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 dont notamment la traduction automatique (TA) dans le contexte de la guerre froide. Ainsi, le premier système de TA a vu le jour suite à un projet commun entre l'université Georgetown et IBM aux Etats-Unis [DOS 55, 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. Ce constat, formulé par le fameux rapport de la commission ALPAC, a eu un impact considérable sur les travaux en TA et dans le domaine du TAL en général. Aujourd'hui bien que le TAL se soit largement industrialisé, l'intérêt pour les traitements linguistiques de base n'a pas cessé. En effet, quel que soit le domaine applicatif visé par un système de TAL moderne, le recours à un module de traitement linguistique de bas niveau comme un analyseur morphologique, syntaxique ou un module de reconnaissance ou de synthèse de la parole, est presque toujours indispensable (voir [JON 11] pour une revue plus complète de l'historique du TAL).

Définition du TAL

Tout d'abord, qu'est-ce que le TAL ? Il s'agit d'une discipline qui se trouve à l'intersection de plusieurs autres branches de la science comme l'informatique, l'intelligence artificielle, la linguistique et la psychologie cognitive. En français, il est plusieurs termes qui désignent des champs proches. Bien que les frontières entre les champs désignés par ces termes ne soient pas toujours très nettes, nous allons tenter d'en offrir une définition sans pour autant prétendre qu'elle fasse l'unanimité dans la communauté. Par exemple, les termes *linguistique informatique* ou *linguistique computationnelle* portent plus sur les modèles ou les formalismes linguistiques développés dans le but d'une implantation informatique. Les termes *industries de la langue, ingénierie linguistique* ou encore *traitement informatique de la langue* concernent plutôt l'édition de logiciels équipés de fonctionnalités liées au TAL. Par ailleurs, le *traitement automatique de la parole* (TALP) désigne une panoplie de techniques allant du traitement de signal jusqu'à l'identification ou la production d'unités linguistiques comme les phonèmes, les syllabes ou les mots. Mise à part la dimension de traitement de signal, le TALP ne présente aucune différence de fond par rapport au TAL. En effet, beaucoup des techniques qui ont été appliquées initialement au traitement de la parole ont trouvé leur chemin vers des applications en TAL dont notamment les chaînes de Markov cachées. Cela nous a encouragés à suivre dans ce livre le chemin unificateur, déjà emprunté par d'autres collègues, qui consiste à mettre dans la même discipline le TAL et TALP (pour un exemple d'approches similaires, voir [JUR 00] qui est toujours le principal livre d'introduction au TAL aux Etats-Unis). Finalement, la *linguistique de corpus* réfère aux méthodes de collecte, d'étiquetage et d'usage des *corpus* à la fois dans les études linguistiques ou de TAL. Comme les corpus occupent une place très importante dans le processus de construction des systèmes de TAL, notamment ceux qui adoptent une approche à base d'apprentissage, nous avons jugé bon, à l'instar d'autres collègues, de considérer la linguistique de corpus comme une branche du TAL.

Dans les sections suivantes, nous allons présenter et discuter les relations qu'entreprend le TAL avec des disciplines connexes comme la linguistique, l'intelligence artificielle (IA), et les sciences cognitives.

TAL et linguistique

Avec la démocratisation des outils de TAL, ces outils font aujourd'hui partie de la trousse de travail de beaucoup de linguistes notamment pour les travaux empiriques à l'échelle d'un corpus. Ainsi, les étiqueteurs en partie de discours, les analyseurs morphologiques et les analyseurs syntaxiques de types variés sont fréquemment utilisés dans les études quantitatives ou pour trouver les données nécessaires pour une

expérience psycholinguistique. Par ailleurs, le TAL offre une perspective nouvelle à l'investigation linguistique et cognitive des langues humaines en ajoutant une nouvelle dimension à ces travaux, celle de la testabilité. En effet, bien des modèles théoriques ont été testés empiriquement à l'aide d'applications de TAL. De l'autre côté, les théories linguistiques restent une source importante pour les systèmes de TAL dont beaucoup en sont inspirés.

TAL et IA

L'IA vise l'étude, la conception et la création d'agents intelligents. Un agent intelligent étant un système naturel ou artificiel doté de capacités perceptives qui lui permettent d'agir dans un environnement donné dans le but de satisfaire ses désirs ou d'atteindre des objectifs préétablis avec succès (voir [MAR 14a] et [RUS 10] pour une introduction générale). Les travaux en IA sont généralement classés dans plusieurs sous-disciplines ou branches comme la représentation des connaissances, la planification, la perception et l'apprentissage. Toutes ces branches sont directement liées au TAL. Cela confère à la relation entre l'IA et le TAL une dimension très particulière. En effet, beaucoup vont jusqu'à considérer le TAL comme une branche de l'IA alors que certains préfèrent considérer le TAL comme une discipline plus indépendante.

Dans le domaine de l'IA, la planification consiste à trouver les étapes à suivre pour atteindre un objectif donné, et ce, à partir d'une description des états initiaux et des actions possibles. Pour un système de TAL, la planification est nécessaire pour effectuer des tâches complexes impliquant plusieurs sources de connaissance qui doivent coopérer pour atteindre l'objectif final.

La représentation des connaissances est importante sur deux plans pour un système de TAL. D'une part, elle fournit les cadres pour représenter les connaissances linguistiques nécessaires pour le bon fonctionnement de tout système de TAL même si l'importance et la quantité de ces informations déclaratives varient considérablement selon l'approche adoptée. D'autre part, certains systèmes de TAL ont besoin d'informations extralinguistiques pour prendre des décisions notamment en cas d'ambiguïtés. Ainsi, certains systèmes de TAL sont couplés à des ontologies ou à des bases de connaissances sous forme de réseau sémantique, de schéma ou de graphes conceptuels.

La perception semble, *a priori*, loin de la langue mais en réalité ce n'est pas le cas lors qu'il s'agit de la langue parlée où le message linguistique est véhiculé par les ondes vocales. Ainsi, le couplage de la reconnaissance vocale, l'équivalent de la perception avec les modules linguistiques de compréhension, est indispensable non

seulement pour la compréhension mais également pour améliorer la qualité de la reconnaissance. Par ailleurs, certains projets actuels portent sur le couplage des modules de compréhension linguistique avec des modules de compréhension de scènes visuelles.

L'apprentissage automatique consiste à construire une représentation suite à l'examen de données préalablement annotées ou pas. L'apprentissage a pu acquérir un intérêt particulier au sein de l'IA à partir des années 2000 notamment grâce aux possibilités qu'il offre pour construire des systèmes intelligents avec un effort minimal comparé aux systèmes symboliques qui nécessite un recours plus intensif à des experts humains. Dans le domaine du TAL, la fréquence d'adoption des approches à base d'apprentissage diffère considérablement suivant les niveaux linguistiques visés. En effet, cela varie entre une domination quasi totale dans les systèmes de reconnaissance de la parole jusqu'à une adoption limitée dans les traitements de haut niveau comme le discours et la pragmatique où le paradigme symbolique est encore dominant.

TAL et sciences cognitives

Tout comme avec la linguistique, la relation entre les sciences cognitives et le TAL a un double sens. D'une part, les modèles cognitifs peuvent servir de support pour un système de TAL et d'autre part, construire un système de TAL suivant un modèle cognitif peut être un moyen pour tester ce modèle. L'apport pratique d'une approche qui mime le processus cognitif reste une question ouverte, car dans beaucoup de domaines construire un système qui s'inspire des modèles biologiques ne s'est pas avéré productif. Signalons par ailleurs, que certaines tâches des systèmes actuels de TAL n'ont pas de parallèles chez les humains comme la recherche d'information à l'échelle des moteurs de recherche ou la fouille de quantités importantes de données textuelles pour en extraire des informations utiles. Le TAL peut être vu comme une extension des capacités cognitives des humains dans le cadre d'un système d'aide à la décision par exemple. D'autres sont très proches comme la compréhension et la génération.

TAL et science des données

Avec la disponibilité de plus en plus de données numériques, une nouvelle discipline a récemment émergé. Il s'agit de la science des données. Elle vise l'extraction, la quantification et la visualisation des connaissances à partir de ces données qui sont principalement textuelles et parlées. Etant donné que dans beaucoup de cas ces données sont en langage naturel, le rôle du TAL au sein du processus d'extraction et de traitement est évident. Actuellement, vu les innombrables

applications dans le secteur industriel tant dans les domaines du marketing que dans les domaines de la prise de décision, la science des données prend une ampleur qui rappelle celle du début d'Internet dans les années 1990. Cela donne au TAL, une dimension applicative confirmée qui n'est pas inférieure à sa dimension scientifique.

Structure du livre

Ce livre a pour objectif de faire un bilan panoramique des travaux tant classiques que modernes dans le domaine du TAL. Il propose une vision unifiée de domaines considérés souvent différents tels que le traitement de la parole, la linguistique computationnelle, le traitement automatique des langues et l'ingénierie des connaissances. Il se veut également profondément interdisciplinaire en considérant, tant 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. Bien entendu, nous ne prétendons pas ici que ce livre couvre la totalité des travaux mais nous avons tenté de balancer entre les travaux francophones, européens et mondiaux. Notre démarche s'inscrit ainsi dans une perspective double celle de l'initiation pédagogique accessible et celle d'un état de l'art d'un domaine, mur certes, mais toujours en évolution.

Par conséquent, ce livre adopte une démarche qui consiste à rendre accessibles les concepts linguistiques et informatiques à travers des exemples soigneusement choisis. Par ailleurs, bien qu'il cherche à donner le maximum de détails sur les approches présentées, il reste néanmoins neutre à l'égard des détails d'implantation pour laisser à chacun une marge de liberté en ce qui concerne les langages de programmation suivant ses préférences personnelles ainsi que ses besoins objectifs.

En plus de l'introduction, ce livre est constitué de quatre chapitres. Le premier porte sur les ressources linguistiques pour le TAL. Il présente les différents types de corpus, leur collecte, ainsi que leurs méthodes d'annotation. Le deuxième chapitre porte sur la parole et son traitement. Tout d'abord, nous présentons les concepts fondamentaux en phonétique et en phonologie ensuite nous passons aux deux applications les plus importantes dans le domaine de la parole : la reconnaissance et la synthèse. Le troisième chapitre concerne le niveau des mots et s'intéresse en particulier à l'analyse morphologique. Enfin, le quatrième chapitre couvre le domaine de la syntaxe. Les concepts fondamentaux et les théories syntaxiques les plus importantes y sont présentés ainsi que les différentes approches d'analyse syntaxique.