TALN 2010, Montréal, 19–23 juillet 2010

Un modèle de caractérisation de la complexité syntaxique

Philippe Blache
Laboratoire Parole et Langage, CNRS & Université de Provence
5, Avenue Pasteur
13604 Aix en Provence - France
blache@lpl-aix.fr

Résumé.

Cet article présente un modèle de la complexité syntaxique. Il réunit un ensemble d'indices de complexité et les représente à l'aide d'un cadre formel homogène, offrant ainsi la possibilité d'une quantification automatique : le modèle proposé permet d'associer à chaque phrase un *indice* reflétant sa complexité.

Abstract.

This paper proposes a model of syntactic complexity. It brings together a set of complexity parameters and represnt them thanks to a unique formal framework. This approach makes it possible an automatic evaluation : a complexity index can be associated to each sentence.

Mots-clés : Complexité syntaxique, analyse syntaxique automatique, parser humain.

Keywords: Syntactic complexity, parsing, human parser.

1 Introduction

L'objectif de cet article est de proposer un modèle unifié de la complexité syntaxique. Il s'agit plus précisément d'identifier un ensemble d'indices de complexité et de les représenter à l'intérieur d'un cadre formel homogène, offrant ainsi la possibilité d'une quantification : le modèle proposé permet d'associer à chaque phrase un *indice* reflétant sa complexité.

L'évaluation de la complexité linguistique est une question aussi ancienne que la syntaxe formelle. L'idée de pouvoir quantifier cet élément est par exemple abordée dans (Chomsky65), en proposant l'estimation d'un indice dérivationnel. On retrouve ce paramètre dans la plupart des études en psycholinguistique (p. ex. (Gibson91)). La complexité y est considérée comme reflétant une difficulté d'interprétation par des sujets humains. Des expérimentations permettent de la mettre en évidence de différentes façons : temps de lecture des mots (cf. (Gibson98), (Vasishth03)), observation des mouvement oculaires lors de la lecture (voir (Lee07)). Un temps de lecture plus long sur un mot ou des retours en arrière du regard sont le signe d'une difficulté. Le problème consiste alors à prédire la complexité, sur la base des propriétés syntaxiques d'une phrase. Bien entendu, il ne peut s'agir là que d'une approximation : les mécanismes d'interprétation sont évidemment également dépendants de facteurs sémantiques ou extra-linguistiques comme la charge mémorielle liée à la fréquence, la saillance, etc. (cf. (Caplan99)). Pour autant, une première modélisation s'appuyant uniquement sur la syntaxe pourra malgré tout constituer une bonne approximation de la

complexité en offrant de plus la possibilité d'une évaluation automatique.

Cet article propose de réunir plusieurs des indices de complexité proposés en psycholinguistique à l'intérieur d'un modèle unique permettant de calculer automatiquement une estimation globale de la complexité d'un énoncé. Après la présentation de quelques facteurs de complexité, nous en proposons une représentation à l'intérieur d'un cadre formel basé sur les contraintes et permettant leur évaluation. Une illustration de l'utilisation de ce modèle appliqué à un cas standard de la complexité syntaxique (complexité comparée des relatives sujet et objet) est ensuite donnée.

2 Les théories de la complexité syntaxique

Nous proposons dans cette section un panorama de quelques approches fournissant des éléments d'évaluation de la complexité, s'appuyant essentiellement sur les dépendances et leurs conséquences. Ces modèles reposent en particulier sur la saturation de la structure (*l'hypothèse de dépendance incomplète*), la complexité informationnelle (*théorie de la localité de dépendance*) ainsi que d'autres paramètres comme le degré d'activation d'une catégorie ou la profondeur de la structure syntaxique.

2.1 Hypothèse de dépendance incomplète

Pour la plupart des théories, deux opérations sont facteurs de complexité dans l'analyse des phrases : l'intégration (insertion d'un élément nouveau à une structure syntaxique existante) et la mémorisation (nombre de dépendances syntaxiques incomplètes).

L'hypothèse de dépendance incomplète (cf. (Gibson91)) repose sur l'idée que la complexité est fonction du nombre de structures incomplètes à mémoriser. L'exemple (1b) présente une structure de ce type, comportant un constituant emboité (la relative) séparant le nom du verbe dont il est sujet. L'exemple (1c), comportant deux constructions de ce type, est considéré comme étant quant à lui trop complexe pour être interprété naturellement. En revanche, l'exemple (1d), comportant le même matériel lexical et le même nombre de relatives, est lui considéré comme plus simple car ne comportant pas de rupture entre les noms et les verbes dont ils sont sujets.

(1) a. *The reporter disliked the editor.*

b. *The reporter [who the senator attacked] disliked the editor.*

c. *#The reporter [who the senator [who John met] attacked] disliked the editor.*

d. *John met the senator [who attacked the reporter [who disliked the editor]].*

Il est donc possible de donner une estimation quantitative de ce facteur en sommant à chaque tête le nombre de dépendances incomplètes. Dans l'exemple (1a), les valeurs suivantes indiquent les coûts pour deux des positions de la phrase :

– Coût au niveau du SN *the reporter* = 1 : le SN sujet dépend du V qui suit
– Coût au SN *the senator* = 3 : les SN *the reporter* et *the senator* dépendent en tant que sujet d'un V à suivre ; *who* dépend en tant qu'objet d'un V à suivre

2.2 Théorie de la localité de dépendance (DLT)

Cette théorie, développée dans (Gibson00) propose de prendre en compte la charge liée au traitement des objets référentiels présents entre entre deux structures syntaxiques, la seconde devant être intégrée à la première. Cette théorie s'appuie sur l'identification de référents du discours, sur le coût de leur intégration et propose de compléter cette évaluation par une quantification du coût de mémorisation des objets traités.

Les référents du discours Plusieurs travaux ont montré l'importance du degré d'accessibilité des référents du discours (notés *DR*). Il s'agit ici de distinguer les référents en fonction de leur statut au moment où ils sont mobilisés pour l'interprétation. Un référent peut être "*nouveau*" (ils sont dans ce cas par exemple introduits par un SN indéfini). Il peut être "*donné*", faisant ainsi référence à un référent déjà utilisé (par exemple introduit par un pronom ou un nom propre). Il peut être également "*accessible*", c'est à dire qu'il a déjà été introduit mais de manière indirecte. Il est ainsi possible d'organiser cette information de façon hiérarchisée. A partir de cette organisation de l'accessibilité, des travaux (cf. (Aissen03)) ont proposé une hiérarchie reposant sur l'aspect plus ou moins défini du SN : *Pronom* > *Nom propre* > *SN défini* > *SN indéfini*. Une simplification de cette observation proposée dans (Gibson98) consiste à dire que les SN indéfinis sont les moins accessibles et nécessitent le plus de ressources pour être intégrés que les pronoms. On considère donc que les objets les plus coûteux sont les référents "nouveaux". Dans le modèle DLT, chaque référent ayant le statut "*nouveau*" représentera une unité de complexité, on les notera dorénavant *DRn*.

Coût d'intégration Le coût d'intégration défini par la DLT consiste à identifier la charge nécessitée par le traitement de nouveaux référents, qu'il s'agisse d'un objet (un SN défini) ou d'un événement (un verbe). Le processus d'intégration consiste, en rencontrant une tête, à rechercher dans la structure syntaxique en cours de construction la tête à laquelle cette nouvelle structure partielle va pouvoir se rattacher (à la manière d'un processus d'adjonction). Le coût de ce processus correspond au nombre de *DRn* séparant ces deux têtes, y compris la dernière.

Dans l'exemple (1b), voici deux exemples de coûts d'intégration :

- Coût à *attacked* = 3 unités. 1 unité pour le *DRn* événement de *attacked* + 2 unités pour les *DRn* entre le début de la relative et l'objet vide (les 2 *DRn* sont *the senator* et *attacked*)
- Coût à *admitted* = 3 unités. 1 unité pour le nouveau R *admitted* + 2 pour les *DRn* (*the senator* et *attacked*) séparant le sujet (*the reporter*) et le V *disliked*.

Coût de mémorisation Le second paramètre utilisé dans l'évaluation de la complexité par DLT est le coût de mémorisation. Il repose sur le nombre minimal de têtes syntaxique nécessaires entre le mot courant et la saturation d'une structure grammaticale complète (une phrase). Par exemple, à la réalisation d'un clitique sujet, on attend la réalisation d'un verbe pour obtenir une structure complète. Dans l'exemple (1b), nous aurons les coûts suivants :

- Coût à *The* = 2 unités. On attend un N pour terminer le SN sujet et un V, tête du SV, pour former un P.
- Coût à *who* = 3 unités. On attend une proposition relative, une position vide objet pour lier *who*, plus un SV principal.

2.3 Autres paramètres

Activation Dans une étude récente, (Vasishth03) décrit une limite des modèles de *dépendance incomplète* et de *localité de dépendance*. Les expériences montrent en effet que le degré *d'activation* d'un mot peut compenser de façon significative la complexité induite par une construction (par exemple l'extraction d'objet). Ainsi, les productions de type (2b) sont traitées plus rapidement que (2a), contrairement à ce que prédisent DI et DLT :

(2) a. *The rat the cat saw died.*

b. *The rat the cat briefly saw died.*

La différence vient de la présence de l'adverbe qui semble activer le verbe. Cette observation est confortée par des comportement identiques pour d'autres modificateurs ou compléments précédant la tête. Un constituant fortement activé sera plus facilement intégré au reste de la structure.

Une analyse comparable est proposée dans (Hawkins10) qui propose un principe de maximisation du traitement on-line, reposant sur le nombre de propriétés non satisfaites à un moment donné de la phrase : on préfère les structures dans lesquelles plus de propriétés sont satisfaites plus tôt.

Profondeur de la structure Plusieurs approches font reposer l'évaluation de la complexité sur la profondeur de l'arbre correspondant. Le modèle le plus simple consiste à compter le nombre de noeuds dans l'arbre (voir par exemple (Ferreira91)). Il est possible de compléter ce paramètre par la prise en compte du type de certains noeuds (e.g. les conjonctions de subordination, induisant un niveau d'emboitement supplémentaire, correspondant donc à un facteur de complexité). Ce type de modèle reste cependant trop superficiel et ne permet pas de rendre compte des effets cités précédemment. On sait en particulier que la complexité doit également prendre en compte le nombre de noeuds énumérés pour construire la structure, pas seulement ceux qui sont construits *in fine* : il s'agit en d'autres termes de prendre en compte l'impact de la stratégie d'analyse sur la complexité (cf. (Abney91)). D'autres approches s'appuient sur la hauteur des piles utilisées par le parser en cours d'analyse, indépendamment de la structure construite (voir par exemple (Schuler09) dont la technique repose sur des modèles de Markov hiérarchisés).

Si la profondeur de la structure ne peut donc à elle seule expliquer les variations de complexité, elle reste tout de même un élément important dans de nombreux modèles psycholinguistiques, à condition d'être complétée par d'autres informations.

Adjacence Dans une étude s'appuyant à la fois sur des temps de réaction et l'analyse de corpus, (Hawkins01) a montré que les syntagmes courts compléments ou adjoints sont plus souvent adjacents de leurs têtes.

(3) a. *The gamekeeper [looked [SP1 through is binoculars][SP2 into the blue but slightly overcast sky]].*

b. *The gamekeeper [looked [SP1 into the blue but slightly overcast sky] [SP2 through is binoculars]].*

En (3a), SP2 est distant de 3 mots de la tête, tandis qu'il est séparé par 5 mots de sa tête en (3b). Le premier type de production est traité dans un temps significativement plus court que le second. Cette observation s'appuie sur le fait que dans le premier cas, la connaissance du fait que le SV est composé de 2 SP est atteinte plus tôt. En d'autres termes, le locuteur atteint en 5 mots (de *looked* à *into*) le fait que 3 syntagmes seront utilisés : SV, SP1 et SP2. En (3b) en revanche, la même information est atteinte en 9 mots.

Hawkins propose un critère formé par le ratio $\frac{nombre\ de\ constituant}{nombre\ de\ mots}$ sur lequel il base un principe (*Early Immediate Constituents principle*) stipulant une préférence pour les ordres linéaires qui maximisent le ratio constituant/mot.

3 Calcul de la complexité : une approche par contraintes

Nous proposons dans cette section une représentation des paramètres de complexité en termes de contraintes. Chacune de ces contraintes peut être interprétée indépendamment de toute théorie. Elles sont ici spécifiées dans le cadre des *Grammaires de Propriétés* (cf. (Blache01)) dans lesquelles l'analyse consiste à évaluer la satisfaction d'un ensemble de contraintes (également appelées propriétés) formant la grammaire. Les contraintes du modèle de complexité reposent sur la satisfaction de certaines propriétés jouant le rôle de pré-conditions permettant l'évaluation du critère.

3.1 Les indices de complexité

Dépendances incomplètes Le principe de cette contrainte repose sur l'évaluation du nombre d'éléments séparant un complément de sa tête. Deux contraintes permettent d'identifier cette situation : la dépendance (notée $\rightsquigarrow$) et la linéarité (notée $\prec$). Soit C_i et C_j deux catégories, avec i et j les indices de leur position dans la chaîne. Si C_j est une tête dont dépend C_i et si C_i précède C_j, alors l'indice DI de dépendance incomplète au moment de l'analyse de C_i est le nombre de dépendances partant de C_i, auquel s'ajoute le nombre de dépendances incomplètes en cours. Cette information se représente par la contrainte suivante :

$$[(C_i \rightsquigarrow C_j) \wedge (C_i \prec C_j)] \Rightarrow [DI[i] \leftarrow (DI[i]+1)] \wedge [DI[j] \leftarrow (DI[j]-1)] \quad (4)$$

En d'autres termes, on incrémente l'indice à chaque mot initialisant une nouvelle dépendance, on le décrémente en rencontrant la cible de la dépendance.

Localité de dépendance On représente ici la partie relevant des coûts d'intégration de cette théorie. Elle consiste à identifier les objets référentiels présents entre une tête et son dépendant situé au début du syntagme. La valeur de l'indice est la cardinalité de l'ensemble de ces référents.

$$[(C_i \rightsquigarrow C_j) \wedge (C_i \prec C_j)] \wedge [\nexists k \mid (C_k \rightsquigarrow C_j) \wedge (C_k \prec C_i)] \Rightarrow DLT[j] \leftarrow |DRn[i,j]| \quad (5)$$
$$\text{avec } DRn_{[i,j]} = \{C_l[+ref] \mid i \leq l \leq j\}$$

Cette contrainte identifie le dépendant C_i le plus à gauche du syntagme projeté par C_j (il n'existe aucun dépendant C_k de C_j qui précède C_i). Elle instancie la valeur DLT au point j de la chaîne. Celle-ci correspond à la cardinalité de l'ensemble des catégories référentielles comprises entre le dépendant le plus à gauche et sa tête.

Activation Le degré d'activation d'une catégorie C_j se mesure en fonction du poids de la relation de dépendance existant entre cette catégorie et une autre catégorie C_i qui la précède et qui dépend d'elle. Nous proposons d'étendre cette mesure en prenant en compte toutes les relations syntactico-sémantiques entretenues par la tête avec des catégories qui la précède. Dans le cadre des *Grammaires de Propriétés*,

il s'agit simplement de l'ensemble des propriétés mettant en relation au moins la tête et une catégorie qui la précèdent. Nous appelons cette mesure la *densité*. Une catégorie est très activée si elle est la cible d'un grand nombre de relations (en termes de graphes, si le degré entrant du sommet est important). Intuitivement, une catégorie sera fortement activée si un grand nombre de contraintes sont elles-mêmes actives et attendent la réalisation de la catégorie pour être satisfaites.

$$(6) \qquad \begin{array}{l} Activ(C_j) = \sum Weight(\mathcal{P}_{C_j}) \\ \text{avec } \mathcal{P}_{C_j} = \{P(x, C_j) \mid (P(x, C_j) \in \mathcal{G}) \wedge (x \sqcup C_i) \wedge (C_i \prec C_j)\} \end{array}$$

La valeur de l'indice d'activation d'une catégorie C_j est donc la somme des poids des propriétés de la grammaire impliquant C_j et une autre catégorie C_i qui la précède. On note que dans l'état de cette définition, c'est la densité qui est prise en compte, en mettant au même niveau toutes les propriétés quel que soit leur type. Il est possible de modifier cet indice pondérant les valeurs en fonction du type de contraintes (permettant par exemple d'augmenter le poids de la relation de dépendance). C'est également à ce niveau que devraient être prise en considération les contraintes de sélection sémantique, augmentant de façon indépendante le niveau d'activation.

Degré d'emboîtement Nous avons vu qu'un des facteurs repris récemment par des approches computationnelles repose sur la profondeur maximale atteinte par la structure syntaxique lors de l'analyse. Dans une approche basée sur les contraintes, il existe plusieurs façons de calculer cet indice. Une solution simple, indépendante des stratégies de parsing, consiste à évaluer les niveaux de dépendance :

$$(7) \qquad [C_i \rightsquigarrow C_j] \Rightarrow Prof(C_i) = Prof(C_j) + 1$$

Dans ce calcul, chaque catégorie C dispose d'un indice de profondeur, donné par la fonction $Prof(C)$, initialisé à 0. En cas de dépendance, cet indice est égal à celui de la catégorie tête incrémenté d'une unité.

Taille des compléments / Adjacence Cette contrainte indique que si deux catégories C_i et C_j dépendent d'une même tête C_k et que C_i est plus grande en nombre de constituants que C_j, alors C_i sera adjacent à C_k. Il s'agit d'une contrainte dynamique dans le sens ou celle-ci sera ajoutée en cours d'analyse à l'ensemble des contraintes de la grammaire lorsque la pré-condition est réalisée.

$$(8) \qquad [(C_i \rightsquigarrow C_k) \wedge (C_j \rightsquigarrow C_k) \wedge (|C_i| > |C_j|)] \Rightarrow C_i \oplus C_k$$

Il est à noter que le principe d'adjacence tel que décrit dans cette contrainte n'implante pas directement le principe EIC (*Early Immediate Constituents*). Ce dernier est en effet destiné à comparer deux constructions, tandis que la contrainte ci-dessus est destinée à évaluer cette propriété de façon indépendante. La prise en compte de cette propriété dans la quantification de la complexité se fera via la satisfaction ou non de ces contraintes.

3.2 Evaluation globale de la complexité

La formalisation des indices de complexité à l'aide de contraintes permet d'en distinguer deux types : indices quantifiés et contraintes dynamiques. Cette distinction recoupe un second niveau d'information : il

est en effet possible de distinguer l'évaluation de la complexité à un instant t de l'analyse de la complexité d'une phrase complète. Nous parlerons dans ce qui suit de *complexité locale* vs. *complexité globale*.

Complexité locale Les premières contraintes correspondent à des indices quantifiés et sont associés à des catégories. Ils permettent de donner une estimation de la complexité au moment d'analyser la catégorie courante. Certains indices ont une valeur proportionnelle à la difficulté d'interprétation : c'est par exemple le cas de la profondeur ($Prof$), de l'indice de dépendances incomplètes (ID) ou de l'indice de dépendance locale (DLT). D'autres indices sont en revanche révélateurs d'une facilitation de l'interprétation, comme le degré d'activation ($Activ$).

On introduit ici la notion de complexité locale permettant de donner une indication de la complexité au moment de l'analyse d'une catégorie C.

$$Loc\_comp(C_i) = Prof(C_i) + DLT(i) + ID(I) - Activ(C_i) \tag{9}$$

Cette valeur est donc formée de la somme des valeurs des indices dénotant une difficulté à laquelle on retranche le degré d'activation. Dans un modèle prenant en compte la fréquence (ou le niveau de marque de la construction), on prendrait en compte cette valeur au titre de la facilitation, comme l'activation. Signalons que dans l'état de cette proposition, il n'est pas possible de proposer une pondération de ces critères, qui sera obtenue par une expérimentation ad hoc. Il est en effet difficile de prédire dans quelle proportion les indices de facilitations peuvent compenser les indices de difficulté.

Complexité globale Le second type de contraintes correspond à des propriétés ajoutées dynamiquement à la grammaire en cours d'analyse. C'est le cas de l'indice sur la taille des compléments entraînant l'adjacence à la tête des compléments. Il s'agit donc au final de contraintes au même niveau que celles de la grammaire. Leur violation entraînera donc, selon les études psycholinguistiques conduites dans ce domaine, un surcroît de complexité. Nous proposons d'élargir cette observation à la prise en compte de la violation de toutes les contraintes. L'idée consiste à dire que la complexité est inversement proportionnelle à la grammaticalité : le nombre de contraintes violées pour une phrase est un facteur indicatif de sa difficulté d'interprétation. Il s'agit donc d'un élément indicatif de sa complexité. La possibilité de quantifier ce que nous avons appelé "*l'indice de grammaticalité (IG)*" permet donc de proposer une valeur qui participera à l'évaluation de la complexité d'une phrase. Sans entrer dans les détails du calcul de IG (cf. (Blache06)), on rappellera simplement qu'il s'agit d'une fonction reposant sur le poids des contraintes satisfaites, violées, leur situation et leur importance relative dans la phrase et dans la grammaire.

Nous proposons une évaluation de la complexité globale d'une phrase se basant sur l'indice de grammaticalité et les complexités locales. L'indice de grammaticalité permet de rendre compte de la structure de l'énoncé et en particulier des contraintes violées. L'expérimentation décrite dans (Blache06) montre que les sujets trouvent plus complexes à interpréter les phrases ayant un indice de grammaticalité plus faible. Nous complétons ce type d'information avec la somme des indices de complexité locale rapportée à la taille de la phrase.

$$Glob\_comp(S_n) = \sum_{i=1}^{n} \frac{Loc\_comp(C_i)}{n} + \frac{1}{IG(S)} \tag{10}$$

4 Application de la complexité locale

Nous illustrons dans cette section à l'aide de quelques exemples, le mode de calcul de la complexité locale et ses effets sur l'identification du niveau de complexité. Dans ces exemples, la grammaire utilisée se limite aux relations de dépendances. Un seul type de propriété étant utilisé, nous n'introduisons pas de poids. Par ailleurs, et pour la même raison, nous ne prenons pas en compte la notion de complexité globale qui repose sur la pondération de la complexité locale par l'indice de grammaticalité de la phrase. En se limitant aux relations de dépendances, cet indice de grammaticalité est en effet directement pris en compte par l'indice d'activation et la profondeur. Une évaluation complète nécessiterait bien entendu la prise en compte d'une grammaire complète. Enfin, le poids respectif de chacun des indices doit être pondéré. En particulier, nous avons vu que l'importance de la profondeur de la structure de l'arbre (ou celle du parser) doit être relativisée.

Les exemples qui suivent ont vocation à illustrer le fonctionnement du modèle proposé plus haut. Les indices des secondes colonnes des tableaux suivants sont ramenés soit au nombre de têtes (c'est le cas des indices *DRn, Intégration*, et *DI*), soit au nombre de mots (*Mémorisation*). Les exemples pris sont ceux traditionnellement utilisés en psycholinguistique et pour lesquels des effets de complexité ont été montré par plusieurs expérimentations (temps de réaction, et mouvement oculaire en particulier). Quelques indices méritent d'être détaillés :

– L'indice *DI* répertorie le nombre de dépendance restant à satisfaire. Chaque tête de syntagme complément d'une autre tête située après incrémente la valeur d'une unité. Par exemple, le SN sujet, au niveau de la tête *reporter*, introduit une dépendance incomplète par rapport au verbe principal qui suit. Le pronom relatif sujet introduit quant à lui une dépendance incomplète pour la même raison et incrémente l'indice d'une unité.
– L'indice *Intégration* comptabilise le nombre de *DRn* séparant deux éléments d'une relation de dépendance. Dans cet exemple, le sujet et le verbe de la principale sont séparés par 3 référents (incluant le référent événement porté par le verbe lui-même).
– L'indice *Mémorisation* s'incrémente pour chaque nouvelle relation de complémentation, spécification ou adjonction attendue en vue de créer une structure syntaxique saturée. Par exemple, le déterminant du SN sujet introduit deux nouvelles relations : une spécification (du déterminant vers le nom) et une de complémentation (du SN qu'il introduit vers le verbe dont il est sujet).
– L'indice *Activation* identifie pour chaque tête le nombre de relations de dépendances introduites par une source antéposée dont elle est cible.
– Le total des coûts somme l'ensemble de ces indices. Rappelons que la valeur de complexité totale de l'énoncé est faite de cette somme diminuée de la valeur d'activation.

Le premier exemple illustre l'application du modèle à une phrase comportant une relative sujet et des objets multiples permettant d'introduire des référents du discours entre le verbe principal et le sujet.

		Le	reporter	qui	adressa	le	photographe	à	l'	éditeur	travaillait	sur	un	bon	sujet
DI	0,86	0	1	2	1	0	1	0	0	1	0	0	0	0	0
DRn	0,86	0	1	0	1	0	1	0	0	1	1	0	0	0	1
Intégration	0,43	0	0	0	0	0	0	0	0	0	3	0	0	0	0
Mémorisation	0,93	2	1	2	2	1	1	1	1	1	1	0	0	0	0
Profondeur	3,93	2	2	3	4	5	5	5	6	6	2	3	4	4	4
Total coûts	7	4	5	7	8	6	8	6	7	9	7	3	4	4	5
Activation	0,64	0	1	0	1	0	1	1	0	1	1	1	0	0	2
COMPLEXITE	6,36														

Le tableau suivant décrit le cas d'une relative objet. Les expériences montrent que ce type de construction est plus complexe que la précédente. Cet effet se manifeste dans plusieurs indices : les dépendances incomplètes sont en plus grand nombre, en particulier suite au fait que les sujets de la principale celui de la relative ainsi que son objet restent à lier au moment du verbe de la relative (cet effet se retrouvant également au niveau du coût de mémorisation). De la même façon, le coût d'intégration est supérieur à la relative sujet, chaque verbe étant séparé de la tête de sa structure locale par deux *DRn*. Au total, malgré une meilleure activation, et conformément aux données des expérimentations, la relative objet est prédite par le modèle comme étant plus complexe à traiter que la relative sujet.

		Le	reporter	que	le	photographe	adressa	à	l'	éditeur	travaillait	sur	un	bon	sujet
DI	1,14	0	1	2	0	3	1	0	0	1	0	0	0	0	0
DRn	0,86	0	1	0	0	1	1	0	0	1	1	0	0	0	1
Integration	0,86	0	0	0	0	0	3	0	0	0	3	0	0	0	0
Mémorisation	1,36	2	1	3	4	3	2	1	1	1	1	0	0	0	0
Profondeur	4,21	2	2	3	5	5	5	6	7	7	2	3	4	4	4
Total coûts	8,43	4	5	8	9	12	12	7	8	10	7	3	4	4	5
Activation	0,71	0	1	0	0	1	2	1	0	1	1	1	0	0	2
COMPLEXITE	7,71														

Le modèle permet par ailleurs de confirmer l'observation de complexité de la relative objet, y compris en comparant des phrases plus différentes. Le tableau suivant révèle cet effet : la phrase, malgré un nombre de mots inférieurs, est prédite comme plus complexe que la première phrase, conformément aux expériences. Signalons au passage que, indépendamment de la structure, la relative objet en anglais peut être encore plus complexe à traiter par un humain en cas d'élision du pronom relatif (*The reporter the senator attacked disliked the editor*).

		Le	reporter	que	le	sénateur	attaquait	détestait	l'	éditeur
DI	1,17	0	1	2	0	3	1	0	0	0
DRn	0,83	0	1	0	0	1	1	1	0	1
Integration	0,83	0	0	0	0	0	3	2	0	0
Mémorisation	1,78	2	1	3	4	3	1	1	1	0
Profondeur	3,33	2	2	3	5	5	5	2	3	3
Total coûts	7,94	4	5	8	9	12	11	6	4	4
Activation	0,67	0	1	0	0	1	2	1	0	1
COMPLEXITE	7,28									

5 Conclusion

Le modèle de complexité décrit dans cet article présente plusieurs intérêts. Il est tout d'abord un outil contribuant à l'explication du fonctionnement du parser humain. Il permet ainsi, en identifiant et localisant les éléments de complexité, de donner un élément de prédiction de la charge cognitive liée à la tâche d'interprétation d'un énoncé. Il devient alors possible de quantifier la complexité de chaque construction syntaxique. Un des phénomènes que ce type de modèle permettra d'examiner est celui de la contribution relative de chaque domaine linguistique (prosodie, morphologie, syntaxe, etc.) au processus d'interprétation : une zone de complexité syntaxique induit vraisemblablement un processus de compensation dans les autres domaines (la prosodie, par exemple, apportant des éléments d'informations complémentaires). Par ailleurs, un modèle de complexité constitue un élément décisif pour certaines théories linguistiques (typiquement la Théorie de l'Optimalité) en offrant la possibilité de hiérarchiser des phénomènes syntaxiques. Enfin, ce type d'information peut constituer un outil efficace pour la contrôle des processus d'analyse

syntaxique automatique, en particulier dans le cadre de stratégies conduisant à une surgénération de structures. Les perspectives de recherches sont nombreuses. L'intégration de ce modèle de complexité à un parser robuste permettra tout d'abord le traitement de données réelles. Il sera ainsi possible de tester sur des sujets humains la difficulté de traitement d'exemples attestés, franchissant un pas supplémentaire vers la description de la langue parlée. Ce type d'outil permettra par ailleurs une étude précise des facteurs de complexité syntaxique pour un sujet humain : il s'agit d'un élément indispensable avant d'envisager des études expérimentales lourdes sur les processus cognitifs, reposant par exemple sur l'utilisation de dispositifs comme l'IRMf.

Références

[Abney91] Abney S. & M. Johnson (1991) "Memory requirements and local ambiguities of parsing strategies", in *Journal of Psycholinguistic Research*, 20(3).

[Aissen03] Aissen, J. (2003) "Differential object marking : Iconicity vs. economy", in *Natural Language and Linguistic Theory*, 21.

[Blache01] Blache P. (2001) *Les Grammaires de Propriétés : des contraintes pour le traitement automatique des langues naturelles*, Hermès Sciences.

[Blache06] Blache P. , B. Hemforth, & S. Rauzy (2006) "Acceptability prediction by means of grammaticality quantification", in proceedings of *COLING/ACL 2006*.

[Caplan99] Caplan, D. & Waters, G (1999) "Verbal working memory and sentence comprehension". Behavioral and brain Sciences, 22.

[Chomsky65] Chomsky, N. (1965) Aspects of the theory of syntax. Cambridge, MIT Press.

[Ferreira91] Ferreira F. (1991) "Effects of Length and Syntactic Complexity on Initiation Times for Prepared Utterances", in Journal of Memory and Language, vol. (30/2).

[Gibson91] Gibson E. (1991) A computational theory of human linguistic processing : memory limitations and processing breakdown, PhD Dissertation, Carnegie Mellon University.

[Gibson98] Gibson E. (1998) "Linguistic complexity : Locality of syntactic dependencies", 1998 vol. 68 (1).

[Gibson00] Gibson, E. (2000) "The dependency locality theory : A distance-based theory of linguistic complexity". In A. Marantz, Y. Miyashita, & W. O ?Neil (Eds.), Image, language, brain : Papers from the first mind articulation project symposium, MIT Press.

[Hale06] Hale J. (2006) "Uncertainty About the Rest of the Sentence", in Cognitive Science 30

[Hawkins01] Hawkins J. (2001) "Why are categories adjacent", in Journal of Linguistics, 37.

[Hawkins10] Hawkins J. (2010) "Processing efficiency and complexity in typological patterns", à paraître in Oxford Handbook of Language Typology (J.J. Song, ed.), Oxford University Press.

[Lee07] Lee Y., H. Lee, P. Gordon (2007) "Linguistic complexity and information structure in Korean : Evidence from eye-tracking during reading", in Cognition, vol. 104 (3)

[Vasishth03] Vasishth S. (2003) "Quantifying processing difficulty in human sentence parsing : The role of decay, activation, and similarity-based interference", in Proceedings of Eurocogsci 03 : The European Cognitive Science Conference 2003

[Schuler09] Schuler W. (2009à "Positive results for parsing with a bounded stack using a model-based right-corner transform", in proceedings of NAACL 2009