TAL & SYNTAXE OBJETS, OBJECTIFS, AMBITIONS ET NOUVEAUX DEFIS

Thomas Lebarbé

TAL & SYNTAXE
OBJETS, OBJECTIFS, AMBITIONS
ET NOUVEAUX DEFIS

Résumé : Le traitement automatique des langues est un domaine à double vocation, les deux étant traitées séparément ou conjointement suivant les équipes de recherche : concevoir des outils informatiques productifs et intégrés aux activités humaines et industrielles (cf. définition de P. Bouillon et al.) ou permettre la constitution de connaissances nouvelles sur le fonctionnement de la langue (cf. préambule de N. Chomsky). La syntaxe, parmi les dimensions linguistiques qui peuvent être étudiées, n'y échappe pas.

Si la finalité scientifique de la syntaxe est la création de connaissances nouvelles sur la syntaxe elle-même, la finalité industrielle est quant à elle de s'intégrer dans un processus plus large, la syntaxe n'est qu'une forme abstraite et formalisée permettant d'accéder et d'interroger les contenus. La multiplicité des objectifs est pour partie responsable de la grande diversité des formalismes, parfois perçus comme antagonistes alors que chacun répond à une perception et un besoin d'interprétation syntaxique de la langue.

La syntaxe en traitement automatique des langues pâtit toutefois d'une approche certes descriptive (décrire structurellement) mais teintée de prescription (l'objet analysé est considéré comme syntaxiquement bien formé). Or les défis actuels ne sont plus l'analyse de formes « parfaites » d'expression : les données produites en lignes (blogs, tweets, etc.), ou le brouillon constituent des terrains d'exploration et de création de nouveaux modèles syntaxiques de la langue, comme un objet certes prescrit mais ne respectant que partiellement ou à la marge les règles que nous savons aujourd'hui reproduire formellement.

Le traitement automatique des langues est souvent présenté selon la définition qu'en donne (Bouillon et al., 1998) : « Le traitement automatique des langues (TAL) a pour objet la création de programmes informatiques capable de traiter automatiquement les *langues naturelles* ».

Nous préférons à « traiter » la notion d'« exploiter ». Le traitement automatique des langues n'a pas pour unique objectif de traiter - dans une interprétation physico-chimique du terme[1]. Le TAL se doit

[1] « Soumettre (une substance) à l'action d'agents physiques ou chimiques, de manière à la modifier, la transformer », définition extraite du Nouveau Petit Robert de la Langue, 2010.

aussi d'exploiter[2]. La fonction du TAL n'est pas uniquement de produire des machines de transformation mais doit permettre de faire ressortir des propriétés de la langue. Nous éludons ici délibérément l'interprétation cognitiviste du terme traitement comme une suite d'opérations mentales (Varela, 1989) qui approche de notre conception mais qui ne nous semble pas être incluse dans la définition de Pierrette Bouillon. L'articulation entre l'informatique et la linguistique doit se faire à double-sens : les propriétés de la langue rendent plus efficaces les outils informatiques ; inversement les outils informatiques apportent à la linguistique des informations, des connaissances nouvelles sur la langue. Cette perception n'est pas nouvelle - les communautés du traitement automatique des langues et de la linguistique de corpus sont depuis longtemps étroitement liées. En témoignent les nombreux chercheurs affiliés aux associations savantes des deux disciplines, telles ATALA[3] et AFLA[4] en France.

Enfin, nous reprenons la notion de « création de programmes informatiques » qui ne nous paraît pas assez poussée. En effet, il ne s'agit pas simplement de créer des programmes mais de concevoir de nouveaux modèles informatiques permettant l'application de modèles langagiers. Le terme utilisé par [Bouillon et al., 1998] a probablement perdu de son sens depuis la rédaction de cette définition, mais nous sommes convaincus que, du moins en recherche, l'informatique ne se réduit pas à une simple ingénierie, à une simple programmation.

Toutefois, le modèle informatique ne prédomine pas, bien au contraire, il est subséquent au modèle linguistique et doit en matérialiser aussi bien les processus que les représentations. Cet exercice exige par conséquent, non pas l'adaptation du modèle linguistique au modèle informatique mais l'adaptation et l'application de modèles informatiques existants voire l'innovation de nouveaux modèles de traitement.

La langue (« les langues naturelles » dans les termes de P. Bouillon), quant à elle, peut être modélisée de deux manières différentes. La dichotomie est parfois présentée sous la forme modèles empiriques / modèles statistiques. A cette terminologie, nous préférons celle de modèles explicites / modèles implicites. Les deux portent le même sens, les mêmes différenciations : des modèles fondés sur une analyse humaine de la donnée langagière versus modèles fondés sur une déduction statistique et automatique des propriétés de la langue.

Ce préambule nous amène à proposer une définition du TAL, tel que nous le percevons, tel que nous le pratiquons :

« Le traitement automatique des langues a pour objet la modélisation linguistique et informatique, fondée sur une analyse de corpus en contexte, afin d'exploiter les langues, résultant en des

[2] « Utiliser d'une manière avantageuse, faire rendre les meilleurs résultats », même source.

[3] ATALA : Association pour le Traitement Automatique des Langues

[4] AFLA : Association Française de Linguistique Appliquée

applications logicielles et / ou des enrichissements de la connaissance des langues. »

Cette articulation entre modélisation linguistique et modélisation informatique pose la question de la relation entre deux disciplines, l'une des sciences dites « dures », l'autre des sciences dites « humaines »[5].

1. LES APPROCHES TALISTES DE LA SYNTAXE

Si l'on en croit le Larousse, la syntaxe est la :

1. « Partie de la grammaire qui décrit les règles par lesquelles les unités linguistiques se combinent en phrases.
2. « Ensemble de ces règles qui sont caractéristiques de telle ou telle langue. »

Il s'agit donc à la fois d'un domaine de la linguistique (en tant que discipline qui s'applique à décrire et expliquer le fonctionnement de la / des langue/s) et d'un ensemble structuré de règles (*a fortiori* formelles) résultant de la première définition.

Cette dichotomie se retrouve à l'identique dans les approches qu'en a le traitement automatique des langues. D'une part une approche analytique de la langue, utilisant les outils numériques à des fins exploratoires. D'autre part une approche applicative, se fondant sur un modèle de langue afin de produire un résultat.

Dans cette seconde approche, la syntaxe ne constitue presque jamais une fin en soi, mais fait partie d'un processus plus global dépendant des fonctionnalités requises d'un outil, d'un domaine d'activité, d'un cœur de métier. Dans ce cas, l'efficacité (en temps) et la pertinence (des résultats) sont des éléments essentiels de l'outil qui vont parfois prendre le pas sur la qualité linguistique du traitement.

1.1. Une dimension industrielle indéniable

Il est peut-être vain d'enfoncer une telle porte ouverte, mais l'infobésité[6] résultant d'une généralisation de l'usage des technologies numériques de communication, crée le besoin de traiter automatiquement, au moins partiellement, la masse d'information à laquelle est confronté le professionnel, quelque soit son domaine d'activité, du moins quand celle-ci implique de se documenter.

Si certaines entreprises se fondent sur des traitements grossiers, de masse, sans modèle de langue explicite – en témoigne le succès probant de Google, aussi bien pour ses outils de recherche d'information que de traduction automatique, d'autres s'appuient sur une représentation de la langue, sur un ensemble de formalismes, qui permettent une analyse plus

[5] Nous n'aborderons pas le débat terminologique où le terme « sciences dures » est opposé, parfois de manière particulièrement désobligeante, au terme « sciences molles », ou « sciences exactes » est opposé à « sciences inexactes »

[6] ou surcharge informationnelle

fine du matériau langagier afin d'en faire ressortir des informations plus précises. D'autres secteurs d'activité quant à eux ne peuvent se dispenser d'un modèle de langue – l'on pense ainsi à des sociétés, plus modestes, telles Synapse[7] qui développe le logiciel de correction Cordial, ou TecKnowMetrix[8] qui s'appuie sur une analyse documentaire fine pour le conseil en gestion de l'innovation. Toutefois, ces dispositifs sont rarement documentés pour des raisons compréhensibles de secret industriel et de concurrence féroce sur le marché (Souque, 2014).

Dans le premier cas, l'identification automatique des structures syntaxiques du texte est essentielle puisque l'objet de l'outil est de détecter les éventuelles erreurs grammaticales de l'écrit pour suggérer une solution. Dans le second, il s'agit d'un pré-traitement dont l'objectif est de permettre une analyse ultérieure des textes, s'appuyant sur ces structures pour identifier les thèmes et les argumentaires au sein de masses de documents scientifiques et techno-juridiques.

Les systèmes de calcul des structures syntaxiques se doivent donc d'être robustes : ils doivent être en mesure de proposer une structure syntaxique unique pour chaque phrase de chaque texte, quitte à ce qu'elle soit partiellement imparfaite.

1.2. La dimension exploratoire du TAL en syntaxe

Mais le TAL n'a pas uniquement comme vocation la construction d'outils à vocation commerciale ou industrielle. Cette perception était déjà présente dans l'introduction de *Syntactic structures* (Chomsky, 1957), bien que trop rarement repris – ce pourquoi nous le citons *in extenso* :

> « The search for rigorous formulation in linguistics has a much more serious motivation than mere concern for logical niceties or the desire to purify well-established methods of linguistic analysis. Precisely constructed models for linguistic structure can play an important role, both negative and positive, in the process of discovery itself. By pushing a precise but inadequate formulation to an unacceptable conclusion, we can often expose the exact source of this inadequacy and, consequently, gain a deeper understanding of the linguistic data. More positively, a formalized theory may automatically provide solutions for many problems other than those for which it was explicitly designed. »

En d'autres termes, construire un système de traitement automatique des langues est un moyen d'accéder à de nouvelles connaissances de la langue, par l'analyse des imperfections mêmes du système. Cette remise en question du modèle linguistique initial permet alors au chercheur de remettre l'ouvrage sur l'enclume, de refaçonner les connaissances

[7] www.synapse-fr.com/
[8] www.tkm.fr/

linguistiques et les soumettre à nouveau au modèle implanté numériquement, dans un cycle empirique de production de connaissances.

Cette approche systématique de *trial-and-error* est probablement à l'origine de la floraison de « *grammars*[9] », dont l'arbre généalogique serait difficile à construire, chacune tentant de s'appuyer sur les faiblesses des autres pour proposer une approche novatrice, un autre regard. L'œil critique verra peut-être dans cette formulation une déperdition d'énergie scientifique alors qu'il s'agit au contraire d'une richesse. La multiplicité des approches, des modèles de langue, des théories linguistiques, participe à la construction de connaissances et à l'évolution même de la discipline.

2. DÉCOMPOSER L'APPROCHE SYNTAXIQUE

Approcher la syntaxe d'un point de vue numérique exige un degré de formalisation important, à la fois d'un point de vue global et d'une façon détaillée. En 2002 (Lebarbé, 2002), nous abordions la conception de systèmes de traitement automatique des langues comme l'enchaînement de trois modèles principaux : linguistique, informatique er représentation (voir fig. 1).

Au commencement, donc, est le problème – en l'occurrence produire une structuration syntaxique d'un ou plusieurs énoncés. Afin de le résoudre, il est nécessaire de concevoir une méthode de résolution du problème en intégrant un modèle du domaine - en l'occurrence un modèle linguistique[10]. Il est alors possible de proposer un modèle informatique[11] existant (ou innovant) qui trouvera son implantation afin de produire un résultat, lui-même conditionné par un modèle de représentation[12].

[9] Nous reprenons ici délibérément le terme anglo-saxon de manière à ne pas être influencés par la connotation prescriptive de sa traduction française.

[10] Le modèle linguistique : il est issu d'une thèorie expliquant le fonctionnement de la langue à analyser. Elle (la théorie) fait parfois appel à d'autres modèles, notamment psycholinguistiques et cognitifs, pour justifier de quelle manière l'humain construit et perçoit des énoncés.

[11] Le modèle informatique : il est issu d'une théorie informatique et est utilisé pour mettre en application le modèle linguistique issu de la théorie linguistique afin de résoudre le problème donné.

[12] Le modèle de représentation : il permet de représenter les résultats des calculs effectués par l'impantation du modèle informatique en fonction du modèle linguistique.

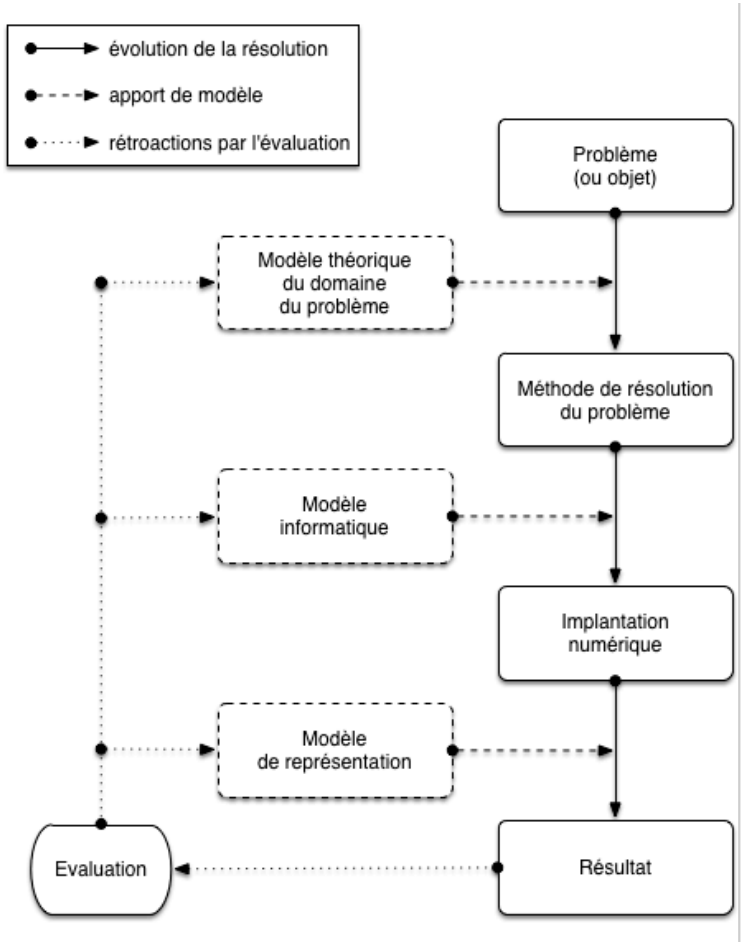


Fig. 1 : Séparation et dépendance séquentielle des modèles

Le processus de modélisation pourrait s'arrêter là – nous serions alors dans une articulation transdisciplinaire : un modèle informatique est apporté à la linguistique afin d'effectuer un traitement donné. L'évaluation, faite au regard d'un attendu défini préalamblement et par le problème donné, permet une rétroaction sur méthodes et modèles qui ont permis de produire le résultat. La séparation des modèles permet alors de distinguer les faiblesses dans l'analyse et l'évaluation des résultats et par conséquent de les corriger, voire de les repenser. C'est autour de cette évaluation que peut alors se construire un échange, un dialogue interdisciplinaire où la modélisation mono-disciplinaire est influencée, enrichie par l'approche de l'autre.

Il est à noter que, selon l'approche, selon la pensée scientifique, cette distinction des modèles et méthodes n'est pas systèmatique. Certaines approches, que l'on pourrait qualifier de « computationnelles » ne distinguent pas modèle linguistique et modèle numérique ne font

qu'un, partant d'un postulat de calculabilité formelle des structures syntaxiques.

2.1 La représentation des résultats

Dans la présentation du numéro de la revue TAL, (Kahane 2000) reprend la définition indirecte de la syntaxe suivante :

> « La phrase est un ensemble organisé dont les éléments constituants sont les mots. Tout mot qui fait partie d'une phrase cesse par lui-même d'être isolé comme dans le dictionnaire. Entre lui et ses voisins, l'esprit aperçoit des connexions, dont l'ensemble forme la charpente de la phrase. [...] Les connexions structurales établissent entre les mots des rapports de dépendance. » (Tesnière 1959).

S'appuyant sur cette notion de connexions et de rapports de dépendance, S. Kahane propose la notion de « grammaire de dépendance » comme étant « toute grammaire formelle qui manipule comme représentations syntaxiques des structures de dépendance. » *De facto*, ce n'est pas la grammaire en tant que processus analytique qu'en tant que résultat dudit processus que cette notion décrit. Du point de vue du TAL, la portée structurante de cette approche fait de la syntaxe une notion manipulable numériquement : des unités délimitées et reliées entre elles.

Se pose ensuite la question de la granularité des unités manipulées. Du point de vue de la machine, l'unité élémentaire manipulable sans aucune injection logicielle de connaissances linguistique est le caractère. Toute forme linguistique d'ordre supérieure devient une construction numérique artificielle, fondée sur des processus et des ressources souvent imparfaites. Ainsi, ne serait-ce que la notion de « mot » est particulièrement hétérogène (Branca-Rosoff, 1995) et se concrétise différemment entre la conception intellectuelle qu'en a le linguiste et la conception numérique que tente de produire la machine[13].

Le TAListe doit donc s'accomoder d'une approximation numérique des notions linguistiques de manière à les manipuler aisément. Ainsi le syntagme (unité fondamentale de la syntaxe) peut revêtir différentes formes, mais la forme « groupe de mots qui, ensemble, constituent une unité syntaxique remplissant une fonction dans la phrase » (Starets, 2000) devient particulièrement complexe à calculer pour l'ordinateur. La notion de syntagme de Chomsky (1957) tente d'y répondre par une opération logique de récursion (NP = NP + PP). D'autres approches préfèrent se fonder sur la notion de mot plein entouré de satellites mots vides (Abney, 1992). Ou encore des agglutinations de mots par pairs proches (McGee Wood, 2000). Toutefois, si ces trois approches

[13] L'exemple stéréotypique est celui de la « pomme de terre » que le linguiste voit comme un mot tandis que l'ordinateur en délimite trois.

s'opposent sur le fond, leurs représentations sont relativement similaires[14] :

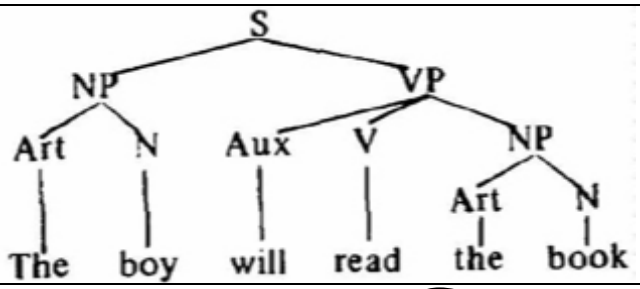
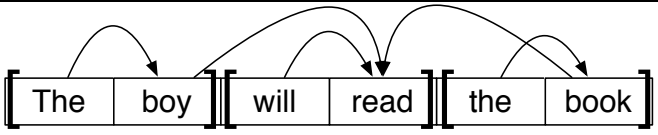
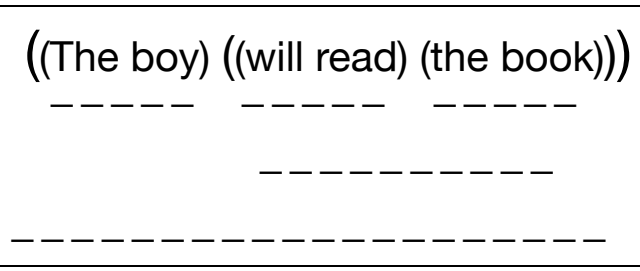
Structure de syntagme récursif	
Structure de chunks	
Structure de grammaire de catégories	

Fig. 2 : proximité des structures syntaxiques

Dans tous les cas, il s'agit de modalités de représentations d'unités imbriquées en lien les unes avec les autres qui, par conséquent, quand elles démontrent le même propos, la même interprétation linguistique, sont transposables.

2.2 La représentation des connaissances linguistiques

De fait, la représentation est avant tout un moyen de valider de manière visuelle aussi bien que numérique la pertinence d'un processus calculatoire dont l'objectif est de valider la théorie linguistique.

Afin d'accéder à ce résultat, une forme de connaissance linguistique doit être inculquée à la machine. Le point de vue le plus courant, développé par Winograd (1975), est de distinguer deux types de connaissances dans un traitement numérique : les connaissances factuelles et les connaissances procédurales.

Il ne s'agit pas ici d'inculquer des « faits des langues » et des « processus linguistiques ou langagiers » à la machine mais de lui fournir la matière et les moyens de produire un résultat. Par essence, les

[14] En nous fondant sur l'exemple donné par J.R. Searl (1972), nous donnons ici nos interprétations des représentations qui, par souci pédagogique, ont été simplifiées.

perceptions linguistique et informatique ne sont pas totalement décorrélés : l'objectif étant de produire une structure syntaxique, les connaissances factuelles s'appuient sur des données, des ressources d'ordre linguistique (par exemple l'attribution d'une ou plusieurs catégories à un mot donné) ; les connaissances procédurales, i.e., les processus par lesquels de nouvelles connaissances factuelles peuvent être inférées à partir de connaissances factuelles existantes, restent elles aussi d'ordre linguistique puisque manipulant des données linguistiques pour produire d'autres données linguistiques.

La syntaxe en TAL se fondant sur la notion de mot (ou token), la ressource factuelle première reste le lexique, qui permet d'accéder à une représentation abstraite des énoncés. Si ces ressources lexicales nécessitent d'être mutualisées, et *a fortiori* normalisées (Seddah et al., 2002), elles dépendent toutefois d'acception linguistiques qui varient d'un modèle à l'autre. Ainsi, pour le français, de nombreux outils se fondent sur un lexique associant une ou plusieurs catégories et sous-catégories à chaque forme. Toutefois, des modèles tels les grammaires de catégories associent à la forme une équation qui permette de la mettre en relation syntaxique avec les autres formes de la phrase (S\NP/NP s'interprétant comme ceci deviendra une phrase – S – s'il trouve un élément nominal qui le suit - /NP – et un élément nominal qui le précède - \NP).

L'enjeu de la désambiguisation est souvent considéré comme préalable à toute structuration syntaxique des énoncés – en témoignent les campagnes d'évaluations telles GRACE (Adda et al, 1998). Toutefois, l'exercice même de désambiguisation se fonde généralement sur une analyse contextuelle qui peut être considérée en soi comme une tâche de structuration syntaxique. Vergne (2001) pousse le paradigme plus loin en montrant que la désambiguisation est congruante à la structuration et par conséquent ne requiert qu'une ressource lexicale légère (mots vides principalement), les autres catgories pouvant être déduites en contexte au fur et à mesure de l'analyse syntaxique.

Le pendant de la ressource lexicale est la ressource procédurale. Il s'agit de donner à la machine, en fonction des données lexicales, le moyen de construire une représentation syntaxique satisfaisante. Ces règles d'analyse dépendent donc de la représentation des résultats, du lexique constitué et du processus d'analyse lui-même. Ainsi, dans l'exemple donné précédemment des grammaires de catégories, le lexique intègre partiellement la procédure, la ressource procédurale consiste alors en un ensemble restreint de méthodes permettant d'appliquer les règles du lexique. A l'inverse, sur des lexiques plus traditionnels, les bases de règles de déduction contextuelle sont plus conséquentes et se fondent généralement sur des suites d'étiquettes contigues pour, d'une part, réduire l'ambiguité et, d'autre part, construire la structure syntaxique. Enfin, même si elles ne peuvent être considérées comme des ressources procédurales, certaines méthodes s'appuient sur des banques d'arbres

aussi exhaustives que possible et tentent de calquer ces arborescences à la suite d'étiquettes que constitue la phrase.

L'importance de l'ingénierie des connaissances dans la constitution de ces ressources ne doit en rien être négligée. La richesse des connaissances factuelles et procédurales est insidieuse : données contradictoires et explosion combinatoire peuvent devenir des obstacles non négligeables pour un traitement efficace et pertinent.

2.3 La représentation des processus

Les connaissances linguistiques factuelles et procédurales s'intègrent dans des processus calculatoires que l'on peut classer dans trois grandes familles selon la perception de la syntaxe et de sa construction.

L'approche *top-down* part de la finalité de l'analyse syntaxique : la phrase. Celle-ci se décompose récursivement en unités jusqu'à atteindre le niveau du mot. Ce principe est induit par les règles déclaratives telles S→NP+VP et possède un potentiel générativiste incontestable.

L'approche *bottom-up* part des unités constitutives de la phrase pour les associer, les relier, afin d'atteindre l'élément suprême qu'est la phrase. Si les définitions lexicales des grammaires de catégories portent les marques de la finalité phrastique (S\NP/NP pour ne redonner que cet exemple), le processus reste bien un processus de montée vers l'entité phrase par agglutinations successives.

Enfin l'approche *left-right* ou longitudinale part d'une hypothèse anthropomorphique. Selon l'idée que «l'esprit aperçoit des connexions, dont l'ensemble forme la charpente de la phrase » (Tesnière, 1959), une déduction contextuelle permet à la machine de construire la structure syntaxique au fur et à mesure de la réception de l'énoncé. Il ne s'agit plus dans ce cas de construire une arborescence comme dans le cas des approches *top-down* et *bottom-up*, mais de délimiter des unités intermédiaires entre mot et phrase et de les relier entre elles.

Chacune de ces approches calculatoires est intimement dépendante du modèle de résolution syntaxique. Les processus ne sont pas interchangeables.

3. LES FORMES IMPARFAITES

Jusqu'à présent, nous avons évoqué le traitement automatique de la syntaxe, aussi bien exploratoire qu'industriel, sans particulièrement évoquer les énoncés ainsi traités.

A ce jour et à notre connaissance, aucun modèle, aucun système n'est en mesure d'analyser l'ensemble des formes « correctes » d'énoncés – par correctes, nous entendons conformes à la grammaire prescriptive. Les échecs des systèmes sur quelques rares énoncés sont la démonstration

de l'imperfection des modèles et, sans pour autant les remettre totalement en question, obligent les chercheurs à les questionner, les affiner.

En ceci, le traitement automatique des langues a permis à la syntaxe de devenir une science expérimentale dotée de modèles (dont elle disposait déjà), de microscopes (les outils) et de boîtes de Petri (les données textuelles). La puissance des outils informatiques dont nous disposons aujourd'hui permet une mise à l'épreuve des modèles sur des volumes de données dépassant toute capacité humaine.

Toutefois, cet essor du numérique, s'il présente un énorme avantage, a aussi apporté des données textuelles imparfaites que le traitement automatique des langues ne peut ignorer (au moins du point de vue industriel) et dont la syntaxe doit s'accomoder.

Les formes de communication numérique écrite représentent un défi pour le traitement automatique des langues. Si les outils grand public les plus répandus se contentent d'un traitement de surface (i.e., le texte est un « sac de mots » (Rastier, 2011)), des traitements plus fins peuvent être plus pertinents et nécessitent alors une structuration syntaxique. Toutefois, contrairement aux textes édités (voire éditorialisés), les énoncés plus spontannés de la communication numériques sont souvent porteurs d'erreurs grammaticales (en témoignent les buzz médiatiques réguliers de tweets truffés de « fautes » de nos politiciens). Or l'erreur grammaticale ou orthographique va soit empêcher l'étiquetage de certains mots, soit causer une attribution de catégorie erronée. En conséquence, l'analyse syntaxique automatique peut échouer. Si ces erreurs vont être noyées dans une analyse de masse, elles risquent de nuire à une analyse plus fine.

Dans un même ordre de problématique, la communauté scientifique des sciences humaines et sociales voit émerger la dynamique des humanités numériques, au sein desquelles, les humanités numériques des textes prennent une place importante. Or ces domaines se fondent sur des sources aussi diverses et variées que des écrits officiels (les minutes du procès de Nuremberg, par exemple) que des écrits en devenir comme les brouillons d'auteurs. Ces derniers portent la trace du processus d'écriture et de création et sont donc des formes imparfaites de la langue. Or ces imperfections (erreurs, hesitations, reprises, reformulations) sont d'un grand intérêt pour comprendre l'auteur. Dans le va-et-vient entre *distant reading* (Moretti 2013) et *close reading* qui permet une approche exhaustive, quantitative et qualitative des corpus, une analyse syntaxique adaptée serait de bon aloi.

Ces nouveaux contextes de formes imparfaites sont autant de nouveaux territoires à explorer pour le traitement automatique des langues

et la syntaxe et soulèvent des questions aussi bien du point de vue des modèles, des ressources que des méthodes calculatoires.

Thomas LEBARBÉ
Université Grenoble – Alpes, Laboratoire LIDILEM

BIBLIOGRAPHIE

ABNEY, S., 1992, « Parsing by Chunks », dans Berwick R.C. et Al. (Eds), *Principle-based Parsing*, Studies in linguistics and philosophy, Kluwer Academic Publishers.

BRANCA-ROSCOF, S., 1995, « Le mot comme notion hétérogène » dans Branca, Ed., *Le mot. Analyse de discours et sciences sociales*, Aix : Publications de l'Université de Provence, « Langues et langage » n°7.

CHOMSKY, N., 1957. Syntactic Structures, The Hague/Paris: Mouton.

GALA N., ZOCK M. Eds., 2013, Ressources lexicales, Linguisticae Investigationes, Supplementa 30, John Benjmains Publishing Company.

KAHANE, S., 2000, « Présentation » dans Kahane, Ed., Grammaires de dépendance, T.A.L.,2000, vol. 41, n°1, pp. 3-8.

McGEE WOOD, M., 2000. Syntax in Categorial Grammar : an introduction for linguists. Supplement to course notes, ESSLLI, Birmingham.

MORETTI, F., 2013. Distant Reading, London/New-York, Verso.

RASTIER, F. (2011), La Mesure et le grain. Sémantique de corpus, Paris : Honoré Champion.

SEARL, J.S., 1972, Chomsky's Revolution in Linguistics, The New York Review of Books, June 29, 1972

SEDDAH, J, JACQUEY, E., 2002, Conceptualisation d'un système d'informations lexicales, une interface paramétrable pour le T.A.L, dans Actes de RÉCITAL 2002, Nancy.

SOUQUE, A., 2014. Modèle de vérification grammaticale gauche-droite. Thèse de l'université Grenoble – Alpes.

STARETS, M., 2000, Théories syntaxiques du français contemporain, Presses de l'Université Laval.

TESNIÈRE Lucien (1959) : Éléments de syntaxe structurale, Kincksieck, Paris.

VERGNE Jacques (2000) : Étude et modélisation de la syntaxe des langues à l'aide de l'ordinateur - Analyse syntaxique automatique non combinatoire, Thèse d'HDR, Université de Caen.

WINOGRAD, T. W., 1975. Frame representation and the declarative-procedural controversy. In D. G. Bobrow et A. Collins (dir.), *Representation and understanding: Studies in cognitive science* (p. 185-210). New York, NY: Academic Press.