forme GATE[3] – exemplaire pour sa modularité et ses interfaces standardisées, ou l'initiative DOBES[4] – exemplaire pour sa gestion de métadonnées linguistiques.

Pour résumer, la dichotomie entre « ressources et instruments » et « méthodologie » a tendance à favoriser la description de l'existant et du provisoire au détriment d'une vision des enjeux scientifiques et technologiques à venir, et brouille quelquefois la clarté du propos. En particulier, certaines thématiques transversales et centrales du domaine se retrouvent éclatées sur plusieurs chapitres et font l'objet de redites : la normalisation (à ce propos, aucune mention du TC 37/SC 4 ?), l'interfaçage des ressources avec des outils, ou la constitution de corpus. Si l'ouvrage représente une bonne introduction aux ressources, outils et pratiques existantes, il n'est pas véritablement convaincant – tout simplement, parce qu'il ne cherche pas vraiment à convaincre. A certains égards, il traduirait presque un constat d'impuissance devant les résultats hétérogènes de l'évolution inéluctablement électronique du français. Pourtant, cette évolution fait émerger de nouveaux enjeux scientifiques et techniques à l'intersection de la linguistique, de l'informatique et des sciences de l'information. C'est cette voie qui mérite d'être creusée, et pourquoi pas en suivant l'invitation à l'action, formulée par B. Habert à la page 10 de l'ouvrage : « *La linguistique se trouve désormais en mesure de recourir à de nouveaux instruments et à des données renouvelées. Il lui faut, dans le même mouvement, en mesurer la nature et l'adéquation à ses objectifs propres, quitte à intervenir pour infléchir l'usage de ces moyens, voire leur conception même*. »

Patrice ENJALBERT, Sémantique et traitement automatique du langage naturel, *Hermès-Lavoisier*, 2005, 410 pages, ISBN 2-7462-1126-2.

par Pascal AMSILI

Université Paris 7 & Lattice
amsili@linguist.jussieu.fr

Cet ouvrage est un recueil de 10 articles, dont plus de la moitié co-écrits par Patrice Enjalbert. L'objectif est double : d'une part présenter les travaux menés depuis 15 ans à Caen sur la question du sens en TAL, et d'autre part illustrer l'actualité des recherches sémantiques en TAL, et proposer méthodologie et état de l'art sur cet aspect.

Sémantique et TAL, quel rapport ?

La question peut paraître saugrenue, mais elle mérite qu'on s'y arrête. On peut dire que le TAL, en tant que domaine de recherche, ne se pose en général pas la

[3] http://gate.ac.uk/
[4] http://www.mpi.nl/DOBES/

question dans les termes de la tripartition habituelle entre *syntaxe* (relation des signes linguistiques entre eux), *sémantique* (relation des signes à leur dénotation) et *pragmatique* (relation des signes à leurs utilisateurs). En effet, le TAL est défini essentiellement par sa dimension applicative : il s'agit de mettre au point outils et méthodes permettant de réaliser des traitements de matériau linguistique, lequel est dans tous les cas porteur de sens (et c'est précisément parce qu'il y a du sens que nous utilisons des applications de TAL). Bien entendu, les traitements envisagés travaillent par définition sur la *forme* du matériau (signal pour la parole, chaîne de caractères pour l'écrit). Il n'y a donc pas d'application de TAL (au sens où nous venons de le définir) qui échappe à la présence de ces trois niveaux : un correcteur orthographique, par exemple, dont on peut penser qu'il s'intéresse essentiellement à la surface (au sens linguistique), ne peut faire abstraction de la sémantique (quand, par exemple, la désambiguisation du sens en contexte permet de désambiguiser syntaxiquement, et ainsi de résoudre un problème d'accord), ni de la pragmatique (puisque la plupart des correcteurs tentent d'intégrer des règles dites d'usage). On peut remarquer d'ailleurs que le TAL ne s'intéresse pas davantage à la syntaxe, ou du moins ne s'intéresse pas à la syntaxe comme fin: déterminer la structure syntaxique d'une chaîne de mots peut être utile pour mener à bien certaines applications, mais, le problème étant notoirement difficile, nombre d'applications se contentent d'approximations sur cette structure syntaxique, ce qui est parfaitement justifié si le traitement visé en est rendu suffisamment efficace.

D'un point de vue différent, on peut considérer la tripartition comme une sorte de guide méthodologique pour élaborer l'architecture d'une application de TAL. On aurait un module syntaxique, un module sémantique, et un module pragmatique. Mais la plupart des applications de TAL ont une architecture tout autre, et certains modules classiques peuvent être vus comme relevant de plusieurs de ces niveaux (l'étiquetage, par exemple, plutôt (morpho-)syntaxique, demande parfois une désambiguïsation des sens en contexte).
A ce point, on pourrait conclure comme le titre provocateur de ce compte-rendu de lecture le suggère, qu'il n'y a pas de pertinence à tenter de rapprocher les deux termes.

Cependant, et c'est ce que montre l'ouvrage dirigé par Patrice Enjalbert, la question devient pertinente si l'on adopte le bon point de vue. D'une part, on peut noter que le TAL puise largement dans les résultats de la linguistique (formelle)[5]. Il est donc naturel de s'interroger sur la façon dont les recherches en sémantique peuvent inspirer les méthodes et les algorithmes du TAL. D'autre part, on peut ouvrir les perspectives, en s'inscrivant non plus dans le TAL au sens étroit, mais dans ce qu'on pourrait appeler la linguistique informatique, ou computationnelle. Si tant est qu'un tel domaine de recherche existe et se distingue de celui de la linguistique formelle, on peut le caractériser de la façon suivante : l'objet d'étude est la langue, et

[5] Bien entendu, il existe des approches du TAL qui ne sont pas linguistiquement inspirées (certaines méthodes probabilistes ou non symboliques). Elles ne sont pas moins légitimes, mais la question traitée ici ne les concerne pas.

les méthodes d'investigation habituelles de la linguistique sont doublées d'un souci de formalisation implémentable. Selon ce point de vue, il devient tout à fait pertinent de s'intéresser à la sémantique. En effet, la sémantique formelle (qu'elle soit lexicale ou non lexicale) élabore des modèles qui peuvent être implémentés, ce qui facilite leur mise à l'épreuve des données. L'ouvrage proposé me semble s'inscrire pleinement dans cette perspective. Il comporte à la fois des discussions sur l'implémentation de modèles linguistiques (2e partie) et des questions sur la façon de « faire entrer de la sémantique » dans des applications standard de TAL (3e partie). C'est la raison pour laquelle il me semble que « sémantique et linguistique informatique » aurait pu faire un meilleur titre. Nous avons donc affaire à un projet pertinent et cohérent.

Ce projet est cependant extrêmement vaste : d'une part, le nombre de modèles sémantiques potentiellement implémentables est important, d'autre part, comme nous le disions plus haut, la sémantique est pertinente à tous les étages des applications TAL. C'est sans doute ce qui explique l'autre parti pris de cet ouvrage : il s'agit d'un ouvrage centré autour des travaux menés à Caen (et à Paris) par une petite équipe de chercheurs, autour de Patrice Enjalbert, Bernard Victorri et Laurent Gosselin. Ce parti pris explique le caractère partiel de l'ouvrage, mais donne aussi une certaine cohérence à l'ensemble.

Résumé des chapitres

L'ouvrage est organisé trois grandes parties : la **première partie**, intitulée "Repères", est destinée aux débutants en sémantique. Le point de vue adopté n'y est pas linguistique, mais déjà orienté vers les problématiques traitées dans la suite. Le chapitre 1 ("sémantique et TALN, première approche") tente de mener une reflexion, sans a priori théorique, sur ce que peut être la construction du sens (et comment on pourrait la simuler). Le chapitre est basé sur de nombreux exemples de textes variés, à propos desquels est menée une reflexion linguistique sommaire pour illustrer la problématique. Le chapitre 2 ("Les paliers de la sémantique") vise à rafiner la reflexion ébauchée au chapitre 1, en considérant successivement le mot, la phrase et le texte. Comme le précédent, ce chapitre est destiné à des débutants en sémantique, et même en linguistique, et on y trouve des définitions de la morphologie, ou des relations classiques de sémantique lexicale, etc.

La **deuxième partie** s'intitule "Modélisation sémantique". Elle comprend quatre chapitres, qui sont chacun organisés autour du modèle d'un phénomène linguistique, et de sa mise en oeuvre dans un système informatique. Le premier chapitre (ch. 3, "polysémie lexicale") est consacré à un modèle de la polysémie lexicale, proposé par Bernard Victorri, et implémenté. La thèse défendue est que les sens des mots peuvent être représentés dans un espace multidimensionel, espace que l'on peut construire en utilisant la relation de synonymie. L'objectif principal du modèle est d'ordre linguistique : il est de permettre la visualisation des positions des mots dans

cet espace. Ce chapitre montre aussi comment un tel modèle mathématique peut être utilisé pour une tâche de TAL, la désambiguïsation des sens en contexte. Il s'agit donc avant tout d'une prise de position linguistique, voire cognitive, concernant le lexique et son organisation, et en second lieu de la façon d'utiliser le modèle en TAL. Le deuxième chapitre (ch. 4, "Calcul de la référence") est quant à lui centré sur une application de TAL, en l'occurrence la résolution automatique des anaphores. Il s'agit du système Calcoref, développé par Michel Dupont. La réalisation de ce système a conduit M. Dupont à élaborer une sorte de modèle, largement inspiré de la notion d'accessibilité proposée par Mira Ariel. Dans le troisième chapitre (ch 5, "Temporalité"), c'est clairement le modèle qui est premier : il s'agit du modèle linguistique de la temporalité proposé par Laurent Gosselin. Les ambitions calculatoires de ce modèle, qui vise à prédire les propriétés aspectuo-temporelles des procès décrits dans un texte, permettent d'envisager une implémentation, laquelle a été réalisée par Cédric Person, co-auteur avec L. Gosselin de ce chapitre. On est clairement ici dans le cadre de l'informatique linguistique : le système construit ne vise pas la réalisation robuste et automatique d'une tâche de TAL particulière, mais plutôt la validation "expérimentale" de la théorie de Gosselin. Le dernier chapitre de cette partie (ch 6, "Sémantique de l'espace et du déplacement"), se situe à une position intermédiaire: il ne présente pas de modèle de la spatialité, mais formule à partir d'observations linguistiques simples un "cahier des charges" que devrait respecter un tel modèle. L'auteur, Yann Mathet, dégage les grands paradigmes de relations spatio-temporelles et en propose une mathématisation (qui a fait l'objet d'une implémentation).

La **troisième partie**, "De la compréhension automatique aux applications documentaires", est d'inspiration plus pratique. Il s'agit essentiellement de passer en revue diverses tâches courantes du TAL, et de montrer ce que peut être, dans ces tâches, un composant sémantique.

Conclusion: l'ouvrage dirigé par Patrice Enjalbert présente d'une façon cohérente les efforts menés depuis une quinzaine d'année par une équipe centrée à Caen, et qui a poursuivi avec cohérence et persévérance la prise en considération du sens dans les applications de TAL. L'expérience menée est intéressante et peut inspirer les chercheurs en TAL intéressés par ce point de vue.