le traitement automatique des langues : des modèles aux ressources

Catherine Fuchs, Benoit Habert

Article paru dans *Le Français Moderne* LXXII : 1, 2004
(F. Fuchs & B. Habert, dir.)

(Version préliminaire)

Introduction
le traitement automatique des langues : des modèles aux ressources

Catherine FUCHS, Benoît HABERT

Pour des non spécialistes, le traitement automatique de la langue (désormais TAL) peut apparaître comme un domaine homogène et unifié. En fait, il n'en est rien [Cori & Léon, 2002]. Depuis la conception de modèles théoriques jusqu'à la fabrication d'outils opérationnels, s'étend une longue chaîne de travaux, dont l'hétérogénéité dans les objectifs, les méthodes et les démarches est manifeste. Pour l'utilisateur (qu'il soit linguiste, ou bien spécialiste d'une discipline le conduisant à travailler sur du texte en langue naturelle), il est essentiel d'avoir conscience de cette hétérogénéité - ainsi que du caractère nécessairement partiel et perfectible des réalisations - afin de pouvoir s'orienter au mieux dans sa quête d'outils utiles et aussi fiables que possible. Pour aider le lecteur à se situer sur ce terrain complexe, nous brosserons à grands traits un bref historique, puis nous présenterons les différents types de ressources numériques actuellement accessibles, avant d'annoncer très brièvement les différentes contributions ici rassemblées. C'est précisément aux nouveaux produits (et aux méthodes associées) offrant des accès renouvelés à certains aspects des textes en langue naturelle, que nous avons choisi de consacrer le présent numéro du Français Moderne.

1 Du traitement automatique des langues aux industries de la langue

L'hésitation entre deux dénominations, *traitement automatique des langues* et *industries de la langue* correspond en fait à une évolution actuelle importante : le reflux des grands projets censés offrir une solution globale aux problèmes du traitement d'une langue et la multiplication des produits, de plus en plus accessibles, d'ingéniérie linguistique [1]. Ces produits et les méthodes associées offrent aux utilisateurs des accès renouvelés à certaines données langagières.

1.1 Rappel

L'objectif du traitement automatique des langues est la conception de logiciels capables de traiter de façon automatique des données exprimées dans une langue (dite « naturelle », par opposition aux langages formels de la logique mathématique). Ces données linguistiques peuvent, selon les cas, être de différents types (textes écrits, dialogues écrits ou oraux, etc.) et de taille variable (du texte entier au mot isolé, en passant par la phrase ou le syntagme). Qui dit « traitement » dit manipulation d'un objet d'entrée aboutissant à la modification de cet objet en un objet de sortie. Selon la nature de l'application, le traitement peut viser à transformer des données linguistiques existantes (à des fins de correction, d'extraction d'information, de résumé, de traduction...) ou bien à en construire (génération de textes à partir d'informations). Le caractère « automatique » du traitement visé impose un certain nombre de contraintes fortes : pour que l'ordinateur puisse effectuer les calculs correspondants, les données linguistiques doivent être appréhendées de façon totalement explicite, cohérente et opératoire - d'où le recours à divers types de formalismes et de techniques informatiques ; à cet égard, il n'est pas inutile de rappeler que l'ordinateur ne sait faire que ce que le concepteur du logiciel lui aura dit de faire ! Précisons enfin que, selon les cas, le traitement peut être automatisé entièrement, ou bien seulement partiellement - auquel cas on parlera plutôt de système « assisté par ordinateur ».

1.2 A l'origine du traitement automatique des langues

Le traitement automatique des langues est né à la fin des années quarante du siècle dernier, dans un contexte scientifique et politique très précis.

D'une part la convergence d'intérêt de plusieurs scientifiques (linguistes, mathématiciens et logiciens) a été à l'origine du courant des « grammaires formelles », au milieu des années cinquante. Leur objectif était de décrire le fonctionnement des langues (conçu comme représentatif, à cet égard, du fonctionnement de l'esprit humain), à la manière d'une machine (et donc grâce à une machine), c'est-à-dire en termes de calculs correspondant au traitement d'informations diverses. Les initiateurs de ce courant ont été, aux U.S.A., Z. Harris (cherchant à caractériser les « structures mathématiques du langage », selon le titre de son ouvrage ultérieur de 1968 : [Harris, 1968]) et N. Chomsky (dans deux articles fondateurs, l'un de 1956 sur la parenté entre théorie des grammaires et théorie des automates [Chomsky, 1956], et l'autre de 1959 sur les propriétés mathématiques de diverses classes de grammaires formelles [Chomsky, 1959] ; voir aussi sa contribution de 1963 reprise dans [Chomsky & Miller, 1968]), et en France M. Gross [Gross & Lentin, 1967] et M.-P. Schützenberger. Ce courant s'originait lui-même assez largement du courant de la « cybernétique » des années quarante (von Neumann, Wiener, Türing, McCulloch) qui s'appuyait sur la logique mathématique (pour décrire le fonctionnement du raisonnement), sur la théorie des systèmes (pour formuler les principes généraux gouvernant tout système complexe) et sur la théorie de l'information de Shannon et Weaver (comme théorie statistique du signal et des canaux de communication).

D'autre part, dans le contexte politique dit de la « guerre froide » entre les U.S.A. et l'URSS, ces deux pays ont été amenés, dès la fin de la seconde guerre mondiale, à s'intéresser (à des fins d'espionnage) à un secteur d'application particulier : celui de la traduction automatique, sur

[1] Cf. l'Association des Professionnels des Industries de la Langue (http://www.apil.asso.fr/).

lequel ont toujours pesé, par ailleurs, des pressions venues du secteur économique. L'histoire de la traduction automatique comporte, schématiquement, trois moments [Fuchs, 1993b]. De la fin des années quarante jusqu'en 1965, c'est l'euphorie des débuts : Weaver propose d'utiliser les techniques du déchiffrage cryptographique pour traduire des textes de façon automatique, et en 1952 se tient au M.I.T. la première conférence sur la traduction automatique ; mais les premiers systèmes américains (systèmes dits « de première génération », fonctionnant mot à mot) s'avèrent décevants, au contraire des systèmes russes beaucoup plus élaborés. En 1965, la commission ALPAC (*Automatic Language Processing Advisory Committee*) s'interroge sur l'utilité de poursuivre les recherches dans ce domaine : dès lors, les crédits sont considérablement réduits et la recherche stagne jusqu'en 1975. Néanmoins de nouveaux systèmes (« de deuxième génération ») voient le jour, tant aux U.S.A. et au Canada qu'en France (au Centre d'Etudes pour la Traduction Automatique de Grenoble), sur divers couples de langues ; contrairement aux précédents, ces systèmes pratiquent une approche indirecte (distinguant un module d'analyse de la langue-source et un module de génération ou *synthèse* de la langue-cible), ils opèrent une stricte séparation entre les connaissances linguistiques et la partie logicielle, et ils calculent la traduction sur la base d'une analyse syntaxique préalable. Depuis 1975, l'augmentation des besoins en traduction ainsi que la sophistication croissante des outils informatiques ont conduit les activités de recherche et développement et la commercialisation de produits à prendre le pas sur les recherches plus fondamentales. C'est ainsi qu'ont été développés de nombreux systèmes d'aide à la traduction (postes de travail pour traducteurs humains) et de traduction assistée par ordinateur, dédiés à des domaines d'application précis (météorologie, notices techniques d'appareils, etc.) : on est ainsi passés du « traitement automatique de la langue » (recherche de systèmes globaux, en grandeur réelle sur toute la langue, fondés sur des théories et des concepts linguistiques) aux « industries de la langue » (développement de systèmes limités du point de vue du domaine d'application et de la couverture de la langue, et visant à répondre à un type particulier de besoin).

A l'heure actuelle, parmi les systèmes commercialisés, on distinguera :

- les traducteurs électroniques de poche, qui ne sont rien d'autre que des dictionnaires contenant quelques dizaines de milliers de mots, d'expressions idiomatiques et de phrases-types de la vie quotidienne pré-enregistrés dans plusieurs langues ;
- les systèmes restreints, robustes et simples, qui atteignent une notable efficacité au prix d'une limitation sur le texte d'entrée (qui doit relever d'un domaine fermé très limité, et se conformer à un vocabulaire et à une syntaxe contraints pré-définis) ;
- les systèmes légers, utilisables sur micro-ordinateurs personnels et contrôlés par l'utilisateur (qui complète son dictionnaire à mesure de ses besoins) : les traductions, qui se font mot à mot, ne sont pas de bonne qualité ;
- les systèmes lourds ou mi-lourds, dédiés aux entreprises effectuant de la veille technologique ou bien aux traducteurs-réviseurs, et dont les performances sont meilleures que celles des précédents, sans être pour autant excellentes.

A côté de la traduction automatique, c'est le domaine du traitement de la parole qui a été présent dès les origines du traitement automatique des langues - là encore, sous l'effet de contraintes extérieures [Lacheret, 1993]. Si les U.S.A. et le Japon se sont particulièrement mobilisés dans ce domaine, l'Europe n'a pas été en reste pour autant (en France, c'est en particulier à Grenoble et à Paris à l'ENST - Ecole Nationale Supérieure des Télécommunications - que les recherches ont été conduites). En synthèse de la parole, dès les années quarante, dans le domaine des télécommunications apparaît la nécessité d'exploiter la redondance du signal de parole en compressant ce signal, afin d'augmenter la capacité des lignes téléphoniques : d'où le développement de techniques de codage de la parole et la mise au point de synthétiseurs (le premier codeur de voix électrique, le VODER, apparaît en 1939). En 1953, une nouvelle technique de synthèse, fondée sur la simulation articulatoire du conduit vocal est présentée au M.I.T. et en 1959, deux synthétiseurs à formants sont réalisés à Stockholm et Edimbourg. Les années soixante-dix verront explorer deux voies de recherche : la reproduction du signal de parole à partir de la simulation fonctionnelle du conduit vocal humain (synthèse par formants) et la simulation de la propagation de l'onde sonore dans le conduit vocal à partir de connaissances physiologiques, articulatoires et mécaniques. En reconnaissance de la parole, les débuts sont marqués, dans les années cinquante, par la réalisation de systèmes de reconnaissance d'unités élémentaires isolées (mots, syllabes, phonèmes, traits) ; puis dans les années soixante les premiers véritables systèmes de reconnaissance de la parole continue voient le jour. Au cours de la décennie suivante, on tentera d'améliorer la performance des systèmes en intégrant des connaissances syntaxiques et sémantiques ; par la suite, les recherches porteront de nouveau sur la question du décodage acoustico-phonétique. Là encore, on voit que les préoccupations des concepteurs ont oscillé, selon les moments, entre la prise en compte des connaissances théoriques (dont les connaissances linguistiques), et la recherche empirique de procédures efficaces.

1.3 Développements ultérieurs

Du côté de l'informatique, deux communautés différentes se sont attelées au traitement automatique de la langue. D'une part les spécialistes en recherche d'information, qui ont participé au développement de systèmes documentaires finalisés (extraction d'informations en vue de l'alimentation de bases de données, résumé automatique, etc.). D'autre part les spécialistes d'intelligence artificielle, davantage intéressés par la conception de grands systèmes (de compréhension ou de génération automatique de textes) et par les questions théoriques qu'elle soulève, dans le contexte épistémologique plus large des sciences cognitives.

Dans le domaine de la parole, on retrouve également, d'un côté la communauté de l'intelligence artificielle (cherchant à modéliser l'ensemble des connaissances nécessaires), et de l'autre des praticiens attelés à la réalisation de systèmes robustes et ergonomiques (notamment pour toutes les applications liées à la communication homme-machine en langue naturelle).

1.4 Confluences récentes

De manière probablement un peu réductrice, on peut repérer dans l'histoire du TAL deux traditions longtemps opposées (la contribution de R. Carré manifeste ce contraste, pour l'oral). La première parie sur le transfert de l'expertise des locuteurs et des linguistes pour disposer des règles nécessaires au fonctionnement de systèmes de TAL [Fuchs, 1993a]. La seconde table, pour obtenir des outils adaptés, sur la mise en évidence automatique, par des techniques d'apprentissage artificiel [Cornuéjols & Miclet, 2002], des régularités présentes dans des échantillons significatifs des données à traiter. Les années quatre-vingt dix sont marquées par la convergence de ces deux approches [Jurafsky & Martin, 2000], auparavant séparées, et par le pas pris par la seconde sur la première, au moins temporairement. L'objectif d'un transfert et d'une explicitation d'expertise (celle du linguiste par exemple) est de plus en plus souvent remplacé par celui d'acquisition à partir de données (on parle d'*apprentissage automatique*). Les règles *a priori* cèdent la place aux régularités. La représentation des connaissances à l'acquisition des connaissances. En termes linguistiques, on pourrait reformuler cela en un déplacement de modèles de la compétence vers

des modèles de la performance, ou encore du système des règles de la langue vers l'observation de régularités dans des corpus textuels.

Les origines de ce déplacement sont multiples. Les approches à règles (dites *symboliques*) peinent à formuler et à savoir utiliser l'ensemble des règles qui seraient nécessaires pour rendre compte de données de taille réelle (par exemple, les règles syntaxiques et sémantiques pour traiter un roman de taille moyenne). Elles sont peu « robustes » : les énoncés déviants les mettent en échec (ou alors, il faut modéliser la déviance et lui donner un statut). A l'inverse, la *recherche d'information* (*information retrieval*) et le traitement de l'oral (*speech processing*) ont montré l'apport de techniques probabilistes et statistiques pour obtenir des systèmes efficaces et adaptables. Par exemple, l'examen des enchaînements de mots (2 ou 3, le plus souvent) et de leurs probabilités dans des corpus dits d'apprentissage suffisamment volumineux permet de développer des modèles efficaces de *reconnaissance de la parole*. Il en va de même pour l'étiquetage morpho-syntaxique, l'attribution à un mot d'une catégorie morpho-syntaxique. Disposer de corpus étiquetés manuellement permet d'*entraîner* un étiqueteur probabiliste. La démarche s'est généralisée progressivement à d'autres domaines du TAL. Dans le même temps, l'accroissement rapide des capacités de stockage et de traitement (il n'est plus rare de pouvoir travailler sur un corpus de 100 millions de mots, soit un millier de romans de taille moyenne ou cinq années du journal *Le Monde*), l'amélioration concomitante des modèles probabilistes ont favorisé la généralisation des approches quantitatives [Manning & Schütze, 1999].

La dénomination *industries de la langue* [Pierrel, 2000] correspond par ailleurs au besoin réel, incontournable, croissant, de disposer d'outils et de méthodes robustes pour traiter la langue sous toutes ses formes : requêtes documentaires (moteurs de recherche) unilingues ou multilingues, extraction d'informations précises, correction orthographique, bases de données textuelles, fouille de données textuelles... Les procédures d'évaluation développées initialement en recherche d'information, dans le cadre en particulier des compétitions internationales en recherche et en extraction d'information (MUC, TREC) organisées par les institutions américaines [D]ARPA et NIST [2]. ont été généralisées à d'autres domaines : reconnaissance d'entités nommées - *Named Entity Recognition*, c'est-à-dire des noms de personnes, d'institution, de lieu ; réponse à des questions factuelles ; étiquetage morpho-syntaxique [Adda et al., 1999]. Un des objectifs implicites est d'encourager le développement d'un marché de modules TAL « enfichables » (*plug-ins*) où l'acheteur puisse choisir son module en fonction des performances comparées des modules disponibles assurant une tâche donnée (il s'agit de pouvoir répondre par exemple à la demande : quel est aujourd'hui le meilleur étiqueteur morphosyntaxique pour le français ?) [Cunningham, 1999, p. 6]. Un autre objectif est de remplacer la question, probablement mal posée parce que trop globale, de l'« efficacité », du succès du TAL en général par des réponses précises sur des tâches déterminées, et plus précisément d'isoler des tâches pour lesquelles on puisse dire que les savoirs et les savoir-faire sont « mûrs», c'est-à-dire robustes, car applicables à des données réelles, volumineuses et souvent « fautives » par rapport à de l'écrit ou de l'oral normé [3].

2 Données annotées et outils d'annotation existants

Différentes ressources numériques sont désormais accessibles aux utilisateurs, en particulier linguistes :

•données : il peut s'agir de corpus, bruts (de simples mots) ou *annotés [Véronis, 2000b]* (des informations morphologiques, syntaxiques, sémantiques, etc. sont ajoutées) ou de dictionnaires électroniques (unilingues ou bilingues) ;

•outils d'annotation : ce sont des programmes qui ajoutent par exemple à des corpus bruts ou déjà annotés de nouvelles informations comme le lemme, la catégorie morpho-syntaxique des mots, les dépendances syntaxiques entre mots, liens de co-référence, etc.

Une recherche spécifique, menée en linguistique, peut conduire à repérer les corpus existants qui permettent de disposer d'occurrences des phénomènes langagiers visés et/ou à appliquer un ou plusieurs outil(s) d'annotation sur un corpus existant ou constitué pour l'occasion. Par exemple, l'examen des types sémantiques des noms qui acceptent le suffixe *-esque* [Corbin et al., 1993][Mélis-Puchulu, 1993][Bartning & Noailly, 1993] peut amener les chercheurs à extraire de la version électronique de plusieurs années du journal *Le Monde* toutes les phrases contenant des mots en *-esque*, à éliminer le *bruit*, c'est-à-dire l'information non pertinente (« presque », « fresque »...). L'étude des conditions de lexicalisation de tels dérivés poussera à recourir aux versions électroniques de dictionnaires (le *Robert*, le *Trésor de la langue française*...). La caractérisation des contraintes phonologiques [Plénat, 1997] bénéfiera du recours à un *phonétiseur*, qui fournit la transcription phonétique d'un mot.

L'évolution rapide des données et des outils rend pratiquement obsolète un inventaire « papier » dès sa parution. On se contentera donc de renvoyer aux points d'entrée suivants :

- •pour les ressources, les sites des deux fournisseurs principaux : LDC (*Linguistic Data Consortium* - http://www.ldc.upenn.edu/) et ELRA (*European Language Resources Association* - http://www.elra.info/) ;
- •pour les associations savantes :
  - •ACL (Association for Computational Linguistics - http://www.aclweb.org/) ;
  - •ATALA (Association pour le traitement automatique des langues - http://www.atala.org/) ;
  - •AFCP Association Francophone de la Communication Parlée - http://www.afcp-parole.org/)
- •pour les revues :
  - •*Computational Linguistics* (voir le site de l'ACL) ;
  - •*TAL* (*Traitement Automatique des langues* ; http://www.atala.org/tal/tal.html) ;
  - •le site d'articles sous formes électroniques lanl.arXiv.org, dans la rubrique *Computation and language*

[2] *[D]efense Advanced Research Projects Agency* : volet Recherche du ministère américain de la défense et *National Institute of Standards and Technology*, agence de la *Technology Administration* du ministère américain du commerce.

[3] La maturité veut d'ailleurs dire désormais que le résultat est utilisable, mais pas forcément tel quel. A l'image de ce qui se passe lorsqu'on trie manuellement les documents rapportés par un moteur de recherche, on admet maintenant qu'il reste le plus souvent à « finir » ce qui est produit par l'application [Cunningham, 1999, p. 9]. Recule d'autant l'idée qu'une application en TAL doive « mimer » le comportement humain. L'essentiel est que soit avantageux le rapport entre ce qui est fait automatiquement et ce qu'il reste à faire manuellement. Ainsi les résumeurs automatiques actuels sont en fait des « écrémeurs » réglables, sélectionnant les phrases les plus probablement lourdes d'information. Ils ne fournissent certes pas des résumés comparables à ceux produits par un humain. Ils permettent néanmoins un survol efficace de l'information.

(http://xxx.lanl.gov/list/cs.CL/recent) ;

•pour un suivi des débats autour des ressources et de leur utilisation, les listes de discussion modérées suivantes :

•Corpora (http://helmer.aksis.uib.no/corpora/) ;
•LN-FR (http://www.biomath.jussieu.fr/LN/LN-FR/).

C'est plutôt une cartographie d'ensemble que nous présentons dans le reste de cette section, allant des précautions à garder en mémoire face aux ressources (section 2.1), aux différents domaines concernés : parole, écrit, multilinguisme (section 2.3), en passant par les acquis actuels selon les niveaux de l'analyse, du mot au texte (section 2.2) [4].

2.1 Critères de présentation et précautions

Sans doute faut-il en préalable distinguer clairement ce qui existe et ce qui est effectivement accessible et utilisable. Données comme outils ont en effet des modes d'existence et d'accessibilité variés, du prototype de recherche non stabilisé ou « privé » au « produit étagère », distribué. Certaines ressources sont accessibles gratuitement (c'est l'univers du logiciel et des ressources dites *libres*), comme WinBrill (http://jupiter.inalf.cnrs.fr/cgi-bin/mep.exe?HTML=mep_winbrill.txt), l'adaptation au français de l'étiqueteur morpho-syntaxique développé par E. Brill [Brill, 1995]. D'autres sont achetables, comme l'étiqueteur morpho-syntaxique Cordial (http://www.synapse.com), parfois à des prix qui dépassent les moyens d'un individu ou d'un centre de recherches en sciences humaines : c'est le cas de dictionnaires sémantiques comme EuroWordNet ou comme le dictionnaire des verbes de J. Dubois et F. Dubois-Charlier [Dubois & Dubois-Charlier, 1997]. Certaines ressources constituent des secrets commerciaux. D'autres voient leur accès restreint pour protéger des données personnelles (comptes rendus d'hospitalisation, par exemple).

Disposer d'une ressource ne garantit pas que son utilisation soit aisée, immédiate. Il peut en effet s'avérer nécessaire de retravailler les données [Habert, 2000][Bonhomme, 2000][Romary, 2000] soit en amont (par exemple, préparer les textes à fournir à un logiciel d'analyse de statistique textuelle pour qu'il puisse les traiter) soit en aval (filtrer la partie des résultats effectivement pertinents, comme le sous-ensemble des phrases comportant un mot se terminant par *-esque* qui est bien un adjectif dérivé avec ce suffixe).

Enfin, le linguiste doit tenir compte de l'adéquation de la ressource et de sa fiabilité. Adéquation : les étiquettes fournies par tel corpus ou par tel outil peuvent ne correspondre que partiellement aux visées de départ. Ainsi, pour qui s'intéresse à l'emploi des temps, en particulier de ceux du passé, l'étiqueteur morpho-syntaxique Cordial ne donne pas directement accès aux temps composés, qui sont considérés comme la suite d'un auxiliaire et d'un participe passé, mais qui ne sont pas identifiés en tant que tels. Fiabilité : les étiqueteurs morpho-syntaxiques ont un taux d'erreur de 2 à 5%, ce qui donne environ un mot mal étiqueté par phrase de longueur moyenne (30 mots). Sur du texte non révisé, comme celui qu'on trouve sur nombre de pages Web, les performances chutent rapidement.

2.2 Niveaux d'analyse et tâches

2.2.1 Découpage en « mots »

Un document est pour l'ordinateur une suite ininterrompue, un flux, de caractères. Ce flux doit tout d'abord être segmenté en « mots » [Laporte, 2000]. La difficulté est que certains caractères sont tantôt séparateurs tantôt non séparateurs : l'apostrophe dans « aujourd'hui » ou le tiret dans « peut-être », mais également l'espace dans les « mots en plusieurs mots » (ou mots composés, unités polylexicales, selon les terminologies). Des dictionnaires électroniques de « mots composés » existent (par exemple, pour le français, ceux du LADL) et permettent de souder par exemple « eaux_usées », tandis que la séquence « eau froide » sera considérée comme composée de deux mots. Des logiciels aident au repérage des suites de mots susceptibles de constituer des « mots en plusieurs mots » [Bourigault & Jacquemin, 2000].

2.2.2 Etiquetage et lemmatisation

L'étiquetage morpho-syntaxique [Habert et al., 1997, ch. 1][Paroubek & Rajman, 2000] consiste à associer à chaque mot en contexte une étiquette plus ou moins détaillée (de la dizaine de parties du discours à plusieurs centaines : l'étiquette porte alors des informations morpho-syntaxiques plus fines). Le taux d'erreur varie entre 2 et 5%. La lemmatisation correspond en général à un « défléchissement ». On remplace une flexion du verbe par l'infinitif, etc. Cette lemmatisation basée sur des connaissances morphologiques est parfois remplacée par une troncation plus grossière des fins de mots pour rapprocher les mots ainsi raccourcis (« sémantique », « sémantème » pourraient être ramenés tous deux à « sémant- »). Les étiqueteurs/lemmatiseurs sont désormais aisément accessibles pour le français (cf. supra Cordial et WinBrill).

2.2.3 Structuration syntaxique

Sont aujourd'hui développés des analyseurs syntaxiques [Abeillé & Blache, 2000] ou *parseurs* (*parsers*) dits robustes [Vergne, 1999]. Le qualificatif signifie qu'ils doivent pouvoir produire des résultats sur du texte tout venant, voire sur du texte mal formé. Cela n'implique pas la capacité à fournir des arbres complets pour l'ensemble des phrases. L'objectif est plutôt soit de retourner des arbres partiels (par exemple les arbres de certains syntagmes, sans rattacher ces arbres à un arbre couvrant la phrase entière) soit de lister les dépendances syntaxiques principales des phrases (par exemple entre un verbe et le nom tête du groupe sujet, ainsi que le nom tête du groupe objet). Le taux d'erreur est plus important que pour l'étiquetage. Les analyseurs syntaxiques ne sont pas aisément accessibles pour le français.

Parallèlement, depuis plus d'une décennie [Marcus et al., 1993], sont constitués des corpus arborés (*tree-banks*), où des annotateurs humains spécialisés attachent à chaque phrase une représentation syntaxique qui peut être très détaillée [Sampson, 1995], éventuellement à partir d'un

[4] Pour une présentation d'ensemble récente, on se reportera à [Mitkov, 2003].

« brouillon » fourni par un parseur [Habert et al., 1997, ch. 2] [5].

2.2.4 Découper, résumer un texte

Un texte aborde souvent plusieurs thèmes. La *segmentation thématique* [Sabah & Grau, 2000] repère les points de passage d'un thème à l'autre. On peut éventuellement regrouper a posteriori les fragments relevant d'un même thème. Certains logiciels d'analyse textuelle commercialisés, comme Alceste [Reinert, 1996], permettent de tels regroupements, qui peuvent faciliter la constitution, pour une étude linguistique, de sous-ensembles plus homogènes au sein d'un corpus. Le « résumé automatique » [et Jean-Luc Minel, 2000] revient le plus souvent à un écrémage : on retient les phrases les plus « lourdes de sens » d'un texte. Ces deux volets du TAL ne donnent pas lieu dans l'immédiat à des outils facilement utilisables en linguistique, même si des applications commercialisées y recourent.

2.2.5 Trouver le sens d'un mot en contexte ou obtenir des dictionnaires sémantiques

L'étiquetage morpho-syntaxique opère déjà une certaine forme de *désambiguïsation sémantique* (par exemple pour « guide » entre le nom et le verbe). On peut aller plus loin et attribuer à chaque mot, en contexte, son sens (en fonction d'un dictionnaire pré-existant). C'est à cette tâche de désambiguïsation sémantique que sont consacrés les articles de R. Martin et J. Véronis dans ce numéro. Des dictionnaires sémantiques ont été développés, essentiellement pour l'anglais avec WordNet - http://www.cogsci.princeton.edu/wn/ - [Slodzian, 1999][Fellbaum, 1998][Fellbaum, 1999] (un dictionnaire limité a été mis au point pour le français dans le cadre du projet européen EuroWordNet). Dans de tels dictionnaires, sont fournies de manière systématique les relations d'hyperonymie, antonymie, méronymie, etc. Des techniques sont étudiées par ailleurs soit pour extraire d'un corpus donné des couples de mots entretenant une relation sémantique donnée, comme l'hyperonymie [Morin, 1999], soit pour regrouper les mots qui sont employés dans des contextes similaires et proposer des ébauches de classes sémantiques [Zweigenbaum & Habert, à paraître]. L'ensemble de ce volet sémantique relève de la recherche. Pour le français, on peut simplement consulter le *Trésor de la langue française informatisé* (http://atilf.inalf.fr).

2.3 Domaines

Nous avons centré implicitement la section 2.2 sur le traitement de l'écrit.

En ce qui concerne l'oral, dénommé *parole* en TAL (voir l'article de R. Carré dans ce numéro), on distingue la *reconnaissance de la parole*, c'est-à-dire la transcription automatique de l'oral vers l'écrit et la *synthèse de la parole*, c'est-à-dire l'émission d'une suite de sons correspondant à des phrases fournies à l'écrit (la contribution de P. Mertens est consacrée à ce volet). La reconnaissance de la parole donne naissance à des produits commercialisés. Elle fonctionne mieux pour un locuteur unique que pour des enregistrements à plusieurs locuteurs. Elle bute également sur la parole superposée (deux locuteurs parlent en même temps).

La dimension multilingue, à l'origine du TAL, recouvre les recherches en *traduction automatique* [Boitet, 2000] et celles en *alignement*. Les traducteurs automatiques commercialisés évitent bien des consultations de dictionnaires et économisent un temps précieux. Mais ils ne fournissent qu'un brouillon de traduction qu'il faut remanier. L'alignement de textes [Habert et al., 1997, ch. 6][Véronis, 2000a], c'est-à-dire la mise en correspondance automatique fine (phrase à phrase voire groupe de mots à groupe de mots) de deux textes traductions l'un de l'autre produit des « mémoires de traduction » imposantes complétant et actualisant l'arsenal des traducteurs. Les aligneurs sont aujourd'hui commercialisés.

3 Organisation du numéro

Les contributions rassemblées dans le présent numéro visent à illustrer les problématiques que nous venons d'évoquer en matière d'évolution des systèmes de traitement automatique de la langue vers des systèmes d'ingéniérie linguistique, à présenter de façon concrète les questions soulevées par la conception de ces systèmes, ainsi que leur utilisation par des linguistes.

La contribution de Geneviève Lallich-Boidin, « Connaître les limites des traitements automatiques pour s'en accommoder », met en évidence, dans le détail et concrètement, les problèmes rencontrés dans l'analyse automatique (évoqués rapidement en section 2.2). Elle favorise un réalisme sain quant aux performances escomptables en TAL, éloigné du défaitisme comme des illusions lyriques.

L'article de Robert Martin, « Etiquetage sémantique du lexique français », dans la ligne de [Martin, 2001], montre comment utiliser les ressources d'un dictionnaire électronique (ici le *Trésor de la langue française électronique*) pour attribuer en contexte à un mot l'acception qui lui convient. De manière complémentaire [6], Jean Véronis, dans « Quels dictionnaires pour l'étiquetage sémantique ? », examine les performances des locuteurs dans le répérage du caractère polysémique d'un mot et dans l'attribution d'un sens en contexte. Il en tire des conclusions sur les caractéristiques que devraient présenter les dictionnaires sémantiques en particulier pour pouvoir être utilisables en désambiguïsation sémantique.

Piet Mertens, dans « Quelques aller retour entre la prosodie et son traitement automatique », se situe en synthèse de la parole. Son article souligne la dimension de modélisation et de mise à l'épreuve des théories linguistique que peut revêtir un système de synthèse vocale. L'article de René Carré (« Traitement de parole et linguistique : interactions ») montre, pour ce qui concerne le traitement de la parole en général, les oscillations et complémentarités entre deux approches, l'une qui consiste à formuler par des règles l'expertise du linguiste, l'autre qui revient à dégager, par des méthodes probabilistes, les régularités fondamentales d'un ensemble de données d'entraînement.

Christiane Marchello-Nizia centre sa contribution (« Linguistique historique, linguistique outillée : les fruits d'une tradition ») sur le rôle des

[5] Voir *The Penn Treebank Project* (http://www.cis.upenn.edu/treebank/home.html), le corpus arboré de l'université de Pennsylvanie, qui a joué un rôle fondateur. Pour une vue d'ensemble du domaine : http://treebank.linguist.jussieu.fr/toc.html. Pour le français, voir : http://www.llf.cnrs.fr/fr/Abeille/.
[6] Nous remercions R. Martin et J. Véronis d'avoir bien voulu intégrer chacun dans son article ses propres réactions aux propositions de l'autre auteur.

corpus et de leur étiquetage dans la formulation et la falsification d'hypothèses pour des états de langue disparus.

L'article qui clôt le numéro, « Bilan et perspectives méthodologiques », délimite le domaine d'utilisation des outils et méthodes du TAL. Il développe les rôles qui peuvent leur être attribués : modélisation ou arsenal descriptif. Il essaie de formuler quelques propositions quant à la formation des jeunes linguistes et à l'évolution des pratiques.

Catherine FUCHS
LATTICE (UMR 8094 CNRS)
ENS 1 rue Maurice Arnoux 92120 Montrouge FRANCE
Catherine.Fuchs@ens.fr
&
Benoît HABERT Groupe Langues Information Représentations (LIR) -
LIMSI (UPR A3251 CNRS)
BP 133, F-91403, Orsay Cedex & université Paris X
habert@limsi.fr http://www.limsi.fr/Individu/habert/

Bibliographie

Abeillé, A., Blache, P. (2000). « Grammaires et analyseurs syntaxiques ». In Pierrel, J.-M. (ed.), *Ingéniérie des langues*, Informatique et systèmes d'information, ch. 2, pages 51-76. Hermès Science, Paris.

Adda, G., Mariani, J., Paroubek, P.., Lecomte, J. (1999). « Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français ». In Amsili, P. (ed.), *Actes de TALN'99 (Traitement Automatique des Langues Naturelles)*, pages 15-24, Cargèse. ATALA.

Bartning, I., Noailly, M. (1993). « Du relationnel au qualificatif : flux et reflux ». *L'information grammaticale*, (58):27-32.

Boitet, C. (2000). « Traduction assistée par ordinateur ». In Pierrel, J.-M. (ed.), *Ingéniérie des langues*, Informatique et systèmes d'information, ch. 12, pages 271-291. Hermès Science, Paris.

Bonhomme, P. (2000). « Codage et normalisation de ressources textuelles ». In Pierrel, J.-M. (ed.), *Ingéniérie des langues*, Informatique et systèmes d'information, ch. 7, pages 173-192. Hermès Science, Paris.

Bourigault, D., Jacquemin, C. (2000). « Construction de ressources terminologiques ». In Pierrel, J.-M. (ed.), *Ingéniérie des langues*, Informatique et systèmes d'information, ch. 9, pages 213-234. Hermès Science, Paris.

Brill, E. (1995). « Transformation-based error-driven learning and natural language processing : A case study in part-of-speech tagging ». *Computational Linguistics*, 21(4):543-565.

Chomsky, N. (1956). « Three models for the description of language ». *I.R.E Transactions on Information Theory*, II(2):113-114.

Chomsky, N. (1959). « On certain formal properties of grammars ». *Information and Control*, (2):137-167.

Chomsky, N., Miller, G. A. (1968). *L'analyse formelle des langues naturelles*. Mathématiques et sciences de l'homme. Mouton/Gauthier-Villars, Paris.

Corbin, D., Dal, G., Mélis-Puchulu, A.., Temple, M. (1993). « D'où viennent les sens *a priori* figurés des mots construits ? variations sur *lunette(s)*, *ébéniste* et les adjectifs en *–esque* ». *Verbum*, (1-2-3):65-100.

Cori, M., Léon, J. (2002). « La constitution du TAL : étude historique des dénominations et des concepts ». *TAL*, 43(3):21-55.

Cornuéjols, A., Miclet, L. (2002). *Apprentissage artificiel. Concepts et algorithmes*. Eyrolles, Paris. Préface de Tom Mitchell. Avec la participation d'Yves Kodratoff.

Cunningham, H. (1999). « A definition and short history of Language Engineering ». *Natural Language Engineering*, 5(1):1-16.

Dubois, J., Dubois-Charlier, F. (1997). « Synonymie syntaxique et classification des verbes français ». *Langages*, (128):51-71. *La synonymie*, Antoinette Balibar-Marbti (resp.).

Desclées, J.-P., et Jean-Luc Minel (2000). « Résumé automatique et filtrage sémantique de textes ». In Pierrel, J.-M. (ed.), *Ingéniérie des langues*, Informatique et systèmes d'information, ch. 11, pages 253-270. Hermès Science, Paris.

Fellbaum, C. (ed.) (1998). *WordNet: an electronic lexical database*. Language, Speech and Communication. The MIT Press, Cambridge, Massachusetts.

Fellbaum, C. (1999). « La représentation des verbes dans le réseau sémantique WordNet ». *Langages*, (136):27-40. Sémantique lexicale et

grammaticale - Yvette Yannick Mathieu (ed.).

Fuchs, C. (ed.) (1993a). *Linguistique et traitement automatique des langues*. Supérieur. Hachette, Paris. avec la collaboration de Anne Lacheret-Dujour et de Bernard Victorri, ainsi que le concours de Laurence Danlos et de Daniel Luzzati.

Fuchs, C. (1993b). « Traduction automatique ». In Fuchs, C. (ed.), *Linguistique et traitement automatique des langues*, Supérieur, pages 193-222. Hachette, Paris.

Gross, M., Lentin, A. (1967). *Notions sur les grammaires formelles*. Gauthier-Villars, Paris.

Habert, B. (2000). « Détournements d'annotation : armer la main et le regard ». In Bilger, M. (ed.), *Corpus. Méthodologie et applications linguistiques*, n 3 in Les français parlés - textes et études, pages 106-120. Champion & Presses Universitaires de Perpignan, Paris.

Habert, B., Nazarenko, A.., Salem, A. (1997). *Les linguistiques de corpus*. U Linguistique. Armand Colin/Masson, Paris.

Harris, Z. (1968). *Mathematical structures of language*. John Wiley, New York. Trad. fr. 1973 *Structures mathématiques du langage*, Paris, Dunod.

Jurafsky, D., Martin, J. H. (2000). *Speech and language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Artificial Intelligence. Prentice Hall, Upper Saddle River, New Jersey.

Lacheret, A. (1993). « Traitement de la parole ». In Fuchs, C. (ed.), *Linguistique et traitement automatique des langues*, Supérieur, pages 173-191. Hachette, Paris.

Laporte, E. (2000). « Mots et niveau lexical ». In Pierrel, J.-M. (ed.), *Ingéniérie des langues*, Informatique et systèmes d'information, ch. 1, pages 25-50. Hermès Science, Paris.

Manning, C. D., Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Marcus, M., Santorini, B.., Marcinkiewicz, M. A. (1993). « Building a large annotated corpus of English: *The Penn Treebank* ». *Computational Linguistics*, 19(2):313-330.

Martin, R. (2001). *Sémantique et automate*. Écritures électroniques. Presses Universitaires de France, Paris.

Mitkov, R. (ed.) (2003). *The Oxford Handbook of Computational Linguistics*. Oxford University Press, Oxford.

Mélis-Puchulu, A. (1993). « Les adjectifs en *-esque* : d'abord des adjectifs construits ». *L'information grammaticale*, (58):33-39.

Morin, E. (1999). « Using Lexico-Syntactic Patterns to Extract Semantic Relations between terms from Technical Corpus ». In Sandrini, P. (ed.), *Proceedings, 5th International Congress on Terminology and Knowledge Engineering (TKE'99)*, pages 268-278, Innsbruck, Austria. TermNet.

Paroubek, P., Rajman, M. (2000). « Etiquetage morpho-syntaxique ». In Pierrel, J.-M. (ed.), *Ingéniérie des langues*, Informatique et systèmes d'information, ch. 5, pages 131-150. Hermès Science, Paris.

Pierrel, J.-M. (ed.) (2000). *Ingéniérie des langues*. Informatique et systèmes d'information. Hermès Science, Paris.

Plénat, M. (1997). « Analyse morpho-phonologique d'un corpus d'adjectifs dérivés en *–esque »*. *French Language Studies*, (7):163-179.

Reinert, M. (1996). « Un logiciel d'analyse lexicale : ALCESTE ». *Les cahiers de l'Analyse des Données*, (4):471-484.

Romary, L. (2000). « Outils d'accès à des ressources linguistiques ». In Pierrel, J.-M. (ed.), *Ingéniérie des langues*, Informatique et systèmes d'information, ch. 8, pages 193-212. Hermès Science, Paris.

Sabah, G., Grau, B. (2000). « Compréhension automatique de textes ». In Pierrel, J.-M. (ed.), *Ingéniérie des langues*, Informatique et systèmes d'information, ch. 13, pages 293-310. Hermès Science, Paris.

Sampson, G. (1995). *English for the Computer. The SUSANNE Corpus and Analytic Scheme*. Clarendon Press, Oxford.

Slodzian, M. (1999). « WordNet et EuroWordNet : questions impertinentes sur leur pertinence linguistique ». *Sémiotiques*, (17):51-70. *Dépasser les sens iniques dans l'accès automatisé aux textes*, B. Habert (resp.).

Vergne, J. (1999). *Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur. Analyse syntaxique automatique non combinatoire*. Habilitation à diriger des recherches en informatique, Université de Caen, Caen.

Véronis, J. (2000a). « Alignement de corpus multilingues ». In Pierrel, J.-M. (ed.), *Ingéniérie des langues*, Informatique et systèmes

d'information, ch. 6, pages 151-172. Hermès Science, Paris.

Véronis, J. (2000b). « Annotation automatique de corpus : panorama et état de la technique ». In Pierrel, J.-M. (ed.), *Ingéniérie des langues*, Informatique et systèmes d'information, ch. 4, pages 111-130. Hermès Science, Paris.

Zweigenbaum, P., Habert, B. (à paraître). « Accès mesurés au sens ». *Mots*, (74).