Domaines et propriétés : une description de la répartition de l'information linguistique

Philippe Blache, Christine Meunier

Domaines et propriétés : une description de la répartition de l'information linguistique

Philippe Blache & Christine Meunier

Laboratoire Parole et Langage
CNRS & Université de Provence
29 av. R. Schuman – 13621 Aix-en-Provence Cedex 1
pb@lpl.univ-aix.fr, meunier@lpl.univ-aix.fr

Abstract

('Abstract' style.) Note that all headings (title and section headings) should use Arial or Helvetica font. All remaining parts of the paper should be written with Times or Times New Roman font. All parts of the text (headings and text) should use single spacing. The 'Abstract' section name should use Arial or Helvetica, 12pt, be bold and centered). The 'abstract' text should use Times, 9pt, regular font. The paragraph should be justified, with 0pt before and after paragraph. Text should be indented by 1cm on both sides.

1 Introduction

L'information linguistique est dispersée et variable, souvent incomplète, et quelquefois instable. L'interprétation d'un message linguistique repose donc sur notre capacité à récupérer ces parties d'information et à les mettre en relation. Les approches classiques s'appuient pour cela sur une mise en correspondance structurelle des différents secteurs d'information : par exemple, correspondance entre un arbre syntaxique et un arbre prosodique [Hirst93], entre une règle syntaxique et une formule logique [Montague73], entre les indices phonétiques et les caractéristiques du lexique [Nguyen01], etc. Dans cette perspective, la notion de domaine peut jouer un rôle extrêmement intéressant : elle consiste à identifier un phénomène en le localisant et à établir un certain nombre de propriétés pouvant le caractériser. Il devient du même coup possible de décrire les relations pouvant exister entre différents secteurs linguistiques. Cependant, une telle approche repose sur une vision modulaire de l'information linguistique : l'identification d'une correspondance directe entre des structures relevant de secteurs différents (phonétique, morphologie, syntaxe, sémantique, pragmatique,...) nécessite au préalable l'analyse de chaque secteur et la construction des structures correspondantes indépendamment les unes des autres. Or, si l'information est dispersée, il est nécessaire d'expliquer comment les différents secteurs interagissent.

La notion de domaine joue un rôle central dans cette explication. Nous considérons qu'un domaine est caractérisé par un ensemble de propriétés qui caractérisent un phénomène linguistique particulier. Une tournure interrogative, qui en français mettra en jeu notamment des caractéristiques morphologiques, syntaxique et prosodique constitue de ce point de vue un domaine. Cette interprétation de la notion de domaine se retrouve ainsi totalement dans la notion de construction proposée par les grammaires de construction (cf. Fillmore98 ou Goldberg95 par exemple). Une construction est en effet présentée comme un ensemble de relations existant entre des objets linguistiques (des catégories ou des traits spécifiant ces catégories). Il est ainsi possible de décrire des propriétés relevant simultanément de plusieurs secteurs de l'analyse linguistique (dans le cas des interrogatives la prosodie et la syntaxe en particulier). Une propriété est donc valide à l'intérieur d'un domaine ou d'une construction. Par exemple, le sujet précède en français le verbe. Dans une construction interrogative, il pourra le suivre, mais cette propriété est locale à cette construction. Cette approche permet de décrire de façon très fine les relations existant entre les différents secteurs et le fait que les propriétés sont interdépendantes sans que l'une ne domine l'autre.

Pendant le traitement d'un énoncé, l'analyse de chacun des secteurs fournit donc un ensemble d'informations de base, l'information finale étant construite en fonction des ces éléments de base mis en relation. Nous proposons pour cela une approche d'analyse à plusieurs niveaux intégrant le traitement des relations entre les différents secteurs de l'analyse linguistique. Les exemples suivants illustrent la difficulté à analyser un domaine sans prendre en compte simultanément l'ensemble des secteurs linguistiques.

2 Dispersion de l'information sur plusieurs secteurs

L'information contenue dans un énoncé est typiquement répartie sur l'ensemble des secteurs de l'analyse linguistique. Elle provient à la fois de la morphologie, la sémantique, la syntaxe, la prosodie ou encore la pragmatique. Il n'est pas possible de penser, nous allons l'illustrer dans la section suivante, que chacun de ces secteurs soit analysé séparément et que la construction de l'information globale (l'interprétation d'un énoncé, sa compréhension) soit le résultat d'une composition des informations contenues dans ces différents secteurs. Au contraire, la notion de domaine (ou de construction) permet de montrer comment ces informations provenant des différents secteurs interagissent à un niveau local. Interpréter un énoncé consiste à identifier un domaine et à vérifier les propriétés spécifiques à ce domaine. L'identification d'un domaine consiste alors à rechercher un ensemble de propriétés spécifiques. Cependant, c'est l'interaction de ces propriétés qui permettra la spécification du domaine. En d'autres termes, il n'est souvent pas possible à l'aide d'un seul des secteurs de l'analyse linguistique de caractériser précisément un phénomène. Les exemples suivants illustrent ces cas où la syntaxe à elle seule ne permet pas l'interprétation totale de l'énoncé. Il s'agit du cas classique d'une ambiguïté globale. Certaines approches tentent de désambiguïser l'énoncé par des heuristiques forçant l'interprétation, nous proposons plutôt de tirer parti de toutes les informations disponibles.

(1) *Marie elle lui parle plus*
(2) *Vous avez su pour Monsieur Paul ?*

Dans cet énoncé, la syntaxe n'indique pas de relation particulière entre "*Marie*" et le reste de la phrase. L'identification de la tournure syntaxique peut alors provenir du contexte pragmatique ou prosodique [Blache02]. Dans le cas d'une relation de coréférence entre "*Marie*" et "*elle*", nous sommes en présence d'une construction disloquée. En revanche, l'association de "*Marie*" à un contour continuatif (descendant) indique plutôt une interprétation vocative (on s'adresse à Marie, il n'y a pas de coréférence entre "*Marie*" et "*elle*"). Cet exemple illustre le fait que l'information syntaxique ne soit pas totalement disponible en prenant seulement en compte la suite de mots. L'information nécessaire à l'interprétation est dans cet exemple répartie sur les secteurs prosodiques, syntaxiques et pragmatiques, elle est partielle à l'intérieur de chaque secteur. Le domaine de la dislocation par exemple n'est donc pas directement identifiable à l'aide d'un seul des secteurs.

De même, l'exemple (2) indique un phénomène de construction particulier. L'interprétation de cet énoncé indique qu'il est arrivé quelque chose de grave à l'objet du discours. Mais si le niveau morphosyntaxique permet de donner les grandes relations, notamment en termes d'agent et de patient, l'interprétation ne peut provenir que de la prise en compte globale de la construction mettant en relation les rôles sémantiques du verbe, les fonctions grammaticales et un cadre d'interprétation particulier provenant de la construction du verbe *savoir* avec un SP en *pour* comme complément unique.

3 Contrôle et interprétation de l'information dans un secteur

Le secteur de l'analyse phonétique est probablement l'un des secteurs où l'impossibilité de construire une grammaire autonome et exhaustive est la plus flagrante. Il est désormais trivial de constater que la production de la parole implique une très forte variabilité des unités sonores du langage. Or, une grande partie de cette variabilité ne gêne en rien l'auditeur pour la compréhension des messages linguistiques. Il est désormais admis que les mécanismes de perception de la parole sont grandement déterminés par un

traitement descendant de l'information où les secteurs sémantiques, syntaxiques, pragmatiques, etc., concourrent à l'identification des unités de bas niveau. La caractérisation et le fonctionnement des unités sonores ne permettent donc pas à eux seuls d'expliquer leur identification. Tant que l'on s'intéresse à la relation de "un-à-plusieurs" (production), la variabilité n'est pas un problème fondamental. En revanche, lorsqu'il s'agit de comprendre le processus de perception ("plusieurs-à-un") le problème de la variabilité de la parole ne peut être contourné (Labov, 1986). Les mécanismes de compréhension sont extrêmement complexes et impliquent la présence d'**indices de différents secteurs répartis** dans le message linguistique. Pour rendre explicite le fonctionnement de chacun des indices identifiés, il est de règle de procéder à un "morcellement" de l'information. Nous allons évoquer ici le problème méthodologique que pose ce morcellement.

La phonétique, comme d'autres disciplines, à bénéficié depuis le début du XXème siècle d'avancées techniques (tel le spectrographe) permettant d'observer les caractéristiques physiques des sons de la parole avec une précision de plus en plus grande. La qualité de ces descriptions a parfois laissé penser qu'il suffisait de donner une description physique précise des sons de la parole pour en donner une définition physique. Toutefois, il est rapidement apparu que les observations physiques faites sur un même sons variaient considérablement. Dans cette perspective, et afin de pouvoir donner une description acoustique pertinente, la "variabilité" devait être éliminée. Pour cela, il fallait éliminer les sources de variation. Autrement dit, il fallait contrôler les situations de production de différentes façons: différentiation des locuteurs, contrôle des situations de production, énoncés stéréotypés, mots isolés, voire prononciation de phonèmes isolés. Ainsi, pour décrire physiquement la voyelle /a/, le plus simple est de faire répéter 20 fois cette voyelle par un même locuteur. Dans ce cas, la variation est minime et il est possible d'attribuer à cette voyelle des caractéristiques acoustiques explicites. Toutefois, que représentent ces caractéristiques? Une forme standard, canonique de la voyelle? Une définition physique invariante? Etant donnée que peu de réalisations sont conformes à ces valeurs standard (Meunier & Floccia 1997), comment l'auditeur reconstruit-t-il la forme supposée "sous-jacente" de la voyelle?

3.1 Indices de segmentation lexicale

(2) un corps beau / un corbeau

On sait qu'il n'existe pas en parole de "blancs" entre les mots. Le signal acoustique est continu. Ce constat est l'un des défis les plus importants concernant l'accès au lexique dans les processus de compréhension du langage: il s'agit en effet de rendre compte de notre capacité à segmenter le signal de parole en unités lexicales là où l'information physique disponible est continue. Plus particulièrement, il est très difficile de mettre en évidence des indices physiques robustes et réguliers sur lesquels l'auditeur pourrait s'appuyer systématiquement. Ce problème est d'autant plus crucial dans les mécanismes d'apprentissage des langues (acquisition ou langue seconde) pour lesquels un schéma descendant (top-down) du traitement de l'information est peu envisageable. Quels sont donc les indices qui permettent à un auditeur de segmenter correctement les suites lexicales ambiguës (2) ? Certains travaux ont mis en évidence la présence d'indices segmentaux et supra-segmentaux caractéristiques de la présence d'une frontière de mot [Vaissière92] [Banel97] [Meynadier99]. Pour mettre en évidence ces indices, il est nécessaire de neutraliser l'information répartie en dehors de ce secteur: ainsi, si l'on fait lire (2) à un locuteur, seuls les indices phonétiques permettront de distinguer les deux séquences (les secteurs syntaxiques, sémantiques, pragmatiques, etc., ne sont alors plus informatifs). On crée ainsi, artificiellement, une situation de communication où l'information n'est plus répartie mais totalement contenu dans un seul secteur. Ainsi, contrôler la production entraîne trois effets majeurs:

- réduire l'information
- réduire la variabilité
- isoler les indices qui sont potentiellement utilisés dans ce secteur.

Toutefois, les indices identifiés dans ces conditions sont-ils présents dans toutes les situations de communication? Et sont-ils traités par les auditeurs? Ces indices font partie des propriétés possibles des

sons de la parole mais il n'est pas *nécessaire* qu'ils soient présents si d'autres secteurs sont fortement informatifs. N'est-il pas également imaginable que ce type de manipulation permette de faire ressortir des indices qui ne seraient jamais utiliser en situation de parole naturelle?

Cela ne nous informe pas sur les processus de compréhension d'un message linguistique dans sa globalité.

4 Domaines et propriétés : une architecture à deux niveaux

Un domaine (ou une construction) se définit par la convergence d'un ensemble de propriétés pouvant provenir d'un ou plusieurs secteurs linguistiques. Nous avons indiqué pourquoi l'analyse d'un phénomène linguistique et plus généralement l'interprétation d'un énoncé ne pouvait être le résultat d'une succession de traitements correspondant à une analyse indépendante des différents secteurs de l'analyse linguistique. En d'autres termes, l'hypothèse de la compositionnalité nous semble fausse car reposant sur une conception modulaire de l'organisation de la langue et de son traitement. Il s'agit donc d'expliquer comment construire une interprétation à partir d'informations dispersées, provenant de secteurs différents et pouvant être mis en perspective très rapidement à un niveau local : plutôt que de construire des analyses totales de chacun des secteurs (phonétique, prosodiques ou syntaxique par exemple), il faut progresser étape par étape en identifiant dès que c'est possible des *domaines* qui constituent en quelque sorte des *points de convergence* d'un ensemble de propriétés ou caractéristiques. Une analyse de ce type repose donc sur une conception fondamentalement parallèle des processus d'analyse dans laquelle les mécanismes ascendants ou guidés par les donnés côtoient des contrôles descendants (provenant de l'identification d'un domaine).

Un tel fonctionnement repose sur la possibilité d'une part d'identifier des propriétés de chaque secteur et d'autre part de les analyser régulièrement de façon à rechercher si à un moment donné, un ensemble de propriétés correspond à un domaine ou une construction. Dans le premier cas, ces propriétés proviennent d'une analyse locale de chaque secteur (phonologique, morpho-syntaxique par exemple). Ainsi, chacun des secteurs est décrit par un ensemble de propriétés qui constitue véritablement la grammaire de ce secteur. Pour ce qui concerne la syntaxe par exemple, une étude détaillée d'une approche de ce type est proposée par les Grammaires de Propriétés (cf. [Blache01]). Mais comme cela a été indique pour la phonétique, il n'est pas pertinent, ni même quelquefois possible d'organiser en système indépendant les propriétés décrivant un secteur. Plus précisément, il est nécessaire dans cette approche que toutes les informations soient décrites de façon à pouvoir être évaluées indépendamment les unes des autres. Nous nous situons alors, comme l'indique [Pullum03] dans une représentation non holistique de l'information. En prenant à nouveau le cas de la syntaxe, toute l'information peut être représentée par des propriétés indépendantes les unes des autres. Les Grammaires de Propriétés proposent ainsi les propriétés suivantes:

- linéarité : ordre linéaire entre les objets
- sélection : nécessité de cooccurrence entre deux ou plusieurs objets
- exclusion : impossibilité de cooccurrence entre deux ou plusieurs objets
- unicité : impossibilité de répétition d'un même objet dans une construction
- dépendance : dépendance syntactico-sémantique entre deux objets d'une construction
- obligation : ensemble d'objets indispensables à la réalisation d'une construction

Chaque objet syntaxique peut être décrit par un ensemble de propriétés de ce type. Une grammaire est formée par un ensemble d'objets ou constructions et correspond donc à un ensemble de propriétés (implantés sous forme de contraintes). Chacune de ces propriétés est évaluable individuellement : elle est vérifiée ou pas. Il n'est donc pas nécessaire, à la différence des approches génératives par exemple, de construire une structure syntaxique globale pour pouvoir donner des indications ou des propriétés syntaxiques d'un énoncé.

Ce type d'approche est généralisable aux autres secteurs linguistiques, toute l'information étant décrite par des propriétés indépendantes. L'idée consiste alors à distinguer deux niveaux dans le processus d'analyse. Le premier repose sur l'identification des propriétés fondamentales de l'énoncé sur la base qui vient d'être

décrite. Ces propriétés identifiées, il devient alors possible de caractériser des domaines qui sont décrits comme étant des ensembles de propriétés.

Il est donc nécessaire de décrire tout d'abord ce que nous appelons des propriétés de premier niveau (ou de premier ordre) qui sont des relations entre les catégories d'un secteur : ordre entre parties du discours, dépendance entre mots, relations entre segments, etc. Mais il est également nécessaire de spécifier des relations entre ces propriétés qui vont par exemple permettre de caractériser un domaine. Cela se fait par des propriétés de second ordre qui spécifient des relations entre les propriétés de premier ordre. Elles permettent également d'indiquer qu'un ensemble donné de propriétés de premier ordre est caractéristique d'un domaine.

Chaque secteur est donc décrit par une grammaire formée par :

- des propriétés de premier ordre indiquant les relations entre les objets définis sur ce secteur (par exemple les catégories en syntaxe, les traits en phonologie, etc.)

  Exemple : Det < N : le déterminant précède le nom
  /u/, /o/ : en français, les voyelles d'arrière sont arrondies

- des propriétés de second ordre spécifiant des informations (typiquement contextuelles) pouvant être déduites de la présence de certaines propriétés de premier ordre

  Exemple : $\{Clit_i \wedge le \wedge être_i\} \Rightarrow Refl_i$: si on a un clitique sujet du verbe être et un clitique *le*, alors un réfléchi s'accordant avec le clitique sujet doit être réalisé.

  C [-arrondi] + V[+arrondi] $\Rightarrow$ C [**+arrondi**] + V[+arrondi]: une consonne non arrondie suivie d'une voyelle arrondie se réalise arrondie (ex: [su])

Ces grammaires de secteurs sont complétées par une grammaire d'interaction entre les secteurs formée de propriétés de second ordre spécifiant des relations entre les secteurs.

Exemple : $\{c_1 \rightarrow c_2\} \neq c_1$ *break* c_2

Du point de vue du traitement, à chaque étape, toutes les propriétés de premier ordre (quel que soit leur secteur) sont évaluées. Le résultat est ensuite utilisé pour l'évaluation des propriétés de second ordre des grammaires de chaque secteur ainsi que de la grammaire d'interaction. Tous les secteurs sont ainsi traités simultanément.

L'intérêt fondamental de cette architecture réside dans le fait qu'il devient possible d'introduire des propriétés spécifiques à un domaine. En effet un domaine, s'il repose sur la convergence de propriétés de premier niveau, est également caractérisé par des propriétés très spécifiques, décrivant par exemple des variations de sens comme celle de l'exemple (2). Le processus consiste donc, une fois le domaine identifié, à vérifier ou introduire les propriétés spécifiquement rattachées à ce domaine.

5 Conclusion

La notion de domaine joue un rôle central dans l'explication des relations pouvant exister entre les différents objets linguistiques, mais également entre les différents secteurs d l'analyse linguistique comme la phonologie, la prosodie, la syntaxe ou la pragmatique. Elle permet en effet d'expliquer comment, dans certains cas, un phénomène linguistique peut être caractérisé par un ensemble de propriétés spécifiques : la convergence de certaines propriétés par exemple prosodiques, morphologiques ou syntaxiques, permet de caractériser un domaine comme celui de l'interrogative en français. Il est alors possible d'expliquer comment ces différentes propriétés interagissent au sein d'un même domaine. L'approche décrite ici, mettant en perspective différentes propositions comme les grammaires de constructions ou les grammaires de propriétés, propose un cadre explicatif reposant sur deux niveaux de représentation de l'information : un niveau constitué par les propriétés de base, propres à chacun des secteurs linguistiques et un niveau supérieur permettant de caractériser des domaines et donc d'introduire de nouvelles propriétés.

Références

Banel, M.H. & Bacri, N. (1997) "Reconnaissance de la parole et indices de segmentation métriques et phonotactiques", *L'Année Psychologique*, 97 (1), 77-112.

Blache P. & Di Cristo A. (2002), "Variabilité et dépendance des composants linguistiques", *in actes de TALN-02*

Blache P. (2001) *Les Grammaires de Propriétés*, Hermès Sciences

Goldberg A. (1995), *Constructions*, Chicago University Press

Hirst D. (1993), "Detaching Intonational Phrases from Syntactic Structure", in *Linguistic Inquiry*, 24:4.

Kay P. & Fillmore C. (1999) "Grammatical Constructions and Linguistic Generalizations", *in Language.*

Marandin & Cori

Montague R. (1973) "The Proper Treatment of Quantification in English", in *Approaches to Natural Language*, Hintikka & al. (ed), Dordrecht.

Meynadier Y., Fougeron C., Meunier C. (1999), "Processing of word initial vowels in French: a production-perception perspective", *Proceedings of the 14th International Congress of Phonetic Science*s, 1083-1086. San-Francisco, USA.

Vaissière, J. (1992) "Rhythm , accentuation and final lengthening in French", in Sundberg, Nord & Carlson (Eds) *Music, Language Speech and Brain*, Wenner-Gren International Symposium Series, 59, 108-120.