Introduction

La langue est un outil central dans notre vie sociale et professionnelle. Il s'agit d'un support pour véhiculer, entre autres, des idées, des informations, des opinions et des sentiments ainsi que pour persuader, demander des informations, donner des ordres, etc. L'intérêt pour la langue d'un point de vue informatique a commencé au début des travaux en informatique elle-même, notamment dans le cadre des travaux dans le domaine de l'intelligence artificielle. En effet, le test de Turing, l'un des premiers tests développés pour juger si une machine est intelligente ou pas, stipule que pour être considérée comme intelligente, la machine doit posséder des capacités conversationnelles comparables à celles d'un humain [TUR 50]. Cela sous-entend qu'une machine intelligente doit posséder des capacités de compréhension et de génération, au sens large de ces termes, d'où un intérêt pour le traitement automatique des langues (TAL) à l'aube de l'ère informatique. Historiquement, le traitement informatique des langues s'est très vite orienté vers les domaines applicatifs, notamment la traduction automatique (TA) dans le contexte de la guerre froide. Ainsi, le premier système de TA a vu le jour après un projet commun entre l'université Georgetown et IBM aux États-Unis [DOS 55] et [HUT 04]. Ces travaux applicatifs n'ont pas été couronnés du succès escompté et les chercheurs se sont vite rendu compte qu'une compréhension profonde du système linguistique est une condition préalable à toute application heureuse.

La vague d'Internet entre le milieu des années 1990 et le début des années 2000 a été un moteur très important pour le TAL et les domaines dérivés, notamment celui de la recherche d'information, qui est passé d'un domaine marginal, limité à la recherche d'information sur le plan d'une grande entreprise, à la recherche d'information à l'échelle d'Internet, dont le contenu ne cesse de s'élargir. Ce développement en matière de disponibilité de données a également favorisé une discipline qui existait déjà à l'état de germe : la science des données (*Data Science*). Située à l'intersection des statistiques, de l'informatique et des mathématiques, la science des données est une discipline qui s'intéresse à l'analyse, la visualisation et le traitement des données

numériques sous toutes leurs formes : images, textes et parole. Le rôle du TAL au sein de la science des données est évident, étant donné que la majorité des informations traitées sont contenues dans des documents écrits ou des enregistrements de parole. Il est ainsi possible de distinguer deux approches différentes mais complémentaires des recherches dans le domaine du TAL. D'une part, il y a les travaux qui visent à la résolution du problème fondamental du traitement de la langue et qui s'intéressent, par conséquent, aux aspects cognitifs et linguistiques de ce problème. D'autre part, de nombreux travaux sont consacrés à optimiser et à adapter les techniques existantes de TAL à des domaines applicatifs divers, comme le domaine médical ou le secteur bancaire.

Ce livre a pour objectif de faire un bilan panoramique des travaux tant classiques que modernes dans les domaines des bases de données lexicales et de la représentation des connaissances pour le TAL, de la sémantique, de l'analyse de discours et des applications du TAL, comme la traduction automatique et la recherche d'information. Il se veut également profondément interdisciplinaire en considérant, autant que possible, sur un pied d'égalité les modèles linguistiques et cognitifs, les algorithmes et les applications informatiques, car nous partons de l'opinion, maintes fois attestée dans le TAL et ailleurs, que les meilleurs résultats ne peuvent être que le fruit du mariage d'une bonne théorie avec une approche empirique bien conçue.

En plus de l'introduction, ce livre est constitué de quatre chapitres. Le premier porte sur le lexique et la représentation des connaissances. Après une introduction aux principes de la sémantique lexicale et des théories du sens lexical, il couvre les bases de données lexicales, les principaux formalismes de représentation des connaissances ainsi que les ontologies. Le deuxième chapitre est consacré à la sémantique. Tout d'abord, nous présentons les approches principales en sémantique combinatoire, comme la sémantique interprétative, la sémantique générative, la grammaire de cas, etc. De plus, nous consacrons une section aux approches logiques de la sémantique formelle utilisées dans le domaine du TAL. Le troisième chapitre porte sur le discours. Il couvre les notions fondamentales en analyse de discours comme l'énonciation, la progression thématique, la structuration de l'information dans le discours, la cohérence et la cohésion. De même, sont présentées différentes approches de traitement du discours comme la segmentation linéaire, l'analyse et l'interprétation du discours et la résolution de l'anaphore. Le quatrième et dernier chapitre est dédié aux applications du TAL. Tout d'abord, nous présentons les aspects fondamentaux des systèmes de TAL, comme les architectures logicielles et les approches pour l'évaluation. Ensuite, nous passons en revue des applications particulièrement importantes dans le domaine du TAL, comme la traduction automatique, la recherche d'information et l'extraction d'information.