Introduction au TALN et à l'ingénierie linguistique, Isabelle TELLIER
Compte rendu, Chapitre 1 et 2

L'introduction du livre d'Isabelle Tellier nous donne d'emblée le thème de l'ouvrage, bien plus explicitement encore que ne le fait le titre : le langage et l'informatique, ainsi que l'interaction de ces deux domaines, avec pour but, ou peut être pour rêve, de donner la parole aux ordinateurs.

S'ensuit alors un bref historique des travaux de recherche s'étant penché sur la question. L'auteur distingue 3 étapes importantes.

La première renvoie à une conférence ayant eu lieu dans les années 1950, réunissant les pionniers du traitement automatique des langues (TAL). C'est alors la traduction automatique qui est à l'honneur, car pouvant avoir un réel potentiel en période de Guerre Froide. Malheureusement, il y aura peu de réussite dans ce domaine.

La deuxième étape se situe dans les années 70 et 80. L'intelligence artificielle se développe et se fait connaître par le biais de « systèmes experts » (des machines qui simulent le raisonnement d'un expert). Mais pour autant, les applications du TAL restent le pus souvent à l'étape de laboratoire, même si cela n'empêche pas le développement de nombreux concepts et modèles dans le domaine.

La dernière étape renvoie cette fois aux importants changements informatique et au développement d'internet dans les années 1990. Elle correspond également à la naissance de l'ingénierie linguistique, qui permet de créer des programmes permettant de travailler sur les nombreux textes disponibles (sur internet notamment) en les classant ou en extrayant des données par exemple.

Nous entrons ensuite dans le vif du sujet : le chapitre 2, intitulé « Traitement automatique du Langage Naturel ».

Après avoir exposé quelques données de bases (comme le nombre de langues dans le monde – environ 5000) et quelques définitions utiles, autant du point du vue linguistique (la distinction entre langue naturelle et artificielle, l'objet d'étude des sciences du langage, ou encore l'intention descriptive du linguiste) qu'informatique (en en définissant les 2 fondements, « le codage des données à l'aide d'éléments discrets et le codage effectif des traitements à l'aide d'algorithmes »), on regroupe ces deux domaines en une seule notion : le TALN (traitement automatique du langage naturel).

Dans un bref historique, on nous expose ensuite les évènements importants de l'histoire de la linguistique, de l'informatique et de ces deux domaines réunis.

En ce qui concerne la linguistique, on peut notamment distinguer, par ordre chronologique : les premières grammaires descriptives fondatrices (ayant déjà cours avant le XVIIIe siècle), puis le règne de la linguistique comparative et historique, aux XVIIIe et XIXe siècle, et la publication du *Cours de linguistique générale* de Saussure (ouvrage publié à titre posthume à partir des notes de cours de ses étudiants). On note aussi l'avènement du Cercle de Prague dans les années 1930-40 (avec notamment Troubetzkoy et Jakobson), Martinet et sa théorie sur la double articulation, une des propriétés fondamentales des langues naturelles, Chomsky et ses travaux sur la grammaire universelle innée ainsi que l'essor de la grammaire générative (toujours avec Chomsky) dans les années 1960.

Pour l'informatique, on retiendra surtout que ses précurseurs était des mathématiciens (comme Pascal ou Leibniz), que l'émergence de la logique booléenne a favorisé l'automatisation des tâches et que c'est en 1945 qu'est défini le plan des constructions des ordinateurs, par Von Neumann.

En ce qui concerne les deux disciplines réunies, leur première association se manifeste avec l'élaboration du test de Turing par le mathématicien anglais du même nom, test élaboré pour « juger de la capacité des machines à penser ». Il vient ensuite la première conférence sur la traduction automatique (déjà évoquée dans l'introduction), ainsi que l'invention du terme « Intelligence

artificielle » (IA), tous les deux dans les années 1950. Dans les années 60, on assiste à la création de deux programmes célèbres, Eliza par Weizenbaum et Student, de Bobrow, programmes capables d'un dialogue basique, suivant un système de mots clés et de situations préalablement définies. C'est aussi à cette période que se développe l'étude des propriétés des langages artificiels et de programmation informatique (avec la théorie des langages formels). Dans les années 1970 et 1980, c'est autour de la sémantique formelle d'être à l'honneur, puis dans les années 1990, celui de la linguistique de corpus.

Il y a plusieurs unités et plusieurs niveaux d'analyses possibles lorsqu'on travaille sur les langues naturelles.

Les unités possibles sont les phonèmes, les morphèmes, et les énoncés, qui correspondent à des données discrètes.

Les niveaux d'analyse possible sont : celui qui contient l'axe syntagmatique et paradigmatique du langage, celui correspondant à la double articulation du langage et enfin celui qui opère la distinction entre sémantique lexicale et propositionnelle.

On peut encore après distinguer quelques notions utiles : l'axe syntagmatique correspond à un assemblage d'éléments dont les possibilités sont nombreuses, voire infinies, et l'association signifiant – signifié semble être le propre de l'homme. De plus, l'auteur fait un lien entre langues naturelles et langages de programmation informatique. En effet, ces derniers possèdent eux aussi la double articulation, et peuvent avoir une sémantique.

Une modélisation informatique des langues naturelles semble donc possible : la difficulté de l'entreprise ne résidera pas alors dans la manière de coder des informations, mais dans la traduction de la combinaison des éléments an algorithme, ainsi que dans e codage de la dimension sémantique.

Une autre difficulté majeure dans la modélisation des langues naturelles, c'est de définir de quelle façon le langage est « traité » par les humains, afin de pouvoir espérer reproduire et adopter cette logique aux machines. On peut diviser le traitement du langage en deux catégories : d'une part les données traitées, de l'autre les traitements appliqués par les locuteurs (si l'on simplifie à l'extrême). Pour autant, l'IA ne cherche plus aujourd'hui à se claquer sur cette démarche pour arriver à ses fins : on tente alors de partir des capacités des machines plutôt que d'essayer de reproduire à tout prix celle des humains.

Le chapitre finit par un bref retour sur le test de Turing (ce qui montre bien l'aspect fondamental de ce test, encore utilisé aujourd'hui pour mesurer l'efficacité des chatbots), en proposant quelques sites regroupant les « meilleurs » chatbots à ce jour (les machines ayant obtenues les scores les plus élevés au test de Turing).