TALN 2004, Fès, 19-21 avril 2004

Densité d'information syntaxique et gradient de grammaticalité

Philippe Blache

Laboratoire Parole et Langage
CNRS – Université de Provence
29, Avenue Robert Schuman
13621 Aix-en-Provence
pb@lpl.univ-aix.fr

Résumé – *Abstract*

Cet article propose l'introduction d'une notion de densité syntaxique permettant de caractériser la complexité d'un énoncé et au-delà d'introduire la spécification d'un gradient de grammaticalité. Un tel gradient s'avère utile dans plusieurs cas : quantification de la difficulté d'interprétation d'une phrase, gradation de la quantité d'information syntaxique contenue dans un énoncé, explication de la variabilité et la dépendances entre les domaines linguistiques, etc. Cette notion exploite la possibilité de caractérisation fine de l'information syntaxique en termes de contraintes : la densité est fonction des contraintes satisfaites par une réalisation pour une grammaire donnée. Les résultats de l'application de cette notion à quelques corpus sont analysés.

This paper introduces the notion of syntactic density that makes it possible to characterize the complexity of an utterance and to specify a gradient of grammaticality. Such a gradient is useful in several cases: quantification of the difficulty of interpreting an utterance, quantification of syntactic information of an utterance, description of variability and linguistic domains interaction, etc. This notion exploits the possibility of fine syntactic characterization in terms of constraints: density if function of satisfied constraints by an utterance for a given grammar. Some results are presented and analyzed.

Keywords – Mots Clés

Syntaxe, analyse, robustesse, contraintes, information linguistique, complexité syntaxique.
Syntax, parsing, robustness, constraints, linguistic information, syntactic complexity.

1 Introduction

Certains phénomènes linguistiques, certaines constructions sont caractérisées par des propriétés très marquées, facilement identifiables. C'est le cas par exemple de la syntaxe pour laquelle des propriétés de forme, d'ordre ou de rection permettent de caractériser des constructions comme les clivées. Il est alors possible de décrire avec une grande précision de telles constructions, mais également de les interpréter facilement. Les clivées sont en effet souvent simples à interpréter même si leur représentation formelle et leur analyse automatique est complexe. Il n'y a en général que peu d'ambiguïté pour ce type de construction. A

l'inverse, il existe d'autres types de constructions, comme les dislocations qui, même si elles sont identifiables, présentent quelquefois plus de problèmes pour leur interprétation.

Il est très difficile de fournir une explication à cette différence de fonctionnement si l'on s'en tient au niveau global de la construction. Il n'est en effet pas possible de dire que les clivées sont plus faciles à interpréter que les disloquées, d'une façon générale, sans en fournir de raison. La liste des constituants de chacune de ces constructions pas plus que leur position dans la structure syntaxique ne nous fournit d'indication sur cette difficulté d'interprétation ou la potentialité d'ambiguïté. La notion de complexité en syntaxe ou de difficulté d'analyse reste encore aujourd'hui un problème difficile à décrire. Certains travaux en psycholinguistique à ce sujet tentent de fournir des explications en termes de distance (cf. en particulier [Gibson00]), mais sans rendre compte des différences très importantes à l'intérieur d'une même construction. Or, à un niveau d'analyse fin, il est possible d'identifier des différences ou du moins de fournir quelques caractéristiques plus régulières pouvant servir d'élément de réponse à ce problème. Il faut pour cela dissocier l'information syntaxique et tenter d'en caractériser plus finement ses propriétés. Il devient ainsi possible non seulement d'indiquer quelles sont les propriétés d'ordre purement syntaxique de la construction (par exemple l'ordre, la restriction de cooccurrence ou l'accord), mais également d'indiquer des caractéristiques générales concernant notamment la quantité d'information syntaxique, son importance, sa fiabilité, etc. Dans certains cas, l'analyse permet de fournir un grand nombre d'informations sur les objets linguistiques (par exemple les catégories) formant une construction. Dans d'autres, ces informations sont au contraire plus rares. Il est alors possible de penser que dans les cas où l'information est rare ou peu "importante", l'interprétation de la construction sera plus difficile que dans ceux où l'information est abondante et fiable. Cette question est au coeur de certaines approches comme les grammaires de construction (cf. [Fillmore93] ou [Goldberg95]) qui proposent une description exploitant en même temps différentes informations provenant de différents domaines linguistiques. Nous proposons dans cet article d'introduire dans la description une notion de *densité d'information*. Celle-ci ne tient pas compte du contenu informationnel porté par la syntaxe mais de sa forme. Il est ainsi possible de fournir un élément d'information quantifiée propre à une construction et donnant des indications sur son interprétabilité. Les constructions à forte densité syntaxique seraient ainsi plus faciles à interpréter que celles à faible densité. L'idée que la quantité d'information facilite l'analyse au lieu de la complexifier est décrite par exemple dans [Vasishth03]. Nous proposons ici de quantifier cette donnée.

Ce type d'approche présente un triple intérêt descriptif, cognitif et computationnel. Il devient en effet possible de fournir des indications sur la complexité syntaxique d'une construction non plus par les seules relations syntaxiques entrant en jeu, mais également par une évaluation de leur importance, notamment du point de vue quantitatif. De plus, la densité permet de fournir un élément de réponse à des questions d'ordre psycholinguistique comme les préférences ou les difficultés d'interprétation. Enfin, du point de vue du traitement automatique, la densité constitue un critère précieux pour identifier des zones où la quantité d'information est plus abondante et plus riche. Ce type de critère peut être utile par exemple dans la recherche d'information, le typage de document ou l'analyse de sa structure discursive. Inversement, le repérage d'une faible densité permet, par exemple dans le cas de systèmes de traitement de langue parlée, d'identifier les cas où des heuristiques particulières devront être appliquées pour permettre l'interprétation.

2 La nature de l'information syntaxique

Plusieurs approches en syntaxe tentent aujourd'hui de se démarquer du cadre génératif pour proposer une vision de l'analyse plus souple et permettant de s'adapter à la réalité des données langagières. La forme des énoncés que nous sommes amenés à traiter ainsi que les processus d'élaboration du contenu informatif montrent en effet que l'information linguistique est dispersée et instable. Dans cette perspective, la vision proposée par les théories génératives stipulant que la langue est un ensemble et la grammaire un processus permettant l'énumération de cet ensemble n'est pas adaptée : la question fondamentale n'est en effet pas de savoir si un énoncé appartient ou non à une langue mais bien d'extraire l'information de cet énoncé, quelle que soit sa forme. Dans certains cas, il est ainsi possible d'associer à l'énoncé analysé une structure syntaxique le décrivant en totalité. Cependant, dans de très nombreux cas, une telle structure n'est pas calculable. De plus la granularité de l'information obtenue dans ce type de démarche est homogène: tous les composants de la structure sont de même niveau. Or, dans de nombreux cas, seules des informations très partielles peuvent être obtenues. C'est le cas de l'exemple (1), tiré de [Mertens93], dans lequel seules quelques indications sont accessibles :

(1) *lundi lavage mardi repassage mercredi repos*

Cet exemple illustre une construction courante, pas seulement en langue parlée, dans laquelle la relation entre les éléments est fournie notamment par la répétition d'un schéma. La prosodie sera dans ce cas d'une grande importance, nous y reviendrons. Pour ce qui concerne le seul niveau syntaxique, nous avons donc une répétition de la séquence `N1[temporel] < N2[action]`. La morphologie et l'ordre linéaire sont les seules caractéristiques de cette construction. Une analyse syntaxique à proprement parler n'est pas possible dans la mesure où aucune relation particulière en dehors de la linéarité ne relie les objets de l'énoncé. Il est en revanche indispensable de spécifier les informations accessibles de façon à permettre l'interprétation. Dans l'exemple (2), nous retrouvons un problème similaire, même si plus d'informations syntaxiques sont accessibles.

(2) *Marie, je la supporte pas*

Dans ce cas, l'énoncé n'est pas interprétable sur la base des seules informations syntaxiques dans la mesure où la relation entre le premier nom et le reste de l'énoncé n'est pas spécifiée. Plus précisément, deux interprétations sont possibles : la première en disloquée, avec une coréférence entre le nom et le clitique, et la seconde en vocatif, sans coréférence. La prosodie dans ce cas jouera un rôle déterminant (cf. [Blache03]). Cet exemple illustre le cas où le domaine syntaxique n'apporte pas suffisamment d'information permettant à lui seul l'interprétation.

D'une façon générale, ces exemples illustrent d'une part la nécessité de représenter tout type d'information syntaxique, y compris en l'absence de véritable relation, et d'autre part l'intérêt de tenter de quantifier cette information. Dans certains cas, l'information syntaxique est en effet plus marquée, plus forte, plus dense que dans d'autres. Nous proposons dans le reste de cet article une définition de cette notion de densité s'appuyant sur une représentation décentralisée de l'information. Dans la prochaine section, nous rappellerons rapidement les principales caractéristiques des Grammaires de Propriétés permettant une représentation de

l'information syntaxique à granularité variable. Nous proposerons ensuite une spécification de la densité en fournissant des exemples tirés de corpus.

3 Une représentation décentralisée de l'information : les Grammaires de Propriétés

Une approche permettant la représentation de l'information linguistique telle que nous l'avons décrite, à savoir instable et dispersée, doit s'appuyer sur une conception non holistique de cette information (cf. [Pullum03]). En d'autres termes chaque type d'information doit être représenté séparément et surtout pouvoir être évaluée séparément. A la différence des règles syntagmatiques dans les approches génératives qui n'ont de valeur que situées dans un processus de dérivation général (donc dans une structure générale), à la différence également de la théorie de l'optimalité (cf. [Prince93]) dans laquelle une contrainte n'a de signification que par rapport aux autres, les propriétés syntaxiques doivent pouvoir être interprétées indépendamment des autres propriétés.

C'est sur ce principe que reposent les *Grammaires de Propriétés* (cf. [Blache01]). Celles-ci distinguent plusieurs types de propriétés syntaxiques, sans relation hiérarchique : aucun ordre d'évaluation n'est imposé et chaque propriété représente une information homogène. La description de la structure syntaxique d'un énoncé est constituée par l'ensemble des propriétés qui peuvent être évaluées. Le tableau suivant récapitule les types de propriétés syntaxiques actuellement utilisées dans les grammaires de propriétés, de nouvelles propriétés pouvant éventuellement être ajoutées (il se peut par exemple qu'une propriété de contiguïté soit utile).

Propriété	*Definition*
Linéarité (<)	Contraintes de précédence linéaire
Dépendance (→)	Relations de dépendance
Obligation (*Oblig*)	Ensemble des catégories obligatoires. Une de ces catégories et une seule doit être réalisée dans un syntagme
Exclusion (≠)	Restriction de cooccurrence entre ensembles de catégories
Exigence (⇒)	Obligation de cooccurrence entre ensembles de catégories
Unicité (*Unic*)	Catégories ne pouvant être répétées dans un syntagme

Une propriété est une contrainte représentant une information spécifique d'une catégorie. La grammaire est un ensemble de contraintes de ce type et chaque catégorie de la grammaire est décrite par un sous-ensemble de ces contraintes. Certaines de ces contraintes sont caractéristiques d'une catégorie et permettront pendant l'analyse de l'instancier. Une analyse consiste, pour un énoncé donné, à évaluer l'ensemble des contraintes. A chaque catégorie sera associé l'ensemble des contraintes que sa réalisation satisfait ainsi que celles qu'elle ne satisfait pas. Ces deux ensembles forment la *caractérisation* de la catégorie. Une telle approche permet donc d'associer une caractérisation à une catégorie, quelle que soit la forme de sa réalisation. Dans les cas où toutes les contraintes sont satisfaites, la catégorie est grammaticale, mais ceci ne lui confère aucun statut particulier. Le fait que des contraintes ne soient pas satisfaites n'empêche en effet pas l'utilisation de la catégorie correspondante. De plus, nous disposons alors d'une approche permettant de spécifier un gradient de grammaticalité.

Plusieurs analyseurs de grammaire de propriétés ont été développés (cf. [Blache02]). Nous présentons rapidement les caractéristiques principales de l'analyseur utilisé pour les expérimentations décrites plus loin. Le schéma général de l'analyse en GP consiste, pour un

ensemble donné de catégories (on parle d'*affectation*) à déterminer sa caractérisation. Pour cela, l'ensemble des contraintes de la grammaire est activé, un sous-ensemble d'entre elles sont évaluables et permettent de commencer à construire une caractérisation. Celle-ci, formée de contraintes satisfaites et non satisfaites, est analysée et en cas de présence de contraintes caractéristiques, la catégorie correspondante est instanciée. Chaque nouvelle catégorie instanciée est alors disponible pour participer à une nouvelle affectation qui sera à son tour évaluée. Le système construit donc de façon incrémentale les affectations (ces ensembles de catégories) qui servent de base à la construction des caractérisations. Le schéma général de l'analyse se présente comme suit :

1. *Construction d'une affectation $\mathcal{A}$ (choix d'un ensemble de catégories)*
2. *Evaluation de la satisfaisabilité de $\mathcal{A}$ parmi l'ensemble total de contraintes*
3. *Analyse de la caractérisation de $\mathcal{A}$, instanciation de la catégorie syntagmatique correspondante*

La caractérisation, au coeur de l'analyse en GP, est un mécanisme de satisfaction de contraintes. Il est possible d'en régler le niveau de relaxation en autorisant ou non la violation de contrainte. Dans le cas où toutes les contraintes doivent être satisfaites, seules des catégories grammaticales sont construites. Cependant, une telle approche n'est pas adaptée à la réalité, en particulier pour de l'analyse de textes tout-venant. Dans l'analyseur utilisé ici, nous avons choisi de relâcher la satisfaisabilité des contraintes de dépendance, d'exigence et d'obligation, ces propriétés étant d'une part les moins filtrantes et d'autre part celles dont la satisfaisabilité peut varier en complétant l'affectation initiale (cf. [Dahl04]). Cette stratégie permet de développer un analyseur déterministe mais tolérant.

L'exemple suivant montre la caractérisation d'une catégorie (la relative "*qu'ils chassaient à cheval*") comportant l'indication de ses bornes dans l'énoncé analysé, son statut de grammaticalité, son affectation et sa caractérisation à proprement parler. Celle-ci indique par exemple le respect des propriétés de linéarité, de l'obligation de réalisation d'un SV avec un relatif sujet ou encore l'unicité de ses constituants.

Catégorie	*Gauche*	*Droite*	*Affectation*	*Caractérisation*
Rel	53	57	ProR:qu; SN:sujet; SV	ProR<SN; ProR<SV; SN<SV ProR[suj]=>SV; SN->SV; SV->ProR Oblig:ProR Unic = {ProR, SN, SV}

4 La densité d'information

Nous avons vu dans la section précédente comment caractériser une catégorie par l'ensemble des propriétés évaluables parmi celles qui la décrivent potentiellement. Le nombre de propriétés évaluées par rapport aux propriétés décrivant la catégorie est intéressant et constitue un premier indice de la "qualité" de l'information. Plus précisément, le nombre de propriétés satisfaites par rapport au nombre total de propriétés décrivant cette catégorie dans la grammaire (noté `dens_sat`) nous fournit une indication brute sur la quantité d'information syntaxique contenue dans la catégorie. L'hypothèse émise est que plus nous disposons d'informations au niveau syntaxique, plus la catégorie construite est fiable. Dans les cas où `dens_sat` est faible, dans la mesure où peu de propriétés ont été satisfaites, la catégorie caractérisée le sera de façon moins fiable que pour une catégorie ayant un rapport élevé. La

fiabilité constitue donc un outil permettant de classer des réalisations de catégories : les catégories très fiables satisfont plus de contraintes que celles qui le sont moins. En d'autres termes, ce rapport représente un élément d'information sur la grammaticalité de la catégorie. Si nous disposons simultanément de l'évaluation des propriétés satisfaites et des propriétés non satisfaites par rapport au nombre total de propriétés, ce *degré de grammaticalité* devient alors très précis. De plus, nous pensons que la fiabilité syntaxique constitue également une facilitation pour l'interprétation. Une catégorie très fiable sera plus facilement interprétable à l'aide des seules informations syntaxiques qu'une catégorie non fiable.

Dans la version de l'analyseur GP présentée plus haut, l'information de fiabilité, ou indication de *densité*, est associée à toute catégorie construite. Cet analyseur étant déterministe, le seuil de tolérance de non satisfaction de propriété est très bas. Le rapport des propriétés non satisfaites sur le nombre de propriété, malgré son importance, n'est donc pas retenu ici. Seule la densité des propriétés satisfaites est utilisée. Cette information est malgré tout intéressante et permet de distinguer plusieurs types de constructions. L'exemple suivant fournit le type d'indications caractérisant une catégorie. En plus des informations données plus haut, nous trouvons donc les indications de densité sous la forme des deux rapports suivants :

- `dens_sat = nb propriétés satisfaites / nb total de propriétés`
- `dens_unsat = nb propriétés non satisfaites / nb total de propriétés`

Rappelons que parmi toutes les propriétés décrivant une catégorie dans la grammaire, pour une réalisation donnée, seul un sous-ensemble des ces propriétés est évaluable. Par exemple, pour un `SN` formé de `/Det N/`, toutes les propriétés mettant en jeu d'autres catégories ne sont pas évaluables. La somme des deux densités n'est donc pas égale à 1, ce qui justifie l'utilisation de ces deux rapports.

Catégorie	*Gauche*	*Droite*	*Statut*	*Dens_Sat*	*Dens_Unsat*	*Affectation*
P	0	24	Vrai	0,375	0,125	SP; SN

Dans cet exemple, la catégorie `P` est formée d'un `SP` et d'un `SN`. La grammaire utilisée stipule une relation de dépendance entre un `SN` et un `SV`, expliquant la présence d'une densité non nulle pour les propriétés non satisfaites. La densité relativement faible de propriétés satisfaites s'explique par le fait qu'une proportion importante des propriétés décrivant la catégorie `P` met en jeu le `SV`.

Un des enjeux des approches formelles de la syntaxe est la capacité à prendre en compte tout type d'énoncé. Il est cependant nécessaire de distinguer un gradient de grammaticalité permettant d'indiquer en quelque sorte le niveau de bonne formation d'un énoncé relativement à une grammaire. Une telle approche permet de n'exclure a priori aucune production et offre l'avantage de fournir une caractérisation, quelle que soit la forme de l'énoncé analysé. La question de la grammaticalité n'est ainsi plus un problème de décision (oui ou non l'énoncé appartient-il à la langue) mais un problème de description pur. La densité joue alors parfaitement ce rôle d'indicateur pour le gradient de grammaticalité. La notion de gradient prend toute son importance lorsqu'on aborde la question de l'interaction entre les différents domaines de l'information linguistique. Il est en effet nécessaire d'une part d'expliquer comment ces domaines interagissent entre eux, mais également de tenter de comprendre pourquoi, dans certains cas, l'interaction est nécessaire plus que dans d'autres. L'hypothèse que nous avons exposée dans [Blache02] et [Blache03], est qu'il existe un seuil d'information nécessaire au-delà duquel l'interprétation d'un énoncé devient possible. Ce seuil est atteint par le cumul des informations provenant des différents domaines. Un domaine peut à lui seul

contenir suffisamment d'information et dans ce cas, les autres domaines deviennent en quelque sorte moins importants. Nous expliquons de cette façon la variabilité plus ou moins grande de tel ou tel domaine en fonction du type de construction. Par exemple, si les informations morpho-syntaxiques sont suffisamment importantes, la prosodie aura une possibilité de variabilité plus grande. Inversement, lorsque la prosodie est caractéristique d'une construction (par exemple un contour ascendant caractéristique de l'interrogation), la syntaxe sera plus variable, ce qui explique la possibilité en français de construire une tournure interrogative avec une forme syntaxique de surface affirmative. De même, dans l'exemple (1) cité plus haut, la densité des relations syntaxiques est faible et l'interprétation est rendue possible par l'identification de la répétition d'un schéma lexical renforcé par la structure prosodique. Plus généralement, dans le cas d'une construction à densité syntaxique faible, le recours à des informations provenant d'autres domaines est probable.

5 Expérimentation

Cette section présente une investigation un peu plus précise portant sur trois corpus limités : un corpus de langue écrite (extrait du journal Le Monde) de 15.420 mots et deux corpus de langue parlée transcrite (entretiens non supervisés, langue spontanée). Ces derniers, compte tenu des difficultés d'analyse, sont beaucoup plus réduits et comportent 523 et 1.923 mots. Il est très difficile d'adopter pour ce qui concerne l'analyse automatique de la langue parlée, une position rendant précisément compte de la production réalisée. La transcription a donc été filtrée de façon à éliminer toutes les informations d'ordre non lexical, en particulier les mots incomplets. En revanche, toutes les répétitions sont maintenues. Il faut par ailleurs préciser qu'il ne s'agit pas dans cette expérience d'évaluer les performances de l'analyseur, mais de comparer des indications syntaxiques construites par un analyseur donné pour une grammaire donnée. Chaque catégorie, lorsqu'elle est construite (et dans la mesure où l'analyseur est déterministe) satisfait les principales contraintes qui la caractérisent. Plus précisément, la stratégie utilisée impose la satisfaction des propriétés de *linéarité*, *d'exclusion* et *d'unicité*. Une catégorie construite par l'analyseur est donc globalement grammaticale et dans la plupart des cas, elle possède une densité de propriétés non satisfaites nulle. Il est important de préciser que la catégorie ainsi construite ce correspond pas toujours à la bonne réponse (erreur d'étiquetage, grammaire partielle, etc.). Cependant, une comparaison brute des résultats sur des corpus différents, si elle ne constitue pas une évaluation des performances de l'analyseur, fournit cependant des éléments d'information intéressants sur l'approche générale.

Nous nous appuyons donc dans ce qui suit les seules indications de densité de propriétés satisfaites. Celle-ci permet notamment comparer la fiabilité de la caractérisation et plus généralement de l'information syntaxique contenue : une densité élevée est en effet révélatrice d'une quantité d'information importante. Le tableau de la figure 3 présente quelques exemples de réalisation du SN pris dans le Corpus A et comportant des densités contrastées :

Densité	*Constituants*
0,034483	ProP
0,068966	Clit
0,103448	N
0,1724138	ProP SA
0,206897	Det SA
0,241379	Det SP Rel
0,275862	Det N

Densité	*Constituants*
0,310345	Det SA SP
0,379310	Det N Rel
0,413793	Det SA N
0,413793	Det N SP
0,517241	Det N Rel SP
0,551724	Det N SA SP
0,655172	Det N SP SA Rel

Figure 3 : *Densités croissantes du SN (Corpus A)*

On constate dans ce tableau des situations différentes selon le type de réalisation. En effet, la densité ne croît pas systématiquement avec le degré de grammaticalité. Les deux densités les plus basses (pronom personnel et clitique) correspondent à des constructions grammaticales dans la grammaire utilisée. Ce phénomène vient du fait qu'une seule catégorie fait partie de la liste des constituants et que seul un petit nombre de contraintes de la grammaire est donc évaluable. Cette même explication est valable pour expliquer la différence de densité entre la construction `/Det N/` et `/Det N SP/` qui est plus dense. De même une réalisation moins grammaticale comme `/Det SA SP/` pourra se voir attribuer une densité plus forte. Le filtrage de ce type d'effet pourrait se faire par la prise en compte de la densité de contraintes non satisfaites. Ce paramètre étant spécifié, la progression de densité correspond bien à une croissance de la quantité d'information, notamment en termes de dépendance sémantique. Les densités les plus élevées correspondent en effet toutes à des constructions grammaticales et complexes.

Il est par ailleurs intéressant d'examiner les résultats obtenus sur les trois corpus, donnés dans la figure 4. Dans ce tableau sont indiqués pour chaque corpus le nombre de mots qu'il contient, le nombre de catégories syntagmatiques construites par l'analyseur, le nombre d'occurrence de chacune de ces catégories, sa proportion dans le corpus et la densité moyenne de ses propriétés satisfaites. La taille des corpus étant limitée, il n'est bien entendu pas question de généraliser trop rapidement les observations. On constate en particulier que le corpus B, très petit, a des écarts relativement importants avec les autres à la fois sur les fréquences, ce qui est normal, mais également sur la densité. Cependant, la moyenne des deux corpus de langue parlée (donnée en figure 5a) se rapproche aussi bien pour la fréquence que la densité du corpus de langue écrite. Il est donc possible d'en tirer quelques informations générales.

	Corpus A			Corpus B			Corpus C		
	Nombre	*Fréquence*	*Dens_sat*	*Nombre*	*Fréquence*	*Dens_sat*	*Nombre*	*Fréquence*	*Dens_sat*
Mots	15 420			523			1 923		
Cat. synt.	10 487			110			728		
P	660	0,062935	0,571860	11	0,100000	0,386363	29	0,039835	0,560344
SA	1 273	0,121388	0,565121	11	0,100000	0,381818	85	0,116758	0,435294
Sadv	314	0,029942	1,000000	5	0,045455	1,000000	37	0,050824	1,000000
SN	3 080	0,293697	0,314140	32	0,290909	0,191810	228	0,313187	0,217332
SP	2 153	0,205302	0,351266	10	0,090909	0,273148	80	0,109890	0,353472
SV	1 912	0,182321	0,391635	24	0,218182	0,333333	160	0,219780	0,350657
Circ	241	0,022981	0,682157	6	0,054545	0,733333	54	0,074176	0,703703
Coord	628	0,059884	0,306187	11	0,100000	0,428571	32	0,043956	0,535714
Rel	226	0,021550	0,750000	0	0,000000	0,000000	23	0,031593	0,708695

Figure 4 : *Résultat d'analyse des 3 corpus*

Sans entrer dans les détails d'une analyse comparée, on constate dans ces tableaux que certaines catégories ont une densité moyenne très élevée, voire maximale (c'est le cas du `SAdv`). Les catégories de ce type sont généralement caractérisées par un petit nombre de constituants possibles et un petit nombre de propriétés dans la grammaire. Il s'agit de catégories relativement stables dans le sens où elles sont peu variables. A l'opposé, la catégorie ayant la plus faible densité moyenne est le `SN`, dans tous les corpus. Cette catégorie est potentiellement grande, elle est surtout très variable dans ses possibilités de construction et

se décrit donc par un grand nombre de propriétés. Par ailleurs, il est intéressant d'examiner la corrélation entre la fréquence et la densité, comme indiqué dans la figure 5b suivante :

Cat	*Fréquence*	*Densité*
P	0,069917582	0,4733535
SA	0,108379121	0,408556
Sadv	0,048139361	1
SN	0,302047952	0,204571
SP	0,1003996	0,31331
SV	0,218981019	0,341995
Circ	0,064360639	0,718518
Coord	0,071978022	0,4821425
Rel	0,015796703	0,3543475

Fig. 5a: *Moyennes*

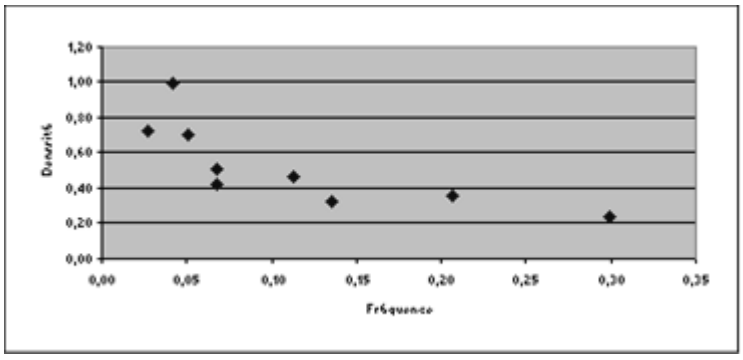

Fig. 5b : *Corrélation densité/fréquence*

On constate dans cette figure une tendance pour les catégories les plus fréquentes à être associées aux densités les plus basses : les trois catégories les plus fréquentes (`SN`, `SV` et `SP`) sont aussi celles ayant les plus faibles densités tandis que les moins fréquentes (`Circ`, `SAdv` et `Coord`) sont celles de densité plus forte. Les éléments d'explication donnés plus haut concernant le nombre de constituants, de propriétés et la combinatoire (donc la variabilité) qui en découle s'appliquent ici. Il est donc possible de dire également que les catégories les plus fréquentes sont celles qui sont décrites par le plus de propriétés. Il est donc nécessaire de pondérer la densité de satisfaction brute décrite précédemment par la densité moyenne de la catégorie concernée. Le calcul de la densité est ainsi ramené pour chaque catégorie à une échelle permettant d'évaluer l'importance de la densité d'une construction donnée en montrant les fluctuations au delà ou en deçà des densités moyennes.

6 Conclusion

La représentation de l'information syntaxique sous forme de contraintes et la notion de caractérisation, décrivant les propriétés syntaxiques d'un énoncé sous la forme d'ensembles de propriétés satisfaites et non satisfaites, qu'on peut en tirer permettent d'introduire la notion de *densité* d'information. Certaines constructions présentent des densités syntaxiques plus élevées que d'autres. Il s'agit d'une part de l'indication d'une quantité d'information variable et d'autre part d'une spécification de la qualité de l'information syntaxique: une densité faible révèle soit une faible quantité d'information, soit une proportion importante de contraintes non satisfaites. La densité constitue donc un outil permettant également la définition d'un *gradient* de grammaticalité utile dans l'analyse de textes tout-venant. Bien entendu, la densité ne peut être définie que pour un ensemble de propriétés donné. Elle n'a donc de valeur comparative que pour cette grammaire. Il est cependant utile de remarquer que les types de propriétés sont généraux et qu'il est sans doute possible (cela reste à démontrer) de représenter à l'aide de propriétés les informations représentées sous un autre formalisme comme les grammaires de dépendance ou les grammaires de construction (voir [Blache04]).

Il s'agit aussi d'un élément d'identification de la complexité syntaxique et, parallèlement, de la difficulté d'interprétation d'un énoncé : une densité faible est associée à une plus grande difficulté d'interprétation. Enfin, la densité permettant de quantifier l'information, elle est un élément quantitatif d'explication de la variabilité : une densité d'information faible est associée à une variabilité plus grande. Cette notion peut donc être exploitée à la fois d'un point de vue

théorique pour identifier les cas où l'interaction entre les domaines linguistiques est nécessaire pour compenser un "défaut" d'information, du point de vue computationnel pour indiquer de façon explicite un seuil de fiabilité de la description construite et du point de vue cognitif en fournissant un élément d'explication de la difficulté d'interprétation.

Références

BALFOURIER J.-M., P. BLACHE & T. VAN RULLEN (2002), "From Shallow to Deep Parsing Using Constraint Satisfaction", in proceedings of *COLING-2002*

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

BLACHE P. & A. DI CRISTO (2002), "Variabilité et dépendances des composants linguistiques", in actes de *TALN-2002*.

BLACHE P. (2003), "Vers une théorie cognitive de la langue basée sur les contraintes", in actes de *TALN-03*

BLACHE P. (soumis), "Constraints: an operational framework for Construction Grammars"

CROFT W. & D. CRUSE (2003), *Cognitive Linguistics*, Cambridge University Press.

DAHL V. & P. BLACHE (2004), "Directly executable constraint-based grammars", *soumis*

FILLMORE C. & P. KAY (1993), *Construction Grammars* (ms), UC Berkeley

GOLDBERG A. (1995) *Construcions: A Construction grammar approach to argument structure*, University of Chicago Press

GIBSON T. (2000) "Dependency locality theory: a distance-based theory of linguistic complexity", in Marantz & al. (eds), *Image, Language and Brain* , MIT Press.

LANGACKER R. (1999), *Grammar and Conceptualization*, Walter de Gruyter.

MERTENS P. (1993) "Accentuation, intonation et morphosyntaxe", in *Travaux de Linguistique* 26

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

PULLUM G. & B. SCHOLZ (2003), *Model-Theoretic Syntax Foundations – Linguistic Aspects*, ESSLLI lecture notes, Vienna University of Technology.

VASISHTH S. (2003) "Quantifying Processing Difficulty in Human Sentence Parsing", in procedings of *Eurocogsci-2003*