine relativement nouveau qui comprend une série des techniques dédiées au traitement permanent et systématique de données et de documents et elles englobent leur collecte, leur signalement, leur analyse, leur stockage, leur recherche et leur diffusion, afin de rendre les informations récupérables pour plusieurs usages, utilisateurs et

objectifs. La documentation et la terminologie ont des caractéristiques communes : elles sont interdisciplinaires, car elles peuvent être appliquées à tous domaines de la science ; et elles sont deux disciplines « pratiques », car elles ont l'objectif de fournir aux utilisateurs documents appropriés. En plus, la documentation fournit également des informations sous forme de publications secondaires, par exemple les bibliographies et les répertoires, ou publications tertiaires, comme les bibliographies de bibliographies. À la base de la documentation, nous trouvons le document qui est une unité d'information qui est décrite par sa forme ou par son contenu. Les documents sont accessibles au public par la forme, c'est-à-dire par les données bibliographiques, et par le contenu, c'est-à-dire par des termes qui forment une sorte d'indexation basée sur des descripteurs comme les mots-clés et les symboles. Pour fournir l'accès aux documents et à l'information en général, les documents sont analysés formellement et sémantiquement afin de les décrire. Pour obtenir une bonne documentation, les scientifiques de l'information (par exemple les indexeurs et les bibliothécaires) mettent en place une série de procédures qui comprennent la description formelle des documents, la description du contenu des documents et leur stockage dans un fichier ou une base de données. La terminologie est utilisée par les scientifiques de l'information pour décrire le contenu des documents et, comme nous l'avons déjà anticipé, les mots-clés ou les descripteurs, qui sont des termes, aident l'indexation de la substance d'un document. La documentation utilise donc la terminologie standardisée pour indexer les documents, en proposant un travail systématique et sans ambiguïté. Pour faire cela, on a besoin d'un thésaurus ou thésaurus documentaire qui est une liste organisée de termes contrôlés et normalisés représentant les concepts d'un domaine de la connaissance (Wikipédia, 2023). La terminologie est donc à la base de la documentation, mais la documentation est aussi importante pour la terminologie, car les terminologues recueillent les termes dans les documents écrits par les spécialistes afin d'acquérir une connaissance d'un sujet et de sa structure conceptuelle et de confirmer la qualité des données. Pour conclure, les documents et la documentation font partie de l'entier processus terminographique : le produit final de ce travail est à son tour un document, qui peut être un glossaire, un dictionnaire ou des lexiques (cf. Cabré 1998, 50–52).

2.5.2 Informatique, ingénierie des connaissances et terminologie

La terminologie est liée aux autres disciplines par deux types de relations : la première est une relation dont la terminologie tire des éléments théoriques pour construire ses principes, par exemple les relations qu'elle entretient avec la linguistique et la logique ; la seconde relation est une relation bilatérale, car la terminologie fournit des éléments aux autres domaines et les derniers font le même, et c'est le cas de l'informatique, des sciences cognitives[15] et des sciences de l'information.

Nous avons déjà vu que la terminologie et l'informatique travaillent ensemble pour créer des outils qui vont changer la méthodologie terminologique, en éliminant les procédures manuelles. En plus, la terminologie est fondamentale à l'informatique et aux sciences cognitives pour le développement de l'intelligence artificielle et des systèmes experts, car elle fournit des concepts qui sont très utiles. À la base il y a donc le concept, qui est le point de rencontre entre la terminologie et les sciences cognitives. Un concept est une unité de connaissance et plusieurs concepts ayant les mêmes caractéristiques forment les systèmes conceptuels, qui se réfèrent à un domaine précis de la connaissance. Les concepts sont le pivot de la théorie des termes, de l'intelligence artificielle, qui s'occupe de la création de systèmes experts, et de la théorie de la connaissance, qui a pour objectif l'étude de la formation de la connaissance, des différents types de connaissances et de la relation entre les connaissances des locuteurs et leur utilisation dans des situations réelles. Ensemble, la théorie de la terminologie et la théorie de la connaissance constituent la base de l'ingénierie des connaissances, qui utilise la connaissance pour construire des systèmes experts dont nous avons déjà parlé : ils sont des logiciels avec des capacités de résolution de problèmes assimilables à celle d'un expert humain dans un domaine donné de connaissances (cf. Cabré 1998, 52–53). Pour être plus précis, l'ingénierie des connaissances consiste :

> À articuler des travaux sur la nature et la représentation des connaissances dans des logiciels et chez les utilisateurs, sur l'analyse des usages, des sources de connaissances et des interactions homme-machine. Mais l'Ingénierie des

[15] « Les sciences cognitives constituent une discipline scientifique ayant pour objet la description, l'explication, et le cas échéant la simulation des mécanismes de la pensée humaine, animale ou artificielle, et plus généralement de tout système complexe de traitement de l'information capable d'acquérir, conserver, utiliser et transmettre des connaissances. […] Les sciences cognitives utilisent conjointement des données issues des six sous-disciplines qui la composent : les neurosciences, la linguistique computationnelle, l'anthropologie cognitive, la psychologie cognitive, la philosophie de la cognition et l'intelligence artificielle. » (Wikipédia, 2024).

> connaissances consiste aussi à adapter des approches de génie logiciel, dans le but de mettre des connaissances à disposition des utilisateurs et au sein d'applications informatiques. Les innovations du domaine comprennent donc des méthodes, des logiciels et interfaces d'aide à la modélisation, ainsi que des représentations conceptuelles ou formelles (cf. Aussenac-Gilles, Nathalie/Charlet, Jean, et Reynaud-Delaître, Chantal 2014, 1).

Au centre de cette presque nouvelle discipline, nous trouvons donc l'identification, la représentation, le traitement, la transformation et le transfert des connaissances et elle s'appuie également sur les éléments des différents domaines qui constituent la terminologie et elle utilise le développement des sciences cognitives pour créer les systèmes experts en utilisant la connaissance. Avec le concept, les systèmes experts représentent également un point de rencontre entre terminologie, intelligence artificielle et ingénierie des connaissances. Un système expert qui veut effectuer les mêmes opérations de l'homme doit posséder les mêmes connaissances qu'un homme acquiert par l'expérience et il doit donc être doté des connaissances nécessaires pour reconnaître et analyser une situation afin d'agir en conséquence (cf. Cabré 1998, 53).

Le domaine de l'ingénierie des connaissances est un domaine très spécifique et particulier qui utilise la terminologie, mais elle est appliquée également à l'informatique plus en général et les deux partagent des échanges en termes de connaissances et outils, comme nous avons déjà affirmé. Il est nécessaire d'ajouter que l'informatique bénéficie de la terminologie, pas seulement en relation à la recherche en linguistique informatique, mais aussi en relation à tous les programmes informatiques qui développent des systèmes basés sur des aspects du traitement du langage naturel : la traduction automatique ou la traduction assistée par machine, l'écriture assistée, les programmes de base de données développés en langage naturel, les interfaces en langage naturel et les systèmes experts basés sur la connaissance utilisent tous la terminologie. Pour faire cela, ils nécessitent un dictionnaire contenant les unités que les ordinateurs comprendront et les systèmes basés sur la connaissance en particulier nécessitent également un dictionnaire pour la désignation des concepts et les thésaurus sont donc privilégiés, car ils sont formés par des termes contrôlés et normalisés représentant les concepts d'un domaine de la connaissance (cf. Cabré 1998, 54–55).

2.6 La terminologie de l'intelligence artificielle et du TAL

Dans le premier chapitre, nous avons exploré l'histoire de la naissance et du développement de l'intelligence artificielle et du TAL pendant les années et, comme nous avons pu comprendre, cette discipline s'est développée principalement aux États-Unis du moins dans les premières années. Le fait que l'intelligence artificielle a été créée et développée aux États-Unis signifie que les premières désignations de concepts ont été faites en anglais. Cependant, il faut souligner que, dans tous les cas, la langue de transmission des connaissances scientifiques est aujourd'hui l'anglais, qui est défini comme *lingua franca*.

2.6.1 Introduction à la *lingua franca*

Originairement la signification du terme *lingua franca* était « langue de contact », car elle désignait la langue utilisée pour le commerce international. En réalité, il s'agit d'une langue qui a une importante fonction véhiculaire : dans notre cas, l'anglais est en effet utilisé pour faciliter la communication entre les locuteurs non natifs, elle ne représente pas seulement une langue employée dans le commerce, mais c'est une langue internationale avec une flexibilité fonctionnelle. La langue anglaise est donc devenue une langue internationale et elle est utilisée pour communiquer dans plusieurs domaines, par exemple, la politique et la diplomatie internationale, le droit international, la finance, les médias et la recherche scientifique. Cependant, l'anglais n'a pas toujours été la *lingua franca*, mais c'était le latin la langue la plus utilisée à l'époque de l'Empire romain et plus tard le français pendant le 17^{e} et 18^{e} siècle (cf. Van Croesdijk 2016, 9–11).

La naissance d'une *lingua franca* remonte donc à la nécessité d'une part de commercer avec plusieurs pays et d'autre part elle servait à la communauté scientifique pour communiquer avec un code international unique. Mais que signifie utiliser l'anglais comme *lingua franca* ? L'anglais est vu comme une source linguistique quasi-exclusive et les mots anglais peuvent être pris d'une autre langue sans adaptation phono-morphologique (sans terminaison vocalique) et ils remplacent souvent des mots autochtones qui peuvent représenter le même concept. Dans un domaine comme la science, qui est international par nature, la communication entre pairs se fait non pas dans la langue nationale, mais dans une langue internationale et il est facile qu'une partie de la terminologie prenne la forme de mots étrangers. Il faut dire que les pays agissent différemment en conséquence : certains pays accueillent favorablement l'entrée des

termes anglais dans leur système linguistique et certains pays ne sont pas très favorable à l'invasion de la terminologie anglaise et nous pouvons présenter deux exemples de deux pays qui adoptent de différentes stratégies : l'Italie et la France. L'italien utilise beaucoup de mots anglais sans adaptation phono-morphologique, même quand il existe un mot italien pour désigner le concept en question. Un exemple est le mot anglais *trend* qui est utilisé à la place du mot italien *tendenza*[16]. D'autres exemples des mots anglais utilisés en italien sont *welfare* pour indiquer *stato sociale*[17], *part time* pour indiquer un contrat de travail à temps partiel ou *deadline* pour indiquer délai de livraison de quelque chose. Pour le français la situation est différente : les Français tendent à ne pas utiliser les termes anglais pour préserver leur langue, car l'anglais peut représenter une menace pour la langue française.

2.6.2 Comment la France fait face aux anglicismes

Tout d'abord identifions-nous le phénomène linguistique des anglicismes. L'anglicisme fait partie des emprunts linguistiques, c'est-à-dire « un procédé par lequel les utilisateurs d'une langue adoptent intégralement, ou partiellement, une unité ou un trait linguistique (lexical, sémantique, phonologique, syntaxique) d'une autre langue » (cf. Boccuzzi 2017, 7). Pour mieux comprendre, nous pouvons penser aux mots « gin », « scout » et « babysitter » qui n'ont pas une origine française, mais ils sont des anglicismes, c'est-à-dire des emprunts anglais qui font partie du lexique français. En français, elles existent de nombreuses typologies d'anglicismes comme les anglicismes lexicaux, les anglicismes sémantiques, les calques, les anglicismes phonétiques, les anglicismes phraséologiques et les anglicismes exotiques (cf. Boccuzzi 2017, 7–8).

Pour freiner l'invasion des anglicismes et pour éviter que les professionnels soient obligés d'utiliser des termes étrangers en générale, les pouvoirs publics français incitent à la création, à la diffusion et à l'emploi de termes nouveaux et le législateur français a décidé le 3 juillet 1996 de prendre le décret n° 96-602 relatif à l'enrichissement de la langue française (cf. Boccuzzi 2017, 3). Le premier article du décret indique :

> En vue de favoriser l'enrichissement de la langue française, de développer son utilisation, notamment dans la vie économique, les travaux scientifiques et les activités techniques et juridiques, d'améliorer sa diffusion en proposant des termes

[16] En français, « tendance ».
[17] En français, « protection sociale ».

> et expressions nouveaux pouvant servir de référence, de contribuer au rayonnement de la francophonie et de promouvoir le plurilinguisme, il est créé une commission d'enrichissement de la langue française.
>
> Cette commission travaille en liaison avec les organismes de terminologie et de néologie des pays francophones et des organisations internationales ainsi qu'avec les organismes de normalisation (Légifrance, 2022).

Ce décret du 3 juillet 1996[18] a institué un dispositif qui a donc l'objectif de créer des termes équivalents pour désigner en français les mêmes concepts et les mêmes réalités qui ont été désignés en anglais. Les domaines les plus touchés par l'invasion des anglicismes sont notamment les sciences, l'économie et la technologie. Les acteurs qui participent à ce dispositif sont plusieurs : l'Académie française, la Délégation générale à la langue française et aux langues de France (DGLFLF), la Commission générale de terminologie et de néologie[19], des commissions spécialisées de terminologie et de néologie (CSTN), des hauts fonctionnaires chargés de la terminologie et de la néologie. Tous ces organes politiques vont recenser les anglicismes (ou les autres emprunts linguistiques) entrés dans le lexique français et ils sont chargés de créer des termes français qui remplacent les mots étrangers. Évidemment, une fois les nouveaux termes français créés, il faut les intégrer dans la langue au niveau pratique : le décret de l'enrichissement de la langue française a établi que les nouveaux termes devront être publiés dans le Journal officiel, puis dans tous les textes légaux et réglementaires et dans toute la documentation concertante les services et les établissements publics de l'État (cf. Boccuzzi 2017, 5). Quelques exemples parmi les plus connus : le terme anglais *computer* n'est pas utilisé en France où les citoyens utilisent l'alternative française « ordinateur » ; le terme *social media manager* a été remplacé par son correspondant français « responsable des réseaux sociaux » ; le terme *email* en français s'appelle « courriel ». Ce ne sont là que quelques-uns des nombreux termes anglais qui ont été bannis du lexique français pour favoriser la langue nationale. Au contraire, en Italie, les termes anglais

[18] Le décret du 3 juillet 1996 n'est pas la première disposition législative concernant la langue française. La loi n° 94-665 du 4 août 1994, dite « loi Toubon », relative à l'emploi de la langue française a ajouté le premier alinéa de l'article 2 de la Constitution française qui, depuis, dit « La langue de la République est le français » et le décret n° 89-403 du 2 juin 1989 a institué un Conseil supérieur de la langue française et une délégation générale à la langue française (Légifrance, 2022).

[19] La néologie est le moyen pour créer de nouvelles désignations qui sont nécessaires dans des domaines particuliers caractérisés par l'émergence de nouveaux concepts (cf. Cabré 1998, 204).

computer, *social media manager* et *email* sont employés par les Italiens : il est rare d'entendre un Italien dire *« Ti invio della posta elettronica »*, il dira *« Ti invio una email »*. L'expression italienne *posta elettronica* serait l'équivalent italien du mot anglais *email*, mais elle n'est pas utilisée largement. Parmi les acteurs chargés par le décret du 3 juillet 1996 nous trouvons l'Académie française. Elle représente la référence pour les questions relatives à l'usage de la langue française et elle est membre de la Commission d'enrichissement. L'Académie française adhère à toutes les étapes du processus d'élaboration des termes, à partir de leur élaboration jusqu'à leur publication dans le Journal officiel. Dans un contexte où l'emploi de termes anglais augmente, l'Académie intègre dans la 9e édition de son Dictionnaire des recommandations à faveur de l'utilisation des termes français (Académie française, s. d.). D'autres acteurs qui sont essentiels pour le dispositif d'enrichissement de la langue française sont les groupes d'experts. Ils sont appelés « collèges » et font partie des ministères dont ils sont au cœur du travail terminologique relatif aux domaines, par exemple, la défense, l'économie, l'environnement et la justice. Les membres de ces collèges peuvent être des professionnels d'un secteur, des journalistes spécialisés, des linguistes, des terminologues ou des traducteurs. Ils analysent les nouveaux concepts qui doivent être désignés en français et proposent des définitions à la Commission d'enrichissement. Il y a également plusieurs organismes à caractère scientifique et technique qui sont associés au dispositif d'enrichissement de la langue française. Ils sont par exemple l'Académie des sciences, les organismes de politique linguistique des autres pays francophones, comme l'Office québécois de la langue française, et les laboratoires universitaires. Ces organismes sont consultés pour la création des termes particuliers en collaborant systématiquement à leur examen (cf. Boccuzzi 2017, 27–29).

2.6.3 Les termes du domaine de l'intelligence artificielle et du TAL

Pendant les dernières années, la France a cherché à remplacer les anglicismes dans plusieurs domaines, y compris le domaine de l'intelligence artificielle qui est riche de termes anglo-américains. Avant la terminologie relative à l'intelligence artificielle était rare en français, mais grâce aussi au décret n° 96-602 relatif à l'enrichissement de la langue française, les Français se sont appropriés du vocabulaire de l'intelligence artificielle. Un rôle important a été joué par un organisme sans but lucratif : c'est

DataFranca[20], un groupe francophone présent au niveau mondial composé de scientifiques, terminologues et traducteurs qui se sont engagés à rédiger des milliers de fiches terminologiques et encyclopédiques des termes relatifs à l'intelligence artificielle. DataFranca a donc comme objectif principal d'assurer le développement et la diffusion du vocabulaire français de l'intelligence artificielle. Le projet de DataFranca est soutenu par les Fonds de recherche du Québec et par l'Office québécois de la langue française et il y a aussi la participation de l'Institut québécois d'intelligence artificielle et de Google (udemnouvelles, 2023).

Comme nous l'avons déjà appris, l'informatique et l'intelligence artificielle se développent rapidement et principalement en anglais. Cependant, la communauté francophone du monde entier veut justement utiliser le français même pour ces domaines. Il est donc important de promouvoir la science et la recherche en français (DataFranca.org 2022, 5). DataFranca a alors publié le livre *Les 101 mots de l'intelligence artificielle : petit guide du vocabulaire essentiel de la science des données et de l'intelligence artificielle* et il a créé une page web dédiée au lexique français de l'informatique qui prend le nom de *Grand lexique français de l'intelligence artificielle* (DataFranca.org, 2024). Cette page offre 8.762 termes français, dont 6.901 avec définition et 1.861 sans définition, 5.777 termes anglais et elle comprend également 206 termes concernant l'informatique quantique, 240 termes sur la cybersécurité et 3.546 termes de la statistique (DataFranca.org, 2024).

Nous avons déjà parlé de la vulgarisation scientifique et le livre *Les 101 mots de l'intelligence artificielle* est un ouvrage de vulgarisation, qui a le but de partager au large public les concepts de l'intelligence artificielle pour en faciliter la compréhension. Le livre a été soutenu par Patrick Drouin, professeur de traduction à la faculté des arts et des sciences de l'Université de Montréal, et par Claude Coulombe, diplômé en physique et en informatique de l'Université de Montréal. Les travaux effectués pour rédiger ce livre ont créé des nouveaux termes, mais comment ? Certains termes anglais ont été traduits presque littéralement en français, un exemple est le terme *chatbot*, qui a été traduit par « agent conversationnel ». Dans d'autres cas, les concepts sont exprimés par des métaphores, par exemple, le terme anglais *data mining* et l'équivalent français « forage de données » utilisent la métaphore de la mine. Pour quelques termes, cependant, les

[20] Pour approfondir, voici le lien vers le site web : https://datafranca.org/.

chercheurs ont décidé de les rendre plus explicites : c'est le cas du terme anglais *transformer* qui est devenu en français « réseau de neurones autoattentif » ou « réseau à autoattention ». Il est important de souligner que le vocabulaire du livre va s'évoluer avec les technologies : par exemple, dans le terme « mégadonnées », le préfixe grec *méga-* signifie « million » et ils seront donc stockées dans des mégaoctets, mais si nous parlons des données stockées dans des téraoctets ou pétaoctets, il ne faut pas parler de mégadonnées : nous devons utiliser le terme « données massives » (udemnouvelles, 2023).

La création de ces outils pour transmettre les termes et les connaissances relatifs à l'intelligence artificielle entraîne des avantages considérables. Le lexique pourra entrer dans tous les milieux de travail en implantant en français les innovations de l'intelligence artificielle et également il pourra aider les organismes publics dans la compréhension du domaine. En plus, tous les citoyens pourront comprendre la science des données et de l'intelligence artificielle, grâce à l'exercice de vulgarisation qui va au-delà de la traduction d'un lexique (DataFranca.org 2022, 5–7).

Dans ce deuxième chapitre, nous avons donc présenté la terminologie et la terminographie, en analysant toutes les théories et les étapes du travail terminographique. Nous avons exploré comment la terminologie peut aider les autres domaines et comment tels domaines sont des sources de connaissances pour la terminologie même, en soulignant l'importance du lien entre l'informatique et la terminologie. En plus, nous avons défini les grandes lignes de la terminologie de l'intelligence artificielle, en présentant les difficultés que la communauté francophone rencontre à cause de l'anglais comme *lingua franca* et des anglicismes et nous avons exposé comment ces termes étrangers sont modifiés pour permettre à la langue française d'être première.

CHAPITRE 3

Présentation théorique et analyse de la méthodologie pour la construction d'un corpus

3.1 Exposé général du corpus

Dans le deuxième chapitre, nous avons parlé brièvement du corpus en mentionnant la linguistique de corpus et la terminologie textuelle et le travail terminographique, mais dans ce chapitre, nous allons l'analyser en détail. Il sera proposé le tableau général relatif au corpus, en incluant la description générale, le rôle et les différentes typologies. En plus, nous pourrons suivre pas à pas la construction des corpus qui ont été utilisés pour le projet européen *YourTerm TECH*, qui constitue le pilier de base de ce mémoire.

3.1.1 Définitions et caractéristiques

La première étape du travail terminologique est la mise en forme d'un corpus. Nous pouvons trouver plusieurs définitions de ce terme : au niveau philosophique, le corpus est un « Recueil réunissant ou se proposant de réunir, en vue de leur étude scientifique, la totalité des documents disponibles d'un genre donné, par exemple épigraphiques, littéraires, etc. » (Centre National de Ressources Textuelles et Lexicales, s. d.). Pour ce qui concerne le domaine de la linguistique, nous avons déjà rencontré une définition plus précise du terme : « Ensemble de textes établi selon un principe de documentation exhaustive, un critère thématique ou exemplaire en vue de leur étude linguistique. » (Centre National de Ressources Textuelles et Lexicales, s. d.). Pour avoir une image complète, voici la définition liée au domaine informatique : « Ensemble de données exploitables dans une expérience d'analyse ou de recherche automatique d'informations. » (Centre National de Ressources Textuelles et Lexicales, s. d.).

Nous avons donc toute une série d'éléments pour bien définir ce concept. Tout d'abord, il s'agit d'une collection ou d'un recueil de documents finis et réels, qui nous renvoient à la notion de complétude et d'authenticité. Puis, il y a la notion de recueil de documents-pièces-énoncés dont la notion d'énoncé : les énoncés composants un corpus peuvent être également appelés données langagières, qui sont des données écrites, par

exemple, journaux et lettres, en formant un corpus écrit, ou des données orales, par exemple, conversations et monologues, en constituant un corpus oral. En plus, les documents d'un corpus peuvent être physiques ou électroniques. La révolution numérique nous a permis d'informatiser les textes et de les stocker sur un support électronique et les corpus ont également fait l'objet du processus de numérisation et nous parlerons donc des corpus électroniques[1] (cf. Dubreil 2006, 62–64). La dernière définition que nous proposons est celle de François Rastier, sémanticien français et docteur en linguistique :

> Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications (cf. Dubreil 2006, 64).

Un corpus présente une structure particulière : elle est formée par deux réalités, qui sont la sélection des données et la mise en conformité et l'enrichissement. Pour ce qui concerne la sélection des données, elle implique leur choix et leur possibilité de bien représenter l'objectif de la recherche du travail terminologique. Cependant, la mise en conformité d'un corpus concerne principalement l'harmonisation du codage de ses caractères, par exemple, l'utilisation du codage UTF-8[2]. L'enrichissement de la structure d'un corpus correspond à l'encodage des informations dans un langage de normalisation qui peut être le métalangage informatique[3] XML et à l'annotation linguistique des parties du discours, c'est-à-dire l'attribution aux parties du discours des marques comme nom, nom propre ou verbe. Un corpus est également caractérisé par une finalité : il veut obtenir un échantillon représentatif[4] d'une langue générale ou d'une langue de spécialité (cf. Dubreil 2006, 64–65).

La représentativité d'un corpus se réfère à la mesure dans laquelle un échantillon d'une langue générale ou spéciale inclut toute la gamme de variabilité d'une population,

[1] Actuellement, le terme « corpus » et le terme « corpus électronique » représentent le même concept (cf. Kida 2013, 134).
[2] « UTF-8 (UCS transformation format 8 bits) est un format de codage de caractères. Il permet de gérer tous les caractères dits unicodes. Chaque caractère est codé sur un ou plusieurs points de code. Chaque point de code est codé sur une suite d'un à quatre octets. Il a été conçu pour être compatible avec certains logiciels originellement prévus pour traiter des caractères d'un seul octet. » (Techno-Science.net, s. d.).
[3] « Langage formel servant à décrire un ou plusieurs langages de programmation. » (Office québécois de la langue française, 2000).
[4] Un échantillon représentatif est une « fraction représentative d'une population ou d'un univers statistique. » (Centre National de Ressources Textuelles et Lexicales, s. d.). Dans notre cas, l'échantillon représentatif est le corpus.

c'est-à-dire les variations d'une langue, d'un langage ou d'un ensemble de phénomènes. Les variations sont distinguées entre variations de situation et variations linguistiques : les variations de situation concernent les éléments « extérieurs » au texte, tandis que les autres concernent exclusivement le texte. Pour mieux comprendre, nous pouvons faire un exemple très simple : si nous travaillons sur un corpus composé d'articles scientifiques, les variations de situation se réfèreront par exemple au domaine, au thème et à l'auteur et les variations linguistiques se réfèreront plutôt à l'ordre stylistique des documents, par exemple, la syntaxe et le lexique. Pour être représentatif, un corpus d'une population langagière doit inclure une série des types de textes et une série de distributions linguistiques d'une population langagière. Toutefois, cela peut présenter des difficultés. Tout d'abord, nous devons toujours garder à l'esprit que la langue est une entité qui est toujours en évolution et il est donc complexe de trouver un échantillon qui soit totalement représentatif d'un phénomène linguistique. Puis, il est fondamental de considérer les types des textes, le nombre de textes, leur longueur et leur accessibilité (cf. Goeuriot 2009, 3–4).

La représentativité du corpus a été étudiée par deux chercheurs : Douglas Biber et Benoît Habert. Selon Biber, la représentativité dépend en premier lieu de la définition de la population que l'échantillon doit représenter. Puis, elle dépend de la gamme des distributions linguistiques qui se trouve dans le corpus : elle doit être équivalente à celle de la population analysée. Ces distributions linguistiques doivent équivaloir à l'ensemble des traits linguistiques, leurs variations et distributions dans un même texte, entre textes et entre différents types de textes. Enfin, un autre élément sur lequel la représentativité dépend est l'échantillonnage des textes de la population langagière. Pour Biber, il est donc nécessaire de faire des répartitions, par exemple, sélectionner un certain nombre de documents pour chaque type de textes de la population ciblée. Habert a une approche différente : il privilégie la production et la réception des documents par rapport aux caractéristiques langagières de la population cible. Selon Habert, la représentativité d'un corpus présente une dimension interne, c'est-à-dire la détermination des différents emplois du langage étudié et une dimension externe qui correspond aux conditions de production et réception des documents. En plus, les documents utilisés pour construire un corpus doivent être catégorisés avant d'être choisis (cf. Goeuriot 2009, 3–5).

Les corpus peuvent être de différentes typologies. Les corpus larges et représentatifs d'un langage sont appelés corpus de références et un exemple est le *British National Corpus* (BNC), qui est une collection de 100 millions de mots de la langue écrite et de la langue parlée en représentant un large échantillon de l'anglais britannique de la fin du 20e siècle ; sa dernière édition a été publiée en 2007 (cf. Burnard 2009)[5]. Puis, il y a le corpus d'études qui sert à observer un aspect particulier d'un langage, un exemple est le corpus créé par Thomas Beauvisage en 2001 qui a le but d'étudier les sous-genres du roman policier. Nous avons déjà vu qu'il existe le corpus écrit composé par de textes écrits et le corpus oral composé de transcription de matériel oral. D'autres types de corpus sont les corpus synchrones qui comprennent des documents relatifs à une période restreinte et les corpus diachroniques qui sont utilisés pour observer l'évolution du langage pendant des périodes différentes. En plus, il existe les corpus ouverts qui sont constamment mis à jour et les corpus fermés. Pour conclure, un corpus peut être monologue ou multilingue, une différence que nous analyserons plus avant dans ce mémoire (cf. Goeuriot 2009, 3).

3.1.2 L'origine du terme « corpus »

Le terme « corpus » est un emprunt latin du terme *corpus iuris* qui signifie « collection de droit romain ». Le terme dérive de *Corpus iuris Iustinianeum* ou *codex Justinianus*, qui est un recueil de textes normatifs et jurisprudentiels voulu par Justinien le Grand, empereur byzantin de 527 à 565 qui voulait réorganiser le système juridique de l'Empire byzantin. La première version du *Corpus iuris* a été écrite en 528 et la seconde version en 533 (Wikipédia, 2024). Pour ce qui concerne la langue française, le terme corpus est utilisé par les expressions *Corpus Domini* et *Corpus Christi*, deux termes qui désignent l'hostie[6] dans les rites liturgiques chrétiens jusqu'au 18e siècle. Puis, dans le 19e siècle, l'usage du terme change : en 1809, le terme indique des données en lettres relatives à une étude scientifique qui est le résultat d'un transfert linguistique depuis le terme allemand *Korpus* ou *Corpus*, attesté en 1787. En 1855, le terme entre à faire partie du domaine de la linguistique, en indiquant un « ensemble de textes établis selon un critère thématique en vue de leur étude linguistique » (cf. Magnani 2017, 3). Cependant,

[5] Pour approfondir ou utiliser le corpus, voici le lien vers le corpus : http://www.natcorp.ox.ac.uk/.
[6] « Petite rondelle mince de pain azyme que le prêtre consacre pendant la messe. » (Centre National de Ressources Textuelles et Lexicales, s. d.).

le terme entre dans les dictionnaires de la langue française seulement en 1968 (cf. Magnani 2017, 3).

3.1.3 Le texte spécialisé à la base du corpus

Nous pouvons dire que l'unité de mesure d'un corpus est le texte, en particulier il s'agit d'un texte spécialisé ou texte de spécialité. Le texte spécialisé est un texte riche en terminologie et il sera la source où le/la terminographe aura la possibilité de détecter les termes utiles pour construire un dictionnaire terminologique auquel il/elle travaille.

Les textes spécialisés sont fondamentaux pour le travail terminographique et pour l'analyse des termes. Tout d'abord, les textes spécialisés sont la preuve de l'existence des termes : ils fournissent les attestations que les termes sont réellement utilisés par les experts. Puis, pendant la phase du repérage des termes, les textes spécialisés offrent des informations sur leur fréquence d'emploi (cf. L'Homme 2020, 133–134). Cela représente un indice essentiel pour la sélection des termes par le/la terminographe : les unités lexicales qui figurent un certain nombre de fois dans les textes spécialisés pourront être des termes. L'indice guide de la fréquence doit tenir en considération deux réalités : le nombre d'occurrences d'un terme et sa répartition (cf. L'Homme 2020, 64). L'occurrence est le nombre total d'apparitions d'un terme dans un texte, tandis que la répartition représente l'ensemble des emplacements où un terme apparaît dans tous les documents d'un corpus, nous pouvons donc savoir si le terme apparaît de manière uniforme ou s'il est localisé seulement dans quelques documents (cf. Labbé/Labbé 2017, 3). Pour mieux comprendre, nous pouvons présenter un exemple très simple : prenons-nous le domaine juridique, dans un corpus composé par des documents juridiques, nous trouvons trois termes, qui sont *idéologie*, *imposition* et *loi*. Pour ce qui concerne la fréquence d'apparition, le terme *idéologie* apparaît 95 fois, le terme *imposition* apparaît 85 fois et le terme *loi* apparaît 1643 fois ; alors que, pour ce qui concerne la répartition, le terme *idéologie* apparaît seulement dans quatre textes du corpus, le terme *imposition* apparaît dans trois textes et le terme *loi* est utilisé dans 33 textes. Si nous analysons seulement la fréquence des termes, nous pourrons dire que les trois mots sont des termes. Cependant, si nous ajoutons à notre étude la répartition, nous dirons que le terme *loi* renvoie à un sens plus central dans le corpus juridique que les autres (cf. L'Homme 2020, 64–65).

Les textes spécialisés fournissent également des éléments définitoires, c'est-à-dire des renseignements relatifs au sens des termes et qui pourront être utiles pour construire

une définition. Parfois, ce sont les auteurs des textes eux-mêmes qui ressentent le besoin de décrire les concepts qui sont centraux pour la compréhension du contenu du texte. Dans les textes spécialisés, nous pouvons aussi trouver les variantes terminologiques : le sens d'un terme peut être exprimé par diverses formes, par exemple, le terme *biocarburant*, qui désigne un type de carburant produit à partir de matériaux organiques non fossiles, peut être substitué par le syntagme « carburant vert », sa variante terminologique. Puis, les textes spécialisés offrent les indices de relations taxinomiques, c'est-à-dire les liens entre les hyponymes et les hyperonymes : l'hyperonyme est un terme général qui englobe plusieurs termes spécifiques, qui seront appelés hyponymes. Il s'agit d'un rapport de hiérarchie entre les termes. Par exemple, le mot *végétal* est un hyperonyme et le mot *tomate* est son hyponyme. Les textes spécialisés permettent aux terminographes la possibilité d'analyser les indices de relations conceptuelles, c'est-à-dire qu'il existe des renseignements utilisés pour la construction des définitions. Ces relations conceptuelles peuvent indiquer la fonction d'un objet ou sa cause et son effet, par exemple, la phrase « Le système d'exploitation est chargé de la gestion des périphériques » explique l'une des fonctions d'un système d'exploitation. Pour conclure, les textes spécialisés offrent d'autres types de renseignements aux terminographes : ils contiennent informations sur les synonymes, les co-hyponymes, les antonymes, les relations méronymiques dont les holonymes et les méronymes, c'est-à-dire la relation entre un mot désignant une « partie » et un mot qui désigne un « tout »[7], et les cooccurrents, c'est-à-dire les affinités sémantiques d'un terme (cf. L'Homme 2020, 134–136).

3.2 Le corpus spécialisé et les autres types de corpus

Parmi les nombreux types de corpus, il faut distinguer le corpus général et le corpus spécialisé. Le corpus général est un corpus composé d'énoncés qui ne relèvent pas nécessairement d'un domaine spécialisé de l'activité humaine. Au contraire, le corpus spécialisé est composé d'énoncés relatifs à un domaine spécialisé de l'activité humaine et il doit être représentatif de la langue de spécialité employée par un certain domaine. Pour construire un corpus spécialisé, il faut faire une sélection rigoureuse des textes spécialisés : ces documents doivent être d'une certaine qualité et ils doivent représenter

[7] Par exemple, le mot *feuille* est un méronyme du mot *arbre* qui est l'holonyme.

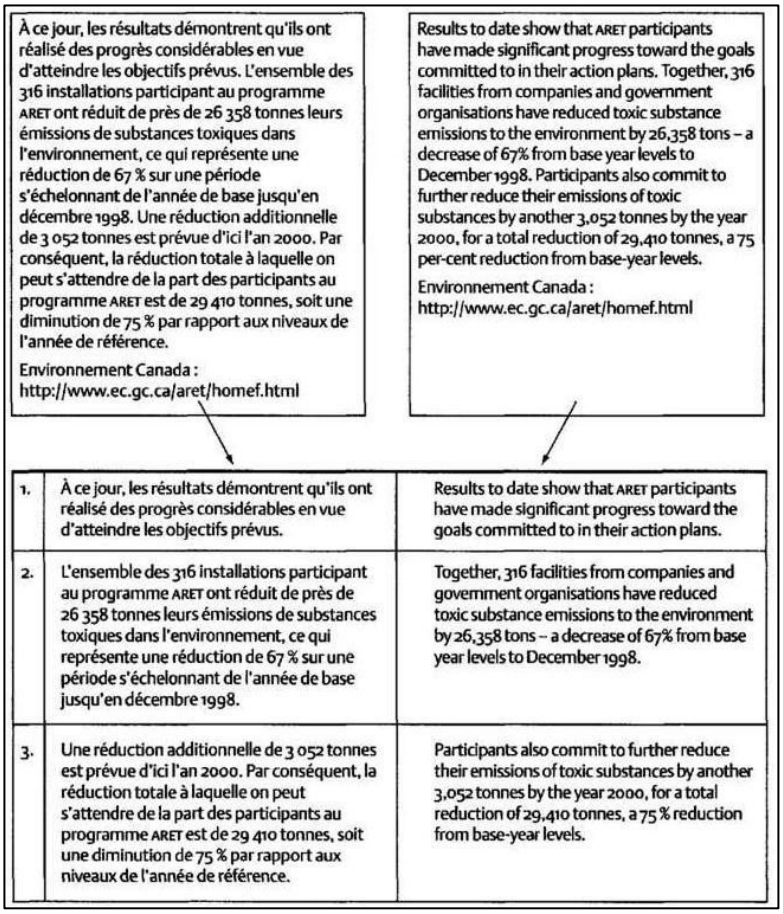
les variétés d'un domaine (cf. Goeuriot 2009, 7). Tout d'abord, il est important d'identifier le domaine de spécialité et les textes doivent donc représenter le domaine exprimé dans les objectifs du travail terminographique. Puis, il faut choisir les textes dans la langue (ou les langues) fixée par les objectifs de travail et si nécessaire il faut inclure les variantes régionales. Toujours lié à la langue, les textes spécialisés qui sont sélectionnés pour construire un corpus spécialisé ne doivent pas être des traductions, car la langue de rédaction doit être celle indiquée dans les objectifs du travail terminographique. Par exemple, si nous travaillons sur un corpus formé de textes en anglais, ils doivent être rédigés en anglais et pas traduits. Un autre élément qui influence le choix des textes spécialisés est le niveau de spécialisation, c'est-à-dire la relation qu'il y a entre l'auteur du texte et les destinataires. Les quatre possibilités sont : expert à expert ; expert à un expert d'un domaine connexe ; didactique, c'est-à-dire des textes qui sont adressés à des spécialistes en devenir ; et expert à non-experts ou vulgarisation. Il faut aussi faire attention au type de documents : il est essentiel de collecter une variété de types de documents, par exemple, articles scientifiques, monographies, manuels, périodiques, sites Web et rapports, lorsque cela est pertinent. Le/la terminographe tiendra également compte de la date de publication pour choisir et inclure des textes récents, car la nouveauté des textes est indispensable, mais il faut toujours suivre les buts du projet terminographique. Puis, il y a le support : la grande majorité des documents d'un corpus sont écrits, car le travail terminologique s'appuie sur l'écrit. Cependant, l'oral est un autre processus de transmission de connaissances : nous parlons des cours, des conférences et des conversations parmi des spécialistes. Les données orales présentent un problème d'accès et il est donc recommandé de les convertir en format électronique. Pour conclure, nous parlons de la taille du corpus qui sera estimée en nombre de mots. La taille idéale est généralement considérée comme comprise entre 200.000 et 500.000 mots, mais il n'existe pas un établi de consensus sur la taille idéale, car les auteurs ont des idées différentes. En tout état de cause, la taille dépend encore des objectifs du projet terminographique. Si nous travaillons sur un corpus électronique, les mots sont appelés mots graphiques et ils ne correspondent pas toujours à une unité lexicale ou à un terme. Les logiciels qui s'occupent d'analyser les corpus informatisés calculent le nombre des mots graphiques en les décomptant en chaînes de caractères délimitées par des espaces ou des symboles (cf. L'Homme 2020, 140–143).

3.2.1 Les corpus multilingues

Nous avons vu qu'il existe le corpus monologue qui est composé par des textes écrits dans une seule langue et le corpus bilingue ou multilingue. Depuis les années 1980, les travaux sur les corpus se sont ouverts également sur les langues européennes et asiatiques : les corpus bilingues ou multilingues sont apparus pour répondre aux besoins de traduction. Ces sont des corpus composés de textes en deux (ou plusieurs) langues et ils représentent des ressources très utiles dans de nombreux domaines, par exemple, la traduction automatique, l'extraction d'informations multilingues et les études comparatives (cf. Goeuriot 2009, 11). Ce type de corpus est principalement utilisé par les terminographes pour rechercher des correspondances interlinguistiques dans plus d'une langue (cf. L'Homme 2020, 146).

Les corpus multilingues sont à leur tour divisés en deux types de corpus : le corpus parallèle et le corpus comparable. Le corpus parallèle est un corpus dont les énoncés sont accompagnés de leurs traductions dans une ou plusieurs langues : ils sont le résultat d'un processus de traduction et les terminographes font la comparaison entre les différentes langues d'étude. Ce type de corpus est très utile pour la création de corpus alignés qui permettent de consulter les segments de traduction de manière plus rapide et facile. Les corpus alignés sont construits grâce à des aligneurs, c'est-à-dire des programmes qui tiennent compte des ponctuations fortes, des limites des paragraphes et de la numération des sections d'un texte pour aligner deux textes. Cependant, ces stratégies adoptées par les aligneurs ne fonctionnent pas toujours, car il y a des phrases sources qui peuvent être traduites par plus d'une phrase cible ou une phrase cible qui traduit plus d'une phrase source. Pour faire face à ces problèmes, certains aligneurs utilisent aussi les noms propres, les chaînes numériques ou l'évaluation quantitative de la longueur des segments. L'alignement automatique n'est pas parfait, mais s'il utilise de bonnes stratégies et des raffinements, il peut fournir des textes alignés presque totalement corrects. Le nombre de langues qui sont étudiées influence l'alignement : dans la majorité des cas, les corpus

alignés sont construits sur deux seules langues, en offrant un « bitexte » qui est plus facile à aligner (cf. L'Homme 2020, 147–148).



À ce jour, les résultats démontrent qu'ils ont réalisé des progrès considérables en vue d'atteindre les objectifs prévus. L'ensemble des 316 installations participant au programme ARET ont réduit de près de 26 358 tonnes leurs émissions de substances toxiques dans l'environnement, ce qui représente une réduction de 67 % sur une période s'échelonnant de l'année de base jusqu'en décembre 1998. Une réduction additionnelle de 3 052 tonnes est prévue d'ici l'an 2000. Par conséquent, la réduction totale à laquelle on peut s'attendre de la part des participants au programme ARET est de 29 410 tonnes, soit une diminution de 75 % par rapport aux niveaux de l'année de référence.

Environnement Canada :
http://www.ec.gc.ca/aret/homef.html

Results to date show that ARET participants have made significant progress toward the goals committed to in their action plans. Together, 316 facilities from companies and government organisations have reduced toxic substance emissions to the environment by 26,358 tons – a decrease of 67% from base year levels to December 1998. Participants also commit to further reduce their emissions of toxic substances by another 3,052 tonnes by the year 2000, for a total reduction of 29,410 tonnes, a 75 per-cent reduction from base-year levels.

Environnement Canada :
http://www.ec.gc.ca/aret/homef.html

1.	À ce jour, les résultats démontrent qu'ils ont réalisé des progrès considérables en vue d'atteindre les objectifs prévus.	Results to date show that ARET participants have made significant progress toward the goals committed to in their action plans.
2.	L'ensemble des 316 installations participant au programme ARET ont réduit de près de 26 358 tonnes leurs émissions de substances toxiques dans l'environnement, ce qui représente une réduction de 67 % sur une période s'échelonnant de l'année de base jusqu'en décembre 1998.	Together, 316 facilities from companies and government organisations have reduced toxic substance emissions to the environment by 26,358 tons – a decrease of 67% from base year levels to December 1998.
3.	Une réduction additionnelle de 3 052 tonnes est prévue d'ici l'an 2000. Par conséquent, la réduction totale à laquelle on peut s'attendre de la part des participants au programme ARET est de 29 410 tonnes, soit une diminution de 75 % par rapport aux niveaux de l'année de référence.	Participants also commit to further reduce their emissions of toxic substances by another 3,052 tonnes by the year 2000, for a total reduction of 29,410 tonnes, a 75 % reduction from base-year levels.

Figure 1 – Exemple d'un corpus aligné (cf. L'Homme 2020, 149)

Il existe de corpus parallèles qui constituent des points de référence pour les terminographes : le corpus *Hansard* est composé d'une série de textes anglais et français basés sur les transcriptions des débats du parlement canadien de 1970 à 1988 et il contient des dizaines de millions de mots ; le corpus *Europarl* groupe des textes du Parlement européen écrits dans 11 langues et il compte plus de 20 millions de mots par chacune langue ; le corpus *Hong-Kong Hansard* a été créé par le *Linguistic Data Consortium*[8] (LDC) et il rassemble les textes en anglais et français issus des discussions, rapports, etc. du parlement de Hong Kong ; le corpus de l'Union des banques suisses (UBS) qui est

[8] Le *Linguistic Data Consortium* est un consortium ouvert d'universités, de bibliothèques, d'entreprises et de laboratoires de recherche gouvernementaux. Il a été créé en 1992 pour remédier à la grave pénurie de données à laquelle étaient alors confrontés la recherche et le développement dans le domaine des technologies linguistiques (Linguistic Data Consortium, s. d.).

utilisé par des organismes qui s'occupent de publier dans plusieurs langues les rapports concernant le développement de l'économie suisse (cf. Goeuriot 2009, 12).

L'autre type de corpus bilingue ou multilingue est le corpus comparable : c'est un corpus composé des deux ensembles de textes. Ces documents peuvent être rédigés dans plusieurs (deux ou plusieurs) langues, ils traitent un seul domaine de spécialité et par conséquent il ne s'agit pas des textes qui sont la traduction des autres. Les énoncés d'un corpus comparable partagent des caractéristiques communes qui déterminent leur comparabilité, par exemple, le niveau de langue ou la variété régionale (cf. L'Homme 2020, 150). Le corpus comparable est distingué entre corpus comparable généraliste et corpus comparable spécialisé. Le premier est composé généralement d'articles de journaux qui s'occupent d'une même période ou d'une même thématique. Le second est composé de documents concernant un domaine scientifique, faisant appel à un langage spécialisé (cf. Goeuriot 2009, 13).

3.2.2 Le corpus annoté : annotation et étiquetage

Un autre type de corpus est le corpus annoté. Cela est un corpus manipulé pour insérer des informations sur la syntaxe, la morphologie ou la sémantique des textes.

Dans la pratique, les terminographes utilisent souvent des corpus bruts, c'est-à-dire des corpus avec des chaînes de caractères non interprétées. Cependant, les textes d'un corpus peuvent être enrichis grâce à des renseignements de nature linguistique, notamment l'annotation de corpus qui sert pour ajouter des informations linguistiques interprétatives. Nous parlerons donc de corpus annoté ou étiqueté (cf. L'Homme 2020, 150–151). Des types courants d'annotation sont la désambiguïsation d'une catégorie de mots et l'étiquetage morphosyntaxique[9], c'est-à-dire l'ajout d'étiquettes indiquant la classe à laquelle appartiennent les parties du discours (noms, verbes, adjectifs, etc.). Ce processus peut être utile pour distinguer les mots qui ont la même orthographe, mais des significations différentes ou une autre prononciation. Certaines terminographes préfèrent travailler sur des corpus non annotés, car ils sont « purs ». Cependant, pour d'autres terminographes, l'annotation sert pour rendre un corpus beaucoup plus utile (cf. Leech, 2004).

L'étiquetage des parties du discours n'est pas la seule méthode d'annotation, mais il en existe d'autres types, selon les différents niveaux d'analyse linguistique des textes.

[9] En anglais, cela s'appelle *part-of-speech tagging* ou *POS tagging*.

L'annotation phonétique a l'objectif d'ajouter des informations sur la prononciation des mots et l'annotation prosodique[10] fournit des informations sur les caractéristiques prosodiques, par exemple, l'accent, l'intonation et les pauses. Un autre type est l'annotation sémantique qui sert pour ajouter des informations relatives à la catégorie sémantique des mots, par exemple, un terme désignant un sport ou un terme désignant un insecte. Puis il y a l'annotation pragmatique qui ajoute d'informations sur les types d'actes de langage, en analysant si un énoncé est un accusé de réception ou une demande de retour d'information ou encore une acceptation. Pour conclure, nous pouvons trouver également l'annotation stylistique qui s'occupe d'informations sur la présentation du discours et l'annotation lexicale complète l'identité d'un lemme des mots dans un texte (cf. Leech, 2004).

La désambiguïsation d'une catégorie de mots et l'étiquetage sont fondamentaux pour réduire l'ambiguïté présente dans les textes et elle rend difficile le traitement automatique (ou semi-automatique) des textes. Un corpus dont les chaînes de caractères ne sont pas étiquetées n'indique pas le bon sens des mots et il ne spécifie pas le rôle joué par les parties du discours, par exemple, le mot « informatique » peut être employé comme nom féminin qui désigne la discipline ou comme adjectif et le mot « programme » peut être employé encore comme nom ou comme verbe. L'étiquetage s'occupe d'éliminer ces ambiguïtés grâce aux étiquettes attachées aux mots d'un texte. Il existe plusieurs méthodes pour expliciter les propriétés des mots, un exemple est l'utilisation d'une barre oblique qui suit le mot et qui indique la partie du discours, comme le montre l'exemple suivant : « Cette$_{/dét}$ technique$_{/nc}$[11] consiste$_{/verbe : conjug}$ à$_{/prép}$ […] ». Certaines étiquettes sont plus précises et elles spécifient, par exemple, les types des déterminants, s'il s'agit d'un article ou d'un adjectif possessif, et les formes verbales, s'il s'agit d'un verbe conjugué ou d'un verbe à l'infinitif (cf. L'Homme 2020, 151–152).

Aujourd'hui, l'étiquetage des textes est automatique grâce aux étiqueteurs. Pour fournir les bonnes étiquettes, les étiqueteurs peuvent commencer par la consultation d'un corpus qui contient les mêmes mots avec les étiquettes. Mais lorsque nous sommes confrontés à des mots qui présentent des ambiguïtés, l'étiqueteur doit effectuer la désambiguïsation afin de retenir une seule étiquette. Pour le faire, il faut examiner le

[10] La prosodie est la « prononciation correcte et régulière des mots selon l'accent et la quantité des syllabes. » (Centre National de Ressources Textuelles et Lexicales, s. d.).
[11] L'expression « /nc » signifie nom commun.

contexte et les étiquettes des mots voisins. Cependant, l'entraînement des étiqueteurs est fondamental et il se base sur des répertoires de mots et de règles qu'ils utilisent. En plus, ils recourent au prétraitement des textes, c'est-à-dire séparer et isoler les mots graphiques. Un autre instrument pour rendre l'étiquetage plus précis est la lemmatisation, c'est-à-dire « […] ramener les formes fléchies des mots variables à une forme canonique » (cf. L'Homme 2020, 154). Dans ce cas-là, les noms sont réduits au singulier, les adjectifs sont réduits au masculin singulier et les verbes à l'infinitif. Le produit de la lemmatisation est donc le lemme et les terminographes peuvent en bénéficier en tant qu'ils/elles ont la possibilité d'analyser toutes les formes des mots (cf. L'Homme 2020, 153–154).

Le logiciel TreeTagger est un outil qui sert pour l'étiquetage et la lemmatisation et il a été développé en 1994 par Helmut Schmid dans le cadre de l'Institut de linguistique informatique chez l'université de Stuttgart[12]. Voici une série d'étiquettes les plus courantes et leur description (Centrum für Informations- und Sprachverarbeitung, s. d.) :

Étiquette	Description
ADJ	adjectif
ADV	adverbe
DET:def	déterminant défini
DET:dem	déterminant démonstratif (ce, cette, ces)
DET:ind	déterminant indéfini (quelque, un, des)
DET:int	déterminant interrogatif (quel)
DET:par	déterminant partitif (du)
DET:pos	déterminant possessif (ma, ta, etc.)
DET:pre	prédéterminant (tout, toute, toutes)
ETR	mots étrangers
KON	conjonction
NAM	nom propre
NOM	nom commun
NOM|NAM:sig	sigle
NUM	numéral
PRO:clo	clitique objet
PRO:cls	clitique sujet
PRO:dem	pronom démonstratif
PRO:ind	pronom indéfini
PRO:int	pronom interrogatif (comment, où, quand)
PRO:pos	pronom possessif (mien, tien, etc.)
PRO:rel	pronom relatif
PRO:ton	pronom tonique
PRP	préposition

[12] Pour approfondir, voici le lien vers le site web du logiciel : https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

PRP:det	préposition+déterminant (au, du, aux, des)
PRT:int	particule interrogative (est-ce que)
SYM	symbole
VER:cond	verbe au conditionnel
VER:futu	verbe au futur
VER:impe	verbe à l'impératif
VER:impf	verbe à l'imparfait
VER:infi	verbe à l'infinitif
VER:pper	verbe au participe passé
VER:ppre	verbe au participe présent
VER:pres	verbe au présent
VER:simp	verbe au passé simple
VER:subi	verbe au subjonctif imparfait
VER:subp	verbe au subjonctif présent
PUN	ponctuation
PUN:cit	guillemets
SENT	fin de phrase

Tableau 1 – Les étiquettes et leur description

En ligne, nous trouvons plusieurs sites web où nous pouvons insérer un texte et puis l'avoir avec les étiquettes par TreeTagger. Voici un exemple pratique et utile d'une phrase tirée de cette mémoire et soumise au logiciel en ligne : « Dans le deuxième chapitre, nous avons parlé brièvement du corpus en mentionnant la linguistique de corpus et la terminologie textuelle et le travail terminographique, mais dans ce chapitre, nous allons l'analyser en détail ».

Mot	Étiquetage	Lemme
<s>		
Dans	PRP	dans
le	DET:ART	le
deuxième	NUM	deuxième
chapitre	NOM	chapitre
,	PUN	,
nous	PRO:PER	nous
avons	VER:pres	avoir
parlé	VER:pper	parler
brièvement	ADV	brièvement
du	PRP:det	du
corpus	NOM	corpus
en	PRP	en
mentionnant	VER:ppre	mentionner
la	DET:ART	le
linguistique	NOM	linguistique
de	PRP	de

corpus	NOM	corpus
et	KON	et
la	DET:ART	le
terminologie	NOM	terminologie
textuelle	ADJ	textuel
et	KON	et
le	DET:ART	le
travail	NOM	travail
terminographique	ADJ	terminographique
,	PUN	,
mais	KON	mais
dans	PRP	dans
ce	PRO:DEM	ce
chapitre	NOM	chapitre
,	PUN	,
nous	PRO:PER	nous
allons	VER:pres	aller
l’	NOM	l’
analyser	VER:infi	analyser
en	PRP	en
détail	NOM	détail
.	SENT	.
</s>		

Tableau 2 – Exemple pratique d’une phrase étiquetée par TreeTagger

Comme nous pouvons le voir, TreeTagger a organisé la phrase en la présentant dans trois colonnes qui indiquent respectivement le mot graphique, la partie du discours et le lemme[13]. TreeTagger commence son analyse par la balise[14] <s>, qui est une balise dite ouvrante caractérisée par les deux signes de comparaison < > qui indique « inférieur » et « supérieur ». La même balise se trouve à la fin de la phrase étiquetée, mais elle présente aussi une barre oblique « / » : </s> ; cela s’appelle balise de fermeture (Ronan HELLO, 2020). Puis nous voyons toutes les étiquettes assignées aux mots, par exemple, l’étiquette du mot « nous » est « PRO:PER », c’est-à-dire un pronom personnel et son lemme est « nous » ou l’étiquette du verbe conjugué « avons » est « VER:pres », qui signifie qu’il s’agit d’un verbe conjugué au présent et le lemme est le verbe à l’infinitif « avoir ». Tous

[13] Pour essayer, voici le lien vers le logiciel : http://corpora.lancs.ac.uk/tree-tagger/.

[14] « La balise, *tag* en anglais, désigne dans le monde de l’informatique et de la programmation une série de caractères destinée à déclencher, de façon automatique, l’exécution d’une action par un programme informatique. À la lecture d’une balise, un programme informatique exécute ainsi instantanément une commande spécifique » (Journaldunet.com, 2019).

ces outils permettent aux terminographes d'étudier les termes sous plusieurs points de vue, en rendant le travail terminographique plus complet.

3.3 La linguistique de corpus

Comme indiqué plusieurs fois précédemment, il existe une relation étroite entre la linguistique et l'informatique. Aujourd'hui nous devons tous traiter une grande quantité de données textuelles sur support électronique et les corpus sont devenus la principale source du traitement automatique du langage. Ainsi, la linguistique de corpus fait son entrée (cf. Condamines 2005, 37). Elle est une branche relativement jeune de la linguistique et elle représente une méthodologie qui est caractérisée par des analyses quantitatives et qualitatives sur l'utilisation de la langue et, pour le faire, elle examine des vastes corpus oraux et écrits disponibles sur support électronique (Johannes Gutenberg-Universität Mainz, 2021).

3.3.1 L'histoire de la linguistique de corpus

Un premier exemple de linguistique de corpus date à l'année 1755, lorsque Samuel Johnson a publié son dictionnaire, le premier qui était basé sur un corpus composé par de fiches de travail accompagnées de citations. Johnson débute ainsi une tradition lexicographique normative qui se base sur des textes authentiques et cette tradition lexicographique est le pilier du *Oxford English Dictionary* (cf. Williams 2006, 152).

Les aspects fondamentaux qui ont caractérisé le développement de la linguistique de corpus sont l'enseignement des langues, en particulier l'enseignement de la langue anglaise, et le contextualisme proposé par John Rupert Firth. Pour ce qui concerne l'enseignement des langues, nous considérons la période entre les deux guerres et les travaux de Harold Palmer au Japon. Il a travaillé sur la théorie et la pratique de l'enseignement de l'anglais comme langue seconde, sur les vocabulaires pour apprenants et sur la collocation en anglais. Pour ce qui concerne le contextualisme, il sera nécessaire pour apprendre une langue. Il est une « […] théorie selon laquelle le sens d'un mot est directement lié au contexte de ce mot dans la phrase » (cf. Encyclopædia Universalis, s. d.). Selon cette théorie donc, les phrases doivent être authentiques et elles doivent faire partie d'un contexte pour bien étudier le lexique et la grammaire des textes. Les travaux de Firth ont été développés par deux étudiants : Michael Halliday et John Sinclair. Halliday a proposé une grammaire systémique et descriptive qui était employée dans la linguistique de corpus contextualiste, en développant principalement l'aspect

grammatical ; tandis que Sinclair s'est occupé de la partie lexicale et donc de l'analyse de corpus contextualiste (cf. Williams 2006, 152–153).

Pendant les années 1960, la linguistique de corpus servait donc pour enseigner l'anglais comme langue étrangère, car les dictionnaires traditionnels ont la fonction d'analyser les seuls mots pris isolément et ils n'étaient pas capables de fournir des renseignements sur comment employer un mot. La linguistique de corpus change le jeu : elle s'ancre dans la linguistique appliquée[15], qui est l'« application des théories, des descriptions, des analyses linguistiques à la pédagogie des langues, à la traduction, aux techniques de communication » (Centre National de Ressources Textuelles et Lexicales, s. d.) et qui a comme objectif principal l'enseignement d'une langue et l'élaboration de dictionnaires. Pour cela, la linguistique de corpus a été développée comme une méthodologie pour découvrir tout ce qui concerne une langue en général et pour soutenir la linguistique appliquée (cf. Teubert 2009).

Toutefois, la recherche dédiée aux corpus a été négligée pendant quelques décennies et elle se développe extrêmement à partir des années 1980, lorsque les corpus et leur utilisation devenaient plus évidents et importants. En plus, les ordinateurs ont commencé à devenir de plus en plus puissants, en stimulant encore le progrès de la linguistique de corpus, qui a pris un essor sans pareil dans les années 2000. Dans l'état actuel des choses, la partie méthodologique de la linguistique de corpus est devenue mature, les langues qui sont étudiées par les linguistes de corpus augmentent de nombre chaque année et la popularité de linguistique de corpus comme discipline est augmentée parmi les chercheurs (cf. Kida 2013, 134).

3.3.2 Linguistique de corpus et informatique

Depuis la révolution numérique, la linguistique de corpus et son progrès sont liés aux outils informatiques et aux ressources électroniques (cf. Williams 2006, 155). L'analyse des corpus avant l'utilisation des ordinateurs était difficile, car il était très facile de commettre des erreurs et l'analyse ne résultait pas complète et reproductible. Grâce aux ordinateurs, les terminographes et les autres chercheurs en linguistique avaient l'opportunité de stocker tous les documents dans des machines (cf. Kennedy 1998, 5). Le premier corpus électronique non diachronique en anglais a été compilé au début des années 1960 aux États-Unis par Nelson Francis et Henry Kučera, deux pionniers de la

[15] En anglais, *applied linguistics*.

linguistique de corpus. Il a été créé pour un projet de recherche sur les langues et il est appelé *Brown Corpus* (cf. Kida 2013, 134).

Aujourd'hui, la linguistique de corpus est donc plus facile, rapide et précise et les corpus informatisés sont toujours plus accessibles et ils réduisent les difficultés liées à la gestion des grandes masses de données. Les tâches terminographiques qui prévoient la recherche, le comptage et le classement des termes sont maintenant plus fiables qu'avant et les probabilités d'apparitions de termes sont également presque sans fautes, en partageant des informations sur les termes plus exacts. Cela est possible aussi grâce aux mathématiques : le traitement automatique du langage et la linguistique ont travaillé ensemble pour nous donner un degré de précision des mesures vraiment considérable (cf. Kennedy 1998, 5).

3.3.3 Emploi actuel de la linguistique de corpus

La linguistique de corpus trouve application dans innumérables disciplines. Tout d'abord, elle est utilisée par la linguistique théorique, la lexicologie et la lexicographie. Puis, il y a d'autres domaines d'études qui en bénéficient, par exemple, la linguistique informatique qui comprend le traitement du langage, la reconnaissance du langage, et la synthèse vocale et encore les sciences de l'information, les systèmes experts, la traduction automatique et le traitement de texte. Les corpus sont devenus importants également pour des domaines comme la psycholinguistique qui se concrétise dans la neuropsychologie, la philosophie du langage et l'analyse du discours. Pour conclure, la linguistique de corpus est essentielle pour l'enseignement des langues (cf. Kida 2013, 140–141).

Les corpus présentent de nombreux avantages, car les linguistes peuvent analyser et étudier plus de matériel dans un délai plus court et avec des résultats plus exacts. Ils sont caractérisés par les traits de la rapidité et de la fiabilité, mais ils présentent d'autres avantages : les corpus informatisés donnent la possibilité aux chercheurs d'accéder aux données ; les données de corpus sont fondamentales pour les locuteurs non natifs qui commencent à étudier une langue ; les corpus permettent aussi d'examiner toutes les variations linguistiques d'une langue grâce aux renseignements qui fournissent, par exemple, la fréquence d'occurrence des éléments linguistiques (cf. Kida 2013, 141).

3.4 Construire un corpus

Nous entrons dans la partie pratique de ce mémoire qui a le but d'analyser la méthodologie et les outils utilisés pour construire les corpus employés afin de rédiger les fiches terminologiques concernant le lexique de l'intelligence artificielle et du traitement automatique du langage.

Quand nous devons commencer à construire un corpus, nous devons tenir compte du type de corpus que nous voulons créer, s'il est un corpus monolingue, parallèle, comparable ou bilingue, etc. Nous devons être sûrs que le corpus reflète le projet terminographique et ses objectifs et nous devons vérifier qu'il puisse être réutilisé ou interchangé par d'autres terminographes ou chercheurs en général. Il faut également décider si le corpus restera « pur » ou si c'est nécessaire l'annotation pour ajouter les informations linguistiques interprétatives des textes et la gestion du travail de construction devra être conçue avec une attention aux moindres détails afin de construire un corpus de qualité. En fait, pour obtenir des résultats qui soient satisfaisants, tout dépend de la bonne qualité du corpus. La qualité d'un corpus peut être influencée par divers facteurs tels que le domaine des textes, qui doivent être bien définis et délimités ; la représentativité des textes qui doit être conforme pour tirer les conclusions ; l'organisation, l'annotation et le contenu du corpus qui doivent favoriser son exploitation. Pour synthétiser les points principaux à suivre pour la construction d'un corpus nous pouvons dire qu'il faut bien définir les attentes, les critères et les procédures à utiliser, il est nécessaire de les décrire dans les résultats de la recherche et il est aussi possible de réévaluer et corriger certains critères en cours de route, s'il devient nécessaire afin de mieux travailler (cf. Marshman 2003, 2–3).

3.4.1 Choix des textes

Dans le cadre d'un projet terminographique, pour construire un corpus composé de textes écrits, nous devons tout d'abord savoir choisir les textes spécialisés ; ensuite, il faut déterminer quels types de texte nous devons chercher ; et puis nous devons fixer le nombre des textes qui feront partie du corpus (cf. Cabré 2008, 38).

Quand nous avons appris comment distinguer les textes de spécialité, nous devons les organiser en fonction de certains critères de classification. Ces critères peuvent concerner le thème dont ils parlent, la dimension disciplinaire, le niveau de spécialisation, les sources, le genre textuel, la classe de texte par rapport à la stratégie discursive, les

langues et la relation entre les textes des langues du corpus si nous choisissons des textes plurilingues. Un corpus spécialisé peut être composé par des textes qui traitent un seul sujet ou plusieurs : dans le premier cas, il s'agira d'un corpus monodisciplinaire et dans le second, il s'agira d'un corpus pluridisciplinaire. Puis, les textes spécialisés sont caractérisés par un certain niveau de spécialisation et le corpus peut contenir des textes avec un niveau de spécialité toujours égal ou des textes qui ont des niveaux différents entre eux. Pour ce qui concerne les sources, les textes composant le corpus peuvent provenir d'un seul type de source ou de plusieurs typologies de sources. En plus, nous devons également tenir en considération le mode de transmission, car les textes d'un corpus peuvent être des textes oraux, écrits ou audiovisuels[16]. Le genre textuel est un autre élément qui nous devons observer : les textes spécialisés peuvent représenter un seul genre, par exemple, des articles scientifiques, ou ils peuvent représenter plusieurs genres, par exemple, un corpus peut être formé par des articles scientifiques, des pages Web, des encyclopédies et des reportages. Toujours lié au genre, les textes peuvent être homogènes ou hétérogènes en ce qui concerne la stratégie discursive, par exemple, un corpus homogène ne comportera que des textes argumentatifs et un corpus hétérogène comprendra des textes argumentatifs, littéraires, descriptifs, explicatifs, etc. En dernier, les corpus peuvent être monolingues, bilingues ou multilingues : les textes qui sont écrits dans plus d'une langue peuvent être mélangés dans le cadre d'un seul thème ou ils peuvent inclure la traduction correspondante dans la deuxième ou la troisième langue, en l'appelant corpus parallèle (cf. Cabré 2008, 39–40).

Combien de textes servent pour construire un corpus afin de représenter l'usage d'une langue ou étudier un domaine d'étude ? Cette question n'a pas une réponse juste, car la taille d'un corpus spécialisé dépend de la finalité du corpus. Si nous devons analyser une langue dans sa totalité, alors nous devons construire un corpus qui en soit représentatif et qui comprenne les variations internes et externes. Ce type de corpus sera appelé corpus de référence. Cependant, si nous devons constituer un corpus pour travailler sur une problématique particulière, la taille du corpus change et elle devra être adéquate aux finalités indiquées : nous pouvons faire un exemple, si nous analyser l'usage d'un pronom

[16] Un exemple d'un corpus audiovisuel est le corpus MATRICE-INA, qui est composé de documents audiovisuels diffusés au cours des 80 dernières années. Il contient environ 30 000 documents audiovisuels : ces documents ne sont pas tous en français et certains ne contiennent pas de la parole, car le corpus comprend aussi de films muets (cf. Antoine/Guinaudeau/Roy 2015, 1–2).

en position enclytique, nous pourrons constituer un corpus de taille moins importante qu'un corpus utilisé pour extraire le lexique d'un domaine de spécialité (cf. Cabré 2008, 40–41).

3.4.2 La fiabilité des sources

Quand les variables que nous prendrons en considération pour la construction d'un corpus spécialisé seront établies, nous procéderons à une série de tâches très importantes : la sélection des sources ; la sélection des textes ; choisir s'il faut utiliser des textes complets ou des fragments ; décider l'architecture de base et l'infrastructure logicielle et matérielle, c'est-à-dire quel système de gestion de corpus textuels employer ; la sélection les bonnes conventions pour la représentation des textes ; et choisir les critères, le langage et le système de balisage structurel (cf. Cabré 2008, 41).

Pour ce qui concerne les sources, nous devons mentionner le Web. En fait, le Web est une des sources les plus utilisées pour la collecte de documents qui composeront les corpus. Dans le Web, nous pouvons trouver de nombreuses typologies de textes telles que les articles de presse, les ouvrages scientifiques, les encyclopédies en format digital, les articles scientifiques et bien d'autres encore (cf. Yapomo 2013, 60). Quand nous pensons au Web et la recherche d'informations, le premier site web auquel nous pensons est presque à coup sûr Wikipédia : c'est une encyclopédie en ligne libre, universelle et multilingue qui est formée par les contributions des milliers de volontaires qui travaillent pour l'enrichir. Cela signifie que chacun peut y collaborer pour offrir des contenus libres, objectifs et vérifiables ; en plus chacun peut également modifier et améliorer ces contenus, sans le besoin de s'enregistrer (Wikipédia, 2024). Sur Wikipédia nous pouvons trouver beaucoup d'informations sur une quantité infinie de sujets, en représentant une colossale base de connaissances libre et accessible à tous. Cependant il faut souligner que l'usage de Wikipédia pour la recherche d'informations n'est pas toujours le bon choix. Tous les articles de Wikipédia doivent provenir de sources fiables : il s'agit des sources qui doivent être publiées avec un processus de publication fiable et elles doivent être écrites par des auteurs considérés comme faisant autorité sur le sujet. En même temps, la fiabilité de ces documents doit garantir l'exactitude des informations, des données publiées et des analyses et il est fondamental que les articles indiquent les diverses sources utilisées. Cette clarification nous fait comprendre que Wikipédia n'est pas une source primaire : c'est plutôt un outil de diffusion des connaissances secondaire et tertiaire. Cela

signifie que les contenus divulgués par Wikipédia peuvent être dignes de confiance si les sources citées et les informations vérifiables utilisées pour les rédiger sont elles-mêmes fiables. La différence entre Wikipédia et une encyclopédie classique se base sur les informations fournies : une encyclopédie traditionnelle apporte des informations cristallisées dans le papier, souvent homologuées et qui n'acceptent pas la discussion ou la possibilité de correction, car elles présentent un savoir qui vient « d'en haut ». Au contraire, Wikipédia est une source accessible à tous, destinée aux lecteurs et construite par les lecteurs et elle ne peut garantir la fiabilité de ses textes à cause de son dynamisme. Les problèmes les plus difficiles à résoudre sont à la fois l'absence d'objectivité et le vandalisme (Wikipédia, 2024).

Pour ce qui concerne la sélection des sources et la sélection des textes pour la construction d'un corpus spécialisé, nous devons donc faire attention à leur fiabilité surtout quand il s'agit des sources et textes en lignes. Pour être certains d'utiliser des sites web fiables, nous devons respecter une série d'étapes : tout d'abord, lorsque nous consultons un site web, nous pouvons lire sa page « À propos » afin de connaître la nature du site ; puis, nous devons nous assurer qu'il ne s'agit pas d'un site satirique ou parodique ; nous devons vérifier les auteurs et les sources, qui doivent être vérifiées et mentionnées ; enfin, nous pouvons examiner comment le contenu du site est écrit, s'il est caractérisé par des informations factuelles ou des opinions et s'il est ouvert à des propos contradictoires, ou s'il en permet qu'une seule lecture des faits (Le Monde, 2022). Une bonne solution est représentée par le choix des sites web qui font autorité, par exemple, les revues scientifiques en ligne qui se trouvent dans des bibliothèques numériques comme JSTOR : c'est une base de données incontournable qui est composée par plus de 12 millions d'articles de revues, de livres, d'images et de sources primaires dans 75 disciplines. JSTOR donne l'accès à un considérable éventail de contenus scientifiques et collabore avec la communauté universitaire afin d'aider une connexion parmi les bibliothèques, les étudiants, les enseignants et les éditeurs (JSTOR, s.d.). Pour la construction du corpus qui est la base de ce mémoire, les documents ont été pris de bibliothèques numériques comme JSTOR, de sites web de sociétés qui s'occupent de la science des données et de sa gestion ou de la transformation digitale et de sites web d'organismes de formation.

3.4.3 L'analyse des systèmes de gestion de corpus

Une fois que les textes qui constitueront le corpus ont été collectés, nous devons choisir quel système de gestion de corpus et quels langages de balisage utiliser. Il existe de nombreux systèmes de gestion, par exemple, AntConc qui est un logiciel d'analyse textuelle. AntConc a été développé par Laurence Anthony, professeur chez l'université de Waseda au Japon, et il est un logiciel gratuit qui nous pouvons télécharger en ligne[17]. Ce logiciel accepte seulement les documents en format TXT, c'est-à-dire les textes bruts[18], tandis que, pour ce qui concerne le type d'encodage des caractères, il accepte les encodages ANSI ou Unicode. L'encodage ANSI (American National Standards Institute) est un jeu de 256 caractères dont chaque caractère est représenté par un octet[19] et l'encodage Unicode est le standard informatique qui rend possible les échanges de textes dans différentes langues, à un niveau mondial actuellement et il est composé de plus de 100.000 caractères. Cela nécessite des multiples octets pour représenter les caractères Unicode : il existe donc différentes solutions techniques pour les représenter, par exemple, le codage UTF-8 dont nous avons déjà parlé (cf. Weiss, s. d.).

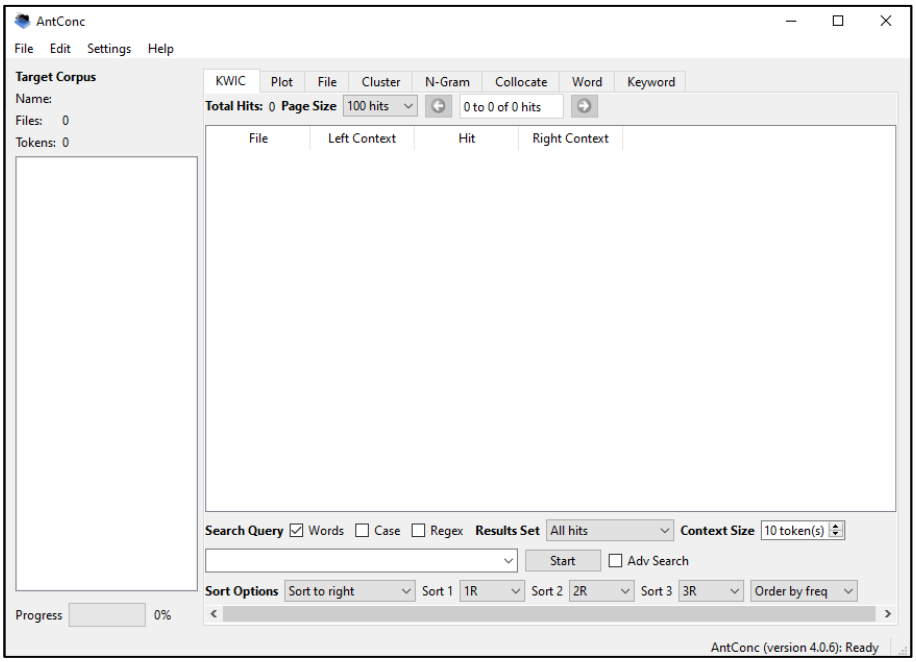


Figure 2 – La première fenêtre du logiciel AntConc

[17] Pour le télécharger, voici le lien vers son site web : https://www.laurenceanthony.net/software/antconc/.
[18] C'est le type de fichier associé au bloc-notes de Windows.
[19] Un octet est une « Unité logique en système codé binaire, constituée de huit caractères binaires ou bits, utilisée dans la plupart des langages machines, et permettant de représenter deux cent cinquante-six caractères alphabétiques » (Centre National de Ressources Textuelles et Lexicales, s. d.).

Les fonctionnalités du logiciel AntConc sont : l'outil « Concordance » qui montre les occurrences d'un mot pivot ; l'outil « Distribution » qui permet de voir les occurrences d'un ou plusieurs mots, représentés par un bandeau et par une barre verticale noire ; la fonction de « Contexte élargi » ; la fonction des « Agrégats » qui produit une liste des groupes de mots contigus contenant un mot pivot ; l'outil dédié aux cooccurrences d'un mot pivot ; et les fonctions qui permettent de rechercher une liste de mots ou expressions et d'effectuer des recherches à base d'expressions régulières (cf. Weiss, s. d.).

Un autre système de gestion de corpus est Sketch Engine[20] : c'est un gestionnaire de corpus et un logiciel d'analyse textuelle qui a été développé par Lexical Computing Limited[21] en 2003. Sketch Engine a pour objectif principal l'accès aux étudiants des langues, comme les lexicographes, les chercheurs en linguistique de corpus ou les traducteurs et à tous ceux qui souhaitent trouver des exemples authentiques de textes selon des requêtes complexes. Une première différence entre AntConc et Sketch Engine est la gratuité, car Sketch Engine est un système de gestion de corpus payant. Une autre différence est les formats des documents qui peuvent être CSV[22], XLSX[23], XML et PDF. Sketch Engine compte 600 corpus rédigés dans plus de 90 langues dont chaque corpus peut contenir jusqu'à 60 milliards de mots et il inclut des corpus en langue générale, mais aussi des corpus en langues de spécialité. Ce logiciel d'analyse textuelle nous permet de bénéficier des multiples fonctions pour analyser un corpus : après le téléchargement du corpus, nous pouvons chercher les collocations et les combinaisons de mots, les synonymes et mots similaires (thésaurus), les exemples d'usage en contexte (concordance), les listes de fréquence et les locutions et les blocs lexicalisés (n-grammes). En plus, nous pouvons extraire la terminologie d'un domaine spécifique ; analyser les tendances d'une langue, comme son analyse diachronique ou les néologismes présents dans son lexique ; analyser les statistiques de l'ensemble du corpus ; créer

[20] Pour approfondir, voici le lien vers le site web : https://www.sketchengine.eu/#blue.
[21] C'est une société de recherche fondée par Adam Kilgarriff en 2003 qui travaille entre la linguistique de corpus et la linguistique informatique et est engagée dans une approche empirique de l'étude du langage (Lexical Computing, s. d.).
[22] La sigle CSV signifie *Comma-separated values*, un format texte ouvert qui représentent des données tabulaires sous forme de valeurs séparées par des virgules (Wikipédia, 2023). En français « Valeurs séparées par une virgule ».
[23] « .xlsx est une extension de nom de fichier pour tableur au format Office Open XML utilisé par Microsoft Office à partir de la version 2007 » (Wikipédia, 2020).

automatiquement un dictionnaire ; et extraire la terminologie bilingue (Université Jean Moulin Lyon 3, 2023).

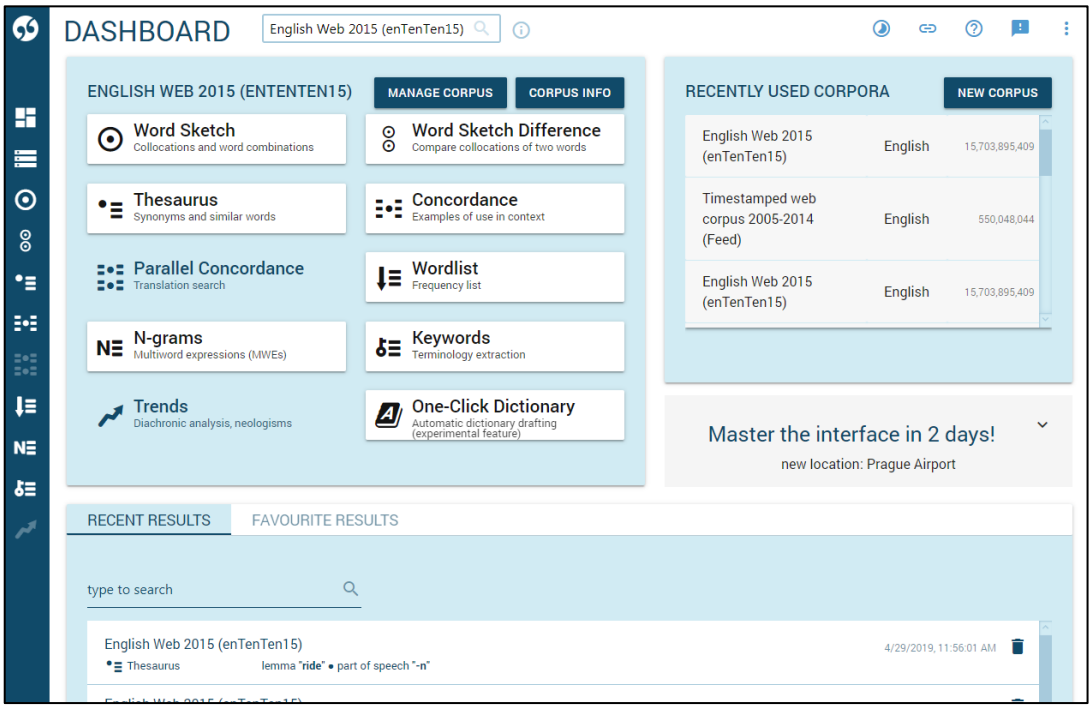


Figure 3 – La première fenêtre du logiciel Sketch Engine

Le dernier logiciel que nous présenterons est TermoStat, un outil d'acquisition automatique de termes en ligne qui a été créé par le Département de linguistique et de traduction de l'Université de Montréal[24]. Le mécanisme de fonctionnement du logiciel est très simple : il est capable d'acquérir les termes d'un corpus sur la base d'une méthode de mise en opposition d'un corpus spécialisé et d'un corpus non spécialisé. En ce moment, la version actuelle en ligne du logiciel prend en charge plusieurs langues : le français, l'anglais, l'espagnol, l'italien et le portugais. Comme pour AntConc, les documents format un corpus qui doit être analysé par TermoStat doivent être en format texte brut (TXT). TermoStat est vraiment précieux pour la tâche d'extraction terminologique : c'est l'une des tâches du travail terminologique qui permet d'étudier les termes extraits d'un corpus sur un domaine de spécialité. Pour le faire, TermoStat fait une comparaison entre un corpus appelé « corpus focus », c'est-à-dire celui en entrée, et un corpus de référence, qui est déjà inclus dans le logiciel de terminologie et est généralement un corpus de langue générale. L'extraction terminologique est basée sur des statistiques : TermoStat examine

[24]Voici le lien vers le site web : https://termostat.ling.umontreal.ca/doc_termostat/doc_termostat_en.html.

dans les deux corpus les termes qui sont plus fréquents dans le corpus focus que dans le corpus de référence et sur la base d'un calcul statistique, le terme qui apparaît le plus dans le corpus focus sera alors un candidat terme, c'est-à-dire un terme qui pourra être objet d'étude (cf. Drouin 2010).

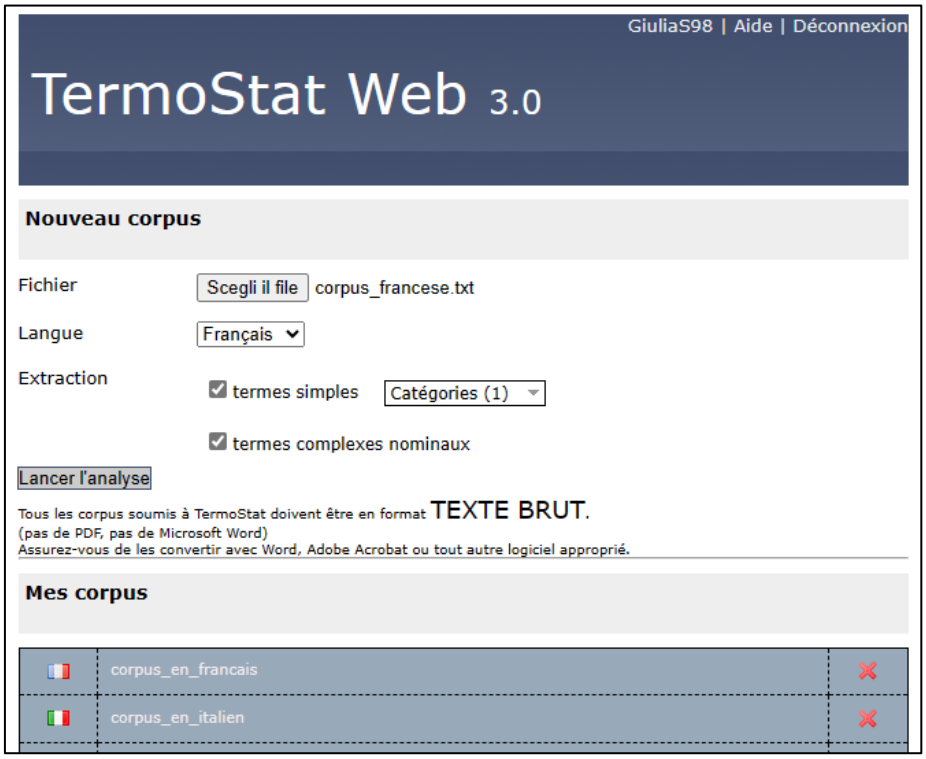


Figure 4 – La première fenêtre du logiciel TermoStat

Dans la figure ci-dessus, nous pouvons voir plusieurs éléments : le logiciel nous fait choisir le fichier, la langue et quels types de termes extraire. En plus, il y a des indications qui nous disent que le format du corpus doit être un texte brut. TermoStat peut donc fournir des listes des termes qui ont été extraits du texte d'analyse. Cette liste propose le candidat de regroupement, c'est-à-dire la variante orthographique avec le moins de modifications relativement à la forme lemmatisée par TreeTagger ; la fréquence des termes ; la matrice qui correspond aux catégories grammaticales de chaque mot candidat (nom, nom adjectif, etc.) ; et le contexte de chaque terme (accessible en cliquant sur le terme lui-même). Le logiciel offre aussi d'autres fonctions qui nous permettent de mieux comprendre les termes. La fonction « Nuage » propose une liste alphabétique des 100 termes dont le score est le plus élevé, qui nous pouvons distinguer des autres par la différence de taille de caractère. La fonction « Structuration » présente un tableau

contenant des candidats et, pour chacun d'eux, la liste des candidats qui l'incluent[25]. Puis il y a la page de décomposition qui offre un tableau où la première ligne est le candidat terme en entrée et les lignes suivantes montrent la tête et l'expansion des candidats termes, l'apposition gauche et droite, l'adjectif qui peut qualifier la tête du candidat et les termes qui forment un terme complexe avec le candidat. La fonction « Graphe » nous permet d'analyser les termes ayant une relation syntaxique partagée avec d'autres en cliquant sur les termes eux-mêmes afin de générer le graphe correspondant. Enfin, la fonction « Bigrammes » présente les bigrammes les plus forts du texte analysé : ils sont composés d'un verbe et d'un nom et accompagnés de leur fréquence et d'un score qui représente la force de l'association entre les mots qui les composent (cf. Drouin 2010).

Pour conclure, la sélection des sources et des textes, la sélection du système de gestion de corpus textuels et la sélection du système de balisage structurel sont fondamentales pour organiser un bon travail terminographique afin de trouver les justes termes à analyser ensuite[26].

3.4.4 Les difficultés

Pendant la construction d'un corpus, nous pouvons rencontrer des problématiques. Il est normal de faire face à des problèmes relatifs par exemple à la classification de textes selon le type de discours qui les caractérise, c'est-à-dire s'il s'agit de textes informatifs, instructifs ou discursifs. Il est également difficile de bien classifier le style d'un texte, la possible relation que le texte a avec un ouvrage étranger, la relation entre les participants, c'est-à-dire entre auteur et destinataire, le niveau de formalité et le statut socio-éducationnel de l'auteur (cf. Marshman 2003, 6).

Il est aussi important de parler de l'équilibre d'un corpus qui fait référence à plusieurs éléments : les sous-domaines d'un champ d'études, la variété d'auteurs, la pluralité de textes et de leurs sources et genres, les diverses situations communicatives et les niveaux de spécialisation. Il existe des critères d'équilibre de corpus, cependant il est difficile de créer un corpus qui soit parfaitement équilibré. Pour concevoir un corpus bien harmonisé, il est donc souhaitable que le/la chercheur(e) fixe tous les objectifs du travail terminographique dès le départ du projet, en les modifiant si nécessaire au cours du

[25] Par exemple, le candidat de regroupement « base » fait partie du candidat terme qui l'inclut « base militaire » ou « base de données ».
[26] Dans le chapitre conclusif de ce mémoire, nous analyserons en détail les corpus que nous avons construits et utilisés pour le projet *YourTerm TECH*, en faisant une analyse qualitative du travail.

travail ; en plus, il/elle devra décrire clairement les procédures suivies dans le rapport des résultats (cf. Marshman 2003, 4).

Comme indiqué précédemment, un corpus bien construit a la possibilité d'être réutilisé. Pour faciliter son utilisation, son partage, et sa modification, il faut employer un système de gestion polyvalent et adapter le corpus à des besoins et objectifs de projets différents que ceux pour lesquels le corpus a été conçu. La réutilisabilité d'un corpus se manifeste donc par une indexation des contenus méticuleuse et complète et par toute la documentation des procédures de recherche utilisées pour trouver les textes composant le corpus (cf. Marshman 2003, 3–4).

Dans ce troisième chapitre, nous avons présenté le corpus et ses caractéristiques, en analysant les différentes typologies de corpus qu'ils existent, sa fonction dans l'étude d'un domaine de spécialité et nous avons fait le point sur l'histoire de la linguistique de corpus et son emploi. Pour compléter le chapitre, nous avons examiné la méthodologie consacrée à la construction d'un corpus, en présentant les nombreuses étapes à suivre afin d'assembler un corpus de bonne qualité.

CHAPITRE 4

Les termes, leur extraction et la compilation des fiches terminologiques

Dans le troisième chapitre, nous avons parlé de l'extraction de termes par des logiciels tels que TermoStat selon le point de vue des corpus. Toutefois, dans le présent chapitre, nous allons analyser en détail le rôle des termes, leur formation, les différentes typologies des termes, comment fonctionne leur extraction, en utilisant aussi des exemples concrets tirés du travail terminographique pour le projet *YourTerm TECH*, et leur stockage. En plus, nous discuterons de la compilation des fiches terminologiques qui sont essentielles pour l'étude des termes d'un domaine de spécialité et, en particulier, pour la réalisation du projet *YourTerm TECH*.

4.1 Les termes et leurs caractéristiques

Nous avons déjà rencontré la définition du terme : c'est une désignation linguistique d'un concept relevant d'un domaine de spécialité. Ils sont contenus dans les dictionnaires spécialisés ou dans les banques de terminologie. Les termes se trouvent à la base de la terminologie et de la terminographie et ils sont essentiels pour le transfert de connaissances.

4.1.1 Terme simple et terme complexe

La plupart des termes ont une nature nominale, c'est-à-dire qu'ils sont formés par des noms ou, pour être plus précis, par des syntagmes nominaux. Des termes composés par de noms sont par exemple *organisme* et *angiographie* et des termes composés par de syntagmes nominaux sont par exemple *système immunitaire* et *gaz à effet de serre*. Nous devons donc faire une première différence : il existe les termes simples et les termes complexes. Le terme simple est la désignation d'un concept par une seule entité graphique, tandis que le terme complexe va désigner de concepts par deux ou plusieurs entités graphiques. Prenons-nous deux exemples du projet *YourTerm TECH* : le terme *algorithme* est un terme simple et le terme *algorithme génératif* est un terme complexe.

Les termes complexes que nous trouvons dans les dictionnaires spécialisés sont plus nombreux que les termes simples, cependant ils portent un sens compositionnel, car nous pouvons comprendre le sens des chaque unité simple qui les composent. Les termes peuvent apparaître également sous forme des sigles ou acronymes, symboles, formules et appellations latines : ces typologies de termes sont aussi appelées unités brachygraphiques (cf. L'Homme 2020, 66–67).

4.1.2 Formation des termes

Les termes sont à la base d'une langue de spécialité, c'est-à-dire une variété diaphasique d'une langue naturelle qui représente un « Ensembles des éléments linguistiques qui caractérisent les moyens utilisés pour communiquer dans une sphère d'activité ou un domaine spécialisé du savoir » (Office québécois de la langue française, 2023). La langue de spécialité est liée à un domaine de connaissance ou à une sphère d'activité spécialisée, qui est dominée par un groupe de locuteurs plus restreint que l'ensemble des locuteurs de la langue dont la langue spéciale est une variété. Cela signifie qu'elle est utilisée pour satisfaire les besoins communicatifs de cette sphère spécialisée afin de garantir une efficacité communicative. Au niveau lexical, une langue de spécialité est composée d'une série de correspondances supplémentaires par rapport aux correspondances générales et communes de la langue et au niveau morphosyntaxique, elle se caractérise principalement par un ensemble de sélections régulièrement récurrentes dans l'inventaire des formes disponibles dans la langue.

Dans une situation de communication spécialisée, il y a le besoin d'utiliser une langue de spécialité et ses termes, mais comment sont-ils créés ? Les mots d'une langue ne sont pas créés de façon anarchique, mais toutes les langues possèdent des règles et des modèles pour leur formation. Pour enrichir le vocabulaire spécialisé d'un domaine, nous pouvons utiliser des ressources existantes, des ressources nouvelles et des ressources d'autres langues (cf. Università degli Studi di Cagliari, s. d., 1).

Pour ce qui concerne l'utilisation des ressources existantes, il existe de nombreuses possibilités de formation de mots. Tout d'abord il y a le cas de l'extension du sens, c'est-à-dire la faculté de donner un sens nouveau à un mot du lexique existant. Il est possible d'agir par métaphore ou par métonymie, en ajoutant de la polysémie à la langue. Grâce à la métaphore, nous pouvons donner à un mot un sens différent en fonction d'une comparaison implicite : cela peut intéresser les mots de la langue générale vers la langue

de spécialité, par exemple, les mots *tête*, *pied* et *bras* qui représentent des parties du corps, mais ils sont aussi utilisés dans le domaine du mobilier, ou encore, les mots *souris* et *fenêtre* qui sont des mots du langage général, mais ils sont employés en informatique. Cependant, la métonymie s'utilise presque exclusivement pour l'enrichissement du lexique de la langue générale : c'est une figure de style, une relation sémantique entre deux mots dont les référents sont liés par une relation de solidarité logique ou physique qui peut autoriser l'emploi d'un mot à la place de l'autre. Voici des exemples : « je bois un verre », « les casques bleus sont intervenus ». Dans la première phrase, il y a une relation entre contenant et contenu, car le mot « verre » se réfère à « un verre d'eau » ; dans la deuxième phrase, l'expression « casques bleus » indique la couleur bleue des casques de la Force de maintien de la paix de l'Organisation des Nations. Une autre modalité d'utiliser les ressources existantes est la conversion ou dérivation impropre qui est un processus de formation de mot par lequel le mot change sa catégorie grammaticale sans aucune modification formelle. Un substantif peut donc devenir un adjectif, par exemple, les termes *un portable*, *un compact*, *un mobile* et un adjectif peut devenir un substantif. Un nom propre peut se transformer en nom commun et le nom propre est dit éponyme, par exemple, le théorème de Pythagore ou la maladie de Parkinson et la maladie d'Alzheimer, qui définissent des maladies en fonction du nom de la personne qui les a découvertes ou qui a découvert la cure, en devenant des noms communs. Même un nom commun peut devenir un nom propre, comme c'est le cas pour la figure rhétorique appelée antonomase qui désigne un personnage par un nom commun ou une périphrase qui le caractérise, par exemple, la « Dame de Fer », expression utilisée pour nommer Margaret Thatcher ; ou un personnage qui rappelle le caractère d'un autre personnage, par exemple, un casanova[1]. Pour terminer, il y a la possibilité d'intégrer des noms de marques renommés dans le lexique d'une langue, par exemple, le sopalin, le scotch et un K-way (cf. Università degli Studi di Cagliari, s. d., 1–2).

Les ressources existantes peuvent également être combinées grâce aux processus de dérivation, composition ou abréviation. La dérivation produit des mots à partir des mots préexistants et elle consiste à ajouter un morphème supplémentaire au mot : ce morphème est un affixe qui est la marque de la dérivation qui peut être ajoutée au début

[1] Un casanova est un homme qui enchaîne les conquêtes amoureuses.

d'un mot, en devenant un préfixe[2], ou à la fin d'un mot sur une base, en devenant en un suffixe[3]. Pour mieux expliquer, les préfixes sont positionnés avant le mot ; ils ne changent ni la flexion ni la catégorie grammaticale d'un mot, mais ils ont une fonction plus sémantique, car ils sont porteurs de sens ; ils sont ordinairement issus du bas latin ou du latin classique et du grec ; ils peuvent se cumuler à d'autres préfixes ; ils s'adaptent à ce qui le suit immédiatement selon le principe d'accommodation phonétique et graphique ; les préfixes les plus utilisés aujourd'hui sont dé(s)-, super- hyper- giga-, non-. De l'autre côté, les suffixes se positionnent après un mot ; ils peuvent changer la classe grammaticale d'un mot avec des exceptions, telles que les diminutifs, les péjoratifs et les collectifs ; ils ne sont jamais autonomes à différence des préfixes qui le peuvent être ; ils sont originaires des multiples langues, par exemple, du latin, du grec, de l'argot, de l'anglais, de l'italien et du provençal ; ils peuvent s'ajouter à un substantif, verbe ou adjectif et également à un nom propre, mais ils ne peuvent pas s'ajouter à un adverbe (sauf quasiment) ; ils peuvent se cumuler à d'autres suffixes ; selon le principe d'accommodation phonétique et graphique, ils partent souvent de la base latine ; les suffixes les plus utilisés aujourd'hui sont -isme, -istique, -ité, -itude. Voici deux listes de préfixes et de suffixes courants.

Préfixes	Origine	Sens	Exemples
a-, an-	grecque	privation, négation	atypique, asymétrique, anormal
anté-	latine	avant	antérieur
anti-	latine	opposition	antithèse, antidote
in-, im-, il-, ir-	latine	contraire, négation	insoluble, imparfait, illogique, irrégulier
dé-, dés-, més-, di-, dis-	latine	contraire, différent ; séparé de, qui a cessé de	désordonné, mésalliance, débranché, débrider, discontinu
dys-	grecque	difficulté, mauvais état	dysfonctionnement
extra-	latine	extrêmement, hors de	extrafin, extraordinaire
ép-, épi-	grecque	position supérieur	épicentre
hyper-	grecque	au-delà de, excès	hypermétrie, hyperfocal

[2] Ce processus s'appelle préfixation.
[3] Ce processus s'appelle suffixation.

hypo-	grecque	au-dessous de, insuffisance	hypothermique, hypotension
inter-	latine	entre	interligne, interférence
intra-, intro-	latine	au-dedans	intramusculaire, introversion
péri-	grecque	autour de	périmètre, périphérie
pluri-	grecque	plusieurs	pluriel
post-	latine	après	postérieur
pré-	latine	devant, avant, en tête de	préhistoire, préfigurer
semi-	latine	à demi	semi-rigide
super-, supra-,	latine	au-dessous	supercarburant, superstructure
trans-	latine	au-delà de, au travers	transférer, transpercer
ultra-	latine	au-delà de	ultrason, ultraviolet

Tableau 3 – Exemples de préfixes en français scientifique

Ce tableau de préfixes représente seulement une partie des préfixes existants (cf. Véron, s. d.). Pour ce qui concerne le projet *YourTerm TECH*, par exemple, il y a d'autres préfixes comme méga- du terme *mégadonnées* ou syn- du terme *synonymie*.

Suffixes	Origine	Sens	Exemples
-ade, -age	latine	collection, produit, action, état	colonnade, feuillage, salade, fromage, ruade, glissade, mariage, promenade, esclavage
-aie, -eraie	latine	plantation, végétaux	chênaie, roseraie
-aille	latine	collectif, péjoratif, action	ferraille, marmaille, racaille, canaille, trouvaille, entaille
-aire	latine	objet, métier, fonction	grammaire, dictionnaire, notaire, propriétaire, bénéficiaire
-aison, -ation, -ison, -oison, -ance, -ence, -ement, -issement, -is,	latine	action, résultat de l'action	trahison, pendaison, natation, salutation, confiance, prudence, déménagement, rajeunissement, roulis,

-ison, -ition			éboulis, semis, guérison, expédition
-al, -el, -âtre, -ance, -ence	latine	qualité	musical, ignorance, noirâtre
-ard, -arde	germanique	objet, personne	brassard, montagnard, campagnarde
-as, -asse, -ace	latine	abondance, collectif, défaut	amas, liasse, populace, bonasse
-at	latine	fonction, lieu	avocat, notariat, pensionnat
-ature, -iture, -ure	latine	fonction, action, résultat de l'action, aspect, produit	législature, capture, aventure, confiture, parure
-aut, -cule, -icule, -ule	latine	diminutif	levraut, corpuscule, veinule, monticule, ridicule
-ée, -etée	latine	abondance, quantité, durée	bordée, cuillerée, soirée, matinée, poignée
-erie	latine	qualité, défaut, collection, action, lieu	effronterie, coutellerie, mutinerie, écurie, boulangerie
-il	latine	lieu, endroit	chenil, fenil
-ille, -iole	latine	diminutif	brindille, faucille, pacotille, bestiole
-in, -ine	latine	diminutif, produit, caractère	enfantin, tambourin, plaisantin
-ique	grecque	qui a rapport à	physique, chimique, mathématique
-ite	grecque	produit, maladie	sulfite, otite, appendicite
-itude	latine	qualité, état	solitude, lassitude
-on, -onne	latine	métier, chose, diminutif	forgeron, bouchon, moucheron, ânon, caneton
-ose	grecque	maladie	arthrose, amiantose, sclérose
-u	latine	qualificatif, abondance d'un état	ventru, charnu, feuillu

Tableau 4 – Exemples de suffixes de la langue française

Cet autre tableau présente les suffixes les plus utilisés (Ministère de l'Éducation de la Saskatchewan, 1999). Pour ce qui concerne le projet *YourTerm TECH*, nous trouvons beaucoup de termes qui comprennent un des suffixes ci-dessus, par exemple, *base de connaissance* (-ance), *linguistique informatique* (-ique), *nettoyage* (-age) et *extraction de mots-clés* (-tion)[4]. Il existe également une typologie de dérivation particulière : elle s'appelle dérivation parasynthétique qui comporte la construction des mots simultanément avec un préfixe et un suffixe. Un exemple est le passage de « lune » à « alunir » : les formes caractérisées seulement par le préfixe *a-lune* ou par le suffixe *lun-ir* n'existent pas. Un autre processus de dérivation avec suffixation est la nominalisation, c'est-à-dire un procédé grammatical qui consiste à transformer en substantif un mot qui était un verbe ou un adjectif. On la rencontre souvent dans les titres de journaux, mais elle est aussi un phénomène typique de la prose scientifique. Des exemples tirés par le projet *YourTerm TECH* sont : *identification de mots* dont le nom « identification » dérive du verbe « identifier » ; *indexation automatique de documents* dont le nom « indexation » dérive du verbe « indexer » ; et *lemmatisation* qui dérive du verbe « lemmatiser » (cf. Università degli Studi di Cagliari, s. d., 2–5).

Le processus de composition est différent : c'est un procédé morphologique qui permet de former de nouveaux mots, surtout des substantifs. La composition peut avoir des schémas variables. Un mot composé peut se former à partir de l'union d'un déterminé et d'un déterminant, par exemple, *ouvre-boîte*, *table à repasser* et *lave-vaisselle*[5]. Il y a diverses possibilités de composition : nom et nom (chou-fleur, timbre-poste) ; nom et adjectif (état civil, cordon bleu) ; adjectif et adjectif (chaud-froid, aigre-doux) ; verbe et nom (vide-ordures, cache-nez) ; verbe et verbe (savoir-faire, laisser-aller) ; pronom et verbe (on-dit, rendez-vous) ; et préposition et nom (après-midi, sans-gêne). Pour ce qui concerne le projet *YourTerm TECH*, nous trouvons des mots composés, par exemple, *algorithme génétique* qui est constitué par un nom et un adjectif et *linguistique informatique* qui est constitué par deux noms. Un mot composé peut être un emprunt direct au grec ou au latin, par exemple, les mots *géographie*, *philosophie* et *philanthrope* ou ils peuvent être formés sur de bases grecques ou latines qui ne sont pas autonomes en français, par exemple, *monoplace* et *anthropologue*. Ce type de

[4] Pour mieux approfondir tous les préfixes et les suffixes de la langues française, voici le lien pour une liste exhaustive : https://www.collegeahuntsic.qc.ca/documents/52d585da-3b25-4c15-920e-4a7541e7bc84.pdf.
[5] Il n'est pas le cas pour toutes les langues, par exemple, l'anglais qui fait le contraire.

composition, connu comme « savante », est très utilisée dans la formation des mots destinés aux vocabulaires de spécialité. Les mots composés peuvent avoir différentes formes : un mot composé peut être soudé, comme *gendarme* et *portefeuille* ; il peut être relié par un trait d'union, par exemple, *sèche-cheveux*, *porte-monnaie* et *arc-en-ciel* ; il ne peut présenter aucun lien graphique comme dans les mots *petit four* et *maison close* ; et il peut être caractérisé par une préposition, par exemple, *machine à laver*, *pomme de terre* et *chemin de fer*. Relativement au sens d'un mot composé, il est possible de le déduire plus ou moins facilement du sens des composants du mot (cf. Università degli Studi di Cagliari, s. d., 5–6).

Un autre type de processus pour la formation des mots est l'abréviation qui est de plus en plus présente en néologie. L'abréviation consiste à tronquer le début d'un mot[6], par exemple, « bus » du mot *autobus* et « auto » du mot *autocar* ou la fin d'un mot[7], par exemple, « promo » du mot *promotion*, « fac » du mot *faculté* et « vélo » du mot *vélocipède*. Le phénomène de l'abréviation est présent surtout dans la langue orale argotique ou familière et dans la langue de spécialité. L'abréviation peut aussi consister à une union de la troncation et de la composition, surtout pour les mots de formation récente (formés après le 1945), par exemple le terme *informatique* de l'union de « information » et « automatique » et le terme *photocopillage* de l'union de « photocopie » et « pillage ». Il existe aussi différentes typologies de ce phénomène de troncation-composition : un mot peut se former à partir d'une apocope et une aphérèse, par exemple, *caméscope* de l'union de « camera » et « magnétoscope » ; il y a la possibilité de rencontrer aussi un mot formé par deux apocopes, comme *modem* de « modulateur » et « démodulateur » ; un mot peut être formé par deux aphérèses, par exemple, *nylon* de « vinyle » et « coton » ; on trouve également des mots composés par une apocope et un mot simple et c'est le cas de *publipostage* de « publicité » et « postage » ou *télécarte* de « téléphone » et « carte » ; et, pour conclure, il est possible d'avoir un mot construit à partir d'un mot simple et une aphérèse, par exemple, le terme *bureautique* de « bureau » et « informatique ». Pour ce qui concerne le projet *YourTerm TECH*, nous avons trouvons un terme qui est formé par deux apocopes : *AdaBoost*, un algorithme d'*adaptive boosting*. Enfin, il y a la siglaison, c'est-à-dire le processus utilisé

[6] Ce processus s'appelle aphérèse.
[7] Ce processus s'appelle apocope.

afin de créer des sigles ou d'acronymes pour économiser l'interlocution. Ce phénomène s'est développé depuis les années 1940 quand nous assistons au développement de la technologie et de la complexité des administrations. Un sigle est formé des premières lettres d'une dénomination ou d'une expression, par exemple, SNFC qui identifie la Société nationale des chemins de fer français. Quant à l'acronyme, c'est un sigle qui inclut parfois une syllabe complète d'un mot et dont la prononciation est donc syllabique, par exemple, UNICEF qui indique le Fonds des Nations unies pour l'enfance (cf. Università degli Studi di Cagliari, s. d., 6–7). Pendant le travail terminographique lié au projet *YourTerm TECH*, nous avons trouvé un seul sigle pour le français : *TF-IDF* ou *méthode de pondération TF-IDF* (sa forme longue) qui correspond au terme anglais *term frequency-inverse document frequency*.

La formation des termes peut également se baser sur des ressources d'autres langues. En particulier, nous parlons des emprunts, qui sont des mots étrangers désignant un concept dont la langue cible ne possède pas un référent. L'emprunt peut être de forme ou de sens : le premier introduit un signifiant nouveau dans le vocabulaire, par exemple, le mot *panino* et la seconde ajoute une acception calquée sur un emploi étranger à un élément déjà existant en français, c'est le cas du verbe *réaliser* qui signifie « accomplir », « effectuer » et aussi « se rendre compte », « se faire une idée » du verbe anglais *to realize*. Puis, il y a le calque qui est un mot traduit d'une autre langue : il existe le calque structural qui est traduit de manière littérale, par exemple, l'expression « été indien »[8] qui est un calque de la locution en anglais *Indian summer* ; et le calque sémantique qui traduit la notion d'un concept, par exemple, le mot « souris » qui indique le dispositif de pointage d'un ordinateur et qui vient de l'anglais *mouse*. Parmi les termes choisis pour le projet *YourTerm TECH*, nous avons trouvé un calque de l'anglais : « tokenisation » qui est la traduction du terme anglais *tokenisation* ou *tokenization* pour indiquer la segmentation d'un texte dans d'unités plus petites. Il existe aussi le xénisme ou pérégrinisme, c'est-à-dire un emprunt qui représente une réalité étrangère qui ne possède pas d'équivalent en français et des exemples sont apartheid[9] qui vient de la langue afrikaans et toundra[10] qui

[8] « Période de beau temps en automne, après une période de froid et avant l'arrivée définitive de l'hiver » (Wiktionnaire, 2023).
[9] « Le mot 'apartheid' évoque immédiatement l'Afrique du Sud. […] C'est un système d'oppression et de domination d'un groupe racial sur un autre, institutionnalisé à travers des lois, des politiques et des pratiques discriminatoires » (Amnesty International France, s. d.).
[10] La toundra est une formation végétale qui caractérise les zones climatiques froides.

correspond au terme russe *мундра* (La langue française, 2024). Dans la plupart des cas, les xénismes ne s'intègrent pas au lexique de la langue cible, en restant fragiles et timidement utilisés (cf. Università degli Studi di Cagliari, s. d., 7–8).

Quand nous parlons des emprunts, il faut aussi discuter de leur intégration dans la langue française et il existe différents cas. Un mot étranger peut être intégré tel quel, par exemple, le mot *salsa* qui désigne une danse latino-américaine. Les emprunts peuvent être adaptés phonologiquement : en général, la langue française est caractérisée par deux tendances, qui sont d'un côté le choix d'une prononciation « à la française » et de l'autre côté l'imitation approximative de la prononciation et de l'orthographe de la langue source. Il existe aussi l'adaptation graphique qui prévoit des modes et degrés différents : l'emprunt maintient son identité totale, c'est le cas des mots *pizzeria*, *sport*, *loft* et *timing* ; l'emprunt subit une francisation graphique partielle, par exemple, les mots *concertos* et *macaronis* ont ajouté la -s finale ; l'emprunt face une réécriture globale, processus qui n'est plus utilisé aujourd'hui, par exemple, le verbe *estropier*[11] qui vient de l'italien *stroppiare*. Nous avons déjà parlé de la condition des mots étrangers dans le deuxième chapitre : dans le monde francophone, les puristes de la langue française s'opposent à l'invasion et à l'emploi des mots étrangers et il existe des organismes qui s'en occupent, par exemple, les commissions spécialisées de terminologie et de néologie et l'Académie française, dont nous avons précédemment discuté (cf. Università degli Studi di Cagliari, s. d., 8–10).

4.1.3 Termes non prédicatifs et termes prédicatifs

Les termes peuvent appartenir à différentes parties du discours. Cependant, le nom est une partie centrale en terminologie. La plupart des termes sont des termes nominaux et souvent les dictionnaires terminologiques ne prennent pas en considération les autres parties du discours, telles que les verbes, les adjectifs et les adverbes. Le nom est préféré en terminologie si nous tenons compte seulement d'une optique conceptuelle, mais si nous voulons intégrer aussi une optique lexico-sémantique, les premières incohérences remontent à la surface. Analysons-nous cet exemple : un dictionnaire spécialisé comprendra les termes *virus* et *protéine*, deux noms, mais il ne comprendra pas les adjectifs *viral* et *protéique*. Les dictionnaires spécialisés et les banques de terminologie

[11] « Priver (quelqu'un) de l'usage normal d'un, de plusieurs de ses membres en le blessant ou en le mutilant » (Centre National de Ressources Textuelles et Lexicales, s. d.).

préfèrent les noms, même si dans les textes spécialisés, noms et adjectifs ou verbes sont utilisés de la même mesure. Par exemple, les textes spécialisés juridiques emploieront le terme de nature nominale *dépôt* et aussi le verbe *déposer* sans aucune distinction. Encore, les textes médicaux utiliseront *sang* et aussi l'adjectif *sanguin*. Le nombre de termes qui n'ont pas une nature nominale est donc nettement inférieur au nombre de termes de nature nominale (cf. L'Homme 2020, 67–69).

Normalement, les entités du monde réel sont désignées par des noms. Selon l'optique conceptuelle, un concept est défini à partir d'un ensemble d'objets qui ont des caractéristiques en commun, par exemple, le concept de « citron » qui désigne un fruit, de forme ovale, jaune et avec du goût amer. Le citron est un objet concret dans le monde réel, cependant il existe diverses unités lexicales qui ne peuvent pas être décrites seulement utilisant l'organisation du monde réel ou l'idée que nous en avons. Il est donc nécessaire de joindre au sens de ces unités lexicales d'autres sens. En particulier, nous parlons des sens complémentaires qui peuvent être des prédicats sémantiques ou d'actants sémantiques : ils servent à compléter le sens d'une unité lexicale. Un exemple est le verbe *léguer* : pour expliquer son sens, il faut faire référence à la personne qui lègue, à la chose qui va être léguée et au bénéficiaire. Dans ce cas, le prédicat sémantique est le verbe *léguer* et les actants sémantiques qui le complètent sont la personne qui lègue, la chose léguée et le/la bénéficiaire. Cela représente une typologie différente de terme : le terme prédicatif, qui a besoin des actants sémantiques pour compléter son sens. Au contraire, les termes non prédicatifs ne nécessitent pas des actants sémantiques, comme pour le terme *citron*. Les termes prédicatifs peuvent avoir un actant sémantique ou plusieurs sur la base des actants qui sont nécessaires pour en définir le sens (cf. L'Homme 2020, 69–71). Relativement au projet *YourTerm TECH*, les termes que nous avons analysés sont presque tous des termes soit pour le français, soit pour l'anglais et l'italien, la seule exception est un verbe que nous avons choisi pour l'italien : le verbe italien *etichettare*, en français « étiqueter » qui identifie l'action pour laquelle les données textuelles sont enrichies d'étiquettes indiquant la classe à laquelle appartiennent les parties du discours d'un texte. Pour la suite, nos termes sont des termes caractérisés par une nature de type nominale, en tant qu'ils sont des noms ou des syntagmes nominaux.

4.1.4 Identifier les termes

Noms, adjectifs, verbes et adverbes peuvent représenter des termes. Génériquement, les adjectifs dérivent de noms ou de verbes, par exemple, l'adjectif *constitutionnel* est dérivé du nom « constitution » et l'adjectif *compilé* dérive du verbe « compiler » ; tandis que les adverbes en terminologie sont les adverbes qui finissent par le suffixe -ment, tel que *cliniquement*. La sélection des termes est essentielle pour un bon travail terminographique : nous devons donc tenir en considération certains critères. Tout d'abord, le sens de l'unité lexicale que nous avons choisie doit être lié au domaine de spécialité que nous devons étudier. Il est évident qu'il sera plus facile d'établir un lien si nous travaillons avec les noms, mais si nous rencontrons des termes prédicatifs, nous pouvons utiliser des critères lexico-sémantiques pour nous aider. Lorsque nous rencontrons une unité lexicale prédicative dont les actants sémantiques présentent des traits qui peuvent confirmer son sens spécialisé, elle peut être spécialisée elle-même. Voici en exemple : dans la phrase « Or plusieurs simulations indiquent que le changement climatique ne devrait pas beaucoup réchauffer (voire refroidir) l'air situé au nord de l'Europe », le verbe *réchauffer*, qui apparaît dans des textes spécialisés relatifs à l'environnement, se présente avec des termes qui sont relevant pour le domaine, c'est-à-dire « changement climatique » et « l'air situé au nord de l'Europe ». Toutefois, il faut tenir compte du fait qu'une unité prédicative possède un sens spécialisé seulement s'elle est combiné avec des actants sémantiques de sens spécialisé. Cela signifie que si l'unité prédicative est accompagnée par des actants sémantiques non spécialisés et qu'elle porte le même sens, elle ne sera pas spécialisée. Ensemble aux actants sémantiques, nous devons également contrôler la parenté morphologique : c'est une autre méthode pour vérifier un sens spécialisé. Si nous avons trouvé un terme qui répond aux critères précédents, il aura des dérivés spécialisés. Voici un exemple : le terme *compilateur* fait partie du domaine de l'informatique et signifie « programme dont la fonction est de convertir un code source en code machine ». Par conséquent, les mots *compiler*, *compilation*, *recompiler* et *compilable* doivent être identifiés comme des termes. Un autre exemple est le terme relatif au domaine de l'nvironnement *pollution* auquel nous ajouterons aussi *polluer*, *polluant* et *dépolluer*. Pour conclure cette partie dédiée aux critères pour identifier les termes, nous devons aussi mentionner la relation paradigmatique. Cela examine les rapports qu'entretiennent des unités linguistiques : un

bon exemple est le terme *interface* pour lequel nous devons considérer les mots *menu* et *fenêtre* qui font partie de l'interface. Un autre exemple sont les deux verbes *accuser* et *défendre*, c'est-à-dire deux antonymes et nous devons donc considérer comme un terme aussi les antonymes (cf. L'Homme 2020, 71–74).

4.1.5 Le terme dans le texte spécialisé

Lorsque les termes sont identifiés, il faut les examiner en relation avec les textes qui les contiennent, c'est-à-dire leur environnement linguistique. Les textes spécialisés présentent des formes linguistiques : certaines peuvent avoir un sens associé au domaine spécialisé et certaines peuvent avoir un autre sens, non spécialisé ou associé à un autre domaine. En plus, une même forme linguistique peut avoir plusieurs sens associés au même domaine de spécialité. Pour faire face à ces difficultés liées au sens, les terminographes peuvent s'appuyer à des tests lexico-sémantiques qui servent pour confirmer les sens propres à un domaine de spécialité. Il existe six critères lexico-sémantiques. Tout d'abord, il y a la substitution par un synonyme qui une stratégie utilisée pour contrôler si nous pouvons substituer à un terme une autre unité lexicale dans un contexte précis. Voici un exemple pour mieux comprendre : dans la phrase « Le clavier constitue le moyen d'entrée classique des données et des commandes », le terme *entrée* peut être substitué par *saisie*. Puis, il y a l'opposition différentielle qui sert pour appliquer un antonyme à une forme qui peut être ambiguë afin de vérifier son sens. Un autre critère est la dérivation morphologique différentielle : elle permet de tester la possibilité de dégager des ensembles de dérivés morphologiques qui correspondent à des sens distincts. Par exemple, dans la phrase « […] le changement de climat et les effets qui en découlent représentent un risque sérieux pour la stabilité politique, économique et sociale », le mot « risque » ne peut pas être associé au verbe « risquer », mais dans la phrase « Ce politicien prend de grands risques qui peuvent nuire à sa carrière », « risques » peut être substitué par le verbe « risquer (gros) ». Puis, le critère de la présence de liens paradigmatiques différentiels examine les oppositions et les parentés sémantiques des unités lexicales sur le plan paradigmatique. Un exemple est le sens du verbe *écrire* dans le domaine de l'informatique : le premier paradigme est représenté par le syntagme « écrire un programme », c'est-à-dire développer ou programmer et le second est représenté par le syntagme « écrire des données sur un support de stockage », c'est-à-dire lire ou adresser. Les derniers critères sont la cooccurrence compatible et la

cooccurrence différentielle : elles permettent d'établir si nous sommes confrontés à de mêmes sens ou à des sens différents en combinant les unités lexicales à des cooccurrents différents. Par exemple, dans les phrases de nature juridique « Cela doit être fait avant le dépôt de l'accusation » et « Précisions d'abord que le mot cautionnement vise toutes les conditions posées à la liberté provisoire ; promesse, engagement de payer avec ou sans caution, dépôt d'argent », il n'existe pas une cooccurrence telles que « dépôt d'argent et de l'accusation », ils ont deux sens différents (cf. L'Homme 2020, 74–80).

Les termes peuvent également subir des variations flexionnelles, c'est-à-dire qu'ils peuvent changer leur forme due au genre et au nombre pour ce qui concerne les noms et due au temps, au nombre, au mode, à la personne et à la voix (active ou passive) pour ce qui concerne les verbes. En plus, les termes peuvent changer en prenant une autre forme très différente de l'original et ils peuvent être scindés. Les expressions relativement longues peuvent se présenter diversement dans un texte ou dans un corpus. Par exemple, l'expression « droits et libertés de la personne » a d'autres formes, à savoir « libertés et droits de la personne », « droits et libertés de la personne humaine » et « droits et libertés individuels ». Cette variation formelle s'appelle variation terminographique et nous devons la distinguer de la synonymie : la variation terminologique est liée aux changements des termes dans les textes spécialisés et elle peut intéresser différentes parties du discours ; tandis que la synonymie représente le rapport entre deux ou plusieurs formes qui ont le même sens en touchant seulement les unités reconnues comme termes et elle s'applique uniquement aux termes qui appartiennent à la même partie du discours. Les variantes terminologiques sont examinées par les terminographes et elles deviennent problématiques lorsqu'elles sont analysées par des traitements automatiques. Il en existe plusieurs : les variantes graphiques concernent des changements minimaux, tels que l'ajout d'un signe diacritique, d'un trait d'union ou d'une alternance majuscule-minuscules ; les variantes flexionnelles concernent les différentes formes fléchies des termes ; les variantes syntaxiques faibles s'occupent de l'emploi des prépositions et des déterminants, par exemple, *siège à bébé* et *siège pour bébé*, *imprimante à laser* et *imprimante laser* et *traitement de parole* et *traitement de la parole* ; et les variantes morphosyntaxiques s'occupent de la variation des parties du discours, en changeant aussi les phrases, par exemple, dans la phrase « L'automate peut être programmé », la verbe peut devenir un adjectif « L'automate est programmable » (cf. L'Homme 2020, 80–83).

À la différence de termes simples, les termes complexes portent à la création de problèmes, surtout si nous parlons des traitements automatiques. Les termes complexes peuvent être composés par des groupes nominaux autonomes insérés dans d'autres groupes nominaux, en formant des termes très longs et difficiles à examiner. Cela concerne le découpage des termes, c'est-à-dire la possibilité de couper les termes complexes longs en des morceaux plus petits. Les termes complexes peuvent aussi être coordonnés ou juxtaposés dans une phrase. Dans ce cas-là, les composantes des termes contenues dans une coordination ou une juxtaposition ne sont pas toujours répétées. Voici un exemple très simple : dans la phrase « L'injection directe de diazoxide dans les artères fémorales, rénales et coronaires élève le débit […] », les termes potentiels sont artères fémorales, artères rénales et artères coronaires, mais seulement le terme complexe « artère fémorale » est une suite graphique, les autres doivent être reconstruites. Les termes complexes rencontrent également des problèmes liés à l'insertion des unités lexicales qui seront difficiles à récupérer et des problèmes liés à l'élision, c'est-à-dire l'omission d'une composante du terme, par exemple, le terme *infection aux voies urinaires* qui peut être transformé en « infection urinaire ». Les terminographes doivent aussi tenir compte de l'ambiguïté de structure : elle concerne la difficulté à établir les liens syntaxiques entre les composantes d'un terme, particulièrement problématique dans un contexte informatique. Pour conclure, les termes complexes peuvent être repris partiellement dans le texte. Cette reprise s'appelle reprise anaphorique : l'anaphore sert dans un texte pour reprendre un concept déjà nommé et représente un phénomène complexe spécialement pour les traitements automatiques. Voici un exemple : dans la phrase « Cette micro-angiographie se traduit par un épaississement des parois vasculaires des vaisseaux de la substance blanche. Ces vaisseaux deviennent tortueux », le terme complexe *vaisseaux de la substance blanche* est repris par « vaisseaux » (cf. L'Homme 2020, 83–89).

4.2 L'extraction de termes

Nous avons parlé de l'extraction de termes dans le deuxième chapitre et dans le troisième chapitre, mais dans ce chapitre nous l'analyserons en détail. L'extraction terminologique fait partie des tâches à réaliser afin de compléter un projet terminographique. Aujourd'hui, grâce à l'informatique et aux technologies appliquées au domaine de la terminologie, il existe des programmes qui sont capables d'extraire les termes d'un corpus : ils sont les extracteurs de termes. Le rôle principal de ces

programmes est « prendre des décisions sur la nature des unités lexicales » (cf. L'Homme 2020, 185). Ils doivent se substituer aux terminographes. De toute évidence, il sera impossible d'éliminer totalement le travail de l'homme, car les extracteurs posent encore des problèmes et des limites, malgré les améliorations et les mises à jour continues. Pour automatiser intégralement l'extraction de termes, nous devons donc travailler encore sur les logiciels et leur entraînement (cf. L'Homme 2020, 185).

4.2.1 Les bases d'un extracteur de termes

L'extracteur de termes s'occupe de trouver des mots graphiques ou des suites de mots graphiques dans un texte ou dans une série des textes (corpus). Ces mots ou suites de mots choisis par l'extracteur sont les candidats-termes qui pourraient correspondre à des unités terminologiques. Imaginons-nous d'examiner un texte ou un ensemble de textes pour y trouver des unités lexicales qui soient des termes : cela sera une tâche facile si nous connaissons le domaine, mais si nous ne le connaissons pas, nous aurons de problèmes au niveau terminologique. La même chose s'applique aux machines, car les outils informatiques tels que les logiciels pour l'extraction de termes rencontrent des difficultés à recevoir les paramètres opérationnels. Ils se basent donc sur d'autres indices. Tout d'abord, ils s'appuient sur la fréquence et sur la répartition d'une unité dans un corpus représentatif d'un domaine de spécialité. Puis, ils vérifient la prédominance de termes nominaux, car les noms sont plus fréquents et la plupart d'extracteurs sont entraînés pour rechercher seulement les noms. Les extracteurs analysent aussi la complexité des termes pour le fait que les termes complexes sont nombreux et un grand nombre des logiciels consacrés à l'extraction de termes recherchent uniquement les termes complexes. Pour conclure, ces logiciels tiennent compte du nombre de séquences qui peuvent coïncider à des termes complexes, car ces derniers sont formés par une séquence finie de parties de discours. Généralement en français les termes complexes sont composés par un nom modifié par un adjectif ou par un syntagme prépositionnel. Les extracteurs utilisent ces indices de manière différente, en donnant des résultats divers (cf. L'Homme 2020, 186–187). Les termes que nous avons analysés pour le projet *YourTerm TECH* sont pour la plupart des termes complexes formé d'un nom et un adjectif ou d'un nom et d'un syntagme prépositionnel.

4.2.2 Comparer des corpus pour extraire les termes

Les termes d'un domaine spécialisé peuvent être choisis grâce à une liste d'exclusion, c'est-à-dire que les terminographes ont la possibilité d'obtenir les concepts les plus importants dans un texte ou un corpus en écartant les mots grammaticaux[12] et d'autres mots fréquents. Cette stratégie-là est très simple et facile à compléter, toutefois, elle utilise l'indice de la fréquence seulement en fonction du nombre d'occurrences d'un mot et il n'est pas toujours le cas pour un mot d'être un terme spécialisé parce qu'il est fréquent dans un texte ou un corpus. Une bonne stratégie consiste à comparer un corpus spécialisé et un corpus de référence : c'est la méthode que nous avons brièvement présentée en parlant des logiciels de gestion de corpus comme TermoStat. La comparaison de ces deux corpus se base sur la fréquence des mots : les possibles termes spécialisés seront les plus fréquents dans le corpus spécialisé que dans le corpus de référence. Ce dernier est un corpus très volumineux et il est composé de textes de plusieurs natures. Normalement, les listes fournies par les extracteurs automatiques de termes comprennent des informations : le nombre d'occurrences des mots dans les deux corpus ; le pourcentage des occurrences des mots par rapport aux autres dans les deux corpus ; et l'évaluation de la fréquence des mots par rapport au corpus de référence. La technique de comparaison de corpus peut présenter des problématiques. La fréquence d'un mot n'est toujours pas indice de son statut terminologique, car il est possible de rencontrer des candidats qui ont une fréquence élevée, mais ils ne désignent pas des concepts pertinents pour le domaine d'étude. En plus, certains extracteurs de termes ne tiennent pas compte de la différence entre terme simple et terme complexe : les listes qui nous sont présentées peuvent contenir des termes simples, mais il n'y a pas une claire indication s'ils sont des parties de termes complexes (cf. L'Homme 2020, 188–192).

[12] Un mot grammatical est un « Mot à sémantisme faible servant essentiellement à assurer des fonctions syntaxiques et appartenant à une liste fermée d'unités. En français, on considère généralement que les prépositions, les conjonctions, les pronoms et les déterminants sont des mots grammaticaux. On oppose ces derniers aux mots lexicaux » (Office québécois de la langue française, 2000).

4.2.3 Les autres techniques pour l'extraction de termes

Nous avons donc vu à quel point il est facile de rencontrer des problèmes afin d'obtenir une bonne liste des candidats termes. Cependant, il existe des techniques différentes qui peuvent aider les programmes informatiques dédiés à l'extraction de termes, en corrigeant certaines imperfections (cf. L'Homme 2020, 192).

Une première technique consiste à rechercher une séquence de mots graphiques qui sont souvent présents dans un corpus spécialisé. Dans ce cas-là, la consultation d'un corpus de référence n'est pas si importante, car le point central est représenté par les textes spécialisés. Cette technique s'appelle calcul des segments répétés : elle permet d'analyser les textes spécialisés afin de trouver des séquences de mots graphiques courants qui se présentent dans ces textes. Les segments répétés peuvent donc constituer des termes complexes, par exemple, dans un texte médical, les séquences *bloc de branche* et *extrasystole ventriculaire* apparaissent plus de deux fois et elles peuvent donc être considérées comme termes complexes. Cependant, il y a également des séquences répétées qui ne sont pas de termes complexes, tels que « s'il s'agit ». C'est donc à ce moment que les terminographes utiliseront une liste d'exclusion pour filtrer les suites de mots graphiques courants afin de choisir les termes complexes destinés à être examinés. La liste d'exclusion permet d'écarter les séquences de mots qui comprennent des déterminants ou des prépositions comme premiers et derniers mots et les verbes conjugués. Pour bien fonctionner, cette technique devra être employée en examinant des corpus de taille raisonnable, car les séquences de mots graphiques qui apparaîtront seulement une fois ne seront pas ramenées (cf. L'Homme 2020, 192–194).

Une autre technique se base sur des calculs statistiques qui sont utilisés pour mesurer le caractère non accidentel de la combinaison de mots graphiques. Cette technique est appelée association forte et pour l'expliquer nous utiliserons le terme *air comprimé*. En examinant un corpus de mécanique de 50.000 mots, nous trouverons le mot « comprimé » 85 fois et le mot « air » 123 fois. L'adjectif succède à « air » pour 81 fois et « air » précède « comprimé » pour 85 fois. Les autres cas dans lesquels ils apparaissent font partie d'autres combinaisons. Par conséquent, il est permis de considérer « air comprimé » un terme complexe par association forte. L'association entre de mots graphiques peut être vérifiée grâce à un calcul statistique, en déterminant si une combinaison de mots est plus fréquente qu'une autre. Ce type de calcul se base sur des

couples, c'est-à-dire des mots graphiques qui ne sont pas obligatoirement contigus et ils sont formés à partir d'une fenêtre. Dans la fenêtre, nous trouverons le nœud, c'est-à-dire un premier mot graphique qui sera contourné d'un nombre fixe de mots à gauche et à droit. Chaque de ces mots sera numérotés de 1 à *n* et de -1 à -*n*. Le calcul statistique prend en considération le nombre de couples où les deux mots *x* et *y* se présentent ensemble ; le nombre de couples où *x* apparaît ; le nombre de couples où *y* apparaît ; et le nombre total de mots du corpus examiné. Alors que nous disposons de toutes ces informations, nous pourrons calculer l'information mutuelle, c'est-à-dire un principe théorique très utilisé en terminologie computationnelle. La probabilité d'un couple est calculée sur le nombre total d'occurrences de *x*, de *y* et de *x*, *y* par rapport au nombre total des mots d'un corpus. Lorsque les calculs sont effectués, chaque couple reçoit un poids qui montre le degré d'association. Si le poids obtenu est plus élevé que trois, les couples peuvent être intéressants. Le calcul de l'information mutuelle est donc utile pour extraire des termes complexes, mais il peut aussi détecter simplement des collocations ou des relations sémantiques (cf. L'Homme 2020, 194–198).

Le calcul des segments répétés et le calcul de l'information mutuelle représentent deux bonnes techniques, car elles sont exploitables sur les textes bruts, mais elles ne sont pas rapportées à des langues spécifiques et elles ne traitent pas les informations linguistiques. Il faut donc recourir à des systèmes de filtrage qui seront rattachés à une langue précise. Une différente typologie de technique pour identifier les termes complexes est l'extraction des séquences de parties du discours. Nous avons déjà constaté que les syntagmes nominaux sont généralement considérés des termes, car ils correspondent à un nombre fini de séquences de parties du discours. Pour trouver ces séquences de parties du discours, il est nécessaire de travailler sur un corpus étiqueté. Dans ce contexte-là, il existe une technique d'extraction de termes qui s'appelle identification de patrons typiques : elle est utilisée pour rechercher de syntagmes nominaux composés de parties du discours régulières. En français, il y a des patrons de base : un nom et un adjectif ; un nom et un autre nom ; un nom, une préposition et un autre nom ; un nom, une préposition, un déterminant et un autre nom ; et un nom, une préposition et un verbe à l'infinitif. Ces patrons seront utilisés pour entraîner les extracteurs afin de découvrir les séquences correspondantes et ils se servent également d'étiquettes. Pour être encore plus précis, les extracteurs recourent à des patrons

spécifiques pour analyser les prépositions et les déterminants. Ils recherchent donc des prépositions admissibles, par exemple, à, de et sur et ils n'admettront pas les articles indéfinis ou les démonstratifs. À ce stade, un des problèmes les plus communs consiste à comprendre si des séquences relevées font partie de syntagmes nominaux plus longs : c'est le problème du découpage du terme (cf. L'Homme 2020, 199–202). Cette technique s'occupe d'identifier si un terme complexe est représenté par un groupe nominal autonome, un groupe nominal suivi d'un adjectif, ou un groupe nominal qui fait partie d'un syntagme prépositionnel rattaché à un nom (cf. L'Homme 2020, 85).

Une technique alternative pour trouver des termes complexes consiste à identifier des frontières de termes : c'est une modalité qui examine quelles parties du discours ne contribuent pas à former des termes, en coupant les textes spécialisés. Les indices principaux qui sont utilisés pour découper les textes sont appelés repères non ambigus et ils sont les signes de ponctuation, les verbes conjugués, les conjonctions de subordination et les pronoms. Les extracteurs qui exploitent cette technique considèrent les mots autour des repères ambigus : par exemple, si nous rencontrons un déterminant précédé d'un verbe ou d'un signe de ponctuation, il représentera une frontière ; ou si nous trouvons un déterminant antéposé à une préposition, il sera considéré une partie d'un terme complexe. Ces combinaisons aident les terminographes à découper les textes et les coupes seront indiquées au moyen du symbole #. Cette technique est parfaite pour souligner spécialement les termes simples (cf. L'Homme 2020, 203–204).

L'identification de patrons typiques et le repérage de frontières entre termes sont deux techniques qui ne présentent pas de problèmes concernant la fréquence, mais elles rencontrent différents types de problématiques. Parfois, les syntagmes nominaux sont formés et ont la même modalité de construction des termes complexes, mais ils ne seront pas de termes. Il existe de syntagmes nominaux qui sont formés d'un nom et d'un adjectif qui n'ont pas un statut terminologique, mais ils correspondent à un des patrons typiques individués (cf. L'Homme 2020, 204–205).

Les logiciels d'extraction de termes peuvent donc utiliser deux différentes catégories de techniques : les techniques statistiques, telles que le calcul des segments répétés et le calcul de l'information mutuelle, et les techniques linguistiques, telles que l'identification de patrons typiques et le repérage de frontières entre termes. Généralement, les extracteurs de termes combinent les deux techniques : une technique

statistique peut recourir à des informations linguistiques minimales et une technique linguistique peut employer un critère de fréquence. Cependant, certains extracteurs de termes résultent plus performants grâce à une approche mixte. Cette typologie de logiciel d'extraction a été proposé par la chercheuse Béatrice Daille et il examine des textes étiquetés ; il trouve des séquences formées d'un nom et un adjectif ou d'un nom et un autre nom ; puis, il crée une liste identifiant des séquences sous forme de couples ; il poursuit avec le calcul de la fréquence et il écartera les séquences qui ont une fréquence inférieure à deux ; enfin, l'extracteur procèdera à calculer d'autres statistiques sur l'évaluation du degré d'association. Le tournant se situe dans la comparaison de la liste produite de l'extraction automatique et d'une liste de référence basée sur les termes de la banque de terminologie *Eurodicautom*[13]. En plus, un groupe d'experts examineront les termes. Cette approche mixte est utile pour écarter des couples statistiques fréquents qui ne correspondent pas à des patrons, ainsi que les collocations et les séquences de mots qui ont une parenté sémantique (cf. L'Homme 2020, 206–207).

4.2.4 Le logiciel TermoStat

TermoStat a été créé pour trouver les termes significatifs dans des corpus spécialisés. Nous avons déjà vu que ce logiciel utilise la comparaison entre un corpus d'entrée ou d'analyse et un corpus de référence. Cependant, nous expliquerons en détail comment il fonctionne.

Avant de poursuivre la présentation du fonctionnement de ce logiciel d'extraction de termes, nous devons le positionner : TermoStat représente l'union de la reconnaissance de parties du discours et de patrons avec la comparaison d'un corpus. Il combine donc des techniques linguistiques à la comparaison de corpus. Tout d'abord, TermoStat s'appuie à l'étiqueteur TreeTagger afin d'attribuer aux mots graphiques des étiquettes indiquant la partie du discours, des informations flexionnelles et le lemme. Puis, l'extracteur en question détecte des séquences de mots graphiques qui correspondent à des patrons typiques. Ensuite, il effectuera des calculs statistiques pour estimer les fréquences dans le corpus d'entrée et dans le corpus de référence. Grâce à ces calculs, nous obtenons trois catégories de mots : la première catégorie comprend les formes très

[13] C'est une ex-base de la Commission européenne qui était géré par le service de traduction. Elle comprend de termes économiques, scientifiques, techniques et juridiques. Aujourd'hui, elle a été substituée par l'IATE (InterActive Terminology for Europe) la nouvelle base terminologique européenne (cf. Lebert, 2010).

fréquentes dans le corpus spécialisé, c'est-à-dire les mots qui pourront être des termes ; le second groupe est formé par les formes banales, c'est-à-dire des formes qui ne sont pas pertinentes pour le corpus spécialisé et pour le corpus de référence ; et le dernier groupe comprend des formes avec une fréquence moins élevée dans le corpus spécialisé, c'est-à-dire des formes significatives pour le corpus de référence (cf. L'Homme 2020, 208).

Pour mieux comprendre le mécanisme de l'extraction de termes par TermoStat, nous présentons un exemple pratique tiré du notre projet *YourTerm TECH*.

Candidat de regroupement	Fréquence	Score (Spécificité)	Variantes orthographiques	Matrice
re	208	221.07	re	Nom
ia	166	196.9	ia	Nom
https	126	172.76	https	Nom
mantique	119	168.52	mantique mantiques	Nom
e	238	156.26	e	Nom
corpus	161	150.27	corpus	Nom
jstor	92	147.99	jstor	Nom
terms	85	142.19	terms	Nom
reque	79	137.02	reque	Nom
to https	78	136.13	to https	Nom Adjectif
langage naturel	80	132.18	langage naturel	Nom Adjectif
grammaire	114	127.39	grammaire grammaires	Nom
this content	66	125.08	this content	Nom Adjectif
this content downloaded	65	124.11	this content downloaded	Nom Adjectif Adjectif
traitement automatique	64	123.14	traitement automatique traitements automatiques	Nom Adjectif
from	84	121.68	from	Nom
this	71	119.97	this	Nom
linguistique	67	118.27	linguistique linguistiques	Nom
pre	56	114.03	pre	Nom
repre	55	114	repre	Nom
syntagme	60	111.05	syntagme syntagmes	Nom

phrase	250	106.52	phrase phrases	Nom
cielle	46	104.07	cielle	Nom
intelligence arti	45	102.91	intelligence arti	Nom Adjectif
traduction automatique	43	100.54	traduction automatique	Nom Adjectif
verbe	117	99.99	verbe verbes	Nom

Tableau 5 – Première partie de la liste des candidats-termes du logiciel TermoStat pour le projet *YourTerm TECH*

Dans le tableau ci-dessus, nous pouvons voir les candidats-termes français avec leur fréquence, spécificité[14], variantes orthographiques et matrice. La liste comprend certains mots qui peuvent être des termes, par exemple, *corpus*, *langage naturel*, *linguistique*, *syntagme* et *traduction automatique*. Malgré un premier coup d'œil sur le tableau et les possibles termes que nous pouvons extraire, les nombreux mots ne font pas partie du lexique de l'intelligence artificielle et du TAL. Des exemples sont *jstor*, *https* et *intelligence arti* : le mot « jstor » n'est pas un terme du domaine de l'intelligence artificielle, car c'est la bibliothèque numérique où nous avons trouvé beaucoup de textes scientifiques sur notre domaine d'étude ; encore, l'entrée « https » n'est pas un terme lié à l'IA, c'est la sigle d'une variante d protocole de communication client-serveur développé pour le Web (*Hypertext Transfer Protocol*)[15], et la même chose s'applique à l'expression « to https » qui est évidemment en anglais et elle se trouve dans nombreuses phrases qui citent *« All use subject to https://about.jstor.org/terms »*[16] ; l'expression « intelligence arti » pourrait être un terme, mais elle n'est pas complète, car il n'existe pas l'adjectif « arti », cela serait « artificielle ». Un mot qui pourrait être un terme est le sigle « ia » qui signifie « intelligence artificielle ». Un autre mot qui ne fait pas partie du domaine pertinent de notre projet est le mot *mantique* : cela signifie « Ensemble des méthodes permettant de prédire l'avenir ou de révéler des informations cachées par des

[14] « Le calcul de spécificité a été proposé par Lafon (1980) afin de cerner le vocabulaire spécifique à un sous-corpus par rapport à l'ensemble d'un corpus. […] Cette approche permet de comparer le comportement des unités lexicales en fonction de critères variables. Nous adaptons légèrement la démarche en fusionnant le corpus de référence et le corpus d'analyse afin de vérifier si le lexique de ce dernier se comporte comme le lexique du premier. Le calcul des spécificités conduit à l'obtention d'un score qui facilite le classement des CT les uns par rapport aux autres » (cf. Drouin 2010).
[15] En français, « protocole de transfert hypertexte ».
[16] En français, « Toute utilisation est soumise à https://about.jstor.org/terms ».

moyens surnaturels », c'est l'art du devin (La langue française, 2024). Nous pouvons ainsi constater que l'extraction automatique de termes n'est pas parfaite, elle présente des problèmes et des limites. Les mots comme *https* et *jstor* ne peuvent pas appartenir au lexique de l'IA, malgré leur fréquence élevée : c'est un des problèmes dont les terminographes doivent être prudents. Le tableau ci-dessus est seulement la première partie de la liste des candidats termes pour la langue française et en procédant à l'examen de la liste, il sera de plus en plus difficile de démonter le statut terminologique des mots. À la différence d'autres logiciels pour l'extraction de termes, TermoStat nous fournit aussi les variantes orthographiques des candidats termes, c'est-à-dire qu'il n'analyse pas les mots un à un, mais il compte également leur variation flexionnelle. Parfois, les extracteurs de termes ne prennent pas en considération la variation flexionnelle, en insérant dans la liste des candidats termes un mot au singulier et le même mot au pluriel : c'est le cas de *machine* et *machines* qui ont été évaluées de manière séparée (cf. L'Homme 2020, 191–192). Nous avons donc présenté les principes selon lesquels les extracteurs de termes fonctionnent et le mécanisme spécifique appliqué par TermoStat pour extraire les termes, en proposant une partie du travail fait pour le projet *YourTerm TECH*.

4.2.5 Les problèmes les plus courants

Nous avons déjà clarifié que les extracteurs de termes peuvent avoir des limites, ils ne sont pas parfaits. Les listes de candidats termes qui sont générées par les extracteurs automatiques contiennent des séquences qui peuvent être des termes, mais il y a également des séquences qui n'intéressent pas les terminographes. Nous pouvons distinguer deux types de problèmes : le bruit et le silence. Le bruit représente les candidats d'une liste qui ne sont pas des termes, ils ne sont pas significatifs pour la terminologie. Le silence se réfère aux termes qui ne sont pas extraits d'un corpus. Pour mesurer le bruit et le silence, il existe deux méthodes d'évaluation : la précision et le rappel. La précision calcule le rapport entre les bons candidats dans la liste extraite et s'elle est élevée, il y aura peu de bruit. Le rappel calcule le rapport entre les bons termes extraits et les possibilités d'un texte et quand le rappel est haut, il y aura peu de silence. Lorsque nous parlons des difficultés des extracteurs automatiques de termes, il faut dire que l'analyse de leur performance n'est pas si facile à approfondir. Toutefois, il existe des méthodes pour l'évaluer. Une première méthode consiste à comparer la liste de candidats termes et

le contenu d'une banque de terminologie ou d'un dictionnaire spécialisé, en mesurant le bruit et le silence. Une autre technique consiste à la validation des termes par les terminographes, en mesurant le bruit, mais pas le silence. La dernière technique consiste à utiliser une liste de référence pour la comparer à la liste de candidats termes, même si nous ne savons pas si cette dernière liste fournit tous les termes d'un texte spécialisé ou non (cf. L'Homme 2020, 214–216).

Les limites qui caractérisent les extracteurs de termes peuvent être liées à leur conception ou à leurs indices. Au moment de la création d'un extracteur automatique de termes, il faut décider s'il localisera seulement les termes complexes ou s'il localisera soit les termes simples, soit les termes complexes. Dans le premier cas, l'extracteur écarte au premier abord les termes simples. Encore, les extracteurs de termes qui sont focalisés seulement sur la recherche des noms et des syntagmes nominaux ne prennent pas en considération les autres parties du discours. D'autres problèmes peuvent découler en relation aux indices utilisés : si l'extracteur applique l'indice de la fréquence, il ne sera pas capable de fournir aux terminographes les termes qui apparaissent une seule fois dans un corpus. Au contraire, si l'extracteur applique l'indice des patrons typiques, il risque d'insérer dans la liste de candidats termes des suites de mots graphiques qui ne sont pas des termes. Les concepteurs d'extracteurs de termes doivent donc faire attention au bruit et au silence. La solution la plus appropriée sera réduire le silence, même si le bruit augmente, car il sera plus facile à contrôler la liste de candidats termes, que rechercher des termes dans des corpus formés par nombreux textes spécialisés (cf. L'Homme 2020, 216–217).

Les complications liées aux extracteurs de termes dépendent aussi de la nature de la langue et de la forme des termes complexes. Cependant, il existe des solutions qui peuvent nous aider. Tout d'abord, certains extracteurs utilisent la technique du découpage du terme, quand ils rencontrent des syntagmes nominaux très longs. Parfois, les extracteurs fournissent simplement la liste de candidats termes à l'utilisateur qui pourra faire un filtrage, ou il découpe les candidats termes selon les informations obtenues sur les autres candidats. Une méthode pour étendre l'extraction au-delà des noms consiste à l'étendre aussi à plusieurs parties du discours, car nous avons vu que les termes peuvent être aussi des verbes, des adjectifs et des adverbes. Un des extracteurs qui compte aussi ces parties du discours est TermoStat. Un autre problème déjà nommé est lié à la

distinction entre les termes complexes et d'autres suites de mots qui n'ont pas un statut terminologique. Pour résoudre cette difficulté, les chercheurs ont créé des mesures statistiques qui permettent d'examiner le statut terminologique d'un candidat terme. Les extracteurs devraient savoir comment déterminer le lien d'un candidat terme avec un domaine de spécialité, autrement connu comme *termhood*, et la stabilité d'une entité linguistique, autrement dite *unithood*. Une autre technique pour distinguer les termes complexes et les autres suites de mots consiste à établir deux listes : une liste contiendra les candidats termes qui seront plus aptes à devenir des termes et l'autre liste contiendra tous les candidats termes. Les extracteurs de termes rencontrent bien d'autres problématiques, mais il n'y a pas toujours une solution (cf. L'Homme 2020, 217–223).

4.3 Le stockage de termes

Aujourd'hui les terminographes utilisent des supports électroniques pour gérer les termes. L'emploi de ces supports remonte aux années 1960 lorsque la quantité des termes et des données terminologiques a commencé à devenir difficile à examiner et traiter. Nous avons déjà nommé les banques de terminologie, mais il existe aussi des ressources spécialisées en ligne qui s'occupent de collecter les termes de nombreux domaines et elles permettent de consulter toutes sortes de données terminologiques (cf. L'Homme 2020, 256–257).

4.3.1 Les bases de données et les documents structurés

Un des supports informatiques les plus utilisés par les terminographes est la base de données : c'est un outil qui permet de collecter et organiser de l'information sur des objets logiquement reliés. Les bases de données s'appuient sur un système de gestion de base de données, généralement abrégé SGBD, qui est formé de programmes finalisés à créer les bases, les gérer, les mettre à jour et les consulter (cf. L'Homme 2020, 260). Les bases de données sont exploitées comme méthode de stockage, de gestion et de récupération de l'information. Les informations contenues dans les bases de données sont donc organisées afin d'être accessibles et bien gérées. Les données peuvent être organisées de manière différente, par exemple, en lignes, dans des colonnes ou tableaux et elles sont toutes indexées pour faciliter leur recherche. Pour garantir de bonnes informations, les données sont mises à jour de manière constante : elles sont complétées ou supprimées lorsque de nouvelles informations sont ajoutées (Oracle, s. d.).

Les bases de données peuvent contenir des agrégations d'enregistrements et des fichiers de données, par exemple, les transactions de vente, les inventaires de produits et les profils de clients. Pour ce qui concerne le stockage, les bases de données sont emmagasinées sous la forme d'un fichier ou d'un ensemble de fichiers. Les informations contenues dans ces fichiers sont divisées en enregistrements qui sont constitués d'un ou de plusieurs champs. Un champ représente une pièce d'information et chaque champ contient des informations qui concernent un aspect ou un attribut de l'objet ou de l'entité décrit par la base de données. En plus, les enregistrements sont organisés en tableaux afin de contenir des informations sur les relations entre les multiples champs présents. Ainsi, les utilisateurs, tels que les terminographes, peuvent facilement rechercher, réorganiser et sélectionner les champs (Oracle, s. d.).

Les bases de données représentent un instrument précieux. Toutefois, les terminographes utilisent de plus en plus les documents structurés. Ces documents sont réalisés à partir de langages de structuration de documents. Un des langages les plus utilisés aujourd'hui est l'*eXtensible Markup Language* (XML) qui dérive du langage *Standard Generalized Markup Language* (SGML). En parlant de termes, les bases de données et les documents structurés sont caractérisés par des points communs : les deux contiennent de nombreuses entrées ; les données d'un document structuré sont distinguées au moyen de balises qui seront équivalents aux champs ; les données peuvent être exploitées au moyen d'une extraction ou d'une recherche. Les balises peuvent être insérées par les terminographes ou par des logiciels appelés éditeurs. Voici un exemple de balise que nous avons déjà examiné : <s> et </s>. Les documents structurés présentent également des différences avec les bases de données : ces différences comprennent la longueur des champs, la présence ou l'absence de champs et la hiérarchisation de champs. Pour ce qui concerne la longueur des champs, la base de données est caractérisée par des champs ayant une taille valable qui est appliquée à l'entière structure. Cependant, les documents structurés peuvent avoir des champs caractérisés d'une longueur très variable. Puis, les bases de données ont une structure stable pour toutes ses entrées et si certains champs restent vides, ils restent aussi réservés. Au contraire, les documents structurés, qui ont toujours des champs réguliers, ont la possibilité d'omettre certaines catégories des données, sans porter préjudice à la cohérence de la structure. Enfin, la hiérarchisation de champs est une particularité des documents structurés. Cette capacité consiste à insérer

dans des champs d'autres champs qui nous appellerons champs subordonnés. La hiérarchisation de champs s'est révélée fondamentale pour d'exploitations précises, car les utilisateurs d'un document structuré peuvent localiser des informations spécifiques sur un terme (cf. L'Homme 2020, 264–267).

4.3.2 Structures et modèles des données

Toutes typologies de données sont réunies dans des catégories de données, qui sont à leur tour réunies dans des structures de données. Les structures de données ont des articulations très différentes et variables. Toutefois, les chercheurs ont essayé de concevoir des règles d'encodage uniforme pour gérer ces structures et pour garantir des échanges de données faciles et rapides. Malgré cela, la présence de disparités est inévitable. Une des premières structures créées sur support électronique s'appelle modèle plat. Ce modèle comprend plus de catégories de données qui seront régulièrement comblées et de catégories qui seront facultatives. La Banque de terminologie du Québec, par exemple, utilise le modèle plat pour encoder et gérer ses données terminologiques. Le modèle typique de cette Banque de terminologie comprend des champs tels que le numéro, la catégorie de données et la description, tandis que l'indication des sources se trouve à la fin de la structure terminologique. Le modèle plat n'est pas parfait : il ne distingue pas certaines catégories de données ; il demande un nombre de champs suffisant ; et il pose des limites, par exemple, il ne sera pas capable d'accueillir plusieurs contextes pour un terme donné (cf. L'Homme 2020, 269–273).

Un autre modèle est le modèle relationnel qui présente des avantages par rapport au modèle plat. Le modèle relationnel fait une division de données en les organisant dans des structures distinctes. Si nous avons des données de nature différente, il faudra les analyser séparément afin de les mieux décrire. Le modèle relationnel est la base sur laquelle s'appuient les bases de données relationnelles qui permettent de regrouper différentes structures à partir d'une relation formelle. Par exemple, si nous avons des données terminologiques et des données bibliographiques à examiner, nous les diviserons dans deux structures : une pour les champs terminologiques et une pour les champs bibliographiques. Leur relation sera représentée par un champ où nous trouvons de l'information commune, par exemple, un champ {source} pour la structure terminologique et un champ {code} pour la structure bibliographique : le champ {source} sera lié au champ {code}. De la même manière, ce modèle s'applique pour les termes

(données linguistiques) et les contextes (données pragmatiques) : nous distinguerons une structure dédiée aux termes et une structure dédiée aux contextes dans lesquels les termes sont utilisés. Un dernier emploi du modèle relationnel est représenté par la distinction des données conceptuelles, c'est-à-dire les termes, et des données linguistiques, c'est-à-dire les informations des termes. C'est une méthode pratique surtout pour les bases de données multilingues (cf. L'Homme 2020, 273–279).

Le dernier modèle que nous présenterons est le modèle hiérarchique. Ce modèle se base sur le principe de l'héritage selon lequel les propriétés spécifiques d'un nœud seront automatiquement attribuées à un nœud fille. Cette technique est utile pour économiser les descriptions de données. Voici un exemple pour comprendre ce principe : dans la classification des insectes, le terme *insecte* présente des propriétés qui seront valables également pour « coléoptère » et « diptère ». Ce modèle est également utilisé pour définir la hiérarchie à héritage multiple, c'est-à-dire qu'une structure présente des nœuds génériques qui se rattachent à d'autres en passant des propriétés à chaque nœud, par exemple, les deux nœuds « chat » et « perroquet » sont des nœuds filles du nœud « animal domestique », mais ils sont aussi deux nœuds génériques. Toutefois, ce modèle doit tenir compte que dans certains cas les nœuds filles ne peuvent pas hériter l'ensemble des propriétés rattachées aux nœuds mères. Il faut donc bloquer l'héritage d'une propriété. Voici un exemple : le nœud « oiseau » ne peut pas attribuer sa propriété « voler » au nœud « émeu ». Il existe aussi des modèles hiérarchiques plus articulés et complexes afin de représenter une grande variété de relations : nous parlons des bases de connaissances terminologiques et des ontologies. La base de connaissance terminologique est une simple base de données terminologique, mais qui comprend les relations qui s'établissent entre les concepts d'un domaine de spécialité. L'ontologie est similaire à la base de connaissance, mais elle va plus loin. L'ontologie est utilisée dans des applications terminologiques pour examiner la représentation explicite et formelle d'une conceptualisation, c'est-à-dire que l'ontologie est construite afin de représenter les connaissances du monde réel selon les conceptualisations. En particulier, l'ontologie s'occupe de la représentation des conceptualisations consensuelles, c'est-à-dire une conceptualisation qui est le résultat d'un consensus entre experts (cf. L'Homme 2020, 280–288).

Pour conclure cette partie dédiée au stockage des termes, nous verrons les logiciels de terminologie. Ces logiciels possèdent des fonctionnalités des systèmes de gestion de bases de données et des documents structurés. Les terminographes qui utilisent les logiciels de terminologie s'occupent de la saisie des données et de leur organisation. Les logiciels de terminologie commerciaux existants ont leurs propres particularités et ils se basent sur des structures préconstruites. En plus, même les organismes publics peuvent créer un service de terminologie en définissant sa gestion et organisation : nous parlons des banques de terminologie. Dans ce cas-là, les terminographes devront enrichir la banque en suivant un protocole établi. La banque de terminologie la plus connue en Europe est le IATE (cf. L'Homme 2020, 292–293). C'est la base de données terminologique de l'UE, utilisée par les institutions et agences européennes depuis 2004. Le projet IATE a le but de fournir une infrastructure web pour toutes les ressources terminologiques de l'UE, améliorant ainsi la disponibilité et la normalisation des informations (iate, 2024)[17].

4.4 FAIRterm et les fiches terminologiques

Dans cette partie finale du chapitre quatre, nous présenterons un outil qui était fondamental pour la rédaction de ce mémoire et pour la réalisation du projet *YourTerm TECH* : FAIRterm. C'est un outil gratuit pour la compilation en ligne de fiches terminologiques multilingues[18]. FAIRterm a l'objectif de fournir et garantir un instrument pour l'organisation optimale des données terminologiques. Il suit des principes « *fair* » établis par l'association *European Open Science Cloud* (EOSC)[19] et se base sur les dernières normes ISO TC/37 SC/3[20] concernant la gestion terminologique afin d'assurer la trouvabilité, l'accessibilité, l'interopérabilité et la réutilisation de la terminologie[21]. L'outil FAIRterm est une initiative développée par le directeur de ce mémoire, Federica Vezzani, professeur assistant en terminologie au département d'études linguistiques de notre université, et par le co-directeur de ce mémoire, Giorgio Maria Di Nunzio,

[17] Pour approfondir, voici le lien vers le site web : https://iate.europa.eu/home.
[18] Pour visiter le site web, voici le lien : https://shiny.dei.unipd.it/fairterm/.
[19] En français, nous l'appellerons Nuage européen pour la science ouverte. Son ambition est de fournir aux chercheurs, innovateurs, entreprises et citoyens de l'UE un environnement pluridisciplinaire fédéré et ouvert dans lequel ils ont la possibilité de publier, trouver et réutiliser des données, des outils et des services à des fins de recherche, d'innovation et d'éducation (European Commission, s. d.).
[20] Pour approfondir, voici le lien vers les normes : https://www.iso.org/fr/committee/48136/x/catalogue/.
[21] Les lettres qui forment le mot anglais *FAIR* correspondent aux termes anglais *Findability*, *Accessibility*, *Interoperability* et *Reusability*

professeur en ingénierie informatique au département d'ingénierie de l'information de notre université (FAIRterm, s. d.). Actuellement FAIRterm donne la possibilité de travailler avec les 24 langues officielles européennes, plus turc, russe, chinois, japonais et coréen et bientôt il sera également possible de travailler avec l'arabe et le géorgien. La ressource FAIRterm propose deux modes de travail : individuel et collaboratif. Dans le premier cas, les étudiants ou les chercheurs qui souhaitent utiliser l'application peuvent accéder à un espace privé afin de remplir des fiches terminologiques et ne consulter que leur travail. Dans le cas d'un travail collaboratif, les utilisateurs peuvent accéder à un espace collaboratif pour remplir des fiches terminologiques et aussi consulter et modifier les fiches de collègues. L'alternative collaborative est très utile au moment où nous rejoignons des projets *YourTerm* (par exemple, le projet *YourTerm TECH* dont nous parlons dans ce mémoire) avec un groupe de participants travaillant sur le même projet (cf. Vezzani/Di Nunzio, 2022).

4.4.1 La fiche terminologique

Nous avons vu que le travail terminographique se compose des différentes étapes et une de ces étapes est la compilation de fiches terminologiques à la fin d'une recherche et de la construction d'un corpus sur un précis domaine de spécialité. Une fiche terminologique est un « Support sur lequel sont consignées, selon un protocole établi, les données terminologiques relatives à une notion » (Office québécois de la langue française, 2005). C'est un document qui contient des informations concernant un terme et aussi sa traduction, facilement accessible et repérable et qui regroupe plusieurs champs contenant les renseignements relatifs à un concept spécialisé.

La fiche terminologique est donc composée par des champs qui ont l'objectif de fournir une bonne compréhension des termes et de nous aider à utiliser les termes correctement. Les champs d'une fiche terminologique réunissent les informations principales sur les concepts d'un domaine et ses termes. Les champs peuvent être groupés en deux séries : les champs concernant les informations sur le concept et les champs concernant les informations sur le terme. La partie des champs dédiée au terme comprend tout d'abord la désignation du concept, puis les informations sur la valeur grammaticale du terme, l'étymologie et l'équivalent dans une ou plusieurs langues (si nous parlons de fiches terminologiques multilingues). En plus, ces champs peuvent contenir des informations sur les multiples formes du terme, ses combinaisons et ses relations avec les

autres termes. Tandis que les champs dédiés au concept indiquent des informations sur le domaine et le sous-domaine, la définition, la représentation du concept et les relations entre d'autres concepts (cf. Pitar 2011, 71–72).

Il existe plusieurs modèles de fiches terminologiques. Une fiche terminologique peut être monolingue, bilingue et multilingue. La fiche terminologique monolingue intéressera seulement un terme dans une seule langue. Voici l'exemple de la fiche terminologique pour le terme *corpus* tiré du Grand Dictionnaire Terminologique :

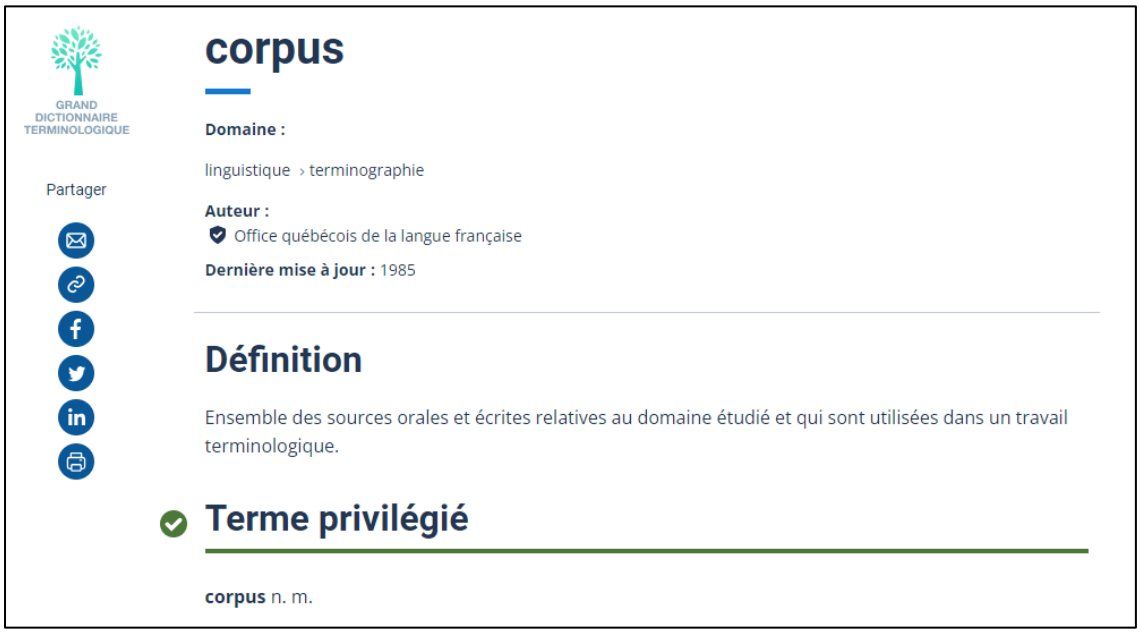


Figure 5 – Exemple d'une fiche terminologique monolingue

Nous pouvons voir que cette fiche comprend tout d'abord le terme d'entrée, le domaine et le sous-domaine, l'auteur de la fiche, l'année de la dernière mise à jour et la définition. La fiche terminologique bilingue ou multilingue présentera également la traduction d'un terme dans une autre langue d'étude (ou dans plusieurs langues d'étude). Voici un exemple tiré de la banque de terminologie européenne IATE, il s'agit de la fiche terminologique bilingue du terme français *apprentissage automatique* et son correspondant anglais *machine learning* :

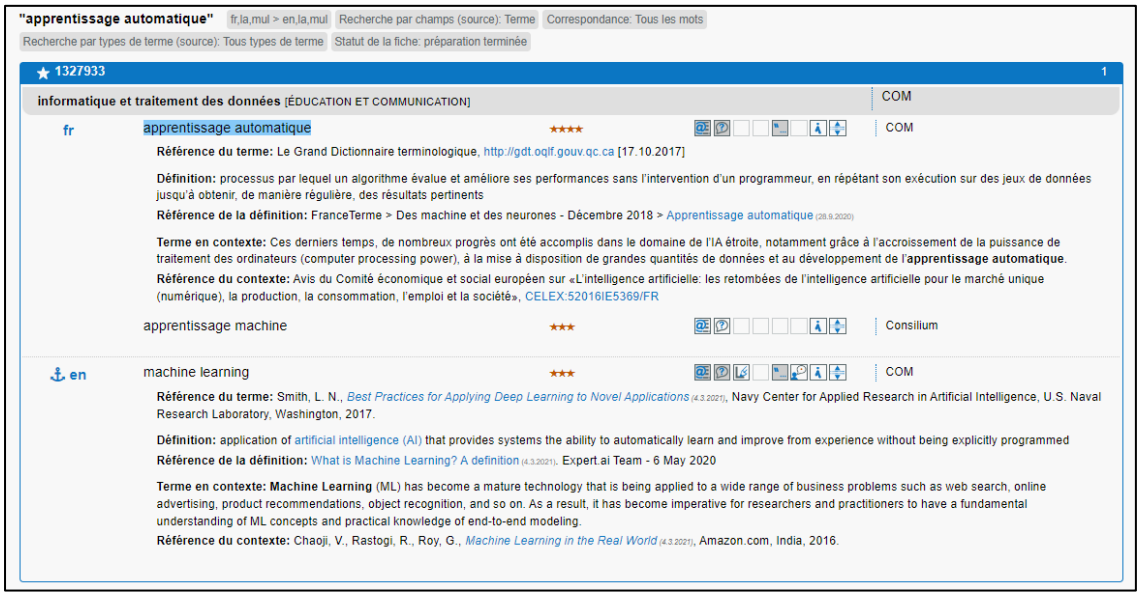


Figure 6 – Exemple d'une fiche terminologique bilingue

À la différence de la fiche terminologique monolingue, cette fiche nous fournit la traduction du terme et toutes les informations reliées au terme dans les deux langues.

4.4.2 La compilation de fiches terminologiques avec FAIRterm

Pour compiler une fiche terminologique sur FAIRterm, nous devons tout d'abord nous authentifier avec nos identifiants personnels. Pour commencer à créer une fiche terminologique, les utilisateurs doivent sélectionner la langue source du terme objet d'étude, le saisir dans la boîte dédiée *Term* et l'ajouter à la base de données (cf. Vezzani 2021, 54). Voici la fenêtre des premiers pas à exécuter pour compiler une fiche terminologique :

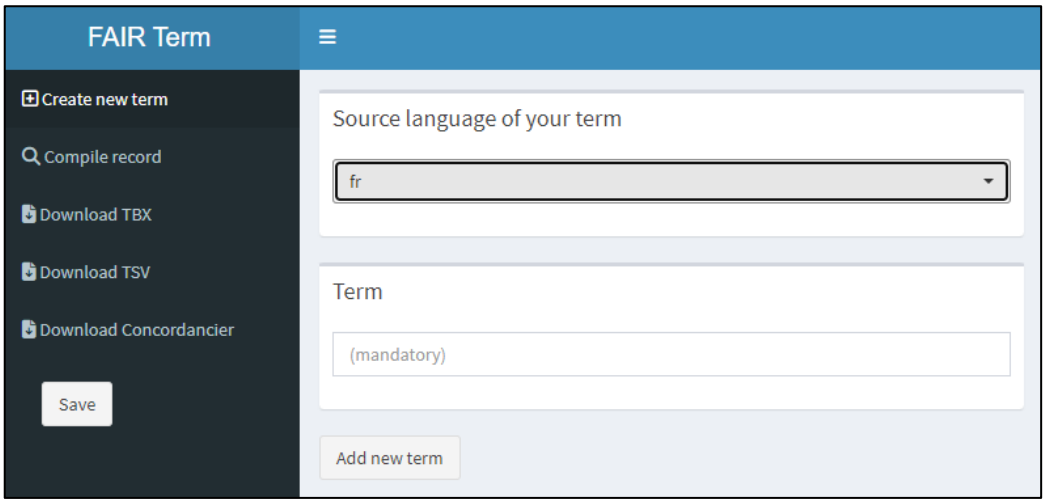


Figure 7 – L'ajout d'un terme sur FAIRterm

Pour passer à la phase de compilation du terme, les utilisateurs doivent cliquer sur la fonction *Compile record*. Dans cette section, FAIRterm demande la sélection de deux langues de travail et permet le choix d'un terme saisi dans la phase initiale de compilation. Ainsi, chaque utilisateur disposera d'une fiche terminologique bilingue qui est caractérisée par la mise en miroir des mêmes catégories de données terminologiques pour le terme source à gauche de l'écran et pour le terme cible à droite de l'écran (cf. Vezzani 2021, 54). Voici la fenêtre de la fonction *Compile record* pour la fiche terminologique du terme « base de connaissance » tiré du projet *YourTerm TECH* :

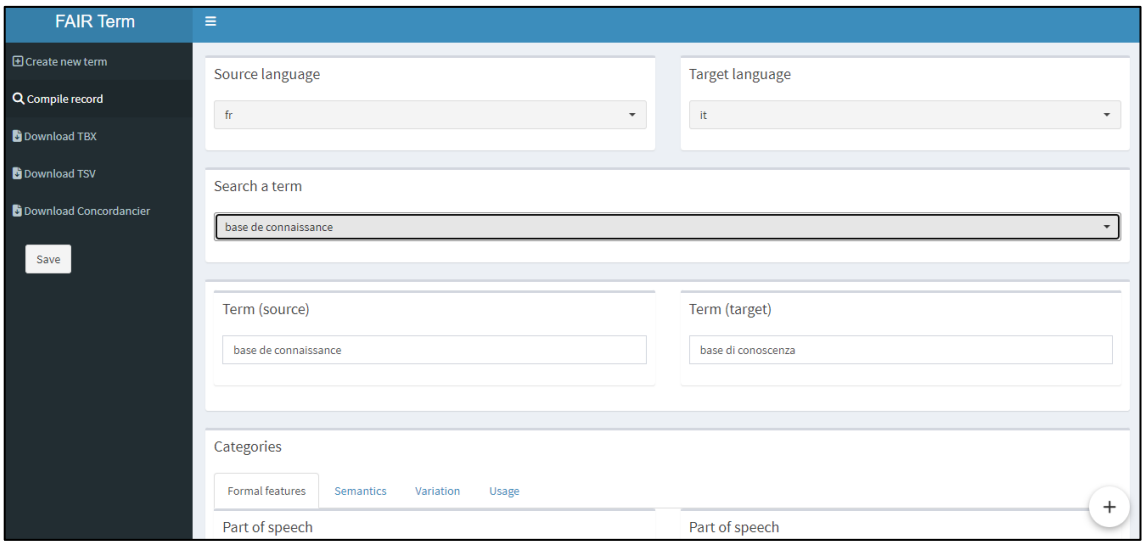


Figure 8 – Fiche terminologique du terme « base de connaissance »

Dans la figure ci-dessus, nous pouvons voir les différentes catégories des informations linguistiques à compiler. Elles sont regroupées en quatre typologies : les caractéristiques formelles, la sémantique, la variation et l'usage.

Tout d'abord, il y a l'entrée, c'est-à-dire le terme faisant l'objet de la description. Puis, dans les caractéristiques formelles, les utilisateurs doivent compiler la partie relative à l'information grammaticale : ils/elles doivent se concentrer sur la partie du discours (nom, verbe, adjectif, adverbe), le genre (masculin ou féminin), le nombre (singulier ou pluriel), la prononciation (en reportant la transcription phonétique du terme), l'étymologie, les dérivés et les notes relatives au terme. Ces champs intéressent particulièrement le terme de nature nominale, mais il est possible d'avoir un verbe comme entrée, comme nous avons vu. Le champ dédié à l'étymologie, à savoir la science qui s'occupe de la recherche de l'origine des mots, présente des problèmes surtout quand les

utilisateurs se trouvent à examiner un terme complexe : dans la majorité de cas, les termes complexes n'ont une étymologie propre, mais ces sont les mots graphiques qui composent les termes complexes qui ont chacun une étymologie. La même chose s'applique au champ de dérivés, car il est difficile de trouver les dérivés d'un terme complexe lorsqu'il peut être lui-même un dérivé. Enfin, les notes relatives au terme servent à ceux qui utiliseront la fiche terminologique dans le cas où le terme objet d'étude possède des caractéristiques particulières nécessaires pour sa complète compréhension.

Dans la section dédiée à la sémantique, les terminographes trouvent comme premier champ à compiler la définition. Elle est l'explication du sens du terme dans le domaine de spécialité et c'est une des parties les plus importantes pour la compréhension du terme, car elle donne une notion ou une image mentale exacte et permet de la différencier d'autres notions. De toute évidence, il est nécessaire d'indiquer la source de laquelle la définition a été prise : les deux boîtes *External cross-ref (definition)* et *Source (definition)*, les utilisateurs de l'outil devront insérer leurs sources et les liens où ils/elles ont récupéré la définition (s'il s'agit des sources en ligne). Toujours lié à la définition du terme, il y a le champ *Notes (definition)* où les terminographes peuvent insérer des informations complémentaires concernant la définition. Dans le cas où la définition n'existe pas ou n'est pas claire ou trouvable, les terminographes la créeront eux-mêmes. Il faudra donc spécifier que la définition est le fruit de la pensée des terminographes, en l'indiquant sur ce champ. Puis, il y a l'analyse sémique qui consiste à décomposer le sens d'un terme dans de petits éléments de sens qui en donnent le sens total. Ces petits éléments sont des atomes de sens et ils s'appellent composants sémantiques. Faisons-nous un exemple pour mieux comprendre : les composants sémantiques du terme « bistouri » seront /outil/ /chirurgie/ /forme de couteau/ /lame/ /incision/. Les sèmes peuvent être de deux types : les sèmes constants et les sèmes contextuels. Les sèmes constants renvoient au sens dénotatif du terme et les autres au sens connotatif du terme. La dénotation correspond au sens objectif d'un terme et la connotation correspond au sens qu'un terme peut prendre en fonction d'éléments sociaux et culturels. Les sèmes constants sont également divisés en sèmes génériques et sèmes spécifiques, tandis que les sèmes contextuels comprennent les sèmes connotatifs qui ne sont pas présents pour tous les termes. Voici un exemple : pour le terme *marguerite*, /fleur/ est un sème générique et /pétales blancs/ est un sème spécifique. Pour faire une bonne analyse sémique, nous partirons de la définition du

terme. Après l'analyse sémique, les terminographes trouvent les champs dédiés aux synonymes, quasi-synonymes, hyperonymes, hyponymes, méronymes et holonymes. Les synonymes sont des termes que l'on peut substituer l'un à l'autre dans un énoncé sans en changer le sens. Trouver les synonymes n'est pas si facile, car certaines fois les termes n'ont pas un synonyme, mais ils ont plutôt un quasi-synonyme : c'est un terme qui n'est pas un synonyme exact, mais il est également employé pour substituer un autre. Voici un exemple tiré du projet *YourTerm TECH* : les synonymes du terme *mégadonnées* sont « big data » ou « données massives » et il n'y a pas de quasi-synonymes ; le synonyme du terme *modèle computationnel* est « dispositif de calcul » et les quasi-synonymes sont « modèle logique », « modèle mathématique » et « modèle statistique ». Puis, il y a les hyperonymes et les hyponymes : c'est une relation qui existe entre un élément sous-ordonné (hyponyme) et un élément superordonné (hyperonyme). Voici un exemple du projet *YourTerm TECH* : le terme *modèles de plongement prédictif de mots* a comme hyperonyme « techniques de traitement automatique du langage » et son hyponyme est « modèle *skip-gram* » qui est une typologie de modèle de plongement prédictif de mots. Pour conclure la partie réservée à la sémantique, les utilisateurs doivent aussi s'occuper des méronymes et des holonymes : un méronyme représente la partie d'un tout et un holonyme est le mot englobant. Un exemple de cette relation sémantique est représenté par le terme *chaussure* qui est l'holonyme de « lacet », « semelle » et « talon » qui sont les méronymes.

Dans la partie dédiée à la variation, les utilisateurs doivent compiler les champs du nom commun et du nom scientifique, des variantes orthographiques, des acronymes, des formes complètes (si l'entrée est un sigle) et des abréviations. Le nom commun représente l'équivalent d'un terme scientifique utilisé au niveau international qui est utilisé dans la langue générale, tandis que le nom scientifique sera l'équivalent scientifique utilisé dans un contexte spécialisé. L'exemple typique qui est fait pour comprendre ces deux champs est le terme *fièvre* qui est le nom commun utilisé pour indiquer le terme spécialisé et nom scientifique *pyrexie*. Les variantes orthographiques concernent des changements minimaux d'un mot, tels que l'ajout d'un signe diacritique. Un exemple tiré du projet *YourTerm TECH* est le terme *interaction homme-machine* dont sa variante orthographique est « interaction homme machine ». Puis, il y a la boîte pour les acronymes, par exemple, l'acronyme du terme *intelligence artificielle* est « IA ». Ensuite,

les utilisateurs trouveront la boîte pour les formes complètes dans le cas où le terme en entrée sera un sigle ou un acronyme et, pour conclure, la section de la variation, il y a les abréviations.

La dernière section d'une fiche terminologique de FAIRterm est réservée à l'usage d'un terme. Dans ce panneau, les terminographes trouvent le domaine, le sous-domaine, le registre, le contexte, les boîtes pour les sources et les collocations. Le domaine et le sous-domaine sont utiles pour comprendre l'environnement linguistique dont les termes font partie et pour savoir quand et comment les termes peuvent être employés. Puis, dans le champ dédié au registre, les utilisateurs devront indiquer si le terme en entrée appartient à un registre spécialisé ou non. Il faut également insérer un contexte : cela sert pour donner un exemple d'une phrase contenant le terme objet d'étude, suivi de la source où les terminographes l'ont trouvé (*External cross-ref* et *Source*). En conclusion, il y a le champ dédié aux collocations, c'est-à-dire une sorte de rapprochement fréquent de deux unités lexicales, par exemple, le terme *algorithme* tiré du projet *YourTerm TECH* est partie de la collocation « écrire un algorithme ».

Dans ce chapitre nous avons donc analysé en détail tout ce qui concerne les termes. Nous avons présenté le système de formation de termes, les différentes typologies de termes, telles que les termes simples et les termes complexes, et nous avons examiné leur extraction, en présentant les techniques pour l'extraction, les difficultés et le logiciel TermoStat que nous avons utilisé pour le projet terminologique. Il était aussi nécessaire de discuter du stockage de termes et de leur encodage et pour conclure, nous avons traité l'outil que nous avons utilisé pour la compilation des fiches terminologiques du projet *YourTerm TECH* : FAIRterm. Nous avons donc présenté ses principes, comment il fonctionne et les catégories des informations linguistiques à compiler pour l'étude d'un terme.

CHAPITRE 5

Analyse qualitative du projet terminographique et considérations finales

Dans ce dernier chapitre, nous présenterons en détail le travail terminographique pour le projet *YourTerm TECH*, en analysant les trois corpus, certains termes des fiches terminologiques compilées grâce à FAIRterm, et nous ferons une analyse totale et qualitative du travail et des outils utilisés.

5.1 Analyse des corpus du projet

Nous avons déjà anticipé dans l'introduction de ce mémoire le cadre du projet *YourTerm TECH*. Il s'agit d'un projet européen qui s'occupe de la diffusion de la terminologie de différents domaines et de son accès.

Tout d'abord, nous avons fait attention à la construction de trois corpus que nous avons utilisés afin d'extraire des termes pertinents pour le projet. Les corpus sont construits en trois différentes langues : français, italien et anglais. Elles sont les langues d'étude choisies pour la réalisation du projet. En général, les documents que nous avons choisis pour les corpus ont été pris de bibliothèques numériques comme JSTOR, de sites web de sociétés, qui s'occupent de la science des données, de la transformation digitale et de l'IA, et de sites web d'organismes de formation.

5.1.1 Le corpus en français

Le corpus français se compose de 16 documents : c'est un groupe hétérogène de documents qui traitent le domaine de l'intelligence artificielle et du traitement automatique du langage. Les documents choisis font partie d'un long processus de sélection et de décision (voir l'annexe E).

Six documents du corpus ont été extraits du site web JSTOR, une bibliothèque numérique en ligne, que nous avons déjà nommé, où il est possible de trouver de nombreux articles de revues, de livres, d'images et de sources primaires dans 75 champs d'études. JSTOR fournit beaucoup de contenus scientifiques en collaboration avec la

communauté universitaire afin d'aider la connexion entre les étudiants et les enseignants, en réduisant les coûts et en augmentant l'espace de stockage (JSTOR, s.d.).

Parmi les six documents trouvés dans JSTOR, trois sont tirés de la revue scientifique *Langages*[1]. C'est une revue qui met à disposition d'une communauté scientifique pluridisciplinaire les résultats des multiples recherches contemporaines, nationales et internationales. *Langages* s'occupe des domaines qui couvrent les sciences du langage, y compris toutes les disciplines connexes, telles que la syntaxe, la sémiotique, la psycholinguistique, le traitement automatique du langage et la traduction, et même d'autres disciplines dans un cadre interdisciplinaire, par exemple, la médecine, la sociologie et le droit. Chaque année, cette revue publie quatre volumes qui sont gérés par un coordinateur qui s'intéresse à solliciter des experts du thème traité en devenant contributeurs. Les volumes édités peuvent se présenter sous plusieurs formes, par exemple, sous forme d'actes de colloques ou de journées d'étude, de travaux fédérés par des projets de recherche, de présentation de ressources tels qu'outils et corpus et de leurs applications linguistiques ; encore, sous forme de bilans disciplinaires ou sous-disciplinaires ; et sous forme d'appel à des contributions thématiques. Puis, les volumes sont soumis à une double expertise : ils sont revisionnés et agréés par un comité scientifique international et multidisciplinaire et ils seront examinés par des spécialistes de la thématique abordée français et étrangers (Armand Colin Revues, s. d.). Cette revue fait partie du panier des revues sous la direction de l'éditeur Armand Colin, signature majeure de l'édition universitaire francophone en lettres, histoire, sciences humaines et sociales, créée en 1870, avec un catalogue de plus de 2000 titres et 15 revues scientifiques. Les publications signées par Armand Colin sont caractérisées d'un haut niveau d'exigence scientifique qui les destine à être des outils de travail incontournables pour les chercheurs, les enseignants et les étudiants (Armand Colin Revues, s. d.). Cette revue nous a fourni trois documents qui se prêtent bien à l'extraction de termes liés à l'IA et au TAL : un document traite ensemble la terminologie, l'ingénierie linguistique et la gestion de l'information ; un autre traite les méthodes de traitement automatique fondées sur les corpus ; et le dernier s'occupe de la linguistique et de l'intelligence artificielle.

[1] Voici le lien pour visiter le site de la revue *Langages* : https://www.revues.armand-colin.com/lettres-langue/langages.

Une autre revue que nous avons prise en considération pour la construction du corpus français est la revue *Cités*[2]. Il s'agit d'une revue scientifique trimestrielle qui s'occupe des grandes transformations des sociétés actuelles d'un point de vue philosophique. Depuis l'année 2000, an de sa naissance, cette revue se concentre sur les réelles demandes du public. Pour faire cela, *Cités* analyse attentivement le monde contemporain, en fournissant des contenus qui sont positionnés entre l'actualité sociopolitique et la réflexion philosophique. Cette revue est particulièrement influencée par une analyse de thèmes qui se base aussi sur le retour au réel et sur l'ouverture au possible. Elle s'engage donc à décrire et étudier les faits sociaux, politiques ou économiques dans toute leur épaisseur, en faisant aussi attention à la dimension critique. *Cités* présente une structure précise : elle est formée par un dossier sur un thème principal et par plusieurs rubriques régulières, par exemple, débat, lexique politique, profils de penseurs politiques contemporains, portraits de villes ou de pays étrangers et recensions. En termes concrets, cette revue s'intéresse aux concepts fondamentaux qui sont affectés par les évolutions contemporaines, notamment, le pouvoir, l'État, l'autorité, le droit, la connaissance, l'identité, l'art et la guerre. En plus, elle s'intéresse également aux thèmes liés à l'actualité culturelle et éditoriale, par exemple, l'impact du développement technologique et ses conséquences, la crise de l'autorité, le rôle des intellectuels, le féminisme, la démocratie et les religions, et encore le corps et la sexualité, l'art contemporain, les figures actuelles de l'Europe, le nouvel ordre international et les transformations apportées par la guerre (puf, s. d.). Dans *Cités*, nous avons trouvé un document qui s'interroge sur l'éthique de l'intelligence artificielle.

La *Revue d'économie financière* (REF) est une autre source que nous avons consultée[3]. Elle a été fondée en 1987 et elle publie quatre numéros chaque année. Cette revue aspire à se confirmer l'instrument de dialogue entre les universitaires, les chercheurs et les professionnels de la finance et de la banque. En plus, elle vise à animer la discussion et la réflexion qui sont fondamentales pour la cohésion de la profession financière et bancaire. La REF s'occupe des sujets concernant l'économie financière, par exemple, l'économie bancaire, la finance d'entreprise, l'histoire financière, la politique monétaire, la finance internationale et bien d'autres. Les numéros publiés sont composés

[2] Voici le lien pour visiter le site de la revue *Cités* : https://www.puf.com/cites.
[3] Voici le lien pour visiter le site de la revue *Revue d'économie financière* : https://www.aefr.eu/fr/numeros?page=1.

des dossiers thématiques et d'une rubrique intitulée *Articles divers*. Les dossiers (ou le dossier) sont gérés par deux directeurs choisis par le comité de rédaction pour leurs compétences sur le sujet traité. Les directeurs choisissent le contenu du dossier thématique et ils s'occupent de contacter les auteurs et puis ils supervisent la qualité scientifique des articles. La rubrique *Articles divers* se compose d'articles non sollicités, acceptés pour publication. Cependant, ils doivent respecter la ligne éditoriale de la revue et les exigences scientifiques communément acceptées. Le monde de la finance est désormais largement internationalisé et la *Revue d'économie financière* s'attache donc à aborder des thèmes internationaux et à favoriser la participation d'auteurs étrangers. En plus, elle publie également en langue anglaise (Association Europe-Finances-Régulations, s. d.). Cette revue nous a fourni un document très intéressant sur l'économie de l'intelligence artificielle.

Une autre revue partie du panier de l'éditeur Armand Colin que nous avons utilisé est la revue *Langue française*. Elle a été créée en 1969 chez Larousse, une maison d'édition française spécialisée dans les ouvrages de référence, comme les dictionnaires. Cette revue a l'objectif de promouvoir les recherches en linguistique française, c'est-à-dire qu'elle veut étudier la linguistique qui s'occupe de la langue française dans sa totalité, en prenant en considération la diversité culturelle, sociale et géographique. En plus, la revue examine également la linguistique française par rapport aux questions de son acquisition, de son apprentissage et de son enseignement en tant que langue maternelle comme en tant que langue étrangère. *Langue française* est une revue trimestrielle et thématique. Ses numéros s'occupent des recherches théoriques et descriptives, en comprenant les faits de langue et de discours, tels que la phonétique, la phonologie, la sémantique, la syntaxique et la morphologique. Les contributions aux recherches se concentrent sur les problématiques contemporaines, théoriques et empiriques. C'est donc une revue qui s'adresse aux enseignants, aux étudiants et aux chercheurs intéressés par l'étude de la langue française dans tous ses aspects (Armand Colin Revues, s. d.). Le document que nous avons trouvé dans cette revue présente les outils informatiques qui peuvent être utilisés par les linguistes.

Tous les documents précédents sont tirés de la bibliothèque numérique JSTOR. Toutefois, nous avons aussi utilisé le moteur de recherche Google, en cherchant des documents qui traitent l'intelligence artificielle et le traitement automatique du langage.

Nous avons trouvé un rapport de recherche du Centre Génie Industriel et Informatique de l'École nationale supérieure des Mines de Saint-Etienne. C'est un des premiers cas d'organismes de formation que nous consultons pour construire nos corpus. L'École des mines de Saint-Étienne fait partie l'Institut Mines-Télécom et elle représente une des plus prestigieuses écoles d'ingénieurs de l'hexagone. Son objectif principal est la formation d'ingénieurs généralistes et de spécialités de haut niveau. En plus, l'École joue également un rôle dominant dans l'accompagnement des entreprises à la transition industrielle (École des Mines de Saint-Étienne, s. d.). Pour ce qui concerne la recherche, l'École des Mines de Saint-Etienne s'engage dans la recherche scientifique de haut niveau afin de relayer la politique économique française et d'accélérer le développement industriel durable. Les publications de l'École sont aussi reconnues par la communauté scientifique internationale (École des Mines de Saint-Étienne, s. d.). Le rapport de recherche que nous avons trouvé traite le traitement automatique du langage naturel appliqué à l'extraction et à la recherche d'informations.

Un autre document qui nous a bien servi est celui du chercheur François Yvon. C'est un chercheur qui travaille principalement sur le traitement du langage naturel et sur la traduction automatique à l'aide de méthodes neuronales et probabilistes. Actuellement, il est affilié au groupe Machine Learning and Deep Learning for Intelligent Access de l'ISIR, l'Institut des Systèmes Intelligents et de Robotique de Paris, un institut qui est sous la double tutelle de Sorbonne Université et du Centre National de la Recherche Scientifique (CNRS) qui est un centre de recherche parmi les plus importantes au monde (Institut Systèmes Intelligents et de Robotique, s. d.). Yvon a été directeur général du CNRS à Orsay et aussi professeur d'informatique à l'Université Paris-Sud et à Télécom Paris (Institut Systèmes Intelligents et de Robotique, s. d.). Yvon a écrit un intéressant document sur le traitement automatique du langage que nous a permis de l'inclure dans notre corpus français.

Ensuite nous avons tous les sites web consultés où nous avons trouvé nombreux de contenus qui nous ont aidés à construire notre corpus. Tout d'abord nous avons le site web Journal du Net : c'est un site créé en 1999, il est édité par CCM Benchmark Group[4]

[4] La société CCM Benchmark fait partie du groupe Figaro, elle compte plus de 34 millions d'internautes. Son objectif est d'accompagner les internautes dans leur vie quotidienne en leur fournissant des conseils. En plus du Journal du Net, CCM Benchmark publie Linternaute.com, Le Journal des Femmes, Comment ça marche et Droit-finances.net (JournalduNet.com, s. d.).

et il est formé par une équipe professionnelle de plus de 14 rédacteurs. L'objectif principal du site Journal du Net est d'informer ses lecteurs en toute indépendance en se concentrant sur une haute qualité et sur la fiabilité. Il s'occupe de vulgariser des contenus qui traitent les thèmes du Cloud[5], du Web3[6], de l'intelligence artificielle, de l'IoT[7] et de la cybersécurité. Dans ce site, nous pouvons trouver gratuitement des articles, analyses, interviews qui sont réalisés par la rédaction en toute indépendance d'annonceurs et de leurs partenaires économiques ; des chroniques d'expert permettant aux professionnels de prendre la parole sur les sujets d'intérêt au sein de leur communauté professionnelle ; une plate-forme de plusieurs de vidéos ; et des contenus encyclopédiques, comme un dictionnaire des mots liés à l'e-business et des chiffres clés de l'Internet (JournalduNet.com, s. d.). Ce site nous a fourni un article qui parle du TAL et de ses techniques.

Un autre site web que nous avons consulté est le site Stat4decision, une entreprise qui s'occupe de la science des données. Elle est caractérisée par une équipe dynamique qui s'occupe de l'intelligence artificielle et de la data science et qui est composée par des data scientists, data engineers, développeurs et formateurs. Ce site offre trois principaux services : la formation, le conseil et le développement. Son objectif consister à accompagner ses clients dans leurs projets data. Pour faire cela, Stat4decision dispose d'un logiciel *open source* où l'équipe combine des connaissances techniques et théoriques avec des compétences affirmées. Le site s'intéresse au transfert de compétences pour toutes les missions effectuées et il propose également la gestion et le suivi à distance des travaux de récurrence pour tous les projets qu'il prend en charge. Dans ce site nous avons trouvé des informations sur le traitement automatique du langage en français (Stat4decision, s. d.).

Une autre source que nous avons utilisée pour la construction du corpus est *Interstices* : c'est une revue de culture scientifique en ligne. Elle a été créée par des chercheurs qui souhaitent diffuser les sciences du numérique. Cette revue en ligne a été

[5] Cela se réfère au *Cloud Computing* qui est une technologie permettant d'utiliser les ressources de serveurs informatiques à distance, via internet (Datascientest.com, 2022).

[6] « Le Web3 est un terme générique qui désigne des technologies, telles que la blockchain, qui décentralisent la propriété et le contrôle des données sur Internet » (Amazon Web Services, s. d.).

[7] « Le terme IoT, ou internet des objets, désigne le réseau collectif d'appareils connectés et la technologie qui facilite la communication entre les appareils et le cloud, ainsi qu'entre les appareils eux-mêmes » (Amazon Web Services, s. d.).

lancée en 2004 et elle est publiée par Inria, c'est-à-dire l'Institut national de recherche en sciences et technologies du numérique. Le suivi scientifique de cette revue est assuré par le comité éditorial, qui comprend des personnes d'Inria, du CNRS, de plusieurs universités et des associations professionnelles du domaine. *Interstices* s'occupe d'un groupe de domaines très vastes : elle s'intéresse à l'environnement, aux algorithmes, à la modélisation et à la simulation, à l'histoire du numérique, à la médecine, à l'interaction humain-machine, à l'intelligence artificielle, aux langages, au traitement d'images et son, à la robotique et aux données (interstices.info, s. d.). Dans cette revue nous avons trouvé des contenus relatifs à la traduction automatique statistique, un thème qui bien est lié au TAL.

Puis, nous avons consulté le site web DataScientest. C'est un organisme de formation spécialisé dans les métiers de la technologie qui travaille en collaboration avec des centaines d'entreprises. DataScientest a été créé en 2017 et, depuis lors, il a formé plus de 8.000 étudiants grâce à une équipe de 50 intervenants qualifiés. Depuis un an, cet organisme a rejoint le Groupe OMNES Education[8] et cette coopération s'engage à développer des programmes en ligne et en alternance sur les thématiques de la Tech, grâce à l'expertise pédagogique et technologique de DataScientest et également au savoir-faire des écoles d'OMNES Education. Ces formations offertes par OMNES Education et DataScientest se déroulent à travers un système composé d'un environnement de programmation complètement en ligne rendu possible par une plateforme technologique, de projets spécifiques et d'un accompagnement professionnel et personnalisé. Les trois principaux thèmes dont cette école s'occupe sont la Data, la Cybersécurité et le DevOps[9] (OMNES Education, 2023). DataScientest s'engage à garantir des contenus constamment mis à jour et à être pionniers du monde de la recherche, en particulier dans l'apprentissage automatique et l'intelligence artificielle (Datascientest.com, s. d.). Ce site nous a fourni d'autre matériel sur le TAL.

[8] OMNES Education est une institution privée d'enseignement supérieur et de recherche interdisciplinaire, implantée dans diverses villes françaises comme Bordeaux, Lyon, Marseille et Paris et dans d'autres villes hors de France, par exemple, Barcelone, Genève, Londres, Monaco et San Francisco (OMNES Education, 2023).

[9] Le DevOps concerne le développement et les opérations. Généralement il est utilisé dans le domaine du développement logiciel et il est aussi adoptée pour la Data Science et l'apprentissage automatique (Datascientest.com, 2021).

Dans la recherche des contenus sur l'IA et sur le TAL, un autre site web que nous avons pris en considération est le site de Myriad : c'est une société de conseil et un éditeur de solutions intelligentes. L'objectif de cette société consiste à accompagner les grandes entreprises dans la gouvernance Data, la qualité, l'analyse et la science des données et également dans la gestion des projets afférents. En plus, elle s'intéresse à l'IA afin d'enrichir les savoirs et les compétences de ses consultants et elle s'engage à répondre de manière globale ou ciblée aux enjeux à elle présentés et à automatiser de manière agile les actifs informationnels des entreprises qui le demandent. Myriad peut compter sur la supériorité de ses activités de recherche et développement, sur son expérience et sur le professionnalisme de ses équipes pour offrir les meilleures pratiques et un service de haute qualité. Toutes les équipes de Myriad travaillent sur la chaîne de valeur de la donnée pour faire face à la constante multiplication des outils et des technologies autour de l'exploitation des données. Enfin, l'un des points forts de cette entreprise est d'utiliser des stratégies adaptées à chaque cas d'usage, en proposant des interventions les plus appropriées par rapport aux attentes des entreprises (Myriad-Data, s. d.). Sur le site web de Myriad, nous avons trouvé un autre document sur le TAL qui nous a aidés à mieux comprendre le domaine et à construire notre corpus.

Nous avons également consulté le site web de l'entreprise Microsoft. C'est une société très connue dans le monde de la technologie et des ordinateurs. Elle a été fondée en 1975 et aujourd'hui elle s'occupe de la création des plateformes et des outils alimentés par l'intelligence artificielle afin de fournir des solutions innovantes qui répondent aux besoins évolutifs liés au domaine en question. Microsoft est engagé dans l'exploitation de l'IA pour la rendre largement disponible de manière transparente et responsable, en permettant à chaque personne et à chaque organisation de la planète d'accomplir davantage (Microsoft, s. d.). Compte tenu de son expertise dans le monde de la technologie et de l'intelligence artificielle, nous avons choisi Microsoft pour ajouter au corpus français des contenus concernant le TAL.

Puis, nous avons trouvé une autre société qui s'occupe de la transformation digitale : le groupe ekino. C'est une entreprise qui est présente à Paris, mais aussi à Bengaluru (Inde), New York, Singapore, Hô Chi Minh-Ville (Viêt Nam), Bordeaux et Hong Kong (ekino., s. d.). Son objectif consiste à transformer leurs clients en termes de digitalisation et pour le faire ekino se concentre à repenser les process et les méthodes de

travail et à aligner une stratégie liée aux technologies de l'information avec la stratégie d'entreprise, tout en exploitant les nouvelles technologies. Le groupe est donc premier dans la stratégie et la feuille de route digitale, l'efficacité opérationnelle, le conseil technologique, la création des nouveaux modèles d'entreprise et également dans la formation. Les principales innovations dont ekino s'intéresse sont l'intelligence artificielle, la réalité augmentée et virtuelle, l'Internet des Objets et les relations homme-machine. Cette entreprise veut donc être un acteur positif du numérique et elle s'engage à développer des solutions qui ont un réel impact sur la société et sur les individus. Le groupe ekino travaille grâce à l'union du design, du consulting, de l'ingénierie, de la data et de l'innovation et grâce à l'aide des multiples équipes qui s'intéressent à la construction des solutions pérennes et cohérentes. Ces équipes suivent les clients de la réflexion stratégique à la réalisation de produit ou service n'importe quel secteur d'activité dans lequel elles doivent travailler. Certaines entreprises qui s'appuient à l'expertise d'ekino sont Renault, Volkswagen, Crédit Agricole et Orange. En plus, ekino s'engage sur la formation des entreprises qui souhaitent faire monter en compétences leurs collaborateurs (ekino., s. d.). Dans les publications d'ekino, nous avons trouvé un article concernant le traitement automatique du langage que nous avons incorporé au corpus.

Pour conclure, nous avons consulté le site web de la revue en ligne *La revue IA*. Le créateur de cette revue est Ilyes Talbi, un ingénieur en intelligence artificielle qui a fondé la revue en 2019 avec l'objectif de remédier au manque de contenus techniques fiables et en français sur l'IA et sur le TAL (La revue IA, 2021). Naturellement, nous avons utilisé un article qui explique le TAL pour l'ajouter au corpus.

5.1.2 Le corpus en italien

Le corpus italien se compose de 15 documents. Comme pour le corpus en français, celui en italien est composé d'un groupe de manière égale hétérogène de documents qui traitent le domaine de l'intelligence artificielle et du traitement automatique du langage (voir l'annexe F).

La première source que nous présenterons est le site web du projet éditorial italien AI4Business. Il s'agit d'un projet consacré complètement à l'intelligence artificielle et il est au service du monde des affaires italiens et des entreprises italiennes privées et

publiques. AI4Business fait partie du NetworkDigital360[10], qui gère le panorama italien de la transformation numérique et de l'innovation entrepreneuriale. Ce projet italien fournit divers produits : des services éditoriaux, des nouvelles, des recherches, des vidéos, des webinaires[11] et des événements organisés pour la mise à jour et la formation de tous les opérateurs qui travaillent avec l'IA. Il s'occupe en particulier des opérateurs qui travaillent dans le commerce, l'administration et le social et qui peuvent évoluer grâce à la diffusion de nouvelles technologies que l'intelligence artificielle offre. L'engagement du projet AI4Business consiste à accompagner les entreprises et les administrations publiques dans la compréhension et la mise en œuvre des technologies innovantes et de les encourager à rencontrer les meilleurs fournisseurs du champ. Dans le site web, nous pouvons avoir accès aux contenus relatifs aux technologies d'IA générative (ChatGPT[12]), à l'IA et ses techniques (apprentissage automatique, apprentissage profond et bien d'autres), à la robotique et à la réalité virtuelle (AI4Business, s. d.). Sur le site web, nous avons trouvé des contenus liés au TAL qui sont essentiels pour l'extraction de termes italiens.

Un autre site qui s'est révélé très utile est le site web d'IBM, dont nous avons déjà parlé. IBM est une multinationale, l'International Business Machines Corporation. La société IBM s'occupe de fournir des infrastructures, des logiciels et des services de conseil grâce à la technologie et à son expertise, afin de favoriser la transformation numérique des grandes entreprises du monde entier. Cette société s'intéresse à l'introduction des solutions d'automatisation intelligentes qui seront fondamentales pour les entreprises en termes d'amélioration du workflow, d'intégration de systèmes et d'un contrôle total sur les opérations à exécuter. IBM est aussi concentré sur les employés numériques personnels qui sont optimisés par l'IA et qui peuvent aider le travail des employés qui s'occupent des tâches les plus répétitives et les plus banales. IBM s'engage également en matière de durabilité, en annonçant un objectif de zéro émission nette de

[10] Le groupe Digital 360 est le plus grand réseau italien de titres et de portails B2B consacrés aux thèmes de la transformation numérique et de l'innovation entrepreneuriale. Son objectif est la diffusion la culture numérique dans les entreprises et les administrations publiques italiennes (AI4Business, s. d.).

[11] « Séminaire se déroulant sur Internet. Le terme webinaire est né d'une combinaison entre « web » et « séminaire ». Un webinaire a pour objectif un travail collaboratif ou un travail d'enseignement à distance. Il regroupe en général la visioconférence, le diaporama et la messagerie instantanée » (Linternaute, 2021).

[12] « ChatGPT est un chatbot d'intelligence artificielle (IA) qui utilise le traitement du langage naturel pour créer un dialogue conversationnel semblable à celui des humains. Le modèle linguistique peut répondre à des questions et composer divers contenus écrits, notamment des articles, des messages sur les médias sociaux, des essais, des codes et des courriels » (LeMagIT, s. d.).

gaz à effet de serre d'ici 2030 dans tous les pays où l'entreprise est présente (IBM, s. d.). Parmi tous les services offerts par IBM, cette multinationale propose des solutions d'IA pour aider les clients à construire « l'entreprise de demain ». Les solutions IBM liées à l'intelligence artificielle comprennent IBM watsonx[13], l'expertise scientifique et les équipes de consultants experts qui s'occupent de la mise en place d'une IA responsable au sein des entreprises (IBM, s. d.). Pour ce qui concerne le projet *YourTerm TECH*, le site web d'IBM était utile pour collecter des données essentielles pour la construction du corpus italien sur l'IA et sur le TAL.

Puis, nous avons consulté le site web d'Oracle. Il s'agit d'une entreprise américaine créée en 1977 par Larry Ellison, Bob Miner et Ed Oates, trois ingénieurs qui ont lancé les Software Development Laboratories en Californie, à Santa Clara, et en 1987, elle devient la plus grande et importante entreprise de gestion de bases de données. Oracle s'occupe donc du développement des systèmes de gestion de bases de données et du *Cloud Computing*. L'entreprise a créé sa base de données sécurisée et insérée dans le cloud, l'*Oracle Database*, et elle a également conçu une base de données à correction automatique, *l'Oracle Autonomous Database*. La mission d'Oracle est claire : aider les utilisateurs à voir les données de manière différente, à ouvrir des perspectives et à débloquer d'infinies possibilités dans ce domaine (Oracle, s. d.). En plus, Oracle a créé *Oracle Cloud Infrastructure*, ou simplement OCI, qui est le premier cloud public conçu pour chaque application. OCI a été pensé pour résoudre les problèmes que les utilisateurs peuvent rencontrer en utilisant les clouds publics existants. Nombreuses entreprises du monde entier se sont appuyées sur cette entreprise américaine, par exemple, Mazda, Alliance Data Systems et Deutsche Bank (Oracle, s. d.). Toutefois, Oracle ne s'occupe pas seulement de la gestion de bases de données, mais aussi de l'IA intégré à ses applications cloud et du TAL. Pour cela, nous avons inséré dans notre corpus italien le document concernant l'IA et le TAL.

Nous avons continué notre recherche et nous avons trouvé le site web du groupe RES. C'est une société italienne qui s'occupe du développement des services et des produits destinés à la gestion et à l'amélioration des données de toute taille, de tout format et pour tous les systèmes d'information. L'acronyme RES signifie « *Research for*

[13] « IBM watsonx est une plateforme d'IA et de données. Elle est dotée d'un ensemble d'assistants d'IA conçus pour vous aider à la déployer et à accélérer son effet au sein de votre entreprise » (IBM, s. d.).

Enterprise Systems »[14] et elle a été fondée en 1987. La société RES est née comme société de conseil et de développement de logiciels afin de fournir aux entreprises un service de gestion de bases de données relationnelles. Au fil des années, RES a commencé à développer des produits spécialisés pour la gestion de systèmes d'information complexes. Depuis 2016, le parcours de développement industriel mis en œuvre a transformé ce groupe qui aujourd'hui opère à la fois en Italie et à l'étranger. Le siège social de RES se trouve à Milan. Puis, il y a d'autres sites italiens à Brescia, Côme et Cuneo, auxquels s'ajoutent Bucarest, New York et San Francisco, et Cuenca (Équateur). RES étudie et crée des solutions logicielles fiables pour améliorer et simplifier la gestion de systèmes d'information des entreprises. La société RES s'occupe donc des technologies de l'information et leurs applications, leur évolution et leur stockage sécurisé et facile (RES, s. d.). Dans le site web, nous avons trouvé des informations sur le TAL lié à la collecte des données et au nettoyage des données.

Un autre site web qui s'est révélé fondamental pour la construction du corpus italien est le site de Seacom. C'est un des premiers pionniers de l'*open source* en Italie. Cette entreprise a été fondée en 1999. Elle est cofondatrice de RIOS, c'est-à-dire la Rete Italiana Open Source, un réseau d'entreprises italiennes spécialisées dans l'*open source* (Seacom, s. d.). Seacom aide donc les entreprises à introduire et utiliser les technologies *open source*, afin de les accompagner sur la voie de l'innovation et de l'amélioration continue. Elle s'occupe de la gestion des données, en développant des architectures informatiques sécurisées et personnalisées ; du DevOps ; et du développement d'applications *open source* basées sur l'intelligence artificielle générative utilisant la recherche vectorielle (Seacom, s. d.). Seacom fournit aussi un espace blog où elle télécharge d'articles et d'informations techniques et dans ce blog nous avons trouvé un article sur le défi du traitement du langage naturel en termes d'interprétation du langage humain pour en extraire des connaissances et de la valeur.

Ensuite, nous avons consulté le site web de DataDeep : c'est un projet de l'entreprise italienne Karon qui est engagé dans le panorama des données et son principal objectif est la sensibilisation des entreprises à la valeur réelle des données. Karon travaille dans le monde des entreprises privées, mais aussi dans la formation et l'école. Son projet DataDeep est consacré à la création des solutions d'analyse des données et d'intelligence

[14] En français : « Recherche pour les systèmes d'entreprise ».

artificielle pour les entreprises (Karon, s. d.). DataDeep est formé par un réseau professionnel de data scientists, data analysts et computer scientists hautement qualifiés dans l'analyse de données et qui travaillent sur le développement de solutions d'intelligence artificielle, en exploitant l'apprentissage automatique et l'IA pour réorganiser les données de production, réduire les coûts et obtenir un avantage concurrentiel (Karon, s. d.). Le projet s'occupe également de la formation et de la vulgarisation des contenus liés à l'informatique, à la gestion des données et à l'IA. Les ressources que le projet offre consistent dans des cours vidéo gratuits, guides sur l'analyse des données et sur l'IA et des articles hebdomadaires sur les mêmes sujets (DataDeep, s. d.). L'article que nous avons choisi pour enrichir le corpus italien concerne le TAL lié à la communication.

Nous avons aussi consulté un blog d'un entrepreneur qui a été impliqué dans le développement des applications et la coordination des entreprises en démarrage et des agences numériques : c'est Luigi Marino. Il a travaillé avec des partenariats importants en accumulant des compétences technologiques et managériales (MarinoLuigi.it, s. d.). Dans son site web est aussi présent un blog où nous pouvons trouver des articles sur l'informatique, l'intelligence artificielle, les nouvelles technologies et bien d'autres thèmes. Pour ce qui concerne le corpus italien que nous avons construit, nous avons utilisé un article sur le TAL et son fonctionnement.

Nous avons parlé précédemment de la bibliothèque numérique JSTOR, une autre est ResearchGate créé en 2008 pour résoudre les problèmes et aider les chercheurs. L'objectif de cette bibliothèque est de connecter le monde de la science et de rendre la recherche accessible à tous. La communauté de ResearchGate est composée par 20 millions de chercheurs qui proviennent de divers secteurs et de plus de 190 pays. Dans cette plateforme, ils/elles se connectent et collaborent avec les autres chercheurs et ils partagent leurs travaux. ResearchGate offre un ensemble énorme de contenus sur plusieurs domaines spécialisés, par exemple, la science, la médecine, l'ingénierie, l'informatique, l'intelligence artificielle, l'apprentissage automatique, etc. (ResearchGate, s. d.). Nous avons ainsi trouvé une étude sur le traitement automatique du langage naturel et sur le développement d'un système de reconnaissance automatique pour l'analyse logique. Les auteurs de ce document sont Marco Maggini et Stefano Meoni. Marco Maggini est spécialisé en ingénierie électronique et informatique et en

systèmes de contrôle. Actuellement il est professeur chez l'université de Sienne au département d'ingénierie de l'information et de mathématiques et ses recherches actuelles traitent l'apprentissage automatique, les réseaux neuronaux, l'exploration du web, les moteurs de recherche et le traitement du langage naturel. Il a été membre du comité de programme de plusieurs conférences et ateliers internationaux et il est l'auteur de plus de 120 publications dans des revues et conférences internationales, en discutant surtout l'apprentissage automatique et l'internet (Università degli Studi di Siena, s. d.).

Puis, nous avons consulté le site web du blog *Osservatori.net digital innovation*. C'est un blog affilié au département d'ingénierie de gestion du Politecnico di Milano (École polytechnique de Milan). Ce groupe de chercheurs est né en 1999 et son objectif depuis sa naissance est de créer une culture dans tous les principaux domaines de l'innovation numérique. Aujourd'hui, le blog représente un point de référence sur le domaine de l'innovation numérique en Italie, intégrant des activités de recherche et de mise à jour continue. L'équipe du blog, formée par 100 professeurs, chercheurs et analystes, s'occupe des questions clés concernant l'innovation numérique. Ces chercheurs et les autres figures professionnelles engagées se proposent de diffuser les connaissances sur les opportunités et les impacts des technologies numériques sur les entreprises, les administrations publiques et les citoyens. Ils/elles se basent sur des modèles interprétatifs qui s'appuient sur des preuves empiriques solides et des espaces de comparaison indépendants, qui regroupent la demande et l'offre d'innovation numérique en Italie. Pour cela, le blog *Osservatori.net* est une source unique d'informations et de données qui donne accès à une plateforme multimédia et interactive qui peut être personnalisée en fonction des intérêts de chacun (Osservatori.net Digital Innovation, Politecnico di Milano, s. d.). Même ce blog nous a fourni des contenus intéressants sur le TAL pour la construction du corpus en italien.

Pour trouver d'autres contenus afin de compléter ultérieurement le corpus, nous avons consulté le site web ATG Intelligence Artificielle ou AGT AI, la division du Groupe Anzani spécialisée dans le domaine de l'intelligence artificielle. Le Groupe Anzani travaille dans le domaine des logiciels et de l'informatique décisionnelle, il a été créé en 1994. Ce groupe a travaillé avec plusieurs multinationales européennes et américaines, grands groupes italiens et aussi avec des petites et moyennes entreprises, en accumulant de l'expérience de haut niveau. L'entreprise dispose de plusieurs sites : en

Italie, à Erba ; en Suisse, à Lugano et à Stabio ; et en Inde, à Bengaluru (ATG Anzani Group, s. d.). En 2019, le groupe a ouvert une succursale à Sondrio dans la Valteline : la division AGT AI. La succursale s'occupe principalement de l'intelligence artificielle appliquée au traitement des données et elle compte déjà plusieurs projets actifs dans différents marchés tels que le *business*, la finance, l'industrie 4.0[15] et le recrutement et la sécurité. En particulier, l'expertise de cette division est divisée en trois domaines liés à l'intelligence artificielle, à savoir, les données, l'imagerie, le texte et l'IA générative. Elle offre plusieurs services, les produits et la possibilité de consulter des études de cas (ATG Artificial Intelligence Division, s. d.). La division nous offre des contenus sur le TAL et comment l'utiliser pour travailler de manière plus efficiente.

Nous avons également cherché des matériaux de cours universitaires et nous avons trouvé un PowerPoint écrit par Salvatore Sorce, chercheur chez le département d'ingénierie chimique, de gestion, informatique et mécanique de l'université de Palerme. Ce document se compose de 20 pages et s'intitule *Introduzione alla Linguistica Computazionale*, en français « Introduction à la linguistique informatique ». Il traite le TAL et comment les machines utilisent le langage humain (Università degli Studi di Palermo, s. d.). Pour cette raison, nous avons décidé de l'intégrer à notre corpus italien, car il est riche en termes spécialisés.

Puis, nous sommes tombés sur le chapitre d'un livre particulier. Le livre s'intitule *L'informatica giuridica in Italia. Cinquant'anni di studi, ricerche ed esperienze*[16]. Ce livre traite de la relation complexe entre l'univers numérique et le droit et il parcourt l'évolution de l'informatique juridique en Italie, en commençant par sa définition progressive du statut épistémologique. Le livre analyse également les institutions qui sont devenues les protagonistes de la recherche dans ce domaine depuis les années 1970, par exemple, sont ensuite rappelés : le CNR, les universités, la Cour de cassation et la Chambre des députés. En plus, il comprend une section spéciale consacrée au ITTIG du CNR (Istituto di Teoria e Tecniche dell'Informazione Giuridica)[17]. Il y a également des sections dédiées aux réalités d'autres pays et aux entretiens avec des

[15] « L'industrie 4.0 révolutionne la façon dont les entreprises fabriquent, améliorent et distribuent leurs produits. Les fabricants intègrent de nouvelles technologies, notamment l'internet des objets (IoT), le cloud computing et l'analytique, ainsi que l'IA et l'apprentissage automatique dans leurs installations de production et dans l'ensemble de leurs opérations » (IBM, s. d.).
[16] En français, « L'informatique juridique en Italie. Cinquante ans d'études, de recherche et d'expérience ».
[17] En français, « Institut de théorie et de techniques de l'information juridique ».

experts non italiens dans le domaine (Biblioteca Centrale Giuridica, 2015). Le chapitre trois parle du traitement du langage naturel et des modèles et applications qui peuvent être appliqués dans le domaine juridique. L'auteur de ce chapitre est Fabrizio Turchi, technologue chez l'Istituto di Informatica Giuridica e Sistemi Giudiziari[18] (IGSG) du CNR de Florence, où il est responsable des systèmes d'information. Il est un expert des normes pour la représentation et la diffusion de documents juridiques en ligne, du développement d'applications web dans le domaine juridique et de l'extraction de connaissances à partir de textes non structurés par l'application du TAL. En plus, il développe d'ontologies libres afin de donner des standards pour garantir l'interopérabilité et l'échange d'informations d'enquête numériques entre différents outils et différentes organisations (Istituto di Informatica Giuridica e Sistemi Giudiziari, s. d.).

Pour rester dans le domaine de la terminologie de l'informatique et de l'intelligence artificielle, nous avons aussi inclus dans le corpus italien un article du site web *Diciamolo in Italiano*, en français « Disons-le en italien ». Le créateur et auteur du site est Antonio Zoppetti. Il travaille dans le domaine de la langue italienne et il est éditeur, auteur et enseignant. En 1993, il s'est occupé de la création de la version CD-ROM du *Devoto Oli*, c'est-à-dire le premier dictionnaire numérique complet mis sur le marché en Italie. Actuellement il est engagé dans la bataille culturelle contre l'abus de l'anglais et d'anglicismes. Pour faire cela, il fait partie de plusieurs projets, parmi lesquelles le site web *Diciamolo in Italiano*, crée en 2017 et qui vise à diffuser une nouvelle culture à travers des données et des réflexions sur le thème de l'écologie linguistique face à l'anglicisation (Diciamolo in Italiano, s. d.).

Pour conclure la liste de sources du corpus italien, nous présentons un document sur l'histoire de l'informatique. L'auteur est Giorgio Casadei, directeur du CRIAD (Centro di Ricerca per l'Informatica Applicata alla Didattica e all'Educazione)[19] de l'université de Bologne. Il travaille sur l'intégration des technologies de l'information dans l'éducation : il s'occupe donc de concevoir des outils et des méthodes d'intelligence artificielle pour la mise en œuvre de systèmes experts afin d'étudier les processus mentaux pertinents dans les processus d'enseignement et d'apprentissage ; puis, il étudie l'utilisation de réseaux de transmission pour l'enseignement à distance ; et il crée aussi

[18] En français, « Institut d'informatique juridique et de systèmes juridiques ».
[19] En français, « Centre d'étude et de recherche en informatique appliquée à l'éducation ».

des systèmes intelligents distribués pour la conception, la mise en œuvre et le test de simulateurs pour l'enseignement (Università di Bologna, s. d.).

5.1.3 Le corpus en anglais

Le corpus anglais se compose de 15 documents, comme celui en italien. Comme pour les autres corpus, celui en anglais est composé d'un groupe hétérogène de documents qui traitent le domaine de l'intelligence artificielle et du traitement automatique du langage (voir l'annexe G).

Le premier document que nous présentons est un essai sur le thème de la linguistique appliquée au traitement du langage naturel. L'auteur est Ted Briscoe, professeur de linguistique informatique chez l'université de Cambridge. Il s'occupe principalement de la linguistique informatique et théorique et du TAL. Ses recherches portent notamment sur les techniques d'analyse syntaxique, l'acquisition d'informations lexicales à partir de corpus textuels et de dictionnaires électroniques, les modèles d'apprentissage des langues humaines, la technologie éducative liée à l'apprentissage des langues et l'évolution des langues. En plus, il a publié plus de 150 articles de recherche et quatre livres et il a travaillé dans de nombreux projets financés par l'UE et le Royaume-Uni (University of Cambridge, s. d.). L'essai se concentre sur les caractéristiques du langage naturel, en prenant comme exemple la langue anglaise.

Puis, nous avons trouvé un article concernant les applications commerciales du traitement automatique du langage naturel. La revue scientifique en ligne où nous avons trouvé l'article s'appelle *Communications of the ACM*. Il s'agit de la principale publication en ligne et imprimée du domaine de l'informatique et des technologies de l'information et elle est considérée la source d'information la plus fiable et la plus compétente pour les professionnels de l'informatique. La revue s'occupe donc des domaines émergents de l'informatique, des nouvelles tendances technologiques et des applications pratiques. Depuis plus de 60 ans, la revue *Communications of the ACM* est fondamentale pour les leaders de l'industrie, car ils/elles ont la possibilité de présenter et débattre des diverses implications technologiques, des politiques publiques, des défis liés au domaine de l'ingénierie et des tendances du marché. Cette plateforme assure des contenus éditoriaux de haute qualité et la vulgarisation des domaines tels que les arts, les sciences et les applications des technologies de l'information. Toutes les recherches, les nouvelles, les opinions et les autres contenus sont publiés avant dans le site web et puis

dans le journal scientifique mensuel, disponible dans plusieurs formats (Communications of the ACM, s. d.). L'acronyme ACM désigne la fameuse organisation américaine Association for Computing Machinery : il s'agit d'un ensemble d'enseignants, des chercheurs et des professionnels de l'informatique qui ont le but commun d'inspirer le dialogue, de partager les ressources et de relever les défis dans le domaine en question. ACM a été fondée à l'origine de l'ère informatique, elle est présente dans le monde entier et compte environ 100.000 membres. Une grande partie de ces membres sont résidents en dehors des États-Unis : il a donc été jugé nécessaire de créer de conseils en Europe, en Inde et en Chine, afin de favoriser les possibilités de mise en réseau qui renforcent les liens entre différents pays et différentes communautés techniques. De cette manière, la capacité de l'organisation à sensibiliser aux questions techniques, éducatives et sociales liées à l'informatique augmente et améliore. Pour ces raisons, elle est considérée comme la plus grande société informatique au monde, en se proposant de promouvoir des normes les plus élevées et de contribuer à la reconnaissance de l'excellence technique (ACM, s. d.). L'article sur les applications commerciales du traitement automatique du langage naturel a été très utile pour la construction du corpus anglais.

Pendant la recherche de contenus à ajouter au corpus, nous avons trouvé un autre chapitre d'un livre sur le domaine qui nous intéresse. Le livre s'intitule *Ancient Manuscripts in Digital Culture: Visualisation, Data Mining, Communication*, en français « Les manuscrits anciens dans la culture numérique : visualisation, exploration de données, communication ». Ce livre traite les progrès et le tournant de l'informatique dans la visualisation des manuscrits juifs et chrétiens anciens, l'exploration de données et la communication. Il rassemble les contributions de dix-sept chercheurs impliqués dans les études bibliques et il présente la diffusion des sciences humaines numériques[20] (Brill, s. d.). Ce livre fait partie de la série *Digital Biblical Studies*, en français « Études bibliques numériques » qui vise à publier les dernières recherches à l'intersection des humanités numériques et des études bibliques, du judaïsme ancien et du christianisme. L'objectif principal de cette série est de démontrer la transformation de la recherche, de l'enseignement, de la cognition et de l'économie de la connaissance dans la culture

[20] « Domaine de recherche et d'enseignement au croisement de l'informatique et des lettres, des arts, des sciences humaines et des sciences sociales, visant à produire et à partager des savoirs, des méthodes et de nouveaux objets de connaissance à partir d'un corpus de données numériques » (FranceTerme, 2019). Les sciences humaines numériques sont aussi appelées humanités numériques.

numérique. En particulier, elle s'occupe de l'étude de la méthodologie et des pratiques des humanités numériques appliquées aux textes, inscriptions, données archéologiques et travaux d'érudition. Les langues qui font l'objet d'étude pour la série sont les langues anciennes, notamment le grec ancien, l'hébreu, le latin, l'arabe, le copte et le syriaque (Brill, s. d.). Le chapitre que nous avons englobé dans le corpus est le chapitre six qui traite l'utilisation du TAL pour la recherche de références textuelles.

Puis, nous avons considéré une autre revue scientifique : elle s'intitule *Science*. Elle a été créée en 1880 grâce à Thomas Edison et elle a été au centre d'importantes découvertes scientifiques. Actuellement, la revue s'occupe de la publication des recherches dans toutes les sciences et ses articles sont parmi les articles les plus cités au monde. Certaines de ces publications les plus importantes sont : le génome humain dans son intégralité pour la première fois ; des images inédites de la surface de Mars ; et les premières études établissant un lien entre le sida et le virus de l'immunodéficience humaine. *Science* est publiée par l'American Association for the Advancement of Science (AAAS), c'est-à-dire la plus ancienne et la plus importante organisation scientifique générale au monde. Cette association est la porte-parole de la science et des scientifiques du monde entier et elle s'engage à communiquer la valeur de la science au public. Pour le faire, elle est au centre de plusieurs activités afin d'aider les gouvernements à formuler des politiques scientifiques et de promouvoir les progrès de l'enseignement scientifique et de la diversité. L'AAAS est à la tête d'autres revues scientifiques en ligne qui font partie de la famille de publications Science : elles sont *Science Translational Medicine*, *Science Signaling*, *Science Immunology*, *Science Robotics* et *Science Advances* (American Association for the Advancement of Science, s. d.). Dans le site nous avons trouvé deux articles qui traitent les progrès du traitement automatique du langage et le TAL en général et qui seront parfaits pour le corpus anglais.

Pendant la recherche, nous sommes tombés sur le site d'une conférence internationale : l'*International Conference on Mechanical, Control and Computer Engineering* (ICMCCE), en français c'est la conférence internationale sur l'ingénierie mécanique, de contrôle et informatique. Dans l'année courante, nous assisterons à la septième édition de cette conférence qui se tiendra à Hangzhou (Chine) du 25 au 27 octobre. Le but est de réunir des universitaires et des experts industriels qui travaillent dans le domaine de l'ingénierie mécanique, du contrôle et de l'informatique au sein d'un

espace commun. La conférence vise à la promotion de la recherche et des activités de développement dans ces domaines et elle s'engage à promouvoir l'échange d'informations scientifiques entre les chercheurs, les développeurs, les ingénieurs, les étudiants et les praticiens travaillant au niveau global (ICMCCE, s. d.). La conférence se tient chaque année et nous avons choisi le document relatif à la 3e conférence qui a eu lieu en 2018. Cette publication parle de l'application du traitement du langage naturel à la traduction automatique et elle peut fournir beaucoup de termes à extraire.

Nous avons continué notre recherche dans les revues en ligne et nous avons trouvé la revue scientifique en ligne *Multimedia Tools and Applications*, en français « Outils et applications multimédias », qui est une revue internationale. Cette revue s'occupe de publier des travaux de recherche qui examinent le développement multimédia, ses outils et ses cas d'applications. Ce journal est reconnu comme le premier dans le domaine du multimédia et son comité de rédaction compte les meilleurs experts mondiaux en matière de multimédia. En plus, il ne s'occupe pas seulement des études de cas, mais aussi d'articles expérimentaux et d'enquête (Springer, s. d.). La revue s'adresse à un public très vaste : aux universitaires, aux praticiens, aux scientifiques et aux ingénieurs qui participent à la recherche et aux applications des systèmes multimédias et tous les articles sont évalués par des pairs. Parmi les thèmes les plus étudiés nous trouvons l'analyse des logiciels d'activation d'applications multimédias, l'hypermédia, les outils de mesure des performances pour le multimédia, les outils de création multimédia, l'analyse de bases de données multimédias et recherche et les outils et applications Web (Springer, s. d.). L'article que nous avons inclus dans le corpus en anglais concerne le traitement du langage naturel, ses tendances et ses défis actuels.

Ensuite, nous avons trouvé le site web d'une entreprise qui se base sur la diffusion équitable de l'intelligence artificielle. C'est l'entreprise allemande Levity. Elle se fonde sur la conviction que l'IA ne devrait pas être un privilège réservé aux entreprises technologiques. Son travail vise à automatiser les flux de travail des entreprises qui utilisent des capacités cognitives, c'est-à-dire semblables à celles de l'homme (Levity, s. d.). Pour ce qui concerne le domaine de l'intelligence artificielle, Levity est un outil qui offre l'entraînement des modèles d'IA sur des images, des documents et des données textuelles (Levity, 2022). Dans le blog du site, nous avons trouvé un article très intéressant

sur le traitement automatique du langage qui est très riche en terminologie et nous l'avons incorporé à notre corpus anglais.

En continuant notre recherche de documents, nous avons trouvé une publication du 1966 sur les ordinateurs et leur capacité de se rapporter à la traduction et à la linguistique. Pour être plus précis, il s'agit d'un rapport écrit par le comité consultatif sur le traitement automatique des langues, l'Automatic Language Processing Advisory Committee (ALPAC). Le thème principal est la traduction automatique et son progrès. Ce comité a été créé en 1964 par le gouvernement des États-Unis et dirigé par John R. Pierce. Son but était l'évaluation du progrès de la linguistique informatique en général et de la traduction automatique. Ce rapport, que nous avons utilisé pour enrichir le corpus anglais, est devenu populaire, car il est riche en scepticisme à l'égard des recherches effectuées dans le domaine de la traduction automatique jusqu'à la fine des années 1960. En plus, il voulait souligner la nécessité d'une recherche fondamentale en linguistique informatique. Par conséquent, le gouvernement américain a décidé de réduire considérablement le financement pour la recherche dans le domaine de la traduction automatique, en marquant ainsi le début du premier hiver de l'IA (Wikipédia, 2024). Ce rapport porte aussi la signature de la société américaine National Academy of Sciences (NAS) ou, en français, Académie nationale des sciences. Il s'agit d'une société privée à but non lucratif créée par le président Abraham Lincoln en 1863. L'Académie est responsable de fournir des conseils objectifs sur des questions liées à la science et à la technologie et elle s'engage dans le progrès de la science en Amérique. Les scientifiques de l'Académie sont élus par leurs pairs en raison de leurs contributions exceptionnelles à la recherche. La NAS a aussi fondé une revue scientifique internationale en 1914 : la revue *Proceedings of the National Academy of Sciences* (PNAS), en français Comptes-rendus de l'Académie nationale des sciences des États-Unis d'Amérique. En plus, l'Académie a fondé deux autres importantes académies : l'Académie nationale d'ingénierie (National Academy of Engineering, NAE) en 1964 et l'Académie nationale de médecine (National Academy of Medicine, NAM) en 1970. Actuellement, les trois entités travaillent ensemble sous le nom de National Academies of Sciences, Engineering, and Medicine[21] (National Academy of Sciences, s. d.). Le rapport parle donc d'un thème qui est étroitement lié au traitement automatique du langage et il s'est révélé utile pour le corpus.

[21] En français, « Académies nationales des sciences, des techniques et de la médecine ».

Notre recherche s'est ensuite arrêtée sur un autre journal en ligne : le *Journal of the American Medical Informatics Association*, dont le sigle est JAMIA. Il s'agit de la première revue à comité de lecture de l'AMIA (American Medical Informatics Association), un comité spécifique consacré à l'informatique biomédicale et à l'informatique de santé. Cette revue scientifique en ligne publie un volume par an avec 12 numéros et s'occupe des articles qui couvrent tous les aspects de ces domaines : nous pouvons trouver des articles sur l'informatique dans les domaines des soins cliniques, sur la recherche clinique, sur la science translationnelle, ou encore sur la science de la mise en œuvre, de l'imagerie, de la santé des consommateurs, de la santé publique et de la politique. Ces articles sont fondamentaux pour ce qui concerne la description des recherches et des systèmes informatiques innovants afin de contribuer au progrès de la science biomédicale et à promouvoir également la santé. En ajoute, le journal permet aux lecteurs et aux chercheurs d'étudier aussi les rapports de cas, des perspectives et des revues sur les développements informatiques les plus importants en matière de mise en œuvre, de politique et d'éducation (American Medical Informatics Association, s. d.). L'article choisi traite l'évolution historique du TAL et ses problèmes courants, en soulignant ensuite les points forts des efforts déployés pour ce qui concerne le langage médical.

Ensuite, nous avons trouvé et consulté le site web WEBIST, c'est-à-dire l'*International Conference on Web Information Systems and Technologies*. L'objectif de cette conférence internationale concernant les systèmes d'information et les technologies du Web est de réunir des chercheurs, des ingénieurs et des praticiens qui sont intéressés aux avancées technologiques et aux applications commerciales des systèmes d'information basés sur le Web, couvrant par exemple les technologies de l'information, le Web sémantique, l'analyse des réseaux sociaux, l'interface homme-machine dans les systèmes mobiles et les interfaces utilisateur (WEBIST, 2024). Puis, nous avons trouvé le compte-rendu de la 4^{e} conférence du 2008 qui s'occupe des avantages de l'exploitation des techniques du TAL pour l'apprentissage en ligne.

Un autre compte-rendu que nous avons inclus dans le corpus est celui de la première conférence du 1985 sur les théories et les méthodologies concernant la traduction automatique des langues naturelles. Ce document s'occupe de la question de la

coopération entre la linguistique et le traitement automatique du langage naturel et entre la linguistique et la traduction automatique (Semantic Scholar, s. d.).

Pour conclure cette partie dédiée au corpus anglais, nous présentons le site web MonkeyLearn. C'est une plateforme d'apprentissage automatique pour l'analyse de textes. En l'utilisant, les utilisateurs peuvent facilement obtenir des données exploitables à partir de textes bruts, par exemple, ils/elles peuvent révéler un sentiment exprimé dans des textes tels que des tweets, des chats ou des articles. La plateforme offre une interface utilisateur graphique qui donne la possibilité de créer et tester des modèles d'apprentissage automatique personnalisés pour résoudre des problèmes particuliers. Ces modèles comprennent des algorithmes entraînés et exécutés instantanément qui peuvent être utilisés sans installer ou déployer de logiciel et ils peuvent être conçus à la volée avec vos données particulières. Pour concevoir un modèle précis, les utilisateurs peuvent utiliser des textes d'un domaine qui ils/elles connaissent. Les modèles de MonkeyLearn sont organisés en deux catégories : il y a les modèles de classification qui prennent du texte et fournissent des étiquettes ou des catégories et des modèles d'extraction qui extraient des données particulières dans un texte. En plus, MonkeyLearn met à disposition des utilisateurs de la documentation et un blog où elle fournit du contenu supplémentaire utile (MonkeyLearn, 2024). Dans le blog nous avons trouvé trois articles que nous avons utilisés pour le corpus en anglais : ils parlent du TAL, de ses techniques et des exemples pratiques qui l'utilisent.

5.1.4 Analyse conclusive du corpus

Nous avons donc examiné toutes les sources des documents que nous avons insérées dans les trois corpus. Toutes les sources mentionnées ont une histoire et elles peuvent être considérées comme sources fiables.

Nous avons déjà vu dans le chapitre trois que pour être certains d'utiliser des sites web fiables, nous devons être très attentifs. Nous devons par exemple lire les pages d'accueil et les pages « À propos » afin de connaître la nature des sites. Puis, nous devons nous assurer qu'il ne s'agit pas d'un site satirique ou parodique. Si nous tombons dans des sites qui utilisent des sources primaires, nous devons vérifier les auteurs et les sources, qui doivent être vérifiées et mentionnées. En plus, nous devons examiner le contenu du site qui peut être caractérisé par des informations factuelles ou des opinions et s'il est ouvert à des propos contradictoires, ou s'il ne permet qu'une seule lecture des faits (Le

Monde, 2022). Une bonne solution est représentée par le choix des sites web qui font autorité, par exemple, les revues scientifiques en ligne. Pendant notre recherche de documents, nous avons rencontré de nombreuses revues scientifiques, par exemple, la revue *Langages*, la revue *Langue française* et la *Revue d'économie financière* pour ce qui concerne le corpus français ; la revue *Science* et la revue *Journal of the American Medical Informatics Association* pour ce qui concerne le corpus en anglais. Ces journaux en ligne représentent tous des sources fiables : ils sont utilisés par la communauté scientifique, ils traitent des thèmes à la fois très complexes, certains ont une histoire ancienne et une tradition éditoriale importante ou ils ont été créés par les gouvernements. Ils ont donc de l'autorité et ils résultent parfaits pour être utilisés dans des projets terminographiques afin d'avoir un corpus riche en terminologie.

Toutefois, nous n'avons pas seulement pris en considération les revues scientifiques, mais aussi les sites web d'entreprises réelles qui s'occupent vraiment du domaine qui nous intéresse. Des exemples sont l'entreprise Stat4decision, qui est une entreprise qui s'occupe de la science de données et Microsoft. Ces deux étaient utiles pour la construction du corpus français. Encore, nous avons IBM, qui est une multinationale qui s'occupe de fournir des infrastructures, des logiciels et des services de conseil afin de favoriser la transformation numérique des grandes entreprises ; Oracle, qui est la plus grande et importante entreprise de gestion de bases de données ; Seacom, entreprise pionnière de l'*open source* ; et DataDeep, qui est le projet de l'entreprise italienne Karon qui est engagé dans le panorama des données. Ces entreprises et leurs contenus nous ont aidés à construire le corpus italien. Elles fournissent des données réelles, car elles travaillent dans le domaine de l'intelligence artificielle et peuvent offrir des contenus de haut niveau.

Nous avons également examiné et attentivement choisi certains blogs, par exemple, le blog La Revue IA, le blog MarinoLuigi.it et le blog de MonkeyLearn. De toute évidence, avant d'inclure les contenus fournis par ces blogs, nous avons fait une analyse sur le statut de ces blogs. Au contraire, nous ne les aurions jamais inclus dans le corpus.

Puis, nous avons utilisé des chapitres de livres sur le domaine, par exemple, le livre sur l'informatique juridique en Italie dont nous avons pris le chapitre trois et le livre *Ancient Manuscripts in Digital Culture: Visualisation, Data Mining, Communication* dont nous avons pris le chapitre six.

Une autre source très importante est représentée par les organismes de formation. Nous avons utilisé, par exemple, les documents de l'École nationale supérieure des Mines de Saint-Etienne ; le blog affilié au département d'ingénierie de gestion de l'École polytechnique de Milan ; et un essai d'un professeur de linguistique informatique chez l'université de Cambridge.

Le principal critère de choix que nous avons utilisé pour la construction du corpus est le critère du domaine : le projet *YourTerm TECH* s'appuie sur la recherche de termes sur le traitement automatique du langage et sur l'intelligence artificielle, nous avons donc effectué des recherches sur ces deux domaines, en trouvant des documents et des sites web qui traitent ces deux domaines d'étude. Les principaux problèmes liés à la construction du corpus se sont présentés lors du choix effectif des documents, en particulier pour ce qui concerne la fiabilité des sources, mais nous avons vu que cet obstacle est franchissable grâce à une méticuleuse analyse de la source même. Tous les documents que nous avons trouvés traitent donc les domaines de l'IA et du TAL directement ou indirectement, par exemple, en parlant simplement du traitement automatique du langage ou en traitant des applications qui se basent sur le TAL et sur l'IA, comme la traduction automatique. Pour conclure cette partie, nous pouvons dire que le processus de recherche mis en œuvre était long et complexe, mais il a été fondamental pour le bon choix des textes et documents afin de construire un corpus d'un bon niveau de spécialité.

5.2 Analyse de l'extraction de termes

Pour ce qui concerne l'extraction de termes, il y a quelques questions à préciser. Nous avons déjà analysé dans la chapitre quatre le processus pour extraire les termes d'un corpus : l'extraction de termes peut se baser sur la comparaison des corpus, sur les techniques statistiques, telles que le calcul des segments répétés et le calcul de l'information mutuelle, et sur les techniques linguistiques, telles que l'identification de patrons typiques et le repérage de frontières entre termes.

Dans le cadre du projet terminographique pour *YourTerm TECH*, nous avons premièrement construit les trois corpus en français, italien et anglais avec Word. Puis, nous avons converti ces trois fichiers Word en trois fichiers texte bruts. Ce passage est fondamental pour ensuite utiliser le logiciel d'extraction automatique de termes que nous avons interrogé, le logiciel TermoStat, car il nécessite des documents en texte brut.

Successivement, nous avons téléchargé les trois fichiers et nous avons lancé l'analyse des fichiers.

5.2.1 L'exigence d'un nettoyage manuel

Nous avons vu comment le logiciel TermoStat fonctionne : il utilise un corpus de référence, qui un corpus de langue générale déjà partie du logiciel, et un corpus d'entrée, qui généralement est un corpus spécialisé. À ce point, TermoStat examine les deux corpus et il extrait les termes les plus fréquents dans le corpus focus, les mêmes qui ne sont pas si fréquents dans le corpus de référence. Sur la base d'un calcul statistique basé sur la fréquence, le logiciel nous fournit donc les déjà nommés candidats-termes, c'est-à-dire les termes qui pourront être l'objet d'étude.

À ce moment-là, nous disposons d'une longue liste de candidats-termes, comme celle que nous avons présentée grâce au tableau 5 dans le chapitre quatre. Les listes fournies par les logiciels d'extraction automatique de termes (non seulement notre liste) ne sont jamais parfaites. Les extracteurs sont capables de garantir aux terminographes des termes qui seront effectivement l'objet d'étude, mais ils ne pourront pas substituer entièrement le travail de l'homme. C'est pourquoi il était nécessaire de procéder à une évaluation manuelle de la liste de candidats termes. Nous avons donc utilisé Excel pour optimiser le travail d'extraction.

Un des objectifs du projet *YourTerm TECH* était la collecte de 150 termes : 50 termes français (voir l'annexe A), 50 termes italiens (voir l'annexe B) et 50 termes anglais (voir l'annexe C). Pour le français, il y a 12 termes simples et 38 termes complexes ; pour l'italien, il y a 9 termes simples et 41 termes complexes ; et pour l'anglais nous avons choisi 17 termes simples et 33 termes complexes. Ces termes seront utilisés pour compiler les fiches terminologiques bilingues sur FAIRterm : les termes français sont traduits en italien ; les termes italiens sont traduits en français ; et les termes anglais sont traduits encore en français. Le fichier Excel est composé de quatre feuilles. Dans la première feuille, nous trouvons tous les 150 termes sélectionnés pour le projet et leur traduction dans les trois langues d'étude, pour un total de 450 termes (voir l'annexe D). Dans la seconde feuille, nous avons inséré les termes en français, dans la troisième il y a les termes anglais et dans la dernière nous trouvons les termes italiens. Au fur et à mesure que les termes étaient vérifiés et saisis dans les feuilles Excel, ils étaient revus manuellement. Les termes qui dans la liste étaient incomplets, les termes qui ne faisaient pas partie du lexique

de l'intelligence artificielle et du TAL, ou les termes fréquents, mais pas pertinents ne pouvaient pas être pris en considération. Nous avons rencontré plusieurs cas de candidats termes que nous avons dû exclure, car ils n'étaient pas appropriés pour notre projet. Nous avions donc besoin de revoir les trois corpus manuellement pour arriver au total de 50 termes pertinents par chaque langue. Ce nettoyage manuel est essentiel pour fournir une extraction de haut niveau, car la seule extraction automatique de termes mis en œuvre par les extracteurs automatiques n'est pas suffisante : il faut toujours que les terminographes procèdent à contrôler les résultats fournis. Examinons quelques exemples.

Pour ce qui concerne le corpus français, TermoStat nous a fourni des termes très valables, par exemple, *apprentissage supervisé*, *expressions régulières*, *lemmatisation*, *ontologie*, *reconnaissance optique de caractères* et *réseaux de neurones*. Ces termes appartiennent au domaine de l'IA et du TAL et ils sont donc pertinents. Pour compléter la liste de termes français, il était nécessaire un nettoyage manuel. Dans les candidats termes, nous avons trouvé l'expression *normalisation effectuée*, mais ce n'est pas un terme pertinent et nous avons donc revu le corpus et nous avons trouvé le terme *normalisation des données*. Un autre exemple est le candidat terme *méth stochastiques*, mais c'est incomplet, le terme à choisir est *méthodes stochastiques*. Puis, dans la liste il y a le syntagme nominal *exemple de reconnaissance vocale* qui n'est pas un terme à proprement parler : le terme sera seulement une partie de ce syntagme, c'est-à-dire *reconnaissance vocale*. De toute évidence, la liste est caractérisée également par de nombreux expressions et mots qui ne sont pas de termes, par exemple, *information manquante*, *nombre de fois*, *acteur incontestable* et *maticien*. Il y a également de candidats qui pourront être des termes, mais qui ne sont pas pertinents pour nos domaines d'étude, par exemple, le mot *gestion*, qui a une fréquence de 39, n'est pas un terme propre de l'IA, ou encore, le mot *étape* avec une fréquence de 22 et le mot *question* avec une fréquence de 73. Parfois nous avons consulté la partie dédiée à la concordance de TermoStat et nous avons de bons résultats, par exemple, dans la liste nous avons le candidat *traitement automatique*, mais c'est incomplet : nous avons cliqué sur le candidat terme et nous avons consulté la concordance où le terme complet est présent *traitement automatique du langage*. La même chose s'applique au candidat terme *méthodes basés* qui est devenu *méthodes basées sur des règles*.

Pour ce qui concerne la liste de candidats termes italiens, nous pouvons dire que le processus n'est pas si différent de celui pour la liste française. TermoStat a fourni des termes pertinents, par exemple, *albero di decisione* (« arbre de décision »), *comprensione del linguaggio naturale* (« compréhension du langage naturel »), *linguaggio di programmazione* (« langage de programmation »), *modelli pre-addestrati* (« modèles préentraînés »), *reti neurali ricorrenti* (« réseaux de neurones récurrents »), *rimozione delle stop-word* (« suppression de mots vides ») et *sistemi conversazionali intelligenti* (« interface utilisateur conversationnelle »). Dans la liste, nous avons également trouvé beaucoup de candidats qui ne pouvaient pas être des termes : *qualche modo* (« en quelque sorte »), *diciannovesimo secolo* (« dix-neuvième siècle »), *cos'è* (« qu'est-ce que c'est »), *ottocento* (« dix-neuvième siècle ») et *due tipi* (« deux types »). Nous avons également trouvé des candidats termes en anglais, par exemple, *chunk identification* qui est devenu *analisi di una proposizione* (« analyse syntaxique de surface ») et *annotation* qui est devenu *annotazione* (« annotation »). Pour arriver à 50 termes, il était aussi nécessaire de faire une recherche manuelle dans le corpus : nous avons ainsi trouvé d'autres termes pertinents, par exemple, *reti convolutive* (« réseaux de neurones convolutifs »), *estrazione delle relazioni* (« extraction de relations »), *matrice di confusione* (« matrice de confusion »), *riassunto del testo* (« résumé de texte ») et *sistema di dialogo* (« système de dialogue »).

En parlant de la liste de candidats termes anglais, nous pouvons dire que nous avons suivi les mêmes passages pour arriver à trouver 50 bons termes anglais. Nous avons dû exclure de nombreux candidats qui n'étaient pas pertinents pour le domaine ou qui n'étaient pas de termes proprement à parler.

L'extraction de termes n'était pas si facile. Les extracteurs automatiques présentent des limites et des problèmes, comme nous avons déjà exploré dans le chapitre quatre. Toutefois, dans le paragraphe précédent, nous avons présenté des exemples pratiques et concrets tirés du projet *YourTerm TECH*. Ils nous aident à mieux comprendre comment fonctionne l'extraction de termes et comment nous pouvons intervenir pour l'améliorer.

5.3 Analyse de la compilation de fiches terminologiques

La compilation de fiches terminologiques s'est déroulée correctement et sans trop d'accrocs, à l'exception de quelques problèmes rencontrés. Comme indiqué plus haut, la compilation de fiches terminologiques a été exécutée grâce à l'emploi de FAIRterm, un instrument fondamental pour l'organisation optimale des données terminologiques. Tout d'abord, nous devons souligner que les 150 termes appartenant au domaine de l'intelligence artificielle et du traitement automatique du langage ne sont pas tous simples à expliquer, surtout sous un point de vue de vulgarisation. Ces deux domaines sont encore des domaines de naissance récente ou, mieux, des domaines en constante évolution et mise à jour. Pour cela leur terminologie peut changer et elle s'enrichit continuellement de nouveaux concepts et désignations. Le travail des terminographes ne s'arrête jamais, car ils/elles doivent étudier de manière constante. Cette mise à jour concerne également les dictionnaires et les banques de terminologie, qui ont besoin d'être revus. Les domaines de l'IA et du TAL ne sont pas de domaines faciles.

La compilation de fiches terminologiques était presque toujours fluide et n'était pas très difficile, cependant nous avons rencontré des problèmes lorsque nous devions compiler certains champs terminologiques, car certains termes ont présenté des problématiques pour ce qui concerne leur description générale.

5.3.1 Quelques cas concrets

Une des parties les plus difficiles à faire face est sans aucun doute l'étymologie de termes. Nous avons indiqué les nombres de termes simples et de termes complexes et ces derniers sont nettement supérieurs en nombre. L'étymologie étudie l'origine des mots et lorsque nous examinons des termes simples, il sera très facile à l'expliquer, car il s'agit d'un seul mot graphique. Si, en revanche, nous sommes confrontés à l'analyse d'un terme complexe, à savoir formé par deux unités lexicales ou plusieurs, en compliquant en outre la situation, nous aurons des empêchements à donner une seule et bonne étymologie. Voici un exemple tiré du projet *YourTerm TECH* : la fiche terminologique du terme *modèles de plongement prédictif de mots*. C'est un syntagme nominal très long qui contient deux syntagmes prépositionnels. Il s'agit donc d'un terme complexe. Son champ de l'étymologie résulte interminable, car il n'existe pas une étymologie unique pour un si long terme, il n'existe pas une origine unique. Quelle est donc la solution ? Afin de fournir une fiche terminologique la plus complète possible, nous pouvons insérer plusieurs

étymologies : dans ce cas-là, nous avons indiqué l'étymologie de chaque mot formant le terme complexe en question. Ce même problème est présent soit dans la langue source que dans la langue cible : le correspondant italien du terme est *rappresentazione distribuita delle parole*.

D'autres problèmes concernent la définition des termes, surtout des termes complexes. C'est le cas de la fiche terminologique du terme *graphes de similarité* et son très long correspondant italien *modello di classificazione basato su grafi per l'elaborazione del testo*, une technique employée par le TAL. Pendant la recherche d'informations sur le terme, nous n'avons pas trouvé une définition adéquate. Par conséquent, lorsqu'il n'existe pas une définition, les terminographes sont obligés de la créer, car c'est une partie essentielle pour comprendre le concept analysé dans la fiche. Nous avons donc trouvé un document PDF en ligne qui s'occupe de cette technique particulière et nous avons pris des morceaux d'informations pour ensuite rassembler une bonne définition du concept. Dans ces cas-là, il faut indiquer dans le champ *Note (definition)* que la définition est le fruit du travail personnel du/de la terminographe. La même problématique s'est présentée pour le terme italien qui est d'ailleurs très long et difficile à définir. En plus, pour la compilation de cette fiche terminologique nous avons aussi utilisé plutôt les algorithmes spécifiques avec un nom propre, par exemple, *TextRank* et *PageRank*, qui sont des algorithmes basés sur les graphes afin d'identifier les mots et les phrases les plus importants. Ils nous ont aidés à mieux comprendre le concept pour créer une définition satisfaisante.

Puis, nous avons rencontré des difficultés dans la compilation du champ relatif à l'analyse sémique, qui consiste à décomposer le sens d'un terme dans de petits éléments de sens appelés composants sémantiques. Nous avons vu qu'il existe plusieurs typologies de composants sémantiques : les sèmes constants, divisés à leur fois en sèmes génériques et sèmes spécifiques, et les sèmes contextuels, qui comprennent aussi les sèmes connotatifs. Les termes simples sont plus faciles à examiner sous un point de vue sémique : nous avons présenté dans le chapitre quatre les deux exemples « bistouri » et « marguerite ». Cependant, les termes complexes sont caractérisés par des composants sémantiques plus difficiles à identifier. Voici un exemple : la fiche terminologique du terme italien *sistemi conversazionali intelligenti* et son équivalent français « interface utilisateur conversationnelle ». Les sèmes de ces termes doivent être très précis en vue de

la possibilité des utilisateurs de cette fiche de reconstruire le sens total du terme à partir de ses atomes de sens. Il y a le risque d'arriver à écrire de nombreux sèmes, en présentant une analyse sémique très longue et parfois déroutante. Il faut donc faire attention au sens du terme en le divisant et aussi il faut bien suivre la définition.

De toute évidence, les problèmes se posent également lorsque nous devons trouver les synonymes ou les quasi-synonymes des termes. Nous avons parlé de la synonymie et de la polysémie et comment la terminologie en s'occupe, mais maintenant nous présenterons des exemples réels. Les langues de spécialité, comme la langue scientifique, se caractérisent par l'utilisation d'une terminologie standardisée afin de rendre la communication entre les experts efficace. Pour cette raison-là, les termes spécialisés ont une caractéristique très importante : la précision. Ce trait implique la présence d'une seule désignation pour chaque concept en termes d'efficacité communicative. Toutefois, ce n'est pas toujours le cas : les termes spécialisés peuvent avoir des synonymes. Un exemple est représenté par la fiche terminologique du terme italien *reti convolutiv*, en français « réseaux de neurones convolutifs ». Ces termes ont des synonymes, à savoir *reti neurali convolutive* et *rete neurale convoluzionale* pour l'italien et « réseau convolutif » pour le français et ils n'ont pas de quasi-synonymes. Quand il est temps d'aborder l'étude des synonymes d'un terme, il faut être prudent : le synonyme d'un terme comporte son emploi dans les mêmes domaines, le même usage et le même sens ; tandis que le quasi-synonyme peut être utilisé à la place du terme, mais les deux ne désignent pas totalement le même concept. La fiche terminologique du terme français *données textuelles* et son correspondant italien *dati testuali* n'indiquent pas les synonymes, mais elle indique les quasi-synonymes : « informations textuelles » et *« materiale testuale »*. Pour ce qui concerne la langue française, « donné » et « information » sont deux concepts différents, mais ils peuvent être interchangés et la même chose s'applique à la langue italienne.

Le dernier champ qui a causé quelques troubles est le champ *Collocation*, où nous insérons des combinaisons d'unités lexicales qui enrichissent l'usage des termes. Un exemple très simple est la fiche terminologique du terme *corpus* : une de ses collocations sera sans doute « construire un corpus ». Un autre exemple est la fiche du terme *dérivation lexicale* : une de ses collocations peut être l'expression « la composition et la dérivation », car les deux apparaissent souvent ensemble. Un exemple d'un terme source italien est *clustering dei documenti* (« partitionnement des documents » en

français) : une possible collocation peut être *« metodi di clustering dei documenti »* (en français, « méthodes de partitionnement des documents »). Ainsi expliquée, la collocation semble ne poser aucun problème, mais en réalité, il faut être prudent, car une collocation incorrecte peut désigner un concept totalement différent. Par exemple, le terme *algorithme* ne peut pas avoir comme collocation « algorithme génératif », car c'est un autre concept, avec une propre définition et de propres traits. La collocation du terme *algorithme* sera « écrire un algorithme ».

Pour ce qui concerne les autres champs, par exemple, les champs relatifs à l'hyperonymie, à la méronymie, aux abréviations, aux domaines et sous-domaines et au contexte n'étaient pas si problématiques. Toutefois, il faut dire que la compilation de fiches terminologiques doit être claire et précise afin de fournir et garantir des données terminographiques qui peuvent être réutilisées par d'autres.

5.4 Considérations finales sur le projet

Le travail effectué pour ce mémoire sur la terminologie de l'IA et du TAL et le travail effectué pour le projet *YourTerm TECH* ont nécessité et englobé de nombreuses tâches et différentes ressources. La rédaction de ce mémoire et la réalisation du projet terminographique nous ont permis d'étudier deux domaines d'étude qui sont réservés le plus souvent aux experts, étudiants et chercheurs ou aux passionnés de ces thèmes en question. En effet, l'intelligence artificielle et le traitement automatique du langage naturel sont deux champs d'études caractérisés de la complexité tant en ce qui concerne la partie scientifique générale, les concepts et les applications, qu'en ce qui concerne sa terminologie et vulgarisation, car les non-experts peuvent trouver difficile comprendre les concepts et le lexique. Les objectifs du projet n'étaient donc pas si faciles à atteindre. Toutefois, les outils utilisés et la méthodologie adoptée nous ont permis de travailler de manière logique et ordonnée.

5.4.1 Les outils employés

Pour atteindre les objectifs fixés au début de ce projet, deux outils étaient essentiels : ils sont TermoStat et FAIRterm.

Le logiciel TermoStat est l'outil qui nous a permis de procéder à l'acquisition automatique de termes en ligne. Ce logiciel est utilisable seulement après enregistrement et son utilisation est toujours gratuite pour des fins de recherche. En se connectant, TermoStat nous fournit une interface graphique très simple et intuitive et, après

l'authentification, nous pouvons commencer à travailler. TermoStat comprend cinq langues de travail : le français, l'italien, l'anglais, l'espagnol et le portugais. Il nous fournit la possibilité de choisir si nous voulons extraire seulement les termes simples ou seulement les termes complexes, ou encore, les deux ensembles, et il nous explique quelques instructions simples sur le format des documents à télécharger. Bien que cet extracteur de termes ne soit pas totalement précis, les résultats qu'il fournit sont tout à fait satisfaisants. En effet, la plupart des termes choisis pour la compilation des fiches terminologiques ont été trouvés grâce à son utilisation et, comme nous l'avons vu, le nettoyage manuel n'a servi que pour quelques questions terminologiques à régler. L'emploi de TermoStat était donc fondamental pour le projet terminologique surtout en termes de rapidité, de commodité et d'analyse des données terminologiques.

FAIRterm est l'application qui nous a permis de compiler des fiches terminologiques de haut niveau. Comme TermoStat, FAIRterm nécessite l'authentification pour pouvoir travailler. Ce logiciel en ligne est très simple à utiliser, son interface graphique est claire et compréhensible, mais pas banal. La structure de fiches terminologiques offertes par le logiciel est vraiment complète, car elle fournit aux utilisateurs toutes les informations qui sont nécessaires pour analyser tous azimuts les termes d'un domaine spécialisé. Les champs des fiches terminologiques permettent aux terminographes de suivre un travail terminographique linéaire et hiérarchisé, afin d'examiner les termes avec ordre et de façon raisonnable. Tous les champs sont compréhensibles et bien définis : les utilisateurs ont la possibilité de naviguer librement sur le site du logiciel sans aucune contrainte en termes de compilation, c'est-à-dire que si une terminographe veut s'occuper premièrement des champs dédiés à la variation (nom commun et nom scientifique, variantes orthographiques, formes complètes, abréviations), elle peut le faire et, plus tard, elle peut retourner aux sections précédentes. Les fiches terminologiques peuvent toujours être modifiées : dans les cas de mise à jour, il sera suffisant d'accéder à la plateforme et de rechercher les termes déjà téléchargés pour les modifier ou corriger. Pour ce qui concerne la traduction spécialisée, FAIRterm peut se révéler un instrument essentiel. Tout d'abord, les champs d'une fiche terminologique FAIRterm garantit la bonne compréhension et utilisation d'un terme, car ils offrent une explication des termes complets. En second lieu, le logiciel et ses caractéristiques favorisent la bonne traduction des termes. En plus, lorsque les fiches terminologiques ont

été enregistrées et stockées dans le logiciel, les utilisateurs ont la possibilité d'exporter les données terminologiques sous forme tabulaire, grâce à la fonction *Download TSV* (*Tab separated values*), et au format standard *TermBase Exchange*, grâce à la fonction *Download TBX*. Cette opportunité très avantageuse permet aux utilisateurs d'importer des données structurées afin de les télécharger dans les systèmes de Traduction Assistée par Ordinateur (TAO)[22] afin de bénéficier des termes déjà examinés et traduits au préalable. Nous pouvons donc apprendre que ce logiciel est une ressource capitale et puissante pas seulement pour le travail des terminographes, mais aussi pour ceux qui s'occupent de traduction spécialisée, en utilisant des données terminologiques précises, complètes et prêt à l'emploi. Pour conclure, l'expérience pratique relative à l'utilisation de FAIRterm s'est révélée exhaustive et performante à tous égards. Comme nous avons déjà indiqué en précédence, l'outil a l'objectif d'offrir des données de la recherche trouvables, accessibles, interopérables et réutilisables, et ce but est pleinement atteint grâce à l'ingéniosité des créateurs du logiciel (cf. Vezzani 2021, 55).

5.4.2 Évaluation de la méthodologie

La méthodologie proposée pour la réalisation du projet *YourTerm TECH* est plutôt simple à mettre en œuvre et caractérisée par de diverses étapes.

Tout d'abord, une des parties les plus importantes pour le travail terminologique est la recherche. Les terminographes doivent se dédier à la recherche de façon curieuse, ils/elles devraient vouloir se nourrir de nouveaux contenus et concepts, afin d'enrichir leur bagage culturel. Voici ce que nous avons fait : compte tenu de notre méconnaissance des domaines de l'IA et du TAL, nous nous sommes lancés dans une recherche approfondie grâce à laquelle nous avons pu comprendre nombreux contenus à l'égard, afin de rédiger ce mémoire et afin de réaliser un bon projet terminographique.

Puis, la collecte des documents pour la création des trois corpus était longue, mais elle s'est révélée essentielle pour le projet. En plus, le contrôle de la fiabilité des sources des documents n'était pas difficile, car nous avions à disposition une série d'étapes pour la vérifier. Le contrôle était donc fluide et ordonné.

Enfin, la compilation de fiches terminologiques a été très utile pour mieux comprendre certains concepts des deux domaines d'étude, car nous avons fait beaucoup de recherches, pas seulement pour trouver les bons documents à fournir à l'extracteur de

[22] En anglais ils s'appellent *computer-aided translation tools* (CAT).

termes, mais aussi pour traduire les termes sources. Nous nous sommes principalement concentrés sur les définitions des termes, qui sont des éléments fondamentaux pour comprendre les concepts et leurs désignations et pour permettre à d'autres chercheurs ou étudiants de les utiliser, en leur fournissant des définitions complètes et sans ambiguïtés. Ce travail de compilation des fiches terminologiques peut apparaître est un travail long et articulé. Cependant, il permet de bien réussir dans le monde de la terminologie.

Dans ce chapitre final, nous avons ainsi fait un résumé du travail effectué pour le mémoire et pour le projet. Nous avons fait une analyse approfondie des trois corpus, en présentant tous les documents qui les composent. Nous avons fourni les motivations pour lesquelles nous avons choisi ces documents, en présentant les sources pertinentes et leur histoire. Puis, l'extraction de termes et la compilation de fiches terminologiques ont été présentées. Nous avons expliqué la modalité de l'extraction automatique de termes et la nécessité d'un nettoyage manuel qui aide les terminographes à rendre encore plus parfait le travail terminographique. En plus, nous avons analysé des cas concrets concernant la compilation de fiches terminologiques. Pour conclure, nous avons fait une évaluation finale totale des travaux pour tirer les ficelles du projet achevé.

Conclusion

Le présent mémoire se concentre sur le domaine de la terminologie de deux domaines particuliers et sur la réalisation d'un projet terminographique spécifique qui soit capable de fournir des données terminologiques satisfaisantes. Les domaines en question sont l'intelligence artificielle et le traitement automatique du langage. Ces deux domaines sont devenus partie intégrante dans la vie de tous et au niveau mondial, car ils représentent des terrains fertiles pour le développement et l'économie dans sa totalité.

Pendant la rédaction du mémoire et surtout pendant la recherche générale, en relation avec l'écriture des chapitres autant qu'avec la réalisation du projet terminographique, nous avons compris à quel point il peut être difficile d'aborder ces deux disciplines scientifiques. Il s'agit en effet des deux champs d'études complexes en termes de compréhension de concepts et de diffusion de connaissances. Il est donc nécessaire de garantir un point de rencontre entre les exigences de compréhension des citoyens et le lexique parfois compliqué de l'intelligence artificielle et du traitement automatique du langage. C'est le fil conducteur qui nous a accompagnés tout au long de ce parcours.

Afin de ne pas perdre de vue notre objectif et pour l'atteindre de la meilleure façon possible, nous avons donc commencé par examiner les domaines de l'intelligence artificielle et du traitement automatique du langage. Il s'agit d'une étape essentielle pour réaliser un projet terminologique digne. Nous avons présenté les deux domaines en fournissant des définitions, un itinéraire historique complet et les caractéristiques les plus importantes à connaître. Il était également fondamental d'expliquer comment ils fonctionnent et quelles sont leurs applications, car ils se sont révélés précieux pas seulement dans le domaine de l'informatique, mais aussi dans d'autres domaines qui peuvent n'avoir aucun rapport avec les disciplines purement liées à la technologie et à l'informatique. L'étude de l'intelligence artificielle et du traitement automatique du langage nous a permis de mieux travailler sur le projet terminographique. De toute évidence, nous ne connaissions que les concepts les plus superficiels et les plus simples, mais la recherche et l'analyse relatives à ces domaines nous ont aidés à les comprendre de manière plus approfondie.

Après avoir consacré nos efforts à l'étude de ces deux domaines, nous nous sommes intéressés à l'autre discipline qui nous a guidés au cours de la création d'une série des termes et des données terminologiques : la terminologie. Nous avons analysé la discipline de la terminologie en présentant en premier lieu sa définition et son histoire. Puis, nous avons présenté la théorie générale qui se positionne à la base de la terminologie, mais aussi les cinq autres théories alternatives qui se proposent de donner quelque chose de plus lorsque nous voulons traiter la théorie de la terminologie. L'analyse de la terminologie nous a portés aussi à examiner la partie pratique, c'est-à-dire la terminographie. Elle a été d'une grande importance pour comprendre comment faire notre projet terminographique. Nous avons également exploré la forte relation qu'il y a entre la terminologie et l'informatique : aujourd'hui, la terminotique permet aux terminographes et aux chercheurs en général de travailler sur des textes ayant un format électronique et elle met à disposition des terminographes de nombreuses ressources informatiques qui leur permettent de systématiser et d'accélérer le travail terminographique. De plus, pour compléter le cadre de la présentation générale, nous avons examiné les nombreuses applications de la terminologie et de la terminographie dans d'autres domaines. Ce sont donc des disciplines qui caractérisent tous les domaines de la vie humaine et toutes ses activités. La terminologie en particulier est considérée comme quelque chose à exploiter afin de diffuser les connaissances de tous domaines, soit dans la communication spécialisée entre experts, soit dans la vulgarisation entre experts et non. Pour terminer, nous avons ajouté une partie dédiée spécifiquement à la terminologie de l'intelligence artificielle et du traitement automatique du langage naturel, en présentant les problèmes les plus grands que nous devons affronter, par exemple, les anglicismes.

Le projet terminographique sur l'intelligence artificielle et sur le traitement automatique du langage s'est composé de plusieurs étapes : la recherche de documents, l'analyse de la fiabilité de documents, la construction de corpus, l'extraction de la terminologie liée aux domaines d'étude, la compilation de fiches terminologiques et l'analyse qualitative totale des travaux. Nous avons donc fait une analyse approfondie du concept de corpus, de la linguistique de corpus et des méthodes employées pour constituer des corpus de haute qualité. Ce passage a été fondamental pour mieux apprendre le corpus, les différentes typologies de corpus existantes, en analysant en particulier le corpus spécialisé qui est composé d'énoncés relatifs à un domaine spécialisé et qui doit

être un bon représentant de la langue de spécialité employée par un certain domaine. En ce qui concerne le projet auquel nous avons adhéré, nous avons travaillé durement pour construire trois corpus spécialisés dans les trois langues d'étude choisies (français, italien et anglais). En ce qui concerne la partie dédiée à la construction d'un corpus, elle était encore plus essentielle, car nous avons pu intégrer nos déjà acquis connaissances sur le sujet et elle nous a permis de concevoir trois corpus qui étaient prêts à utiliser pour l'étape suivante du projet : l'extraction automatique de termes. Le corpus est un outil précieux pour les terminographes qui veulent s'occuper d'un domaine spécialisé qu'ils/elles ne connaissent pas. Un corpus permet de s'approprier de nombreuses connaissances et informations qui seront de fondamentale importance pour réaliser un projet terminographique. Toutefois, pour faire cela, les terminographes doivent travailler très attentivement : il faut choisir les justes textes ; il faut bien les organiser ; et il faut contrôler méticuleusement les sources et leur fiabilité.

L'autre passage fondamental pour la réalisation du projet et pour une bonne rédaction du mémoire était l'extraction de termes. Pour mieux comprendre ce sujet et pour travailler de manière plus efficiente, nous nous sommes d'abord concentrés sur l'étude minutieuse des termes. Nous avons compris comment ils peuvent être formés, les différentes typologies de termes existantes et comment les identifier. Toutefois, la partie consacrée à l'extraction automatique de termes était la plus importante, car elle nous a permis de bien examiner comment elle fonctionne. Une grande contribution à l'extraction terminologique est représentée par un logiciel d'extraction automatique de termes qui nous a aidés pendant le projet : le logiciel TermoStat. Une fois les corpus prêts, nous avons pu les télécharger sur TermoStat, qui nous a fourni les candidats termes qui pourront devenir des termes objet d'étude grâce à une série de méthodes, par exemple, l'analyse de la fréquence. TermoStat était capable de sélectionner de nombreux candidats termes qui étaient représentatifs du domaine de l'intelligence artificielle et du domaine du traitement automatique du langage. Toutefois, tous ces candidats termes n'étaient pas de vrais termes : certains étaient incomplets, certains n'étaient pas de termes tout à fait pertinents pour les deux domaines cités, et certains ont été écartés parce qu'ils étaient trop simples. Le logiciel TermoStat a donc joué un rôle de premier plan dans l'étape de l'extraction des termes. Toutefois, la sélection automatique de termes faite par l'extracteur en ligne a été analysée : un nettoyage manuel effectué par l'être humain sur

une liste des candidats termes est toujours conseillé. Dans le cadre du projet *YourTerm TECH*, le nettoyage manuel nous a permis de compléter la liste de termes à analyser et à les revoir de manière approfondie afin d'être sûr de présenter des termes pertinents et valables.

Lorsque nous avons notre liste de termes révisés, nous avons pu passer à la compilation de fiches terminologiques. Ce travail de compilation est essentiel pour l'étude de la terminologie de tous domaines spécialisés, car il permet de créer des données terminologiques pertinentes, mais aussi de les insérer dans des ressources terminologiques qui font autorité, par exemple, les banques de terminologie comme le projet européen IATE. Ces ressources terminologiques ne seront pas utilisées seulement par les terminographes, mais tous y auront accès, à savoir, les étudiants, les enseignants, les traducteurs et les chercheurs. Le logiciel que nous a permis de compiler les fiches terminologiques pour le projet *YourTerm TECH* est FAIRterm, un instrument qui s'est avéré infaillible dans le cadre d'un projet de terminographie. Grâce à FAIRterm, nous avons compris l'importance de chaque nuance des termes : tous les éléments liés à la sémantique d'un terme, à sa variation et à son usage sont utiles pour la compréhension d'un concept, surtout si ce dernier est particulièrement compliqué. De toute évidence, la compilation de fiches terminologiques pour des termes du domaine de l'intelligence artificielle et du domaine du traitement automatique du langage était considérablement difficile pour certains aspects, car certains concepts sont vraiment épineux. Toutefois, le travail de recherche que nous avons effectué avant nous a aidés à franchir cet obstacle.

Après avoir complété chaque étape du travail terminographique, nous avons pu faire nos considérations finales. À la fin, nous avons compris l'importance de la terminologie et également sa complexité. Lorsque nous sommes confrontés à l'étude d'un domaine spécialisé, il faut avoir une base théorique solide, et aussi une bonne méthodologie. Un bon travail terminologique exige une collecte exhaustive de textes spécialisés qui représentent le domaine d'étude, un pratique et efficient système de gestion de corpus, l'intervention de l'être humain qui ne doit pas manquer et des données terminologiques bien décrites.

Un aspect important qui caractérise les travaux des terminographes est la possibilité qu'ils/elles ont de partager leurs efforts. Avoir accès aux données terminologiques est essentiel pour tous. Pour cette raison, le projet *YourTerm TECH* proposé par *TermCoord*

et ses objectifs nous a inspirés à fournir un ensemble de termes solide concernant l'intelligence artificielle et le traitement automatique du langage, car il s'agit des deux disciplines compliquées et parfois inaccessibles en termes de compréhension. Ainsi, le grand public pourra se connecter au monde informatique. Les experts, les experts d'autres domaines, les chercheurs et aussi les passionnés (non-experts) auront l'opportunité de tirer profit de ces ressources pour une variété d'objectifs.

Annexes

Annexe A – Liste de termes français pour FAIRterm

Termes simples	Termes complexes
algorithme	agent intelligent
compréhension	algorithme génératif
corpus	algorithme génétique
lemmatisation	analyse logique
mégadonnées	apprentissage non supervisé
morphologie	apprentissage supervisé
nettoyage	arbres syntaxiques
ontologie	automatisation des processus intelligents
prétraitements	base de connaissance
synonymie	collecte (des données)
TF-IDF	données textuelles
tokenisation	étiquetage par transformation
/	expressions régulières
/	extraction de mots-clés
/	graphes de similarité
/	identification de mots
/	indexation automatique de documents
/	intelligence artificielle
/	intelligence artificielle distribuée
/	interaction homme-machine
/	linguistique de corpus
/	linguistique distributionnelle
/	linguistique informatique
/	méthodes basées sur des règles
/	méthodes stochastiques
/	modèles de classification binaire
/	modèles de concepts

/	modèles de plongement prédictif de mots
/	modèles statistiques de langage
/	normalisation des données
/	petites données
/	processus de digitalisation
/	reconnaissance optique de caractères
/	reconnaissance vocale
/	réseau bayésien
/	réseaux de neurones
/	réseaux de neurones artificiels
/	traitement automatique du langage

Annexe B – Liste de termes italiens pour FAIRterm

Termes simples	Termes complexes
annotazione	albero di decisione
derivazione	algoritmo AdaBoost
etichettare	analisi del sentimento
flessione	analisi di una proposizione
interpretabilità	analisi lessicale
lemma	analisi morfologica
linguistica	analisi semantica
metadati	analisi sintattica
microprocessore	apprendimento di trasferimento
/	apprendimento semi-supervisionato
/	assistente virtuale
/	classificazione testuale
/	clustering dei documenti
/	comprensione del linguaggio naturale
/	controllo ortografico
/	dati non strutturati
/	estrazione delle relazioni

/	grammatica generativa
/	indagine empirica
/	linguaggio di programmazione
/	linguaggio umano
/	matrice di confusione
/	modelli pre-addestrati
/	modelli statistici
/	modello computazionale
/	modello sequenza-sequenza
/	moderazione dei contenuti
/	programmazione lineare intera
/	reti convolutive
/	reti neurali profonde
/	reti neurali ricorrenti
/	riassunto del testo
/	ricerca vocale
/	riconoscimento dei pattern
/	rilevamento dello spam
/	rimozione delle stop-word
/	risoluzione delle coreferenze
/	sintesi automatica della voce
/	sistema di dialogo
/	sistemi conversazionali intelligenti
/	training set

Annexe C – Liste de termes anglais pour FAIRterm

Termes simples	Termes complexes
ambiguity	anaphora resolution
chatbot	bag of words model
grammar	Chomsky hierarchy
inference	cognitive computing
lexicon	conditional random fields
n-gram	data mining
parser	data set
perceptron	deep learning
phonology	discriminative method
polysemy	entity linking
pragmatics	generative method
semantics	information extraction
stem	information retrieval
stemming	language detection
stopword	machine learning
syntax	machine translation
token	multitask learning
/	named entity recognition
/	natural language generation
/	neural architecture search
/	neural language model
/	noisy data
/	parts-of-speech tagging
/	pruning method
/	question answering
/	reinforcement learning
/	Smart Search
/	support vector machine
/	text mining

/	topic model
/	web scraping
/	word error rate
/	word sense disambiguation

Annexe D – Les 150 termes et leur traduction

Français	Italien	Anglais
agent intelligent	agente intelligente	intelligent agent
algorithme	algoritmo	algorithm
algorithme génératif	algoritmo generativo	generative algorithm
algorithme génétique	algoritmo genetico	genetic algorithm
analyse logique	analisi logica	semantic role labeling
apprentissage non supervisé	apprendimento non supervisionato	unsupervised learning
apprentissage supervisé	apprendimento supervisionato	supervised learning
arbres syntaxiques	alberi sintattici	parse trees
automatisation des processus intelligents	automazione intelligente dei processi	Intelligent Process Automation
base de connaissance	base di conoscenza	knowledge base
collecte (des données)	raccolta dei dati	data collection
compréhension	comprensione (del contenuto)	comprehension
corpus	corpus	corpus
données textuelles	dati testuali	textual data
étiquetage par transformation	apprendimento basato sulla trasformazione	transformation-based tagging
expressions régulières	espressioni regolari	regular expressions
extraction de mots-clés	estrazione di parole chiave	keyword extraction

graphes de similarité	modello di classificazione basato su grafi per l'elaborazione del testo	graph-based ranking algorithm
identification de mots	riconoscimento della parola	word recognition
indexation automatique de documents	indicizzazione automatica dei documenti	search engine indexing
intelligence artificielle	intelligenza artificiale	artificial intelligence
intelligence artificielle distribuée	intelligenza artificiale distribuita	Distributed Artificial Intelligence
interaction homme-machine	interazione uomo-computer	human-computer interaction
lemmatisation	lemmatizzazione	lemmatization
linguistique de corpus	linguistica dei corpora	corpus linguistics
linguistique distributionnelle	semantica distribuzionale	distributional semantics
linguistique informatique	linguistica computazionale	computational linguistics
mégadonnées	megadati/big data	big data
méthodes basées sur des règles	sistemi basati su regole	rule-based systems
méthodes stochastiques	modelli stocastici	stochastic systems
modèles de classification binaire	classificazione binaria	binary classification
modèles de concepts	database semantico-lessicale	lexical database
modèles de plongement prédicatif de mots	rappresentazione distribuita delle parole	word embedding
modèles statistiques de langage	modelli di linguaggio	statistical language model
morphologie	morfologia	morphology
nettoyage (de données)	pulizia dei dati	data cleaning

normalisation des données	normalizzazione (dei dati)	normalization of data
ontologie	ontologia	ontology
petites données	small data	Small Data
prétraitements	pre-elaborazione	pretreatment (of textual documents)
processus de digitalisation	digitalizzazione	digitization
reconnaissance optique de caractères	riconoscimento ottico dei caratteri	optical character recognition
reconnaissance vocale	riconoscimento vocale	speech recognition
réseau bayésien	rete bayesiana	Bayesian network
réseaux de neurones	reti neurali	neural network
réseaux de neurones artificiels	reti neurali artificiali	artificial neural networks
synonymie	sinonimia	synonymy
TF-IDF	funzione di peso tf-idf	term frequency-inverse document frequency
tokenisation	tokenizzazione	tokenization
traitement automatique du langage	trattamento automatico del linguaggio	natural language processing
arbre de décision	*albero di decisione*	decision tree
AdaBoost	*algoritmo AdaBoost*	AdaBoost
analyse des sentiments	*analisi del sentimento*	sentiment analysis
analyse syntaxique de surface	*analisi di una proposizione*	chunking
analyse lexicale	*analisi lessicale*	lexical analysis
analyse morphologique	*analisi morfologica*	morphological decomposition
analyse sémantique	*analisi semantica*	semantic analysis
analyse syntaxique	*analisi sintattica*	syntactic analysis
annotation	*annotazione*	annotation

apprentissage par transfert	*apprendimento di trasferimento*	transfer learning
apprentissage semi-supervisé	*apprendimento semi-supervisionato*	semi-supervised learning
assistant personnel intelligent	*assistente virtuale*	virtual assistant
classification de textes	*classificazione testuale*	text classification
partitionnement des documents	*clustering dei documenti*	document clustering
compréhension du langage naturel	*comprensione del linguaggio naturale*	natural language understanding
correcteur	*controllo ortografico*	spell checker
données non structurées	*dati non strutturati*	unstructured data
dérivation lexicale	*derivazione*	morphological derivation
extraction de relations	*estrazione delle relazioni*	relationship extraction
étiqueter	*etichettare*	to tag
flexion	*flessione*	inflection
grammaire générative	*grammatica generativa*	generative grammar
recherche empirique	*indagine empirica*	empirical research
interprétabilité	*interpretabilità*	interpretability
lemme	*lemma*	lemma
langage de programmation	*linguaggio di programmazione*	programming language
langage humain	*linguaggio umano*	human language
linguistique	*linguistica*	linguistics
matrice de confusion	*matrice di confusione*	confusion table
métadonnées	*metadati*	metadata
microprocesseur	*microprocessore*	microprocessor
modèles préentraînés	*modelli pre-addestrati*	pre-trained models
modèles statistiques	*modelli statistici*	statistical models
modèle computationnel	*modello computazionale*	computational model

modèle séquence à séquence	*modello sequenza-sequenza*	Sequence to Sequence model
modération de contenu	*moderazione dei contenuti*	content moderation
programmation linéaire en nombres entiers	*programmazione lineare intera*	Integer Linear Programming
réseaux de neurones convolutifs	*reti convolutive*	convolutional neural network
réseaux de neurones profonds	*reti neurali profonde*	deep neural networks
réseaux de neurones récurrents	*reti neurali ricorrenti*	recurrent neural network
résumé de texte	*riassunto del testo*	text summarization
recherche vocale	*ricerca vocale*	voice search
reconnaissance de formes	*riconoscimento dei pattern*	pattern recognition
détection de spams	*rilevamento dello spam*	spam detection
suppression de mots vides	*rimozione delle stop-word*	stopword removal
résolution de coréférence	*risoluzione delle coreferenze*	co-reference resolution
synthèse vocale	*sintesi automatica della voce*	text-to-speech
système de dialogue	*sistema di dialogo*	dialogue system
interface utilisateur conversationnelle	*sistemi conversazionali intelligenti*	conversational user interface
jeu de données d'apprentissage	*training set*	training data
ambiguïté	ambiguità	*ambiguity*
résolution d'anaphores	risoluzione delle anafore	*anaphora resolution*
modèle du sac de mots	modello della borsa di parole	*bag of words model*
agent conversationnel	chatbot	*chatbot*
hiérarchie de Chomsky	gerarchia di Chomsky	*Chomsky hierarchy*

informatique cognitive	informatica cognitiva	*cognitive computing*
champs aléatoires conditionnels	campi casuali condizionali	*conditional random fields*
exploration de données	estrazione di dati	*data mining*
jeu de données	insieme di dati	*data set*
apprentissage profond	apprendimento profondo	*deep learning*
modèle discriminatif	modello discriminativo	*discriminative method*
annotation sémantique	collegamento di entità	*entity linking*
modèle génératif	modello generativo	*generative method*
grammaire formelle	grammatica formale	*grammar*
inférence	inferenza	*inference*
extraction d'information	estrazione di informazioni	*information extraction*
recherche d'information	reperimento dell'informazione	*information retrieval*
		intent monitoring
reconnaissance de langue	riconoscimento della lingua	*language detection*
lexique	lessico	*lexicon*
apprentissage automatique	apprendimento automatico	*machine learning*
traduction automatique	traduzione automatica	*machine translation*
apprentissage multi-tâches	apprendimento multitasking	*multitask learning*
reconnaissance d'entités nommées	riconoscimento di entità denominate	*named entity recognition*
génération automatique de textes	generazione del linguaggio naturale	*natural language generation*
recherche automatique d'architecture neuronale	ricerca dell'architettura neurale	*neural architecture search*
modèle de langue neuronal	modello neurale del linguaggio	*neural language model*
n-gramme	n-gramma	*n-gram*

donnée bruitée	dati rumorosi	*noisy data*
analyseur syntaxique	parser	*parser*
étiquetage morpho-syntaxique	analisi grammaticale	*parts-of-speech tagging*
perceptron	percettrone	*perceptron*
phonologie	fonologia	*phonology*
polysémie	polisemia	*polysemy*
pragmatique	pragmatica	*pragmatics*
méthode d'élagage	metodo di potatura	*pruning method*
système de questions-réponses	sistemi di Question Answering	*question answering*
apprentissage par renforcement	apprendimento per rinforzo	*reinforcement learning*
sémantique	semantica	*semantics*
recherche intelligente	ricerca intelligente	*Smart Search*
radical	radice	*stem*
racinisation	stemming	*stemming*
mot vide	stopword	*stopword*
machines à vecteurs de support	macchine a vettori di supporto	*support vector machine*
syntaxe	sintassi	*syntax*
fouille de textes	estrazione di testo	*text mining*
token	token lessicale	*token*
modèle thématique	topic model	*topic model*
moissonnage du Web	web scraping	*web scraping*
taux d'erreur de mots	tasso di parole errate	*word error rate*
désambiguïsation lexicale	disambiguazione del senso della parola	*word sense disambiguation*

Annexe E – Bibliographie du corpus en français
Chakchouk, Moez, *Réflexion éthique sur l'intelligence artificielle*, dans Cités, n° 80, L'intelligence artificielle : enjeux éthiques et politiques, Presses Universitaires de France, 2019, 91–100, https://www.jstor.org/stable/10.2307/27069022.
Chaudiron, Stéphane, *Terminologie, ingénierie linguistique et gestion de l'information*, dans Langages, n° 157, La terminologie : nature et enjeux, Armand Colin Revues, 2005, 25–35, https://www.jstor.org/stable/41683540.
Chrimni, Walid, *Qu'est-ce que le NLP (Natural Language Processing) ?*, La revue IA, 2021, https://larevueia.fr/quest-ce-que-le-nlp-natural-language-processing/.
Cori, Marcel, *Des méthodes de traitement automatique aux linguistiques fondées sur les corpus*, dans Langages, n° 171, Construction des faits en linguistique : la place des corpus, Armand Colin Revues, 2008, 95–110, http://www.jstor.com/stable/23906378.
Crochet-Damais, Antoine, *Natural language processing (NLP) : définition et techniques*, JournalduNet.com, CCM Benchmark Group, 2022, https://www.journaldunet.fr/web-tech/guide-de-l-intelligence-artificielle/1501887-natural-language-processing-nlp/.
De Goursac, Alex, *Le natural language processing – Au cœur de l'interaction humain-IA*, Myriad-Data, 2017, https://myriad-data.com/wp-content/uploads/2019/07/nlp.pdf.
Dialoguer avec les robots : le natural language processing, Microsoft, 2022, https://experiences.microsoft.fr/articles/intelligence-artificielle/natural-language-processing/.
Fabien, Maël, *Traitement Automatique du Langage Naturel en français (TAL / NLP)*, Stat4decision, 2019, https://www.stat4decision.com/fr/traitement-langage-naturel-francais-tal-nlp/.
Gong, Li, *La traduction automatique statistique, comment ça marche ?*, 2013, interstices.info, https://interstices.info/la-traduction-automatique-statistique-comment-ca-marche/.
Introduction au NLP (Partie I), ekino., 2018, https://www.ekino.fr/publications/introduction-au-nlp-partie-i/.
Mateu, Jean-Bernard et Pluchart, Jean-Jacques, *L'économie de l'intelligence artificielle*, dans Revue d'économie financière, n° 135, Technologies and Changes in

the Financial Sector, Association Europe-Finances-Régulations, 2019, 257–272, https://www.jstor.org/stable/10.2307/26891441.
Pottier, Bernard, *Linguistique et intelligence artificielle*, dans Langages, n° 87, Sémantique et intelligence artificielle, Armand Colin Revues, 1987, 21–31, https://www.jstor.org/stable/41682107.
Robert, Jérémy, *Natural Language Processing (NLP) : Définition et principes*, Datascientest.com, 2020, https://datascientest.com/introduction-au-nlp-natural-language-processing.
Silberztein, Max, *Les outils informatiques au service des linguistes*, dans Langue française, n° 203, Les outils informatiques au service des linguistes, Armand Colin Revues, 2019, 7–14, https://www.jstor.org/stable/10.2307/26834993.
Tannier, Xavier, *Traitement automatique du langage naturel pour l'extraction et la recherche d'informations*, École nationale supérieure des Mines de Saint-Etienne, Rapport de recherche 2006-400-006, 2006, https://www.researchgate.net/publication/237243896_Traitement_automatique_du_langage_naturel_pour_l'extraction_et_la_recherche_d'informations.
Yvon, François, *Une petite introduction au Traitement Automatique des Langues Naturelles*, Orsay, Université Paris-Saclay, 2007, https://perso.limsi.fr/anne/coursM2R/intro.pdf.

Annexe F – Bibliographie du corpus en italien

Altobello, Giulio, *Natural Language Processing: cos'è, come funziona e applicazioni*, AI4Business, NetworkDigital360, 2022, https://www.ai4business.it/intelligenza-artificiale/natural-language-processing-tutto-quello-che-ce-da-sapere/.
Casadei, Giorgio, *Storia dell'informatica*, Università di Bologna, (s. d.), https://www.cs.unibo.it/casadei/appunti.pdf.
Cos'è l'elaborazione del linguaggio naturale (NLP)?, IBM, (s. d.), https://www.ibm.com/it-it/topics/natural-language-processing.
Cos'è l'elaborazione del linguaggio naturale?, Oracle, (s. d.), https://www.oracle.com/it/artificial-intelligence/what-is-natural-language-processing/#whymanageddb.

Cos'è il Natural Language Processing (NLP) e come funziona, Osservatori.net Digital Innovation, Politecnico di Milano, 2024, https://blog.osservatori.net/it_it/natural-language-processing-nlp-come-funziona-lelaborazione-del-linguaggio-naturale.

Creazzo, Annalisa et Fieromonte, Martina, *La costruzione di un modello di Natural Language Processing: dalla raccolta alla pulizia dei dati*, RES Group, 2021, https://res-group.eu/articoli/la-costruzione-di-un-modello-di-natural-language-processing-dalla-raccolta-alla-pulizia-dei-dati.

Esposito, Massimo, *Linguaggio naturale e intelligenza artificiale: a che punto siamo*, AI4Business, NetworkDigital360, 2019, https://www.agendadigitale.eu/cultura-digitale/linguaggio-naturale-e-intelligenza-artificiale-a-che-punto-siamo/.

Fabbri, David, *La sfida del Natural Language Processing: leggere e interpretare il linguaggio umano per estrarne conoscenza e valore*, Seacom, 2021, https://www.seacom.it/la-sfida-del-natural-language-processing-leggere-e-interpretare-il-linguaggio-umano-per-estrarne-conoscenza-e-valore/.

Maggini, Marco et Meoni, Stefano, *Natural Language Processing: studio e sviluppo di un riconoscitore automatico per l'analisi logica*, Università degli Studi di Siena, 2008, https://www.researchgate.net/publication/238706019_Natural_Language_Processing_studio_e_sviluppo_di_un_riconoscitore_automatico_per_l%27analisi_logica.

Manuelli, Riccardo, *NLP e Machine Learning: l'avanguardia della comunicazione*, DataDeep, 2022, https://datadeep.it/2022/08/03/nlp-e-machine-learning-lavanguardia-della-comunicazione/.

Marino, Luigi, *Cos'è il Natural Language Processing (NLP) e come funziona*, MarinoLuigi.it, 2023, https://www.marinoluigi.it/cose-il-natural-language-processing-nlp-e-come-funziona/.

NLP: cos'è e come può aiutarti ad essere più efficiente, ATG Artificial Intelligence Division, 2020, https://atgartificialintelligence.com/nlp-cose-e-come-puo-aiutarti-ad-essere-piu-efficiente/.

Sorce, Salvatore, *Introduzione alla Linguistica Computazionale*, Università degli Studi di Palermo, (s. d.), https://sites.unipa.it//sorce/didattica/sei1213/SEI1213_01_Linguistica_Computazionale_intro.pdf.

Turchi, Fabrizio, *Natural Language Processing: modelli e applicazioni in ambito giuridico*, dans Peruginelli, Ginevra et Ragona, Mario, L'informatica giuridica in Italia. Cinquant'anni di studi, ricerche ed esperienze, Collana ITTIG-CNR, Serie "Studi e documenti", n° 12, Napoli, ESI, 2014, 521–534, http://www.ittig.cnr.it/EditoriaServizi/AttivitaEditoriale/CollanaSeD/sed-12/Turchi.pdf.

Zoppetti, Antonio, *Il disastro della terminologia informatica italiana di fronte all'inglese*, Diciamolo in italiano, 2018, https://diciamoloinitaliano.wordpress.com/2018/05/14/il-disastro-della-terminologia-informatica-italiana-di-fronte-allinglese/.

Annexe G – Bibliographie du corpus en anglais

Briscoe, Ted, *Introduction to Linguistics for Natural Language Processing*, University of Cambridge, 2013, https://www.cl.cam.ac.uk/teaching/1314/L100/introling.pdf.

Church, Kenneth et Rau, Lisa, *Commercial Applications of Natural Language Processing*, dans Communications of the ACM, vol. 38, n° 11, 1995, 71–79, https://doi.org/10.1145/219717.219778.

Graham, Brett, *Using Natural Language Processing to Search for Textual References*, dans Hamidović, David/Clivaz, Claire et Bowen Savant, Sarah, Ancient Manuscripts in Digital Culture: Visualisation, Data Mining, Communication, vol. 38, n° 11, chap. 6, Brill, 2019, 115–132, https://www.jstor.org/stable/10.1163/j.ctvrxk44t.11.

Hirschberg, Julia et Manning, Christopher, *Advances in natural language processing*, dans Science, vol. 349, n° 6245, American Association for the Advancement of Science, 2015, 261–266, DOI: 10.1126/science.aaa8685.

Hong, Changchun et Zong, Zhaorong, *On Application of Natural Language Processing in Machine Translation*, 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 2018, DOI: 10.1109/ICMCCE.2018.00112.

Joshi, Aravind, *Natural Language Processing*, dans Science, New Series, vol. 253, n° 5025, American Association for the Advancement of Science, 1991, 1242–1249, https://www.jstor.org/stable/2879169.

Khurana, Diksha/Koli, Aditya/Khatter, Kiran et Singh, Sukhdev, *Natural language processing: state of the art, current trends and challenges*, dans Multimedia Tools and Applications, vol. 82, Springer, 2022, 3713–3744, https://doi.org/10.1007/s11042-022-13428-4.
Kleinings, Hanna, *What Is Natural Language Processing (NLP) & How Does It Work?*, Levity, 2022, https://levity.ai/blog/how-natural-language-processing-works.
Language and Machines – Computers in Translation and Linguistics, n° 1416, Washington D. C., The National Academies Press, 1966, https://nap.nationalacademies.org/resource/alpac_lm/ARC000005.pdf.
Nadkarni, Prakash/Ohno-Machado, Lucila et Chapman Wendy, *Natural language processing: an introduction*, dans Journal of the American Medical Informatics Association, vol. 18, n° 5, Oxford University Press, 2011, 544–551, DOI: 10.1136/amiajnl-2011-000464.
Pérez-Marín, Diana/Pascual-Nieto, Ismael et Rodríguez, Pilar, *About the benefits of exploiting natural language processing techniques for e-learning*, dans Proceedings of the Fourth International Conference on Web Information Systems and Technologies, SciTePress, 2008, 472–475, DOI: 10.5220/0001532304720475.
Raskin, Victor, *Linguistics and Natural Language Processing*, dans Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Colgate University, Hamilton, New York, 1985, 268–282, https://aclanthology.org/www.mt-archive.info/70/TMI-1985-Raskin.pdf.
Roldós, Inés, *10 Examples of Natural Language Processing in Action*, MonkeyLearn, 2021, https://monkeylearn.com/blog/natural-language-processing-examples/.
Wolff, Rachel, *Natural Language Processing (NLP): 7 Key Techniques*, MonkeyLearn, 2021, https://monkeylearn.com/blog/natural-language-processing-techniques/.
Wolff, Rachel, *What Is Natural Language Processing*, MonkeyLearn, 2020, https://monkeylearn.com/blog/what-is-natural-language-processing/#:~:text=In%20natural%20language%20processing%2C%20human,the%20same%20way%20as%20humans.

Bibliographie

Aussenac-Gilles, Nathalie/Charlet, Jean, et Reynaud-Delaître, Chantal, *Ingénierie des connaissances*, dans Marquis, Pierre/Papini, Odile, et Prade, Henri, Représentation des connaissances et formalisation des raisonnements, vol. 1, chap. 20, Panorama de l'intelligence artificielle, Cépaduès Editions, 2014, 500–537, https://www.irit.fr/publis/MELODI/AussenacCharletReynaudDelaitre_book-cepadues-chap-IC-2014.pdf.

Boccuzzi, Celeste, *L'enrichissement de la langue française : Anglicismes et Recommandations Officielles*, Università degli Studi di Bari Aldo Moro, 2017, https://www.uniba.it/it/docenti/boccuzzi-celeste/attivita-didattica/Dossier_20172018_EAMCA_bc_FNL.pdf.

Bolasco, Sergio, *Statistica testuale e text mining: alcuni paradigmi applicativi* [Statistiques et exploration de textes : quelques paradigmes d'application], dans Quaderni di Statistica, vol. 7, 2005, https://www.labstat.it/home/wp-content/uploads/2015/03/BOLASCO.pdf.

Cabré, Maria Teresa, *Constituer un corpus de textes de spécialité*, dans Cahier du CIEL 2007-2008, Academia.edu, 2008, 37–56, https://www.academia.edu/3194418/CONSTITUER_UN_CORPUS_DE_TEXTES_DE_SPECIALITE.

Cabré, Maria Teresa, *Terminology: theory, methods, and applications* [Terminologie : théorie, méthodes et applications], trad. Janet Ann DeCesaris, Amsterdam/Philadelphia, John Benjamins Publishing Co, vol. 1, 1998, https://www.academia.edu/38151335/_M_Teresa_Cabre_Terminology_Theory_Methods.

Condamines, Anne, *Linguistique de corpus et terminologie*, dans Langages, 39e année, n° 157, 2005, 36–47, DOI : https://doi.org/10.3406/lgge.2005.973.

Curley, Robert, *The Britannica Guide to Inventions that Changed the Modern World* [Le guide « Britannica » des inventions qui ont changé le monde moderne], New York, Britannica Educational Publishing, 2010, http://www.e4thai.com/e4e/images/pdf2/The%20Britannica%20Guide%20series/The.Britannica.Guide.to.Inventions.that.Changed.the.Modern.World.pdf.

Dahl, Veronica, *An Introduction to Natural Language Processing: the Main Problems* [Introduction au traitement du langage naturel : les principaux problèmes], dans Triangle: Language, Literature, Computation, n° 1, Simon Fraser University & GRLMC-Universitat Rovira i Virgili, 2010, 65–78, 388958-Text de l'article-562893-1-10-20210616 (1).pdf.

DataFranca.org, *Les 101 mots de l'intelligence artificielle : petit guide du vocabulaire essentiel de la science des données et de l'intelligence artificielle*, 1e édition, Gérard Pelletier, 2022.

Diagne, Abibatou, *La reconceptualisation et l'adaptation d'expression en terminologie culturelle*, dans Revista Digital Internacional de Lexicología, Lexicografía y Terminología, Sénégal, n° 5, 2022, https://revistas.unc.edu.ar/index.php/ReDILLeT/article/view/39959.

Dubreil, Estelle, *La dimension argumentative des collocations textuelles en corpus électronique spécialisé au domaine du TAL(N)*, HAL open science, Paris, CNRS, 2006, https://theses.hal.science/tel-00486063/.

Foley, Robert/Martin, Lawrence/Mirazón, Marta, et Stringer, Chris, *Major transitions in human evolution* [Les grandes transitions de l'évolution humaine], dans Philosophical Transactions: Biological Sciences, vol. 371, n° 1698, Royal Society, 2016, 1–8, http://dx.doi.org/10.1098/rstb.2015.0229.

Gaudin, François, *La socioterminologie*, dans Langages, 39e année, n° 157, 2005, 80–92, DOI : https://doi.org/10.3406/lgge.2005.976.

Goeuriot, Lorraine, *Découverte et caractérisation des corpus comparables spécialisés*, HAL open science, Paris, CNRS, 2009, https://theses.hal.science/tel-00474405/document.

IBM, et Morning Consult, *IBM Global AI Adoption Index 2022* [L'indice mondial d'IBM sur l'adoption de l'IA en 2022], Watson, 2022, https://www.ibm.com/downloads/cas/GVAGA3JP.

Kennedy, Graeme, *An Introduction to Corpus Linguistics* [Introduction à la linguistique de corpus], Longman, 1998, https://coehuman.uodiyala.edu.iq/uploads/Coehuman%20library%20pdf/English%20library%D9%83%D8%AA%D8%A8%20%D8%A7%D9%84%D8%A7%D9%86%D9%83%D9%84%D9%8A%D8%B2%D9%8A/linguistics/%27An%20Introduction%20to%20Corpus%20Linguistics%27%20-%20Kennedy%20Graeme.pdf.

Kida, Ireneusz, *Introduction to corpus linguistics* [Introduction à la linguistique de corpus], University of Silesia, 2013, 133–144, file:///C:/Users/Utente/Downloads/INTRODUCTION_TO_CORPUS_LINGUISTICS.pdf.

Konieczny, Sébastian, et Prade, Henri, *L'intelligence artificielle – de quoi s'agit-il vraiment ?*, Toulouse, Cépaudès-Èditions, 2020.

L'Homme, Marie-Claude, *La terminologie : principes et techniques*, dans Paramètres, 2e édition, Canada, Les Presses de l'Université de Montréal, 2020.

L'Homme, Marie-Claude, *Sur la notion de « terme »*, dans Meta, vol. 50, n° 4, Les Presses de l'Université de Montréal, 2005, 1112–1132, DOI : https://doi.org/10.7202/012064ar.

Labbé, Cyril et Labbé, Dominique, *La répartition du vocabulaire*, HAL open science, Paris, CNRS, 2017, https://hal.science/hal-01621060/document.

Laurent, Antoine/Guinaudeau, Camille et Roy, Anindya, *Analyse du corpus MATRICE-INA : exploration et classification automatique d'archives audiovisuelles de 1930 à 2012*, Matrice Memory, 2015, https://www.matricememory.fr/wp-content/uploads/2015/03/Laurent14j2.pdf.

Magnani, Eliana, *Qu'est-ce qu'un corpus ? Compte-rendu de la journée d'études*, HAL open science, Paris, CNRS, 2017, https://shs.hal.science/halshs-01610087/document.

Manuel d'annotation pour les corpus du projet PERCEO, Centrum für Informations- und Sprachverarbeitung, (s. d.), https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/spoken-french-annotation-manual.pdf.

Marshman, Elizabeth, *Construction et gestion des corpus : résumé et essai d'uniformisation du processus pour la terminologie*, Observatoire de linguistique Sens-Texte, 2003, https://olst.ling.umontreal.ca/static/pdf/terminotique/corpusnormes.pdf.

Matignon, Laëtitia, *Introduction à la robotique*, Caen, Université de Caen, 2011, https://perso.liris.cnrs.fr/laetitia.matignon/index/coursL1robotique.pdf, [dernière consultation : 11.12.2023].

Oakley, Kenneth, *Man the tool-maker* [L'homme, le fabricant d'outils], Londre, Jarrold and sons, 1961, https://ignca.gov.in/Asi_data/18399.pdf.

Park, Seungbae, *Extensional Scientific Realism vs. Intensional Scientific Realism* [Réalisme scientifique extensionnel vs. Réalisme scientifique intensionnel], dans Studies in History and Philosophy of Science, 2016, 46–52, https://philsci-archive.pitt.edu/15615/1/Extensional.pdf.

Pitar, Mariana, *La fiche terminologique – expansion et application*, dans Scientific Bulletin of the "Politehnica" University of Timişoara, Transactions on Modern Languages, vol. 10, n° 1-2, 2011, 70–83,https://www.researchgate.net/publication/322466200_La_fiche_terminologique-expansion_et_applications.

Procédés de formation des mots, Università degli Studi di Cagliari, (s. d.), https://web.unica.it/unica/protected/410453/0/def/ref/MAT407468/#:~:text=Situation%20de%20sp%C3%A9cialisation%20%3A%20des%20sp%C3%A9cialistes,les%20%C3%A9poques%20et%20les%20domaines.

Quarta, Alessandra, et Smorto, Guido, *Diritto privato dei mercati digitali* [Le droit privé des marchés numériques], Le Monnier Università, 2020.

Spataro, Michela, et Furholt Martin, *Detecting and explaining technological innovation in prehistory* [Détecter et expliquer l'innovation technologique dans la préhistoire], Leiden, Sidestone Press, 2020, https://www.sidestone.com/openaccess/9789088908248.pdf.

Temmermann, Rita, *Training Terminographers: the Sociocognitive Approach* [Formation des terminographes : l'approche sociocognitive], Bruxelles, Euralex, 2000,

453–460, https://euralex.org/elx_proceedings/Euralex2000/053_Rita%20TEMMERMANN_Training%20Terminographers_the%20Sociocognitive%20Approach.pdf.

Teubert, Wolfgang, *La linguistique de corpus : une alternative*, Semen, 2009, https://journals.openedition.org/semen/8923.

Van Croesdijk, Anouk, *L'anglais en France : langue internationale neutre ou sujet de discussion ? Attitudes des Français par rapport au rôle de l'anglais comme langue internationale*, Utrecht University Student Theses Repository, 2016, https://studenttheses.uu.nl/bitstream/handle/20.500.12932/23226/Attitudes%20linguistiques%20-%20Role%20international%20de%20l%27anglais.pdf?sequence=2.

Vezzani, Federica et Di Nunzio, Giorgio Maria, *FAIRterm Web Application: A Practical Guide* [Application Web FAIRterm : un guide pratique], University of Padua, 2022, https://yourterm.eu/wp-content/uploads/2022/05/FAIRterm-presentation-update.pdf.

Vezzani, Federica, *La ressource FAIRterm : entre pratique pédagogique et professionnalisation en traduction spécialisée*, dans Synergies Italie, n° 17, Gerflint, 2021, 51–64, https://gerflint.fr/Base/Italie17/vezzani.pdf.

Vušović, Olivera, *Approches théoriques de la terminologie et nature de termes : quelques considérations*, vol. 10, n° 4, Université du Monténégro, 2014, 83–90, https://www.revista-studii-uvvg.ro/wp-content/uploads/2014/12/I9.pdf.

Weizenbaum, Joseph, *ELIZA – A Computer Program For the Study of Natural Language Communication Between Man And Machine* [ELIZA – Un programme informatique pour l'étude de la communication en langage naturel entre l'homme et la machine], dans Communications of the ACM, vol. 9, n° 1, A. G. OETTINGER, 1966, 36–45, https://web.stanford.edu/class/cs124/p36-weizenabaum.pdf.

Williams, Geoffrey, *La linguistique et le corpus : une affaire prépositionnelle*, Texto, revue de linguistique en ligne, 2006, 151–158, http://www.revue-texto.net/Parutions/Livres-E/Albi-2006/Williams.pdf.

Yapomo, Manuela, *Construction de corpus multilingues : état de l'art*, Les Sables d'Olonne, TALN-RECITAL 2013, ATALA, 2013, 56–68, https://hal.science/hal-01073648/document.

Yvon, François, *Une petite introduction au Traitement Automatique des Langues Naturelles*, Orsay, Université Paris-Saclay, 2007, https://perso.limsi.fr/anne/coursM2R/intro.pdf.

Sitographie

À propos d'IBM, IBM, (s. d.), https://www.ibm.com/fr-fr/about, [dernière consultation : 11.06.2024].

À propos d'Oracle, Oracle, (s. d.), https://www.oracle.com/fr/corporate/, [dernière consultation : 11.06.2024].

À propos de IATE, iate, 2024, https://iate.europa.eu/home, [dernière consultation : 03.06.2024].

À propos de Wikipédia, dans Wikipédia. L'encyclopédie libre, Wikimedia Fondation, 2024, https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:%C3%80_propos_de_Wikip%C3%A9dia, [dernière consultation : 13.05.2024].

À propos, Datascientest.com, (s. d.), https://datascientest.com/a-propos, [dernière consultation : 11.06.2024].

About [À propos], Levity, (s. d.), https://levity.ai/about, [dernière consultation : 15.06.2024].

About JSTOR [À propos de JSTOR], JSTOR, (s. d.), https://about.jstor.org/, [dernière consultation : 14.05.2024].

About LDC [À propos de LDC], Linguistic Data Consortium, The Trustees of the University of Pennsylvania, (s. d.) https://www.ldc.upenn.edu/about, [dernière consultation : 18.04.2024].

About Science & AAAS [À propos de *Science* et d'AAAS], American Association for the Advancement of Science, (s. d.), https://www.science.org/content/page/about-science-aaas, [dernière consultation : 14.06.2024].

About the ACM Organization [À propos de l'organisation ACM], ACM, (s. d.), https://www.acm.org/about-acm/about-the-acm-organization, [dernière consultation : 14.06.2024].

About the journal [À propos du journal], American Medical Informatics Association, (s. d.), https://academic.oup.com/jamia/pages/About, [dernière consultation : 15.06.2024].

About Us [À propos], Communications of the ACM, (s. d.), https://cacm.acm.org/about-us/, [dernière consultation : 14.06.2024].

About us [À propos], Lexical Computing, (s. d.), https://www.lexicalcomputing.com/lexical-computing/, [dernière consultation : 14.05.2024].

About us [À propos], ResearchGate, (s. d.), https://www.researchgate.net/about?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6ImluZGV4IiwicHJldmlvdXNQYWdlIjoicHVibGljYXRpb24iLCJwb3NpdGlvbiI6Imdsb2JhbEZvb3RlciJ9fQ, [dernière consultation : 13.06.2024].

About. Progetto editoriale [À propos. Projet éditorial], https://www.ai4business.it/about/, AI4Business**,** NetworkDigital360, (s. d.), [dernière consultation : 14.05.2024].

AI History: gli anni '80 e i sistemi esperti [Histoire de l'IA : les années 1980 et les systèmes experts], Klondike, Milano, 2021, https://www.klondike.ai/ai-history-anni-80-sistemi-esperti/, [dernière consultation : 06.12.2023].

Aims and scope [Objectifs et champ d'application], Springer, (s. d.), https://link.springer.com/journal/11042/aims-and-scope, [dernière consultation : 15.06.2024].

ALPAC, dans Wikipédia. L'encyclopédie libre, Wikimedia Fondation, 2024, https://en.wikipedia.org/wiki/ALPAC, [dernière consultation : 15.06.2024].

Anatomie d'une balise HTML, Ronan HELLO, 2020, https://ronan-hello.fr/series/html/structure-balise-html, [dernière consultation : 23.04.2024].

Ancient Manuscripts in Digital Culture Visualisation, Data Mining, Communication [Les manuscrits anciens dans la culture numérique : visualisation, exploration de données, communication], Brill, (s. d.), https://brill.com/flyer/title/34930?print=pdf&pdfGenerator=headless_chrome, [dernière consultation : 14.06.2024].

Artificial Intelligence (AI) Market […] : Global Opportunity Analysis and Industry Forecast, 2022-2030 [Marché de l'intelligence artificielle (IA) [...] : Analyse des opportunités mondiales et prévisions industrielles, 2022-2030], Next Move Strategy Consulting, 2023, https://www.nextmsc.com/report/artificial-intelligence-market, [dernière consultation : 08.12.2023].

Balise : définition, traduction, dans Journaldunet.com, CCM Benchmark Group, Paris, 2019, https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203619-balise-definition-traduction/, [dernière consultation : 22.04.2024].

Burnard, Lou, *What is the BNC?* [Qu'est-ce que le BNC ?], British National Corpus, University of Oxford, 2009, http://www.natcorp.ox.ac.uk/corpus/index.xml, [dernière consultation : 25.04.2024].

Ce que les clients et les partenaires d'IBM disent du potentiel de watsonx, IBM, (s. d.), https://www.ibm.com/fr-fr/watsonx/resources/client-quotes, [dernière consultation : 11.06.2024].

ChatGPT, LeMagIT, (s. d.), https://www.lemagit.fr/definition/ChatGPT, [dernière consultation : 11.06.2024].

Chi siamo [Qui sommes-nous ?], RES, (s. d.), https://res-group.eu/gruppo-res/, [dernière consultation : 12.06.2024].

Chi siamo [Qui sommes-nous ?], Seacom, (s. d.), https://seacom.it/chi-siamo/, [dernière consultation : 12.06.2024].

Chi sono [À propos], Diciamolo in Italiano, (s. d.), https://diciamoloinitaliano.wordpress.com/chi-sono/, [dernière consultation : 13.06.2024].

Chrimni, Walid, *Qu'est-ce que le NLP (Natural Language Processing) ?*, La revue IA, 2021, https://larevueia.fr/quest-ce-que-le-nlp-natural-language-processing/, [dernière consultation : 11.06.2024].

Cités. Présentation, puf, (s. d.), https://www.puf.com/cites, [dernière consultation : 07.06.2024].

Cohen, Dan, *Cloud Computing : Qu'est-ce que c'est ? Comment ça fonctionne ?*, Datascientest.com, 2022, https://datascientest.com/cloud-computing, [dernière consultation : 10.06.2024].

Comma-separated values, dans Wikipédia. L'encyclopédie libre, Wikimedia Fondation, 2023, https://fr.wikipedia.org/wiki/Comma-separated_values, [dernière consultation : 14.05.2024].

Comment juger de la fiabilité d'un site ?, Le Monde, Société Éditrice du Monde, 2022, https://www.lemonde.fr/les-decodeurs/article/2022/12/20/comment-juger-de-la-fiabilite-d-un-site_5067739_4355771.html, [dernière consultation : 14.05.2024].

Concordancier - Définition, Techno-Science.net, Yvelines, (s. d.), https://www.techno-science.net/definition/10980.html, [dernière consultation : 05.03.2024].

Contextualisme, dans Encyclopædia Universalis, (s. d.), https://www.universalis.fr/dictionnaire/contextualisme/, [dernière consultation : 26.04.2024].

Corpus juris civilis, dans Wikipédia. L'encyclopédie libre, Wikimedia Fondation, 2024, https://fr.wikipedia.org/wiki/Corpus_juris_civilis, [dernière consultation : 09.04.2024].

Corpus Linguistics [Linguistique de corpus], Johannes Gutenberg-Universität Mainz, 2021, https://www.english-linguistics.uni-mainz.de/corpus-linguistics/, [dernière consultation : 25.04.2024].

Corpus, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/definition/corpus, [dernière consultation : 21.02.2024].

Corpus, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/etymologie/corpus#, [dernière consultation : 09.04.2024].

Cosa facciamo [Notre mission], Osservatori.net Digital Innovation, Politecnico di Milano, (s. d.), https://www.osservatori.net/it/chi-siamo/conosciamoci/cosa-facciamo, [dernière consultation : 13.06.2024].

Curriculum di Giorgio Casadei [Curriculum de Giorgio Casadei], Università di Bologna, (s. d.), https://www.cs.unibo.it/casadei/curriculum2.htm, [dernière consultation : 14.06.2024].

Décret n° 96-602 du 3 juillet 1996 relatif à l'enrichissement de la langue française, Légifrance, Secrétariat général du Gouvernement, 2022, https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000378502/, [dernière consultation : 04.04.2024].

Diachronie, dans Dictionnaire en ligne Larousse, Paris, (s. d.), https://www.larousse.fr/dictionnaires/francais/diachronie/25130, [dernière consultation : 28.02.2024].

Digital Biblical Studies [Études bibliques numériques], Brill, (s. d.), https://brill.com/flyer/serial/DBS?print=pdf&pdfGenerator=headless_chrome, [dernière consultation : 14.06.2024].

Document d'information. Tableau des suffixes, Ministère de l'Éducation de la Saskatchewan, Programme d'études, Niveau secondaire, Écoles fransaskoises, 1999, https://www.k12.gov.sk.ca/docs/francais/fransk/fran/sec/prg_etudes/stratl12b.html, [dernière consultation : 20.05.2024].

Drouin, Patrick, *Présentation du logiciel*, Observatoire de linguistique Sens-Texte, 2010, https://termostat.ling.umontreal.ca/doc_termostat/doc_termostat_en.html, [dernière consultation : 24.05.2024].

Échantillon, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/definition/%C3%A9chantillon, [dernière consultation : 16.04.2024].

Eleyehou, Denis, *DevOps : qu'est-ce que c'est ? Principe, avantages, formation*, Datascientest.com, 2021, https://datascientest.com/devops, [dernière consultation : 11.06.2024].

Estropier, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, s. d., https://www.cnrtl.fr/definition/estropier, [dernière consultation : 21.05.2024].

Été indien, dans Wiktionnaire. Le dictionnaire libre, Wikimedia Fondation, 2023, https://fr.wiktionary.org/wiki/%C3%A9t%C3%A9_indien#Anglais, [dernière consultation : 21.05.2024].

European Open Science Cloud (EOSC), European Commission, (s. d.), https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en, [dernière consultation : 03.06.2024].

Évolution, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/definition/%C3%A9volution, [dernière consultation : 13.10.2023].

Fabrizio Turchi. Dirigente tecnologo [Fabrizio Turchi. Technologue principal], Istituto di Informatica Giuridica e Sistemi Giudiziari, (s. d.), https://www.igsg.cnr.it/persone/fabrizio-turchi/, [dernière consultation : 13.06.2024].

Facts About Microsoft [Quelques faits sur Microsoft], Microsoft, (s. d.), https://news.microsoft.com/facts-about-microsoft/#About, [dernière consultation : 11.06.2024].

FAIR Terminology, FAIRterm, (s. d.), https://shiny.dei.unipd.it/fairterm/, [dernière consultation : 04.06.2024].

Fiche terminologique, dans Office québécois de la langue française, 2005, https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/8355434/fiche-terminologique, [dernière consultation : 04.06.2024].

François Yvon, Institut Systèmes Intelligents et de Robotique, (s. d.), https://www.isir.upmc.fr/personnel/yvon/?lang=en, [dernière consultation : 08.06.2024].

Geoffrey Leech, *Developing Linguistic Corpora: a Guide to Good Practice. Adding Linguistic Annotation* [Développer des corpus linguistiques : un guide de bonnes pratiques. Ajouter une annotation linguistique], AHDS literature, languages and linguistics, 2004, https://users.ox.ac.uk/~martinw/dlc/chapter2.htm, [dernière consultation : 19.04.2024].

Grand lexique français de l'intelligence artificielle, DataFranca.org, 2024, https://datafranca.org/wiki/Accueil, [dernière consultation : 08.04.2024].

Grand lexique français de l'intelligence artificielle, DataFranca.org, 2024, https://datafranca.org/wiki/Cat%C3%A9gorie:GRAND_LEXIQUE_FRAN%C3%87AIS, [dernière consultation : 08.04.2024].

Guerre froide, dans Wikipédia. L'encyclopédie libre, Wikimedia Fondation, 2024, https://fr.wikipedia.org/wiki/Guerre_froide, [dernière consultation : 22.11.2023].

Guillot, Agnès, *Histoire de la robotique : des automates aux premiers robots*, FUTURA, 2003, https://www.futura-sciences.com/tech/dossiers/robotique-robotique-a-z-178/page/2/, [dernière consultation : 11.12.2023].

Histoire de l'intelligence artificielle, Conseil de l'Europe, (s. d.), https://www.coe.int/fr/web/artificial-intelligence/history-of-ai, [dernière consultation : 04.12.2023].

Home [Page d'accueil], ATG Artificial Intelligence Division, (s. d.), https://atgartificialintelligence.com/, [dernière consultation : 13.06.2024].

Home [Page d'accueil], DataDeep, (s. d.), https://datadeep.it/, [dernière consultation : 12.06.2024].

Home [Page d'accueil], ICMCCE, (s. d.), http://www.icmcce.com/, [dernière consultation : 14.06.2024].

Home [Page d'accueil], Karon, (s. d.), https://www.karon.it/, [dernière consultation : 12.06.2024].

Home [Page d'accueil], MarinoLuigi.it, (s. d.), https://www.marinoluigi.it/, [dernière consultation : 12.06.2024].

Home [Page d'accueil], Springer, (s. d.), https://link.springer.com/journal/11042, [dernière consultation : 15.06.2024].

Home [Page d'accueil], WEBIST, 2024, https://webist.scitevents.org/Home.aspx, [dernière consultation : 15.06.2024].

Hostie, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/definition/hostie, [dernière consultation : 09.04.2024].

Humanités numériques, FranceTerme, 2019, https://www.culture.fr/franceterme/terme/EDUC120, [dernière consultation : 14.06.2024].

Hyperlien, dans Dictionnaire en ligne Larousse, Paris, (s. d.), https://www.larousse.fr/dictionnaires/francais/hyperlien/41050, [dernière consultation : 22.11.2023].

IBM, dans Wikipédia. L'encyclopédie libre, Wikimedia Fondation, 2024, https://fr.wikipedia.org/wiki/IBM, [dernière consultation : 08.12.2023].

Il gruppo [Le groupe], ATG Anzani Group, (s. d.), https://www.anzanigroup.com/il-gruppo/, [dernière consultation : 13.06.2024].

Informatique linguistique, DataFranca.org, 2024, https://datafranca.org/wiki/Informatique_linguistique, [dernière consultation : 12.04.2024].

Inspired by data. Powered by technology. Human by design. [Inspiré par les données. Alimenté par la technologie. Humain par sa conception.], ekino., (s. d.), https://www.ekino.fr/, [dernière consultation : 11.06.2024].

Instructions Conditionnelles avec If, dans INF1563 Programmation I, Université du Québec en Outaouais, (s. d.), http://w3.uqo.ca/adavoust/cours/instructions_conditionnelles_avec_if.html, [dernière consultation : 30.11.2023].

Intelligence artificielle : opportunités et risques, Parlement européen, 2023, https://www.europarl.europa.eu/news/fr/headlines/society/20200918STO87404/intelligence-artificielle-opportunites-et-risques, [dernière consultation : 08.12.2023].

Intelligence, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/definition/intelligence, [dernière consultation : 22.11.2023].

Internet, Institut national de la statistique et des études économiques, Paris, 2020, https://www.insee.fr/fr/metadonnees/definition/c1864#:~:text=Ensemble%20de%20r%C3%A9seaux%20mondiaux%20interconnect%C3%A9s,de%20communication%20commun%20(IP), [dernière consultation : 22.11.2023].

ISO 1087 : 2019 (fr). Travail terminologique et science de la terminologie — Vocabulaire, ISO, 2019, https://www.iso.org/obp/ui/fr/#iso:std:iso:1087:ed-2:v1:fr:term:3.4.2, [dernière consultation : 22.04.2024].

Kenneth Oakley, dans Wikipédia. L'encyclopédie libre, Wikimedia Fondation, 2023, https://en.wikipedia.org/wiki/Kenneth_Oakley, [dernière consultation : 13.10.2023].

Kleinings, Hanna, *What Is Natural Language Processing (NLP) & How Does It Work?* [Qu'est-ce que le traitement automatique du langage naturel (TALN) et comment fonctionne-t-il ?], Levity, 2022, https://levity.ai/blog/how-natural-language-processing-works, [dernière consultation : 15.06.2024].

L'école DataScientest, OMNES Education, 2023, https://www.omneseducation.com/nos-etablissements/datascientest/, [dernière consultation : 11.06.2024].

L'École, École des Mines de Saint-Étienne, (s. d.), https://www.mines-stetienne.fr/lecole/, [dernière consultation : 08.06.2024].

L'essentiel à savoir sur une base de données, Oracle, (s. d.), https://www.oracle.com/fr/database/definition-base-de-donnees/, [dernière consultation : 31.05.2024].

L'IA, c'est quoi ?, Conseil de l'Europe, (s. d.), https://www.coe.int/fr/web/artificial-intelligence/what-is-ai, [dernière consultation : 24.11.2023].

La Machine de Turing, JP Zanotti, 2023, https://zanotti.univ-tln.fr/turing/, [dernière consultation : 04.12.2023].

Langages. Revue internationale des sciences du langage, Armand Colin Revues, (s. d.), https://www.revues.armand-colin.com/lettres-langue/langages, [dernière consultation : 06.06.2024].

Langue de spécialité, dans Office québécois de la langue française, 2023, https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/26560898/langue-de-specialite, [dernière consultation : 17.05.2024].

Langue française. Présentation, Armand Colin Revues, (s. d.), https://www.revues.armand-colin.com/lettres-langue/langue-francaise, [dernière consultation : 08.06.2024].

Langue, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/definition/langue, [dernière consultation : 19.12.2023].

Le nostre soluzioni [Nos solutions], Seacom, (s. d.), https://seacom.it/soluzioni/, [dernière consultation : 12.06.2024].

Lebert, Marie, *Les meilleurs dictionnaires de langues en ligne*, dans ActuaLitté, 2010, https://actualitte.com/article/80838/distribution/les-meilleurs-dictionnaires-de-langues-en-ligne, [dernière consultation : 27.05.2024].

Libri in vetrina [Livres en vedette], Biblioteca Centrale Giuridica, 2015, https://www.giustizia.it/giustizia/page/en/bcg_recensione_libro?contentId=BNA1117451#, [dernière consultation : 13.06.2024].

Linguistics and Natural Language Processing [Linguistique et traitement automatique du langage naturel], Semantic Scholar, (s. d.), https://www.semanticscholar.org/paper/Linguistics-and-Natural-Language-Processing-Raskin/a215a0d8ca3d4cbaf0a6cc624c921f4b27d88ea7, [dernière consultation : 15.06.2024].

Linguistique de corpus, dans Linternaute.com, CCM Benchmark Group, Paris, 2021, https://www.linternaute.fr/dictionnaire/fr/definition/linguistique-de-corpus/, [dernière consultation : 12.04.2024].

Linguistique, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/definition/linguistique, [dernière consultation : 25.04.2024].

Mantique, dans La langue française, 2024, https://www.lalanguefrancaise.com/dictionnaire/definition/mantique, [dernière consultation : 25.05.2024].

Marco Maggini. Presentazione [Marco Maggini. Présentation], Università degli Studi di Siena, (s. d.), https://docenti.unisi.it/it/maggini, [dernière consultation : 13.06.2024].

McCarthy, John, dans Treccani. Vocabolario online, Rome, (s. d.), https://www.treccani.it/enciclopedia/john-mccarthy/, [dernière consultation : 25.11.2023].

Métalangage, dans Office québécois de la langue française, 2000, https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/8369343/metalangage, [dernière consultation : 15.04.2024].

Minsky, Marvin Lee, dans Treccani. Vocabolario online, Rome, (s. d.), https://www.treccani.it/enciclopedia/marvin-lee-minsky/, [dernière consultation : 27.11.2023].

Mission [Notre mission], National Academy of Sciences, (s. d.), https://www.nasonline.org/about-nas/mission/, [dernière consultation : 15.06.2024].

Mot grammatical, dans Office québécois de la langue française, 2022, https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/26559677/mot-grammatical, [dernière consultation : 24.05.2024].

NEUMANN, John von, dans Treccani. Vocabolario online, Rome, (s. d.), https://www.treccani.it/enciclopedia/john-von-neumann_%28Enciclopedia-Italiana%29/, [dernière consultation : 04.12.2023].

Nos expertises, ekino., (s. d.), https://www.ekino.fr/expertises/, [dernière consultation : 11.06.2024].

OCR – reconnaissance optique de caractères, Microsoft, 2023, https://learn.microsoft.com/fr-fr/azure/ai-services/computer-vision/overview-ocr, [dernière consultation : 13.01.2024].

Octet, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/definition/octet, [dernière consultation : 14.05.2024].

Pourquoi les clients choisissent-ils OCI, Oracle, (s. d.), https://www.oracle.com/fr/cloud/why-oci/, [dernière consultation : 11.06.2024].

Présentation, Institut Systèmes Intelligents et de Robotique, (s. d.), https://www.isir.upmc.fr/isir/presentation/, [dernière consultation : 08.06.2024].

Prosodie, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/definition/prosodie, [dernière consultation : 20.04.2024].

Prototypique, dans Linternaute.com, CCM Benchmark Group, Paris, 2021, https://www.linternaute.fr/dictionnaire/fr/definition/prototypique/, [dernière consultation : 03.04.2024].

Qu'est-ce que l'apartheid ?, Amnesty International France, (s. d.), https://www.amnesty.fr/focus/apartheid, [dernière consultation : 21.05.2024].

Qu'est-ce que l'industrie 4.0 ?, IBM, (s. d.), https://www.ibm.com/fr-fr/topics/industry-4-0, [dernière consultation : 13.06.2024].

Qu'est-ce que l'Internet des objets (IoT) ?, Amazon Web Services, (s. d.), https://aws.amazon.com/fr/what-is/iot/, [dernière consultation : 10.06.2024].

Qu'est-ce que le Web3 ?, Amazon Web Services, (s. d.), https://aws.amazon.com/fr/what-is/web3/, [dernière consultation : 10.06.2024].

Qui sommes nous ?, JournalduNet.com, CCM Benchmark Group**,** (s. d.), https://www.journaldunet.com/magazine/static/1418511-qui-sommes-nous/, [dernière consultation : 10.06.2024].

Qui sommes-nous ?, Armand Colin Revues, (s. d.), https://www.revues.armand-colin.com/qui-sommes-nous, [dernière consultation : 07.06.2024].

Qui sommes-nous ?, interstices.info, (s. d.), https://interstices.info/qui-sommes-nous/, [dernière consultation : 10.06.2024].

Qui sommes-nous ?, Myriad-Data, (s. d.), https://myriad-data.com/#engagement-rse, [dernière consultation : 11.06.2024].

Qui sommes-nous ?, Stat4decision, (s. d.), https://www.stat4decision.com/fr/a-propos/qui-sommes-nous/, [dernière consultation : 10.06.2024].

Recherche et innovation, École des Mines de Saint-Étienne, (s. d.), https://www.mines-stetienne.fr/recherche/, [dernière consultation : 08.06.2024].

Revue d'économie financière, Association Europe-Finances-Régulations, (s. d.), https://www.aefr.eu/fr/numeros?page=1, [dernière consultation : 07.06.2024].

Robotique, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/definition/robotique, [dernière consultation : 11.12.2023].

Sciences cognitives, dans Wikipédia. L'encyclopédie libre, Wikimedia Fondation, 2024, https://fr.wikipedia.org/wiki/Sciences_cognitives, [dernière consultation : 01.04.2024].

Signorelli, Andrea Daniele, *Le origini dell'Intelligenza Artificiale* [Les origines de l'intelligence artificielle], Il Tascabile, 2017, https://www.iltascabile.com/scienze/origini-intelligenza-artificiale/, [dernière consultation : 01.12.2023].

Sketch Engine, Université Jean Moulin Lyon 3, 2023, https://bu.univ-lyon3.fr/sketch-engine, [dernière consultation : 14.05.2024].

Soffer, Virginie, *Le vocabulaire incontournable de l'intelligence artificielle*, udemnouvelles, 2023, https://nouvelles.umontreal.ca/article/2023/01/23/le-vocabulaire-incontournable-de-l-intelligence-artificielle/, [dernière consultation : 05.04.2024].

Solutions d'intelligence artificielle (IA), IBM, (s. d.), https://www.ibm.com/fr-fr/artificial-intelligence?lnk=flathl, [dernière consultation : 11.06.2024].

Synchronie, dans Dictionnaire en ligne Larousse, Paris, (s. d.), https://www.larousse.fr/dictionnaires/francais/synchronie/76126, [dernière consultation : 28.02.2024].

Système expert, dans Dictionnaire en ligne Larousse, Paris, (s. d.), https://www.larousse.fr/encyclopedie/divers/syst%C3%A8me_expert/95444, [dernière consultation : 28.11.2023].

Tarnoff, Ben, *Weizenbaum's nightmares: how the inventor of the first chatbot turned against AI* [Les cauchemars de Weizenbaum : comment l'inventeur du premier chabot s'est retourné contre l'IA], The Guardian, 2023, https://www.theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai, [dernière consultation : 06.12.2023].

Technologie, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/definition/technologie, [dernière consultation : 22.11.2023].

Tecnologia [Tecnologie], dans Treccani. Vocabolario online, Rome, (s. d.), https://www.treccani.it/enciclopedia/tecnologia/, [dernière consultation : 22.11.2023].

Ted Briscoe: Short Biography [Ted Briscoe : brève biographie], University of Cambridge, (s. d.), https://www.cl.cam.ac.uk/~ejb1/short-bio.html, [dernière consultation : 14.06.2024].

Terminologie, Académie française, (s. d.), https://www.academie-francaise.fr/la-langue-francaise/terminologie, [dernière consultation : 05.04.2024].

Terminology without borders [Terminologie sans frontières], Knowledge Centre on Interpretation, (s. d.), https://knowledge-centre-interpretation.education.ec.europa.eu/fr/node/22952, [dernière consultation : 22.06.2024].

The Cybernetics Thought Collective: A History of Science and Technology Portal Project – Warren S. McCulloch [Le collectif de pensée cybernétique : Projet de portail sur l'histoire des sciences et des technologies – Warren S. McCulloch], University of Illinois Board of Trustees, 2014, https://archives.library.illinois.edu/thought-collective/cyberneticians/warren-s-mcculloch/, [dernière consultation : 01.12.2023].

Théorie de l'information, Techno-Science.net, Yvelines, (s. d.), https://www.techno-science.net/glossaire-definition/Theorie-de-l-information.html, [dernière consultation : 18.12.2023].

Thésaurus documentaire, dans Wikipédia. L'encyclopédie libre, Wikimedia Fondation, 2023, https://fr.wikipedia.org/wiki/Th%C3%A9saurus_documentaire, [dernière consultation : 30.03.2024].

Toundra, dans La langue française, 2024, https://www.lalanguefrancaise.com/dictionnaire/definition/toundra, [dernière consultation : 21.05.2024].

Transistor, dans Centre National de Ressources Textuelles et Lexicales, Nancy, ATILF, (s. d.), https://www.cnrtl.fr/definition/transistor, [dernière consultation : 11.12.2023].

TreeTagger – a part-of-speech tagger for many languages [TreeTagger – un marqueur de parties du discours pour de nombreuses langues], Centrum für Informations- und Sprachverarbeitung, (s. d.), https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/, [dernière consultation : 22.04.2024].

TreeTagger, Corpora.lancs.ac.uk, (s. d.), http://corpora.lancs.ac.uk/tree-tagger/, [dernière consultation : 22.04.2024].

Turing, Alan Mathison, dans Treccani. Vocabolario online, Rome, (s. d.), https://www.treccani.it/enciclopedia/alan-mathison-turing/, [dernière consultation : 04.12.2023].

UTF-8, Techno-Science.net, Yvelines, (s. d.), https://www.techno-science.net/glossaire-definition/UTF-8.html, [dernière consultation : 15.04.2024].

Véron, Muriel, *Tableau de préfixes courants en français scientifique (liste non exhaustive)*, UNIT, Fondation UNIT, (s. d.), http://ressources.unit.eu/medias/filipe/pdf/tableau_prefixes.pdf, [dernière consultation : 20.05.2024].

Vulgarisation scientifique, Université de Sherbrooke, (s. d.), https://www.usherbrooke.ca/langue/le-francais-en-outils/vulgarisation-scientifique#:~:text=Le%20but%20de%20la%20vulgarisation,d%27une%20certaine%20culture%20scientifique, [dernière consultation : 08.04.2024].

Walter Pitts, dans Dictionary of the Philosophy of Mind, Washington University in St. Louis, (s. d.), https://home.csulb.edu//~cwallis/artificialn/walter_pitts.html, [dernière consultation : 01.12.2023].

Webinaire, dans Linternaute.com, CCM Benchmark Group, Paris, 2021, https://www.linternaute.fr/dictionnaire/fr/definition/webinaire/, [dernière consultation : 11.06.2024].

Weiss, Frédéric, *AntConc, logiciel d'analyse textuelle*, École normale supérieure de Lyon, (s. d.), http://cid.ens-lyon.fr/ac_article.php?fic=antconc.php, [dernière consultation : 14.05.2024].

What is MonkeyLearn? [Qu'est-ce que *MonkeyLearn* ?], MonkeyLearn, 2024, https://help.monkeylearn.com/en/articles/2174206-what-is-monkeylearn, [dernière consultation : 15.06.2024].

Wikipedia:Attendibilità di Wikipedia [Wikipedia:Fiabilité de Wikipedia], dans Wikipédia. L'encyclopédie libre, Wikimedia Fondation, 2024, https://it.wikipedia.org/wiki/Wikipedia:Attendibilit%C3%A0_di_Wikipedia#:~:text=Wikipedia%20%C3%A8%20una%20fonte%20a,degli%20aspetti%20positivi%20e%20negativi, [dernière consultation : 14.05.2024].

XLSX, dans Wikipédia. L'encyclopédie libre, Wikimedia Fondation, 2020, https://fr.wikipedia.org/wiki/XLSX, [dernière consultation : 14.05.2024].

Résumé en italien

Questa tesi è dedicata alla terminologia di due aree di studio che stanno diventando sempre più importanti negli ultimi anni: l'intelligenza artificiale e l'elaborazione del linguaggio naturale. L'elaborato segue una struttura ordinata per tematiche e divisa in cinque capitoli. Ogni capitolo è a sua volta suddiviso in più paragrafi e sottoparagrafi che hanno la funzione principale di analizzare accuratamente alcune tematiche. L'obiettivo principale della tesi consiste nell'analisi di 150 termini relativi all'intelligenza artificiale e all'elaborazione del linguaggio naturale in tre lingue, ossia il francese, l'italiano e l'inglese.

Le motivazioni di fondo che ci hanno portato a scegliere questo progetto sono molteplici. Prima di tutto, grazie al corso di traduzione specializzata francese è stato possibile entrare in contatto con il mondo della terminologia, comprendendone l'importanza. In particolare, lo studio della terminologia dell'intelligenza artificiale e dell'elaborazione del linguaggio naturale ci permette di dimostrare quanto sia importante avere a disposizione dei dati terminologici completi, date anche le difficoltà che si possono incontrare quando ci si avvicina a questi due domini. Di conseguenza, grazie alla partecipazione al progetto *Terminology without borders* dell'unità di coordinamento terminologico del Parlamento europeo, *TermCoord*, abbiamo la possibilità di aderire al miglioramento della comunicazione relativa a diversi settori specialistici adattando la loro terminologia alle esigenze dei cittadini.

Nel primo capitolo ci siamo occupati esclusivamente di definire i due domini che dobbiamo analizzare. Siamo quindi partiti dall'intelligenza artificiale presentando le varie invenzioni dell'uomo dalle origini ad oggi. L'intelligenza artificiale è la disciplina che riunisce scienze, teorie e tecniche per poter far sì che una macchina imiti le capacità cognitive di un essere umano. Abbiamo continuato presentando la storia dell'intelligenza artificiale, dalla sua nascita agli sviluppi più recenti. L'intelligenza artificiale in quanto disciplina nasce nel 1956, nel New Hampshire, quando due ricercatori, John McCarthy e Marvin Minsky, tennero un convegno dedicato allo sviluppo di sistemi intelligenti. Gli scienziati che appoggiavano le idee di McCarthy e Minsky sostenevano la teoria secondo la quale le macchine potessero riprodurre l'intelligenza umana e che per fare ciò dovessero ricevere istruzioni chiare e inequivocabili per eseguire gli algoritmi richiesti.

Nascono quindi i sistemi esperti, un insieme di software capaci di risolvere problemi come un esperto umano farebbe. Con il passare degli anni, dopo aver superato periodi di inattività e problemi tecnici ed economici, l'intelligenza artificiale evolve: si introduce l'apprendimento automatico e poi l'apprendimento profondo, due tecniche di programmazione che permettono alle macchine di imparare dall'esperienza. Ad oggi, quindi, l'intelligenza artificiale ha fatto passi da gigante, infatti, essa fa e farà sempre più parte delle nostre vite e sarà centrale in molte attività. Non a caso, l'utilizzo di sistemi di intelligenza artificiale si fa sempre più strada anche in settori in cui nessuno penserebbe mai di introdurla. Oltre a discipline quali la robotica, l'informatica, l'ingegneria e l'industria, essa viene applicata anche alle scienze umane e sociali. Una branca particolarmente interessante dell'intelligenza artificiale è l'elaborazione del linguaggio naturale, ossia lo studio dell'interazione tra macchina e linguaggio naturale. Le sfide maggiori che affronta l'elaborazione del linguaggio naturale riguardano diverse tecnologie, come il riconoscimento vocale o la generazione del linguaggio naturale. Questo perché il linguaggio umano può essere ambiguo, poco chiaro, e la macchina non sempre riesce a comprenderne le sfumature e i contesti.

Una volta presentati i due domini, siamo potuti passare all'analisi del "collante" che unisce sia il progetto terminologico sia la tesi stessa: la terminologia. Il secondo capitolo è infatti dedicato totalmente alla presentazione della terminologia e della terminografia. La terminologia è la scienza che si occupa della teoria e del quadro concettuale dello studio dei termini. Essa si occupa della standardizzazione di vocabolari specialistici al fine di garantire comprensibilità e univocità. La terminografia, invece, rappresenta la parte pratica, interessandosi a una serie di attività, quali la raccolta, la compilazione e la gestione dei termini. Proseguendo, vengono introdotte la teoria generale della terminologia proposta da Eugen Wüster, per la quale prevale un approccio di tipo concettuale, e tutti gli altri approcci alternativi a quello di Wüster, quali l'approccio socioterminologico. In seguito, abbiamo esaminato la terminografia e le fasi del lavoro terminologico che sono: la costruzione di un corpus, l'estrazione dei termini, la raccolta dei dati relativi ai termini, la loro analisi e sintesi, la loro codifica, l'organizzazione dei dati terminologici e la loro gestione. Queste tappe sono fondamentali per realizzare un buon progetto terminologico. Terminologia e terminografia sono influenzate dall'informatica. Grazie a quest'ultima e al suo progresso, infatti, oggi i terminografi

hanno a loro disposizione più risorse tecnologiche che li aiutano nei loro progetti, come testi in formato elettronico e strumenti che velocizzano il lavoro terminografico. L'analisi del nostro "collante" prosegue con la presentazione dei domini che lo applicano. La terminologia viene, infatti, utilizzata da più discipline, ad esempio, la comunicazione specializzata, la documentazione e la gestione della conoscenza, la traduzione specializzata, ma anche l'ingegneria della conoscenza e l'informatica. Infine, dato il progetto a cui abbiamo aderito, ci siamo occupati di discutere della terminologia dell'intelligenza artificiale e dell'elaborazione del linguaggio naturale. Oggi la lingua di trasmissione del sapere scientifico è l'inglese, definita lingua franca per il suo ruolo comunicativo e centrale. Il lessico dell'intelligenza artificiale e dell'elaborazione del linguaggio naturale è infatti ricco di termini inglesi, gli anglicismi, ossia termini o costruzioni della lingua inglese che vengono recepiti e usati in un'altra lingua: un facile esempio è il termine *business*. Per quanto riguarda il mondo francofono, per evitare l'uso di termini stranieri, le autorità pubbliche incoraggiano la creazione, la diffusione e l'uso di nuovi termini in lingua francese. In Francia, ad esempio, nel 1996 il legislatore ha deciso di emanare un decreto sull'arricchimento della lingua francese con l'obiettivo di creare termini equivalenti per designare in francese gli stessi concetti e realtà che sono stati designati in inglese. Questo rappresenta uno scoglio nell'analisi dei termini dell'intelligenza artificiale e dell'elaborazione del linguaggio naturale, ma anche un'opportunità, in quanto, creando un lessico adeguato in francese, chiunque potrà comprendere tali domini, grazie a un'opera di divulgazione che va oltre la traduzione.

Nel terzo capitolo ci siamo occupati di uno strumento essenziale per la realizzazione del progetto: il corpus. Per comprendere accuratamente il corpus e le sue caratteristiche e funzioni ci siamo concentrati inizialmente sul quadro teorico. Un corpus è un insieme di testi selezionati e organizzati per facilitare le analisi linguistiche. Questa raccolta di documenti, che devono essere finiti e reali, deve essere caratterizzata da completezza e autenticità. Questo perché il corpus deve rappresentare un campione reale e significativo di una lingua. I dati linguistici che compongono un corpus possono essere di vario tipo: dati scritti o orali, dati fisici o elettronici. Questo ci fa capire che i documenti di un corpus possono essere, ad esempio, riviste, capitoli di libri, saggi, ma anche monologhi e conversazioni. Inoltre, oggi i corpus sono perlopiù in formato digitale, servendosi della tecnologia e dell'informatica essi sono diventati uno strumento accessibile a tutti e sempre

più performante. Nel nostro caso ci siamo forniti di un corpus specialistico, ossia rappresentativo di un dominio in particolare e della lingua speciale (scientifica) che lo caratterizza. Alla base troviamo quindi il testo specialistico. I testi specialistici si differenziano dai testi di lingua generale per la loro ricchezza terminologica, fornendo la prova dell'esistenza dei termini utilizzati dagli esperti. In più, essi forniscono ai terminografi altre informazioni terminologiche, ad esempio, la frequenza con cui i termini vengono utilizzati. Esistono poi diversi tipologie di corpus: il corpus monolingue, che è costituito da testi in una sola lingua; il corpus bilingue o multilingue, che sono composti da testi in due o più lingue e utilizzati principalmente per cercare corrispondenze interlinguistiche in più di una lingua. Un altro tipo di corpus molto importante è il corpus annotato ossia un corpus che viene manipolato per inserire informazioni sulla sintassi, la morfologia o la semantica dei testi che lo compongono. Proseguendo con l'analisi teorica dei corpora, abbiamo presentato la linguistica dei corpora, branca della linguistica che rappresenta una metodologia per l'analisi quantitativa e qualitativa dell'uso della lingua. Infine, abbiamo parlato dei metodi per costruire un corpus di buona qualità. Prima di tutto, è necessario che i testi selezionati per costruire il corpus siano testi specializzati e soprattutto affidabili. Le fonti sono una parte importante per capire se un testo è autorevole. In seguito, è fondamentale scegliere il giusto software di analisi e trattamento dei corpora. Alcuni esempi sono AntConc, Sketch Engine e TermoStat. Tutto questo va preso in considerazione per poter costruire dei corpora qualitativamente adeguati a realizzare progetti terminografici che possano fornirci una terminologia soddisfacente.

Con il quarto capitolo, ci avviamo sempre di più verso la parte pratica del progetto. Questo capitolo infatti è dedicato all'estrazione della terminologia, sia dal punto di vista teorico, sia dal punto di vista pratico. Per cominciare, ci siamo soffermati nel dettaglio sui termini. Secondo l'Organizzazione internazionale per la normazione (ISO), un termine è la designazione linguistica di un concetto all'interno di un settore specialistico (ISO 1087: 2019). Esistono due principali tipologie di termini: i termini semplici, formati da una sola unità lessicale, come "interpretabilità", e i termini complessi, formati da due o più unità lessicali, come "linguaggio di programmazione". In seguito, abbiamo esaminato come si formano i termini e le principali tecniche sono le neoformazioni derivazionali o composizionali. Tuttavia, i termini possono essere creati anche a partire dalla rideterminazione semantica di parole del lessico generale o del lessico di altre

scienze. I termini possono essere anche formati da sigle o abbreviazioni e da denominazioni eponime da nomi propri. Inoltre, la formazione di nuovi termini può anche partire da prestiti o calchi da lingue straniere. Nella maggior part dei casi, i termini sono formati da sostantivi e spesso i dizionari tengono in considerazione solo i termini formati da nomi, tendendo a tralasciare i termini che possono essere rappresentati da verbi, aggettivi o avverbi. Questo avviene perché normalmente le entità del mondo reale sono indicate con dei nomi, tuttavia, ciò non significa che le altre parti del discorso non possano avere uno status terminologico rilevante. Lo studio approfondito del termine ci ha fatto capire quanto sia importante l'esame dei termini in relazione ai testi che li contengono, cioè al loro ambiente linguistico. Per questo i terminografi possono utilizzare dei controlli lessico-semantici per confermare il significato specifico di un termine appartenente ad un campo specialistico della conoscenza. Un esempio di questi testi è la sostituzione tramite sinonimi, una strategia utilizzata per verificare se è possibile sostituire un'altra unità lessicale a un termine in un contesto specifico.

Dopo questa breve ma attenta introduzione al concetto di termine, siamo passati all'estrazione automatica dei termini. Si tratta di una delle tappe più importanti nel processo terminologico e oggi, grazie alle nuove tecnologie, i terminografi possono usufruire degli estrattori terminologici, ossia dei programmi informatici dedicati a questa attività di estrazione. L'estrattore terminologico elabora automaticamente una lista di unità lessicali o sequenze di unità lessicali in un testo o in un corpus. Questa lista contiene quindi i possibili termini, dei candidati, che potrebbero corrispondere a unità terminologiche reali. Le tecniche che si trovano alla base di questi estrattori sono diverse. Una di questa prevede il confronto tra due corpus, un corpus di riferimento e un corpus specializzato. In questo caso i termini di un campo specializzato possono essere selezionati tramite una lista di esclusione, escludendo le parole grammaticali (senza senso proprio) e altre parole frequenti. Questa strategia è molto semplice, tuttavia utilizza l'indice della frequenza solo in relazione al numero di occorrenze di una parola quando non è scontato che una parola sia un termine specializzato perché è frequente in un corpus. Altre tecniche sono il conteggio di unità lessicali frequenti in corpora specializzati e l'identificazione di pattern tipici. L'estrattore terminologico che abbiamo usato per il progetto è TermoStat, un estrattore che unisce il confronto tra un corpus di riferimento e uno specializzato e i vari altri metodi di estrazione. TermoStat fornisce una lista di

possibili termini indicando la loro frequenza, l'analisi delle specificità, che permette di estrarre il linguaggio peculiare relativamente alle singole parti di una partizione (cf. Bolasco 2005, 40), le varianti ortografiche e l'indicazione della classe grammaticale. Nonostante lo sviluppo attuale degli estrattori terminologici, essi presentano comunque dei problemi e dei limiti. Infatti, quando i terminografi si trovano davanti a una lista di possibili termini creata da un estrattore, dovrebbero sempre procedere con una pulizia e un'analisi manuali. Alcuni termini, infatti, possono non essere pertinenti per il dominio studiato; oppure, possono essere presenti delle sequenze di unità lessicali che non sono termini a prescindere, perché incomplete o senza uno status terminologico appropriato.

Quando la lista dei termini sarà completa e ripulita, si può passare alla fase successiva: la compilazione delle schede terminologiche, passo imprescindibile per fornire la diffusione di dati terminologici soddisfacenti. Sempre nel quarto capitolo, quindi, ci siamo occupati di spiegare come funziona la compilazione di schede terminologiche e abbiamo presentato lo strumento che abbiamo utilizzato in questo processo: FAIRterm. La scheda terminologica è il supporto dove i terminografi registrano i dati terminologici relativi a un concetto seguendo degli standard prestabiliti. Questo documento contiene le informazioni su un termine ed è composto da più campi. Le schede terminologiche possono essere monolingue e quindi prevedere il solo termine in analisi, o bilingue (e multilingue), ossia prevedere anche l'inserimento della traduzione del termine. Lo scopo di una scheda terminologica è quello di fornire una buona comprensione dei termini per poterli usare correttamente. Per quanto riguarda il nostro progetto, abbiamo usufruito di FAIRterm, uno strumento creato per l'organizzazione ottimale dei dati terminologici. FAIRterm segue i principi equi stabiliti dall'associazione *European Open Science Cloud* (EOSC), al fine di garantire la reperibilità, l'accessibilità, l'interoperabilità e il riutilizzo della terminologia. La scheda terminologica alla base di FAIRterm offre ai terminografi una solida base per un'analisi completa dei termini specialistici: le informazioni terminologiche sono raggruppate in quattro tipologie, che sono caratteristiche formali, semantica, variazione e uso. Ogni gruppo contiene poi molteplici campi in cui inserire le informazioni più varie, ad esempio, nel gruppo "semantica" possiamo inserire informazioni relative ai sinonimi, ai meronimi e agli iperonimi di un termine; o ancora, nella sezione "variazione", possiamo inserire se il termine ha una sigla o se ha varianti ortografiche; mentre nella sezione "uso", possiamo

inserire il dominio e il sotto-dominio, il contesto nel quale abbiamo trovato il termine (un esempio di frase contenente il termine) e il registro. Si tratta quindi uno strumento completo e facile per completare il processo terminologico e per aiutarci a fornire le conoscenze adeguate alla diffusione della terminologia relativa a un certo dominio.

In relazione al progetto *YourTerm TECH*, bisogna riconoscere che non è stato facile portare a termine certi passaggi. Anzitutto, la costruzione dei corpora si è rivelata complicata in relazione all'affidabilità delle fonti, impedimento superato grazie all'analisi accurata dei singoli documenti che abbiamo anche presentato nel capitolo finale in quanto relativo all'esame qualitativo di tutto il progetto. Altre difficoltà sono state riscontrate nella compilazione di alcune schede terminologiche. Il dominio dell'intelligenza artificiale e quello dell'elaborazione del linguaggio naturale sono due discipline complesse, ricche di una terminologia talvolta complicata e molto ricercata. Per questo motivo, la compilazione delle schede termologiche non si è riscontrata troppo facile. Alcuni campi terminologici come l'etimologia e la definizione sono stati compilati incontrando delle difficoltà. Parlando delle definizioni dei termini, ad esempio, non sempre potevamo contare su una definizione prestabilita e disponibile. In questi casi, grazie alla ricerca e alle informazioni raccolte in seguito, siamo riusciti a costruire autonomamente le definizioni, indicando ovviamente il fatto che la definizione fosse stata creata o manipolata dell'autore della scheda. I problemi maggiori sono spesso legati ai termini complessi, che sono formati talvolta da lunghe sequenze di parole e da sintagmi preposizionali o aggettivali. Il pattern più comune riscontrato è un sintagma nominale modificato da un sintagma preposizionale o da un altro sintagma nominale.

In conclusione, tutto il progetto terminologico e la stesura della tesi ci hanno aiutato a capire ancora di più l'importanza della terminologia, partendo da una solida base teorica e proseguendo con un buon metodo di lavoro. Il progetto *YourTerm TECH* è stato fondamentale per capire come poter agevolare l'accesso alla terminologia dell'intelligenza artificiale e dell'elaborazione del linguaggio naturale (e di tutti agli altri domini esistenti), per garantire la diffusione delle conoscenze tenendo conto delle esigenze comunicative di tutti.