Vers une approche cognitive du traitement automatique des langues

Philippe Blache

Vers une approche cognitive du traitement automatique des langues

Philippe Blache
Laboratoire Parole & Langage
CNRS - Université de Provence
29 Avenue Robert Schuman
13621 Aix-en-Provence
blache@lpl.univ-aix.fr

Parmi les activités humaines, le langage occupe un statut particulier, par sa dimension sociale, sa relation à l'élaboration de la pensée, sa participation à la construction de connaissances, etc. La compréhension de son fonctionnement reste un des enjeux importants de la science. Si nous sommes aujourd'hui capables de décrire (souvent très précisément) les mécanismes constitutifs des différents composants du langage que sont le lexique, la syntaxe, la sémantique ou la pragmatique, en revanche l'explication du fonctionnement global de la production et la perception du langage, décrivant les interactions entre ces composants reste difficile. La démarche classique consiste à examiner un sous-ensemble de phénomènes, et en proposer une modélisation. Le traitement automatique des langues naturelles joue un rôle important dans cette démarche. Il constitue un véritable terrain d'expérimentation permettant de spécifier des mécanismes, élaborer des modèles et les tester du matériel naturel. Il s'agit d'un certain point de vue de la paillasse du biologiste avec ceci de particulier que ces expérimentations conduisent régulièrement à la création d'outils.

Nous présentons dans cet article quelques aspects fondamentaux du traitement automatique des langues. Plutôt qu'une présentation technique qu'il sera possible de trouver ailleurs, nous aborderons plutôt les enjeux actuels du TAL et comment il devient possible d'aborder, grâce à l'évolution des théories linguistiques en même temps que des technologies disponibles, la question de la cognition. Cette question se pose désormais avec acuité, au moment où ce domaine de recherche aborde désormais le traitement de la langue en situation naturelle, notamment dans un contexte de communication.

1. Quelques aspects du traitement automatique des langues

L'information linguistique est répartie en différents secteurs (appelés niveaux dans une conception hiérarchisée de l'information, que nous récusons ici) que sont le lexique, la morphologie, la syntaxe, la sémantique et la pragmatique. Chacun de ces secteurs fait, dans une perspective de traitement automatique, appel à des techniques ou des informations particulières. Cette vision modulariste est de fait adaptée à l'approche classique compositionnelle de l'élaboration du sens : le sens du tout est fonction du sens des parties. Il est donc d'usage en TALN de distinguer les traitements en fonction du secteur auquel ils correspondent : analyse lexicale, syntaxique, sémantique, pragmatique. Selon les théories utilisées, les techniques appliques ou encore les applications visées, le rôle joué par chacun de ces secteurs pourra être différent, plus ou moins important. Certaines approches pourront par exemple n'utiliser que des informations de niveau lexical, d'autres au contraire, s'intéresser à l'organisation du discours quasi indépendamment des autres domaines. Par ailleurs, cette approche modulariste entraîne également une vision séquentielle du traitement de l'information, partant du lexique pour arriver au sens. Une telle approche est réductrice, ne correspond pas à la réalité cognitive du traitement de l'information linguistique : les théories linguistiques modernes proposent en effet une vision

intégrée, reposant sur l'interaction des domaines. Cette conception est également à l'œuvre en TALN et un nombre croissant de techniques proposent une architecture parallèle du traitement.

Pour autant, les problèmes spécifiques à chacun des secteurs doivent être identifiés et traités. Nous décrivons dans cette partie leurs principaux aspects. Il ne s'agit pas ici de fournir une présentation technique des différents traitements (on pourra pour cela se référer aux nombreux ouvrages de présentation du TAL, en particulier Jurafsky), mais plutôt d'en présenter les fondamentaux de façon à décrire l'évolution du traitement de chacun de ces domaines, et les mettre en perspective avec le propos de cet article : comment le TALN permet-il aujourd'hui d'intégrer une approche cognitive du traitement de la langue.

1.1. La catégorisation

Le problème de la catégorisation est fondamental dans toute activité cognitive : quelles sont les unités de base sur lesquelles le traitement de l'information s'appuiera ? Il s'agit dans tous les cas d'une question complexe qui nécessite d'une part la spécification du type d'information contenu dans une catégorie et d'autre part la définition des propriétés caractéristiques permettant de reconnaître une catégorie.

Pour ce qui concerne le traitement automatique des langues, la première tâche à effectuer est celle de la catégorisation. Nous n'abordons pas dans cet article les aspects concernant le traitement de la parole, pour lequel la première tâche, elle aussi de catégorisation, concerne l'identification des phonèmes. Pour le TAL, la première étape d'analyse lexicale consiste à associer à chaque forme (chaque mot d'un texte, chaque unité d'un énoncé) une catégorie (classiquement appelée partie du discours) et une description plus ou moins précise de ses caractéristiques.

La question de la catégorisation, au niveau le plus bas, se traite en recherchant la forme dans un lexique (les principaux lexiques du français sont des lexiques de formes). Cette opération permet de retourner un ensemble de catégories candidates, en même temps que les informations associées. Des techniques de désambiguïsation probabilistes peuvent être appliquées dès cette étape, sur la base du contexte, avec un taux de succès se situant, en fonction des techniques et des types de textes, autour de 95%. Il est important de noter que ce type de performance est valable pour le traitement de matériel écrit, mais chute de façon importante pour la catégorisation d'énoncés oraux. Dans ce cas, le premier problème vient de la rareté de corpus oraux étiquetés, permettant un apprentissage. De plus, l'oral est sujet à des phénomènes variés, notamment de disfluence, compliquant cette tâche.

Les informations associées à la forme lexicale sont, a minima, d'ordre morpho-syntaxiques et concernent par exemple les informations de temps, de genre, de nombre, etc. Des informations plus précises peuvent également être fournies par ce niveau d'analyse, par exemple d'ordre syntaxique concernant les relations que le mot entretient avec son contexte. Il s'agit typiquement pour un verbe de fournir le type de compléments avec lequel il se construit, pour un adjectif d'indiquer sa possibilité de précéder le nom etc. Les informations peuvent enfin également être d'ordre sémantique (on parlera alors de sémantique lexicale). Il s'agit pour chaque objet de lui associer une description des éléments de sens qu'il contient. Pour un verbe, il 'agira en particulier de la structure argumentale, permettant de préciser le rôle sémantique spécifique de chacun de ses compléments. Pour un déterminant, il s'agira par exemple de préciser ses fonctions notamment en termes de quantification, etc.

Ressources : La tâche de catégorisation s'appuie sur des lexiques électroniques rassemblant toutes ces informations, généralement associées à des formes. Un des lexiques les plus couvrants du français a été développé au LPL (disponible sur le site du CRDO : http://crdo.fr/), il contient plus de 450.000 formes lexicales. A côté de ces ressources, des corpus étiquetés et corrigés manuellement permettent d'acquérir des informations contextuelles sur la base desquelles le processus de désambiguïsation sera déclenché. Ces corpus ne sont pas toujours librement disponibles. Un des effets positifs des campagnes d'évaluation organisées régulièrement pour le traitement du français est la production de ce type de ressource. La campagne GRACE a ainsi permis, en comparant plusieurs étiqueteurs, de construire un tel corpus (cf. Adda99). Plusieurs étiqueteurs du français sont aujourd'hui disponibles, en particulier l'étiqueteur WinBrill ou celui du LPL (également disponible sur le site du CRDO, cf. supra)

1.2. Le niveau syntaxique

L'analyse syntaxique automatique (en anglais parsing) a longtemps été considéré comme le cœur du traitement linguistique, à la fois du point de vue théorique que technique. C'est le premier problème auquel s'est attaqué le TALN. Depuis que l'informatique existe, il est devenu nécessaire développer des langages artificiels de programmation, décrits par une grammaire et pour l'analyse desquels il a fallu développer des techniques de traitement, en particulier d'analyse syntaxique, à la base de tout compilateur. Ces mêmes techniques ont rapidement été utilisées pour tenter de traiter les langues naturelles. L'idée de départ est que les langues naturelles peuvent être au moins partiellement décrite par une grammaire context-free et que nous disposons pour cela de techniques appropriées. Les premières grammaires et les premiers analyseurs ont ainsi vu le jour.

Le problème posé est simple : étant donné un énoncé (une séquence de mots), peut-on dire s'il est grammatical, en lui associant au passage une description sous la forme d'une structure syntaxique (en général un arbre). Nous avons besoin pour cela d'une grammaire et d'un mécanisme l'utilisant. Dans le cas des grammaires context-free, le principe repose sur la recherche d'une dérivation permettant, à partir d'un symbole de départ de la grammaire représentant la phrase, d'indiquer les différentes étapes permettant de parvenir jusqu'à l'ensemble de mots de la phrase à analyser, chacune de ces étapes reposant sur l'utilisation d'une règle syntagmatique.

Il existe de nombreux algorithmes permettant d'effectuer cette opération, certains pouvant être sophistiqués. Parmi les plus utilisés, on peut citer Earley et le CKY, du nom de ses auteurs (Cocke, Kasamy, Younger). Sans entrer dans les détails (consulter pour cela les ouvrages d'introduction comme Jurafsky ou Gardent), le premier algorithme s'appuie sur la génération à chaque étape du processus d'un ensemble d'items représentant la chaîne à analyser, la situation de l'analyse (ce qui a été analysé et ce qui doit l'être) et les informations spécifique à cette situation (notamment les règles utilisées). Cet algorithme permet de façon très simple de générer l'ensemble des solutions qui sont autant d'arbres possibles pouvant ête associés à la phrase dans le cas où celle-ci porte une ambiguïté syntaxique. Il est à noter que cet algorithme, initialement conçu pour les grammaires syntagmatiques simples a été adapté pour la prise en copte de formalismes plus récent et de plus haut niveau comme le formalisme DI/PL (cf. Shieber85). Le CKY quant à lui introduit une notion intéressante pour l'analyse non-déterministe : la possibilité de réutiliser des parties d'analyse déjà effectuées. Là encore, cet algorithme a donnée lieu à de nombreuses adaptation pour tenir compte de l'évolution des formalismes.

A côté de ces approches symboliques, des techniques probabilistes pour l'analyse syntaxique ont également été proposées. La plus simple consiste à guider le processus d'analyse en associant des probabilités aux règles syntagmatiques. On parle ici de grammaires syntagmatiques probabilistes

(PCFG, cf. Carpenter, Carroll, …). Cette technique constitue une réponse simple et efficace au contrôle du non-déterminisme. Des technique plus sophistiquées ont plus récemment été proposées, permettant l'assemblage direct de parties entières d'analyse, en s'appuyant non plus sur une grammaire, mais sur des ensembles d'arbres partiels (cf. Bod99).

Une évolution majeure dans la représentation de l'information syntaxique a été l'introduction dans les années 80 de la notion de traits, permettant d'associer à chaque catégorie une description précise de ses caractéristiques lexicales, morphologiques, sémantiques, syntaxiques, phonologiques, etc. Ce changement profond a reposé sur l'introduction d'un mécanisme approprié à la gestion des traits et qui a profondément modifié les techniques d'analyse : l'unification (cf. Kay83). Toutes les théories linguistiques, tous les formalismes intègrent désormais cette dimension.

Les traits permettent tout d'abord de fournir une description extrêmement précise des propriétés linguistiques. Au-delà de la simple représentation des connaissances, cet aspect est tout à fait fondamental car il a ouvert la porte à la possibilité d'intégrer au lexique toutes sortes d'information, y compris de niveau syntaxique. Cette caractéristique, appelée la lexicalisation, permet d'associer à chaque entré lexicale des propriétés génériques de sa catégorie, mais également des caractéristiques propres, par exemple sur le type de complément construit.

De plus, le mécanisme d'unification a permis d'introduire une notion particulièrement importante dans le traitement des langues : les contraintes. L'analyse dans ce type d'approche repose en effet sur la vérification des valeurs des traits de chaque catégorie. Il s'agit en d'autres termes de vérifier que l'objet qu'on est en train de construire est compatible avec son contexte. D'une part chaque catégorie est décrite par ses traits propres et d'autre part, le contexte (les catégories voisines) va spécifier des relations (ou contraintes) entre catégories. Les phénomènes d'accord sont un exemple caractéristique de ce type de contraintes. Dans ce type d'approche, l'unification devient donc un mécanisme essentiel de l'analyse syntaxique. Aujourd'hui, des formalismes largement répandus s'appuient de façon quasi exclusive sur les traits et les relations qu'ils entretiennent entre eux. C'est le cas en particulier de HPSG qui considère l'analyse syntaxique non plus comme un processus de dérivation, mais plutôt comme un mécanisme de satisfaction de contraintes (cf. Sag03).

Ressources : Les nombreux travaux sur l'analyse syntaxique automatique et, comme pour le traitement du lexique, les différentes campagnes d'évaluation notamment pour l'analyse syntaxique du français (campagnes Easy et Passage, cf. Vilnat 04, Paroubek08, Villemonte08) ont permis de constituer des ressources. Il s'agit de corpus de textes et de transcription de l'oral segmentés manuellement en chunks (unités syntaxiques non récursives). Les analyseurs syntaxiques obtiennent sur ce type de tâche d'excellents résultats (avec un f-score autour de 93%). Là encore, un système de segmentation en chunk sur la base dur formalisme proposé pour la campagne Passage, est disponible au LPL via le CRDO. A côté de ces ressources automatiques, il existe des ressources construites manuellement : il s'agit de banques d'arbres syntaxiques (appelées treebanks). Pour le français, un projet de constitution de treebank est en cours depuis plusieurs années (cf. Abeillé03).

1.3. Les nouveaux enjeux du TAL

Aujourd'hui, l'évolution des ressources et des techniques permet d'envisager un traitement plus précise et plus efficace des données textuelles. L'intégration d'informations sémantiques aux ressources lexicales est un chantier avançant rapidement, notamment grâce à des techniques d'acquisition automatique. Parallèlement, la création et l'utilisation d'ontologies offre également de nouvelles possibilités en particulier dans le cadre de la recherche d'information. Ce domaine est sans doute celui où le TAL a fait des progrès les plus spectaculaires ces dernières années et de nombreuses applications ont ainsi vu le jour, permettant des traitements très efficaces, dans des masses de données volumineuses.

Un des enjeux majeurs qui est cependant encore devant nous est le traitement du langage dans son contexte d'utilisation, et en particulier le traitement de parole spontanée. Les progrès mentionnés précédemment permettent désormais d'aborder ce problème. Tout d'abord, les techniques d'aide à la création et l'annotation de corpus permettent la constitution de ressources de haut niveau, intégrant des informations variées. Nous commençons en effet à disposer de corpus audio et vidéo, de conversations spontanées comprenant des annotations sur tous les niveau d'information : phonétique, prosodie, syntaxe, discours, gestes, etc. Ce type de ressource est extrêmement précieux d'une part pour permettre l'identification et la description précise des phénomènes observés, et d'autres part parce qu'elles peuvent servir de base pour l'entraînement de systèmes stochastiques.

Plusieurs tâches sont ainsi à accomplir pour avancer dans cette direction. Il faut tout d'abord, sur la base des corpus existants, développer des systèmes d'étiquetage adaptés à l'oral spontané. Ce type de production est caractérisé par la présence de disfluences (hésitations, reprises, répétitions, bribes, etc.) qui perturbent les techniques classiques. De nouveaux étiqueteurs sont en cours de développement pour prendre en compte ces phénomènes.

La seconde étape consiste à mettre au point un système de segmentation de la parole en unités pertinentes (correspondant aux phrases pour l'écrit). Cette tâche n'est pas triviale, mais les techniques stochastiques, sur la base d'informations prosodiques et morpho-syntaxiques, permettent d'entrevoir des solutions. Aucun système n'existe à ce jour effectuant ce type de traitement de façon efficace, mais le problème devrait être réglé rapidement.

L'étape d'analyse syntaxique est évidemment plus difficile à traiter. Quelques approches ont été proposées s'appuyant sur une étape préliminaire de transformation de l'énoncé produit en phrase canonique (en particulier en éliminant les bribes, répétitions, etc.). Ce type de technique n'est pas satisfaisant d'une part car elle et sujette à de nombreuses erreurs et d'autre part parce qu'elle s'éloigne d'un modèle cognitif plausible. L'objectif de la linguistique moderne doit être en effet de proposer un modèle théorique cognitivement fondé en même temps que tes techniques efficaces. Il convient pour cela de disposer de modèles adaptés au traitement de l'oral et de développer des techniques adaptées à leur mise en œuvre. Ces modèles et techniques devront en particulier rendre compte des différents domaines et de leurs interactions.

2. L'organisation de la construction du sens

Le traitement automatique des langues naturelles a un enjeu essentiel : permettre d'accéder à l'information contenue dans un message. Il s'agit pour nous de comprendre comment le sens s'élabore de façon à parvenir à son décodage et en proposer un traitement automatique. Nous sommes là au cœur des sciences cognitives : comment cette activité humaine par excellence qu'est la communication via le langage fonctionne-t-elle, et est-il possible d'en proposer une simulation via des processus artificiels ?

La linguistique moderne, nous y reviendrons dans la section suivante, nous apprend que l'information permettant l'interprétation d'un énoncé est répartie au travers des différents niveaux (on parle plutôt de domaines) sur lesquels repose le signal linguistique : morphologie, syntaxe, prosodie, etc. Mais elle est également répartie sur les différentes modalités utilisées dans la communication humaine, en particulier le geste et la parole (pouvant être complétés dans la communication homme-machine par d'autres types de médias artificiels). Chacun de ces domaines est donc porteur d'une partie de l'information. Traditionnellement, on considère pertinente l'hypothèse compositionnelle selon laquelle la construction du sens d'un énoncé est un processus incrémental, auquel chacun des domaines contribue en apportant une partie de l'information : l'information est une composition du sens de ses parties. Cette approche est intéressante du point de vue du traitement automatique car simple : chaque domaine (pouvant ici être conçu comme un module), construit une structure fournissant sa partie d'information, il suffit de les assembler pour construire le sens global.

Cependant, une telle conception ne permet pas de rendre compte de nombreux phénomènes dans lesquels le sens ne résulte pas d'un processus compositionnel au sens classique du terme, mais sur un processus de plus haut niveau : l'interaction des domaines. Dans ce cas, on considère qu'un même type d'information est répartie sur plusieurs domaines et l'interprétation n'est possible qu'en les prenant en compte simultanément. L'exemple suivant en est une illustration :

(1) *Marie je la supporte pas*

La structure syntaxique de cette phrase ne permet pas à elle seule de décider s'il existe une relation de coréférence entre « Marie » et le pronom « la ». Dans le cas où cette relation existe, il s'agit d'une construction disloquée, dont l'interprétation est que le locuteur ne supporte pas Marie. En revanche, sans relation de coréférence, il s'agit d'une construction vocative, le locuteur s'adressant à Marie en lui parlant de quelqu'un d'autre et lui disant qu'il ne supporte pas cette autre personne. Si la syntaxe ne permet pas de choisir entre l'une ou l'autre de ces interprétations, la prosodie le peut : un contour intonatif ascendant sera associé ici à la dislocation tandis qu'un contour plat indiquera un vocatif. Dans ce type de phénomène, nous voyons bien que l'interprétation est rendue possible en prenant en compte simultanément la prosodie et la syntaxe, c'est leur interaction qui produit du sens.

D'une façon plus générale, en étudiant la langue en situation, en particulier dans un contexte de communication, ce type d'interaction entre domaines est systématique et l'information provient en particulier de la convergence entre parole (prosodie, syntaxe, etc.), gestes et contexte. De plus en plus de travaux portent ainsi sur l'étude de la communication multimodale (homme-homme, homme-machine). De façon encore plus critique que dans le cas de l'étude d'une modalité isolée, c'est l'interaction des différents domaines qui constitue ici le cœur du processus.

Dans une perspective visant à la description des mécanismes de production et de perception en condition naturelle, et donc cognitivement fondée, il est donc indispensable de proposer des mécanismes permettant le traitement de chacun de ces domaines, mais également de leur interaction. Du point de vue plus spécifique du traitement automatique, il s'agit alors de proposer une architecture plus parallèle que celle traditionnellement adoptée qui repose sur une séquence de traitements.

Cette conception de l'organisation de l'information a des conséquences directes non seulement sur la compréhension de l'élaboration du sens, mais également sur la réalisation concrète du langage. En particulier, une des caractéristiques du langage est la grande variabilité de sa

réalisation d'un locuteur à un autre ou chez un même locuteur. Cette variabilité porte sur tous les domaines évoqués précédemment. La prosodie est bien entendu le premier domaine qui vient à l'esprit lorsqu'on évoque cette question : l'intonation, mais également les pauses, la durée des phonèmes, etc. peut connaître une très grande variabilité intra ou inter-locuteurs. Cependant, dans de nombreux cas, la prosodie semble être assujettie à des contraintes fortes contrôlant voire empêchant toute variabilité, y compris dans des langues non tonales. C'est le cas par exemple des constructions interrogatives en français qui, en l'absence dispositifs comme la présence de pronom interrogatif ou la reprise du sujet par un clitique doivent impérativement être réalisées par un contour intonatif ascendant. De même, au niveau lexical, un même message pourra être véhiculé à l'aide de mots différents, certains pouvant être référentiels, d'autre pas (les pronoms). Le niveau syntaxique est lui également sujet à une grande variabilité : un même message pourra être véhiculé à l'aide de structure syntaxique différentes, plus ou moins marquées. Ce phénomène peut être illustré sur la base de l'exemple précédent : reprenant l'exemple donné plus haut, on suivant illustre ce phénomène :

(1a) *C'est Marie que je supporte pas*
(1b) *Marie je supporte pas*

La première réalisation s'appuie sur une construction syntaxique explicite, une clivée, exprimant sans ambiguïté les relations grammaticales : l'objet clivé, « Marie », ne peut être interprété que comme l'objet du verbe supporter et est donc ici, en termes de rôle sémantique le « patient » de cette action. Cette même interprétation peut être donnée à la seconde réalisation, utilisant un dispositif totalement différent et beaucoup moins explicite quant à la réalisation des rôles syntaxiques et sémantiques évoqués plus haut.

Peu d'hypothèses permettent d'expliquer les conditions de cette variabilité. Une observation est cependant possible si l'on reprend l'idée de la dispersion de l'information au travers de différents domaines : lorsque l'un des domaines véhicule suffisamment d'information, les autres acquièrent un degré de variabilité important. En reprenant l'exemple précédent, le lexique et la syntaxe de 1a permettent de parvenir à une interprétation unique, sans ambiguïté. Dans ce cas, et l'observation des données réelles de cette construction le montre, la prosodie pourra être variable. En revanche, dans le cas du second exemple, le lexique et la syntaxe ne suffisent pas à eux seuls pour construire une interprétation stable. La prosodie ici jouera un rôle prépondérant et perdra sa variabilité : seul un contour ascendant pourra être ici réalisé. De même, en appliquant à une phrase un contour intonatif ascendant à la fin, on obtiendra en français sans ambiguïté une interprétation interrogative. Du coup, la structure syntaxique acquiert une grande variabilité, pouvant aller jusqu'à l'absence totale de marqueur interrogatif spécifique au niveau morho-syntaxique ou syntaxique.

L'explication que nous donnons à ce phénomène est alors simple : le niveau d'information d'un énoncé doit atteindre un certain seuil pour permettre son interprétation. Chaque domaine, nous l'avons vu, contribue à l'apport d'une partie de l'information. Dès que ce seuil est atteint, les phénomènes de variabilité apparaissent. Ainsi, dans le cas où la syntaxe véhicule à elle seule suffisamment d'information pour atteindre ce seuil, les autres domaines deviennent variables.

Nous obtenons là une théorie quantitative de la variabilité qui repose donc sur notre capacité d'identifier et mesurer cette information véhiculée dans chacun des domaines. C'est l'un des enjeux de modèles linguistiques actuels, débouchant sur la possibilité d'une évaluation automatique de ces niveaux.

Avant d'examiner plus précisément cet aspect du problème, il est utile d'analyser plus précisément ses enjeux théoriques.

3. Un détour par l'évolution des théories linguistiques

Les théories linguistiques ont longtemps adopté un point de vue formel, ne prenant pas en compte la langue comme un objet d'étude, mais plutôt comme un ensemble de données validant a posteriori des hypothèses. Là où les sciences du réel partent de l'observation des données et tentent d'en fournir une modélisation, la linguistique a adopté un point de vue théorique, expliquant la langue à travers les mécanismes qui la supportent et non pas pour elle-même. Les approches génératives en particulier, qui ont dominé la seconde moitié du XXème siècle, ont proposé une vision de la langue en tant qu'ensemble de séquences de mots pouvant être générés par une grammaire à l'aide d'un mécanisme : la dérivation. Cette conception a le mérite d'être claire et efficace par exemple en termes de traitement automatique : ce qui est généré fait partie de la langue, ce qui ne l'est pas est en dehors. Nous avons donc une procédure décidant de la grammaticalité d'un énoncé en lui associant au passage une structure syntaxique (à partir de laquelle il sera par exemple possible de construire une représentation sémantique).

Cette conception repose fondamentalement sur l'hypothèse de l'existence d'une grammaire universelle (ensemble de principes valides pour toutes les langues) qui constituerait selon Chomsky (dans ces premiers écrits) un véritable organe mental dont dispose de façon innée chaque individu. La grammaire est ici conçue comme un système complet avec un état initial élaboré. L'acquisition de la langue est un processus de raffinement de ce système initial par l'ajout progressif de nouvelles règles, enrichissant au fur et à mesure le système initial (qui est indépendant de l'environnement). Dans cette architecture, différents modules interagissent pour élaborer un énoncé et permettre son interprétation : phonétique, phonologie, morphologie, syntaxe, sémantique, etc. Ces modules sont indépendants et contribuent séparément à l'opération d'interprétation : chacun transmet à l'autre une structure, le tout étant, dans les approches génératives, dominé de fait par la syntaxe.

Dans le même temps que ces approches théoriques se développaient, la linguistique descriptive, rejetant de son côté toute forme de modélisation, n'a pas construit d'alternative permettant d'expliquer le fonctionnement de la langue. De plus, la linguistique descriptive s'est essentiellement attachée à décrire les langues dans une perspective de fait normative. La tradition grammairienne, de même que les travaux sur la typologie des langues reposent dans la plupart des cas sur l'observation d'une partie seulement de l'objet d'étude : la langue normée. Rares ont été les travaux portant effectivement sur la langue parlée, en situation, prenant en compte le contexte de production et de perception.

Nous sommes ainsi parvenus à une situation de quasi-blocage dans lesquelles finalement les aspects cognitifs de la langue étaient relégués au second plan.

La situation a cependant évolué rapidement ces dernières années. Les limites des approches théoriques, ne permettant pas de prendre en compte la langue dans sa globalité en tant qu'ensemble d'usages, certains étant normés, d'autres moins, ont été rapidement identifiés et des solutions proposées. Chomsky le premier avait pointé cette limite, en indiquant que la grammaticalité n'était finalement pas une notion binaire, mais qu'il existe un phénomène d'échelle illustré par le fait que les locuteurs peuvent associer à un énoncé un degré de grammaticalité :

certains énoncés sont parfaitement grammaticaux, d'autres pas du tout et d'autres encore sont entre les deux, se rapprochant plus ou moins de l'un des pôles. En tout état de cause, il convient pour une théorie de rendre compte des ces phénomènes, sans se limiter à l'analyse des seuls énoncés bien construits. C'est par exemple ce que propose de faire la Théorie de l'Optimalité (cf. Prince93) en proposant un mécanisme d'identification de la structure optimale, celle-ci pouvant ne pas satisfaire toutes les propriétés (ou contraintes) exprimées dans la grammaire.

Est ainsi progressivement apparue l'idée que les mécanismes linguistiques pouvait ne pas être conçus comme un processus d'énumération, comme dans les approches génératives, mais plutôt une recherche de modèle. Dans une perspective logique, nous passons d'une vision reposant sur théorie de la preuve (appuyée sur une démarche syntaxique de la construction de la preuve) à une théorie des modèles (reposant sur une approche sémantique). Cette distinction a été décrite en particulier par Pullum qui distingue d'un côté la syntaxe générative énumérative (Generative-Enumerative Syntax, GES) et de l'autre la syntaxe basée sur la théorie des modèles (Model-Theoretic Syntax, MTS). Il s'agit dans ce second type d'approche, plutôt que de rechercher la grammaticalité d'un énoncé, d'en décrire les caractéristiques, quelle que soit sa forme. Nous sommes ainsi capables de rendre compte de la langue dans sa globalité, avec tous ses usages, et pas seulement d'un de ses sous-ensembles, la langue normée. Ce type d'approche ouvre la porte à une nouvelle conception de l'organisation et du fonctionnement de la langue. Il devient en effet possible d'une part de rendre compte de tout type de production, et notamment la production orale, et d'autre part de prendre en compte simultanément ces différentes sources d'information évoquées précédemment. Les phénomènes linguistiques sont caractérisés par un ensemble de propriétés syntaxiques, sémantiques ou encore prosodiques qu'il convient de décrire simultanément. C'est le projet d'un certain nombre de théories linguistiques, en particulier celui des Grammaires de Construction, notées CxG (cf. Fillmore99). Cette théorie décrit en effet les constructions comme des lieux de convergence d'un ensemble de propriétés. Une construction est par exemple une tournure syntaxique particulière (l'inversion sujet-verbe, la coordination) ou une unité particulière (les syntagmes, les unités lexicales) ou des phénomènes de restriction lexicale. Ces constructions sont décrites par un ensemble de propriétés lexicales, syntaxiques ou sémantiques, toutes situées au même niveau, évitant ainsi l'écueil d'une architecture hiérarchisée et séquentielle dans laquelle chaque module fonctionne séparément et construit une représentation avant de la transférer à un autre module.

Les travaux actuels en CxG portent essentiellement sur l'interaction syntaxe-sémantique et offrent une solution très cohérente en permettant une relation directe entre ces deux domaines. Mais les CxG ne se limitent pas `a cette question : il est bien entendu possible (et même souhaitable) de décrire une construction `a l'aide d'autres types de propriétés prosodiques (utiles par exemple pour la description de phénomènes d'extraction) ou pragmatiques. Ces aspects sont explorés notamment dans le cadre de « *Embodied Construction Grammar* » (cf. [Bergen05]).

4. Vers une modèle cognitif : les contraintes

Les avancées technologiques du TALN, en même temps que l'évolution des théories linguistiques permettent d'envisager désormais des modèles computationnels cognitivement fondés pour le traitement du langage. Ces modèles permettent en effet de commencer à prendre en compte la langue dans tous ses usages, dans des situations et des contextes naturels. Ils s'appuient pour cela sur une conception décentralisée de l'information linguistique, basée sur l'interaction des différents domaines entrant en jeu dans l'élaboration du sens. Une telle organisation s'inscrit dans le cadre des approches basées sur les modèles. Il convient d'en proposer un cadre d'implantation,

tirant parti de la flexibilité de l'approche. Les représentations basées sur les contraintes, à la fois du point de vue théorique, mais également pour leur mise en œuvre informatisées, sont un bon candidat pour constituer un tel cadre.

4.1. Contraintes et TAL

Les contraintes doivent être considérées comme l'expression de propriétés. Elles jouent en informatique un rôle double, à la fois de filtrage, mais également d'instanciation. La programmation logique par contraintes, en particulier, par sa visée déclarative, a montré comment la description d'un problème sous forme de contrainte correspond à son traitement. Dans cette approche, la résolution d'un problème repose tout d'abord sur l'identification des objets puis la spécification des relations qu'ils entretiennent entre eux. Ici, les propriétés internes, mais également exocentriques de chaque objet sont représentées (implantées) par des contraintes. Chaque contrainte est porteuse d'une information particulière, l'ensemble des contraintes forme un système, c'est l'interaction des contraintes portant sur les mêmes objets qui conduit à la solution.

Dans cette approche, deux aspects sont fondamentaux. Tout d'abord, la granularité de l'information représentée par les contraintes est variable : les objets sur lesquels portent les relations peuvent être simples ou complexes, de même que les relations portant sur ces objets peuvent être de niveau différent. Par ailleurs, toutes les contraintes sont au même niveau, l'ensemble des contraintes forme un système. Elles sont de plus évaluables indépendamment les unes des autres. Nous avons donc là un cadre parfaitement adapté à la nécessité de représenter une information partielle et dispersée, mais également parallèle. Il n'est pas nécessaire, dans la mesure où toutes les informations sont représentées sous forme de contrainte, de déterminer un fonctionnement modulaire par domaine, comme cela est proposé dans les approches classiques, et notamment les approches génératives. Le seul processus utilisé est celui de la satisfaction de contrainte. L'analyse d'un énoncé consiste à vérifier pour un système de contraintes donné (constituant la grammaire du langage) et pour un énoncé donné les contraintes qui sont satisfaites. L'état du système de contrainte après évaluation constitue ainsi une description précise de l'énoncé à analyser. De plus, le processus de satisfaction permet comme cela a été souligné plus haut, de générer de nouvelles informations. Enfin, toutes les contraintes pouvant interagir, nous disposons ainsi d'une prise en charge directe de l'interaction existant entre les domaines, telle que décrite précédemment.

4.2. Modélisation et jugement de locuteurs

Parmi les problèmes identifiés mais non résolus par les approches classiques du traitement des langues (aussi bien du point de vue formel que computationnel), se trouvent les phnomènes d'échelle : comment peut-on rendre compte du fait que certains énoncés sont plus acceptables, plus complexes ou plus facilement interprétables que d'autres. Ces phénomènes ont à voir avec la notion de quantité, mais également de qualité d'information véhiculée par l'énoncé. Nous avons vu plus haut comment ces phénomènes interviennent pour expliquer la notion de variabilité dans le langage. L'intérêt des approches telle que celles décrites ici réside dans le fait que nous commençons à être capables de quantifier certains phénomènes. En particulier, pour ce qui concerne la syntaxe, plusieurs approches ont récemment vu le jour permettant d'expliquer et de prédire dans une certaine mesure les jugements d'acceptabilité des locuteurs.

Ces approches réside toutes sur la représentation de l'information basée sur les contraintes. Il s'agit en particulier de la *Linear Optimality Theory* (cf. Keller00), des *Weighted Constraint Dependency*

Grammars (Menzel00) et des Grammaires de Propriétés (Blache 05). Plusieurs idées fondamentales sont partagées par ces théories :

- Toutes les informations syntaxiques peuvent être représentées par des contraintes
- Toutes les contraintes peuvent être violées
- La grammaticalité est inversement proportionnelle au nombre de contraintes violées

Ces approches reposent donc, d'un point de vue technique, sur la capacité à relâcher des contraintes. Dans les deux premières théories (LOT et WCDG), il est ainsi possible de calculer un coefficient de grammaticalité (appelé degré d'harmonie en LOT), basé sur le nombre de contraintes violées, en tenant compte de leur importance respective représentée par un poids.

En GP, cette approche reposant sur l'idée que la violation de contrainte est cumulative est complétée par d'autres paramètres. Dans la mesure où toutes les informations sont représentées sous forme de contraintes et où, en G¨, elles sont indépendantes les unes des autres, il est également possible de tenir compte du nombre de contraintes satisfaites, de la quantité d'information produite, ainsi que d'autres phénomènes de compensation de violation de contraintes (comme la position de la violation dans l'arbre syntaxique). Au total, il est possible d'obtenir un modèle permettant de donner une évaluation plus précise de la grammaticalité, tenant compte d'un ensemble varié de critères. Nous avons montré que cet indice de grammaticalité est corrélé avec les jugements d'acceptabilité produits par des locuteurs (cf. Blache06).

Les contraintes fournissent donc un modèle computationnel efficace pour une première quantification de la grammaticalité et, plus généralement, de la qualité de l'information syntaxique d'un énoncé. Ce modèle constitue ainsi un élément de réponse à la question de la complexité syntaxique : à quelle condition une phrase est-elle plus difficile à traiter qu'une autre ? Il s'agit ainsi d'une approche ouvrant des perspectives totalement nouvelles en psycholinguistique et plusieurs projets sont prévus, permettant d'évaluer plus précisément le rôle cognitif des contraintes : un première expérience utilisant l'électro-encéphalographie est en cours. Elle devra confirmer l'effet de la violation de contrainte sur le traitement syntaxique en mesurant la réaction des sujets face à la variation contrôlée d'une contrainte. Il sera également possible avec cette expérience de valider le poids relatifs des contraintes qui en GP sont affectés aux types de contraintes.

4.3. Acquisition

Le fait que chaque contrainte peut être évaluée indépendamment du reste du système est un atout fondamental de ce type d'approche. Il n'est en effet pas nécessaire de disposer d'un système structuré complet pour pouvoir traiter l'information. Pour ce qui concerne le langage, il n'est donc pas nécessaire de disposer d'une grammaire complète pour produire ou percevoir le langage. En phase d'acquisition, il est alors possible d'imaginer que l'enfant – ou l'apprenant - dispose d'un ensemble de contraintes (un système) représentant sa compétence. Ces contraintes n'étant pas nécessairement structurées, la compétence de l'apprenant peut donc être diversifiée : très spécialisée et efficace pour certain phénomènes et au contraire superficielle pour d'autres. Il est possible de cette façon d'expliquer un processus d'acquisition non homogène de systèmes comme la flexion verbale : l'usage de certaines personnes, à certains temps se mettant en place avant d'autres. De la même façon, certaines constructions (discours indirect, passif) peuvent être acquises plus rapidement que d'autres, sans qu'une notion de complexité syntaxique ne l'explique. Il n'est pas possible de rendre compte de ce type de phénomènes en termes de raffinement

incrémental d'un système homogène, la grammaire. Au contraire, chaque usage, chaque compétence peut être décrit de façon indépendante du système complet. Il est donc possible pour un même apprenant d'avoir des niveaux de compétence différents pour des phénomènes différents : l'usage du style narratif incluant un certain nombre de stéréotypes, y compris le passé simple, apparaît très tôt chez l'enfant pour raconter des histoires, sans que le système de flexion verbal n'ait été totalement acquis.

Une explication basée sur les contraintes permet de rendre compte de ce type de phénomène. En effet, les contraintes étant indépendantes les unes des autres, il est toujours possible d'ajouter une nouvelle contrainte ou un nouvel ensemble de contraintes au système. Chaque contrainte pouvant être évaluée indépendamment des autres, il est ainsi possible de stipuler des informations de niveau et de précision différentes à l'intérieur d'un même système. Les contraintes agissent donc dans une certaine mesure comme des processus de facilitation ou d'inhibition.

5. Conclusion : les contraintes comme modèle cognitif

Le modèle computationnel des réseaux de neurones domine encore aujourd'hui les sciences cognitives pour plusieurs raisons. Tout d'abord il repose sur l'idée que l'information se construit sur la base d'un ensemble d'informations élémentaires, ou plus précisément sur l'interprétation d'un état stable d'un système complexe formé d'unités simples. Par ailleurs, la notion d'apprentissage permet également de proposer un modèle pour l'acquisition, mais aussi la spécialisation de sous-réseaux. Enfin, la métaphore elle-même de neurones est bien entendu séduisante pour quiconque cherche à modéliser l'activité cognitive. Concrètement, ce type de technique s'avère très efficace pour des tâches de reconnaissance de forme (typiquement, pour ce qui concerne le sujet qui nous intéresse, la reconnaissance de la parole). Mais les problèmes plus complexes (et hétérogènes) comme la compréhension mettent en œuvre des réponses à des niveaux variés, introduisant une variation de la granularité de neurones, certains pouvant être de « haut niveau » et exécutant de fait des tâches de traitement de l'information. La question se pose alors de avoir si nous sommes dans le même type d'approche consistant à proposer un traitement complexe par addition de tâches simples et homogènes.

Nous pensons que les modèles basés sur les systèmes de contraintes constituent une approche alternative de modèle de traitement cognitif qui, sans s'appuyer sur une métaphore physiologique, permet de proposer des réponses notamment aux problèmes d'acquisition. L'approche repose sur l'idée qu'un système de contraintes est formé d'informations pouvant être de granularité très variables à l'intérieur d'un même système et qui interagissent. Plusieurs systèmes de contraintes peuvent cohabiter avec la possibilité (mais pas la nécessité) d'interagir. Il devient dans cette perspective possible de représenter chaque domaine de l'information par un ensemble de contraintes plus ou moins riche, permettant d'expliquer des phénomène variés comme la prise en compte de source d'informations hétérogènes, d'information dégradée, d'impact relatif de certains types d'information par rapport à d'autre, de densité d'information, etc.

Au-delà des aspects propres au traitement automatique des langues, les approches basées sur les contraintes peuvent donc constituer un véritable modèle cognitif général.

[Aarts07] Aarts B. (2007) Syntactic Gradience, Oxford University Press.
[Abeillé03] Abeillé A., L. Clément, and F. Toussenel (2003) "Building a treebank for French", in A. Abeillé (ed) *Treebanks*, Kluwer, Dordrecht.

[Adda99] Adda G. , Joseph Mariani, Patrick Paroubek, Martin Rajman, Josette Lecomte, (1999) "L'action GRACE d'évaluation de l'assignation de parties du discours pour le français", revue *Langues*, 2:2.

[Bard96] Bard E., D. Robertson & A. Sorace (1996) "Magnitude Estimation of Linguistic Acceptability", Language 72:1.

[Bergen05] Bergen B. & N. Chang (2005) "Embodied Construction Grammar in Simulation-Based Language Understanding", in Jan-Ola Ostman and Miriam Fried (Eds.), *Construction Grammars: Cognitive grounding and theoretical extensions*, Benjamins

[Blache05a] Blache P. & J.-P. Prost (2005) "Gradience, Constructions and Constraint Systems", in H. Christiansen & al. (eds), *Constraint Solving and NLP*, Lecture Notes in Computer Science, Springer.

[Blache05b] Blache P. (2005) "Property Grammars: A Fully Constraint-Based Theory", in H. Christiansen & al. (eds), *Constraint Solving and NLP*, Lecture Notes in Computer Science, Springer.

[Blache06] Blache P. (2006) "A Robust and Efficient Parser for Non-Canonical Inputs", in proceedings of *Robust Methods in Analysis of Natural Language Data*, EACL workshop.

[Chomsky75] Chomsky N.. (1975) *The Logical Structure of Linguistic Theory*, Plenum Press

[Croft03] Croft W. & D. Cruse (2003) *Cognitive Linguistics*, Cambridge University Press.

[Foth05] Foth K., M. Daum & W. Menzel (2005) "Parsing Unrestricted German Text with Defeasible Constraints", in H. Christiansen & al. (eds), *Constraint Solving and NLP*, Lecture Notes in Computer Science, Springer.

[Fillmore98] Fillmore C. (1998) "Inversion and Contructional Inheritance", in Lexical and *Constructional Aspects of Linguistic Explanation*, Stanford University.

[Kay99] Kay P. & C. Fillmore (1999) "Grammatical Constructions and Linguistic Generalizations: the what's x doing y construction", *Language*.

[Karlsson95] Karlsson F., A Voutilainen, J. Aikkile & A. Antilla (eds) (1995) Constraint grammar: a language independent System for parsing unrestricted texts, Berlin: Mouton de Gruyter.

[Keller00] Keller F. (2000) *Gradience in Grammar. Experimental and Computational Aspects of Degrees of Grammaticality*, Phd Thesis, University of Edinburgh.

[Keller03] Keller F. (2003) "A probabilistic Parser as a Model of Global Processing Difficulty", in proceedings of *ACCSS-03*

[Menzel98] Menzel W. & I. Schroder (1998) "Decision procedures for dependency parsing using graded constraints", in S. Kahane & A. Polguère (eds), Proc. ColingACL Workshop on *Processing of Dependency-based Grammars*.

[Paroubek08] Paroubek P. , I. Robba, A. Vilnat and C. Ayache (2008) "EASY, Evaluation of Parsers of French: what are the Results?", in proceedings of *LREC08*

[Pereira80] Pereira, F. & David D. Waren (1980 "Definite Clause Grammars for Language Analysis", in *Artificial Intelligence*, 13

[Prince93] Prince A. & Smolensky P. (1993) Optimality Theory: Constraint Interaction in Generative Grammars, *Technical Report* RUCCS TR-2, Rutgers Center for Cognitive Science.

[Sag03] Sag I., T. Wasow & E. Bender (2003) *Syntactic Theory. A Formal Introduction*, CSLI.

[Schröder02] Schröder I. (2002) *Natural Language Parsing with Graded Constraints*. PhD Thesis, University of Hamburg.

[Sorace05] Sorace A. & F. Keller (2005) "Gradience in Linguistic Data", in *Lingua*, 115.

[Villemonte08] Villemonte de la Clergerie E., O. Hamon, D. Mostefa, C. Ayache, P. Paroubek and A. Vilnat (2008) "PASSAGE: from French Parser Evaluation to Large Sized Treebank", in proceedings of *LREC08*

[Vilnat04] Vilnat A., P. Paroubek, L. Monceaux, I. Robba, V. Gendner, G. Illouz, M. Jardino (2004) "The Ongoing Evaluation Campaign of Syntactic Parsing of French: EASY", in proceedings of *LREC04*.