La linguistique entre science et ingénierie

Gabriel G. Bès

La linguistique entre science et ingénierie

Gabriel G. Bès

GRIL
Université Blaise-Pascal
34, avenue Carnot
F-63037 Clermont-Ferrand cedex
Gabriel.Bes@univ-bpclermont.fr

RÉSUMÉ. *Une grille d'analyse est proposée pour caractériser sur le plan épistémologique des travaux relevant de la formalisation ou du traitement informatique des langues. Elle porte sur les trois éléments-clés d'une science de l'empirie : le domaine des observations, des hypothèses et du test de corroboration. Dans le premier, on se donne comme objet de base des expressions, suites de codes sur support magnétique, et l'Observateur. Les hypothèses sont analysées selon leurs possibilités de calcul, d'extension et de portabilité linguistiques, et selon leurs visées descriptives ou explicatives. Le test de corroboration les confronte avec les observations. Des travaux qui font référence dans l'approche texto-algorithmique et dans l'approche symbolique sont analysés selon cette grille. Des perspectives sont proposées à partir d'éléments d'une évaluation provisoire des deux approches.*

ABSTRACT. *An analytical grid is proposed in order to epistemologically characterize studies concerning the formalization and/or computational processing of natural languages. The grid concernes the three basic elements of an empirical science: the domains of observations, of hypotheses and of the corroboration test. In the first domain, the basic objects are expressions, i.e. sequences of codes on magnetic support, and the Observer. Hypotheses are analyzed in terms of the possibility of calculating on them, and of linguistically extending or porting them. The corroboration test confronts hypotheses and observations. Reference works of the text-algorithmic and of the symbolic approaches in language studies are analyzed with respect to the proposed grid. Some perspectives are proposed from elements of a partial evaluation of the two approaches.*

MOTS-CLÉS : *épistémologie, méthodologie scientifique, linguistique informatique ou computationnelle, TALN, formalisation du langage.*

KEYWORDS: *philosophy of science, methodology of science, computational linguistics, NLP, language formalization.*

1. Introduction et objectifs

Les sciences du langage constituent un domaine à la fois très éclaté et très actif où la communication entre ses composantes est faible. On ne parle pas de *linguistique* mais des *sciences du langage*, ce qui traduit la diversité des points de vue adoptés pour aborder l'objet d'étude. Notre objectif est de proposer une grille d'analyse épistémologique du domaine visé : travaux relevant de la formalisation des langues ou de leur traitement informatique ou des deux.

La caractérisation d'un domaine exige que l'on adopte un ou plusieurs points de vue pour l'examiner. Or ici nous rencontrons à nouveau une diversité de points de vue, selon ce qu'on croit être la méthodologie, l'épistémologie, la philosophie ou l'histoire des sciences. Nous avons emprunté à (Auroux, 1998) des éléments-clés pour construire notre grille. Celle-ci est construite autour des trois domaines de base d'une science de l'empirie : le domaine de l'observation, le domaine de l'hypothèse et le domaine du test de corroboration, qui met en rapport les deux précédents. Or, les notions correspondant à ces trois domaines, trop larges, peuvent être utilisées de manière très différente. La section 2 précise les choix méthodologiques empruntés à (Auroux, 1998) et la section 3 l'utilisation que nous en faisons.

Dans les sections 4 et 5 nous décrirons, en fonction de la grille, des travaux très différents, volontairement choisis en tant que tels. Ils diffèrent par leurs objectifs, par leur conception et par le moment où ils ont été proposés. La section 4 est consacrée à l'approche texto-algorithmique, proche de l'ingénierie, et la section 5 à l'approche que nous disons « symbolique » et qui est censée exprimer une connaissance. Les travaux dits ici symboliques[1] sont parfois reconnus comme étant « théoriques », les références étant la grammaire générative chomskienne, les grammaires d'unification, les grammaires catégorielles et la sémantique vériconditionnelle. Les travaux classés ici dans l'approche texto-algorithmique traitent, avec une certaine efficacité, de données larges. Les mots-clés qui renvoient à ces travaux sont *grammaire de surface, analyse robuste, linguistique de corpus*. L'objectif n'est ni de comparer les travaux observés ni d'exprimer des jugements de valeur sur eux, mais de les appréhender à partir de la grille.

La grille et son application relèvent d'une démarche que l'on peut qualifier de « rationaliste », si on comprend par là que l'on s'intéresse au premier chef à la validation des descriptions et hypothèses. Mais elle est appliquée aux hypothèses ou descriptions effectivement produites, et non, ou non seulement, aux déclarations générales qui les accompagnent. L'application de la grille se limite à établir des constats sur des travaux

1. Dans la tradition chomskienne, on dirait « algébriques », cf. (Miller et Chomsky, 1963) où l'expression est utilisée pour distinguer modèles stochastiques et modèles algébriques. Ce choix terminologique a largement persisté, cf. (Cori et Marandin, 2001). Comme tout choix terminologique, il introduit des pièges conceptuels. La théorie du codage relève des modèles stochastiques, mais elle a aussi des propriétés algébriques, cf. (Chomsky et Miller, 1963). Par ailleurs, le terme est fortement connoté « compétence chomskienne », dont le présent travail entend se démarquer. Nous utilisons ainsi « symbolique », comme dans (Habert *et al.*, 1997). En revanche, « formel », largement répandu, est employé comme dans (Chomsky, 1963).

précis. Cependant, si la grille venait à être considérée comme adéquate, elle pourrait être utilisée comme point de départ de développements relevant de la sociologie ou de l'histoire des sciences ; ces développements éventuels ne sont pas visés dans ce travail. Dans la section 6, nous nous limitons à proposer des perspectives à partir des éléments d'un bilan provisoire.

2. Les choix méthodologiques de base

Pour expliciter les choix à partir desquels on observera des travaux précis, notre référence est (Auroux, 1998), travail qui situe la linguistique par rapport au rationalisme, à l'empirisme et à la normativité. Ce travail dépasse notre objectif, car il est concerné par des questions de métaphysique (Auroux, 1998, p. 3) et, plus généralement, par des questions relevant de la philosophie du langage. Mais il le fait dans le cadre du *réalisme épistémologique*, ce qui a permis de délimiter nos choix.

Le réalisme épistémologique se démarque d'une vision normative de l'épistémologie : « L'épistémologie a perdu trop de temps sur ces questions abstraites de scientificité qui conduisent à déterminer *a priori* ce qui est de la science et ce qui n'en est pas, selon des critères qui ne peuvent être que normatifs. Nous leur préférons un point de vue résolument réaliste » (Auroux, 1998, p. 8). Ce point de vue réaliste s'organise à partir d'un axiome : « les sciences sont des phénomènes sociaux qui existent et peuvent être objet d'analyses très différentes dans leurs points de vue et leurs finalités » (Auroux, 1998, p. 8-9). L'épistémologie doit donc d'abord être descriptive. Le réalisme épistémologique « consiste à partir de la réalité des sciences telles qu'elles se pratiquent et telles qu'elles évoluent dans le temps. Il n'y a aucune raison pour que les objets réels, historiques et culturels que sont les disciplines scientifiques, correspondent à des types idéaux » (Auroux, 1998, p. 137-138).

La description des objets réels se doit d'expliciter la manière de les observer. Les points suivants ont guidé la manière d'observer proposée dans notre grille.

RATIONALISME MINIMUM ET EMPIRICITÉ. Le rationalisme est minimum, car il ne se pose pas le problème de la validité ou de la nécessité ou des fondements de la connaissance scientifique (Auroux, 1998, p. 125). En revanche, il asserte le caractère véridictionnel des hypothèses en linguistique. Auroux (1998, p. 69) affirme ainsi la possibilité de mathématiser et formaliser la théorie linguistique et il pose que toute grammaire portant sur la chaîne parlée « énonce des généralités à valeur prédictive » (*ibid.*, p. 40). Dans ce cadre, toute « assertion grammaticale peut être *discutée et contestée* [italiques dans le texte, cf. (Auroux, 1998, p. 176)], moyennant un *test* [toujours en italiques dans le texte] et tout particulièrement, un test de corroboration » (Auroux, 1998, p. 179-180).

VISION POPPÉRIENNE. Seule la falsification permet de conclure (Auroux, 1998, p. 175) ; un universel peut au plus être corroboré ou falsifié (Auroux, 1998, p. 41). L'utilisation de l'induction est question de stratégie (Auroux, 1998, p. 53) et la déduction ne suppose pas l'adoption d'un rationalisme philosophique (Auroux, 1998, p. 53-54).

DONNÉES. L'objet empirique possède, vis-à-vis du sujet connaissant, le statut de *donnée*. Il doit être accessible par des protocoles définis et renouvelables, et « exister indépendamment du dispositif cognitif » (Auroux, 1998, p. 144-145).

Nos choix de base permettent de justifier la description que nous visons des travaux décrits, description volontairement modeste, de « bas niveau » épistémologique pourrait-on dire, et en tout cas étrangère à tout questionnement philosophique, mais faite à partir de choix clairs[2]. Le réalisme épistémologique invite à décrire des productions attestées. Si l'on accepte les points précédents, on se donne le droit d'interroger ce qui, dans ces productions, est présenté comme hypothèse, ce qui est censé être une donnée et comment on applique le test de corroboration.

Notons que la vision poppérienne, conjointement avec le réalisme minimum, clarifie l'alternative entre induction et déduction. On accepte, bien entendu, que seule la falsification est concluante, de telle manière que l'on s'abstiendra d'admettre qu'une description est « vraie » parce que « fondée » sur beaucoup d'observations, mais pour autant on ne renonce pas à l'induction comme stratégie de recherche.

Notons enfin que le test de corroboration, conjointement avec le statut accordé aux données, écarte le relativisme cognitif, justement critiqué par (Sokal et Bricmont, 1997), cf. en particulier les critiques faites à Feyerabend (*ibid.*, p. 77-89). Mais l'exigence d'objectivité n'exclut pas la possibilité d'opérer avec des observations de type différent, pourvu qu'elles satisfassent les exigences sur les données.

3. La grille d'analyse

Notre grille introduit des précisions sur les domaines de l'observation, de l'hypothèse et de leur mise en relation. En parallèle, il y a trois objets de base : des expressions sur support magnétique associées à des observations, un système d'hypothèses, que nous supposons informatisé ou informatisable, un test de corroboration qui va mettre en rapport les résultats déductibles du système d'hypothèses avec les expressions associées aux observations.

La spécification de la grille reprend l'exigence d'explicitation du domaine de l'observation et du type d'hypothèses qui peuvent être avancées, exprimée dans (Bloomfield, 1957) et dans (Bloch, 1948), travaux qui ont méthodologiquement marqué la linguistique structurale américaine. Mais la grille ne reprend pas le type d'hypothèses et d'observations qui y sont proposées, cf. (Bach, 1965).

---

2. Nous ne prétendons pas épuiser le contenu de (Auroux, 1998). Nos choix portent sur des éléments des chapitres 1 et 2 consacrés respectivement au rationalisme et à l'empirisme, et nous ne prenons pas position sur la normativité de la linguistique (chapitre 3), nous limitant à observer qu'elle ne serait pas un obstacle à son caractère véridictionnel (Auroux, 1998, p. 224). Nous adhérons à bon nombre des analyses faites par (Auroux, 1998) des travaux relevant de l'épistémologie de la linguistique : analyses de (Katz, 1981), de (Katz et Postal, 1991), de (Milner, 1989), de (Itkonen, 1978), du prétendu rationalisme de Chomsky (Auroux, 1998, p. 17, 81) et de la revendication chomskienne de galiléisme (Auroux, 1998, p. 181, 270-273). Sur des points précis, nous indiquerons d'autres convergences et des divergences.

Notons que nous proposons UNE grille pour la description des productions linguistiques dans le domaine visé, mais non LA grille. Elle est conditionnée par les choix méthodologiques de base et par les choix sur les limites et conditions qui seront introduites dans les trois domaines sur lesquels elle porte.

3.1. *Le domaine de l'observation*

On se donne des *expressions* (désormais *expr*), qui sont des suites de codes (par exemple Ascii ou Unicode) sur support magnétique correspondant à des caractères perceptibles par un être humain. Les expressions sont nécessaires, mais elles sont loin d'être suffisantes. On a besoin d'un *Observateur*. Un Observateur (désormais *Ob*) est une classe d'êtres humains doués de capacités d'observation définissables selon des points de vue opérationnels, et capables d'exprimer le résultat de leur activité spécifique dans un langage adéquat[3]. N'importe quel utilisateur natif des expressions d'une langue n'est pas de ce fait un Ob ; on devient Ob lorsqu'on a réussi à bien pratiquer – c'est-à-dire de manière systématique et intersubjective – le point de vue défini pour l'observation que l'on souhaite faire, et seulement ce point de vue-là, et qu'on est capable d'exprimer le résultat de l'observation dans un langage adéquat [4].

Grâce à l'activité de Ob, on obtient la paire <*expr, Obs*>, où *expr* est une suite de codes et *Obs* est le résultat de l'activité de Ob exprimée dans un langage adéquat. Les résultats obtenus par Ob doivent être intersubjectifs : par là, on entend que tous ceux qui relèvent de la classe obtiennent des résultats identiques ou, au moins, avec des divergences qui ne sont pas significatives pour l'utilisation qu'on en fera.

Notre Ob ne doit pas faire plus que ce qui est requis d'habitude pour l'observation des données linguistiques. La différence est qu'il doit le faire de manière explicite et en reconnaissant de manière tout aussi explicite qu'il n'y a pas une seule manière d'observer : ce que l'on obtient comme observation est un objet construit que l'on caractérise par les conditions qui ont présidé à sa construction. Nous soulignons que, dans tous les cas, un Observateur est requis.

La paire <*expr, Obs*> satisfait les conditions sur les données requises par les choix méthodologiques de base (cf. §2). Ce n'est pas le cas des contenus mentaux, pourtant souvent invoqués. Sur ce point, nous adhérons à la position radicale de Kamp et Reyle (1993, p. 10-11) : « *[...] the only access which the theorist seems to have to the language of thought is via the languages we speak. Looking into people's heads [...] is an option that is simply not available.* » L'exclusion des contenus mentaux est justifiable par les exigences sur les données et, tout particulièrement, par l'exigence d'accessibilité à l'objet que l'on étudie dans l'empirie à l'aide de protocoles définis et renouvelables[5].

---

3. Sur la nécessité d'exprimer les observations dans un langage, cf. (Granger, 1992).
4. Nous ne suivons pas entièrement sur ce point (Auroux, 1998, p. 145).
5. Nous acceptons le *hypothesis non fingo* newtonien (cf. (Auroux, 1998, p. 47)) et la remarque de Auroux (1998, p. 158) sur le décalage par rapport aux neurosciences de la grammaire générative ; de (Changeux, 2002) on conclut à l'impossibilité actuelle de cette accessibilité.

Par ailleurs, notons que notre Ob doit porter son attention sur des expressions, et non sur le comportement d'utilisation effective d'une langue[6] : parler, écrire, mémoriser un texte, acquérir la langue... Nous nous abstenons de discuter sous quelles conditions on peut obtenir ce type d'observations plus larges, tout en satisfaisant les exigences sur les données de la section 2.

Si l'on voulait caractériser le domaine observationnel de la grille dans une terminologie chomskienne, on dirait qu'elle n'est concernée ni par la performance ni par l'acquisition du langage, mais on ne dirait pas, pour les raisons évoquées, qu'elle porte sur le contenu mental que les chomskiens nomment « compétence ». (Marandin, 1993) évoque une « vulgate » qui est en fait ce que les chomskiens nomment le « modèle de la performance[7] ». Ce « modèle » pose que le comportement locutoire serait déterminé par des procédures qui s'appliquent à une connaissance (la compétence), procédures effectives et connaissance étant séparées dans le cerveau. Or, le domaine observationnel de notre grille reste en deçà de ces affirmations[8]. Mais nous observons que la séparation des « données » (au sens informatique, c'est-à-dire en tant que source d'informations déclaratives auxquelles accède une machinerie algorithmique) et des algorithmes qui les utilisent n'est pas uniquement motivée par les idées des tenants de la « vulgate » invoquée : nous renvoyons à (Gazdar et Mellish, 1989) pour une justification de cette séparation sur le plan formel, argumentation reprise par (Cori et Marandin, 2001, p. 61-63). Ce qui veut dire qu'une description d'un ensemble de $<expr, Obs>$ sous forme d'une grammaire ou de tout autre dispositif formel n'implique ni n'exclut, selon notre grille, que ce dispositif représente quelque chose qui soit dans la tête de quelqu'un[9].

Puisque, dans le domaine de l'observation, il y a des *expr*, on trouve, à la base, la question des *expr* « bien formées[10] », c'est-à-dire celle de déterminer quelles *expr* il faut observer. Des capacités spécifiques doivent être attribuées à Ob pour ce faire.

---

6. Sur ce point nous nous écartons donc de la discussion de (Auroux, 1998, p. 299-305).
7. (Marandin, 1993, p. 9) pour caractériser la vulgate renvoie au schéma de (Chomsky, 1963).
8. De ce fait, nous ne suivons pas (Marandin, 1993) dans l'idée de considérer le dispositif de l'analyseur comme se substituant à l'empirie et nous ne le suivons pas non plus dans ses affirmations selon lesquelles un analyseur associé à un modèle de grammaire « se trouve, de fait, dans la place et dans le rôle de l'instance de performance postulée dans le cadre de la G[rammaire] G[énérative] T[ransformationnelle] » (*ibid.*, p. 10), ou qu'il est « un avatar de l'appareil de la performance » (*ibid.*, p. 15) dans le même cadre.
9. Dans la tradition post-chomskienne des grammaires d'unification, on observe les deux attitudes. GPSG (Gazdar *et al.*, 1985) se met à l'écart d'une interprétation psychologique (« ... we feel it is possible, and arguably proper, for a linguist (*qua* linguist) to ignore matters of psychology... ») alors que LFG revendique la « réalité psychologique » des descriptions, cf. (Bresnan et Kaplan, 1982). Selon notre grille, on peut décrire une formulation LFG, mais on ne tiendrait pas compte des justifications psychologiques.
10. Nous utilisons *« bien formées »* pour exprimer que la notion se décline de bien de manières différentes ; cf. par exemple *recevabilité* de (Martin, 1978), ou *énoncés, énonçables* de (Desclés, 1978). Dans le cadre chomskien, elle concerne *grammatical* utilisé dans le domaine de l'observation dans (Chomsky, 1957, p. 13) comme synonyme de *acceptable*, alors que dans (Chomsky, 1965, p. 10-11) *grammaticalité* et *acceptabilité* sont distinguées, toujours dans le domaine de l'observation. Dans (Chomsky, 1964, p. 80), *grammatical* et *degrés de grammaticalité* restent dans le domaine de l'observation, alors que dans (Chomsky et Halle, 1968, p. 4) *grammar* est utilisé « *with a systematic ambiguity [...] to refer both to the system of rules in the mind of*

En particulier, il faudra spécifier, d'une manière ou d'une autre, quelles sont les *expr* bien formées. Or, sur cette question de base, deux types de capacités, formellement différentes, peuvent être attribuées à Ob.

Pour situer le problème, on part de l'idée que l'on a V, le vocabulaire d'expressions élémentaires, et que l'on peut obtenir V^+, l'assemblage d'expressions formées à partir de V. On accepte que dans V^+ il y ait des expressions bien formées et d'autres qui ne le soient pas. On suppose enfin qu'il y a des *textes* qui sont des suites d'énoncés bien formés qui respectent des contraintes particulières, permettant de distinguer un *texte* d'une suite d'énoncés quelconques, sans aucun lien entre eux.

Une capacité minimale est accordée à Ob lorsqu'on suppose qu'il sait reconnaître des textes et les différencier des suites d'énoncés qui ne sont pas des textes. On suppose aussi qu'il peut accéder à la source qui produit le texte et qu'il sait caractériser son domaine dénotationnel, c'est-à-dire qu'il sait caractériser de manière très large le type de contenu du texte (par exemple, textes de documentation technique, textes de tel journal, ou textes de tel journal portant sur les informations économiques, etc.). Un Ob ainsi caractérisé ne doit donc pas savoir faire plus, sur cette question de la bonne formation, que ce que sait faire un documentaliste ou un bibliothécaire lorsqu'il classe des documents par leur origine ou par leur contenu [11].

Une autre capacité beaucoup plus forte est attribuée à Ob lorsqu'on spécifie un critère de bonne formation que Ob doit pratiquer. Dans ce cas, Ob est censé utiliser un prédicat unaire `bf` (de bonne formation) avec OUI/NON comme valeurs possibles de son argument, pour classer des expressions comme étant bien ou mal formées. Si l'on soumet à Ob ainsi caractérisé une expression de V^+, il sera en état, en s'exprimant par `bf`, de lui associer soit OUI soit NON. Il peut ainsi définir, en principe, l'ensemble récursif d'expressions bien formées [12].

Si l'on n'attribue à Ob que des capacités analogues à celles d'un documentaliste, sa tâche sera très allégée. Il n'est pas censé utiliser `bf` et il n'est donc pas nécessaire d'expliciter des critères de bonne formation. Soit $C^t = \{txt_1 \ldots text_n\}$, où *txt* abrège

---

the speaker-hearer [...] and to the theory that the linguist constructs as a hypothesis ». On a institué ainsi dans la terminologie de base la non-distinction entre ce qui relève du domaine de l'hypothèse et ce qui relève de l'observation, ce qui conduit à la métonymie, malheureuse selon notre grille, signalée par (Marandin, 1993, p. 12), qui par ailleurs rappelle une autre dichotomie « possible de langue » et « possible événementiel » de (Milner, 1989). Notons que les notions sont présentées à partir de quelques exemples, ce qui mène à une faible stabilité d'utilisation (Chevalier, 1976), la pratique effective ajoutant encore à la confusion. Par exemple (Kamp et Reyle, 1993) utilisent (p. 470) *grammatical* et *ungrammatical* (p. 446) mais aussi *less natural* ou *less felicitous* (*ibid.*, p. 226 note 22) ou *gibberish* (*ibid.*, p. 228).

11. C'est le type d'expertise qui permet, par exemple, de caractériser les différents registres des corpus Brown, LOB et Susanne dans (Habert *et al.*, 1997).

12. Nous disons en principe, car on sait que V^+ n'est pas fini. Mais on peut se donner des limites à la réitération des structures et à l'observation des types de structures, ce qui introduit des simplifications contrôlées sur l'objet de l'étude et le rend accessible en pratique. Le vrai problème pour Ob vient de la note 10. Notons cependant que même les textes du journal *Le Monde* présentent parfois des fautes d'accord, et, de ce fait même, des énoncés mal formés.

texte. On a attribué à Ob la capacité de définir C^t, car il peut déterminer l'appartenance de txt_i dans C^t. Or, dans le cas des langues vivantes, à un moment *t*, on n'aura pas dans C^t tous les textes possibles qui satisfont les critères d'appartenance à C^t. En revanche, dans *t* on aura, grâce à la supposition introduite sur la caractérisation de *texte*, un ensemble d'énoncés bien formés (l'union des énoncés dans les éléments de C^t), sans que Ob ait eu à statuer sur eux, en utilisant `bf`. Mais ce sera un ensemble récursivement énumérable d'énoncés bien formés et non un ensemble récursif. Le dispositif d'utilisation de C^t par un Ob avec des capacités réduites fonctionne comme une machine de Turing : étant donné C^t à un moment *t*, elle s'est arrêtée sur OUI, ce qu'elle fera par supposition sur tous les énoncés de txt_i dans C^t, mais elle ne se sera pas arrêtée sur les énoncés d'un txt_j qui, à un moment *t*, ne serait pas dans C^t. Par conséquent, étant donné V^+, on sera dans l'impossibilité de principe de définir la partition entre expressions bien et mal formées, et on ne pourra pas, ayant cerné dans le moment *t* un sous-ensemble d'expressions bien formées, accéder par l'observation aux énoncés virtuels bien formés et non attestés dans le moment *t*.

On ne peut rien demander à Ob qui n'implique pas d'avoir recours aussi à lui dans l'une ou l'autre des deux perspectives précédentes ou dans un dosage des deux. Si l'on veut associer des expressions à une structure quelconque, même pauvre, alors Ob doit intervenir avec des capacités autres que celles portant sur la détermination de la bonne formation : il doit acquérir le point de vue à pratiquer, apprendre à l'exercer sur les *expr* et notamment à l'exercer sur les *expr* considérées d'une manière ou d'une autre comme bien formées, et apprendre à exprimer les résultats dans un langage adéquat. Ob agit ainsi, dans tous les cas, comme une fonction telle que *Ob(expr)* = <*expr, Obs*>, où *Obs* est exprimé dans un langage adéquat et sa sémantique dépendra du type d'observation que Ob doit pratiquer. On peut imaginer que Ob observe des aspects syntaxiques, phonologiques ou sémantiques des expressions, qu'il se limite à des énoncés ou qu'il observe des relations entre énoncés dans un texte et que, dans ce cas, ces relations soient du type schéma argumentatif ou anaphore textuelle, mais, dans tous les cas, en plus de rendre compte de la bonne formation, il doit acquérir une manière d'observer.

3.2. *Le domaine de l'hypothèse*

Nous caractérisons un système d'hypothèses (désormais *SH*) au moyen d'une sous-grille d'analyse comportant trois éléments : type de formalisme utilisé ; possibilité d'extension et de portabilité linguistiques ; visée descriptive ou explicative. Le point de départ pour caractériser le type de formalisme utilisé est la citation suivante de King (1999) :

> [...] a theory is a collection of hypotheses, and a formalism is a mechanism whereby that theory can be explicitly and unambiguously expressed. For example, Maxwell's equations of electromagnetism constitute a theory. The formalism in which that theory is expressed is vector analysis. Note, vector analysis does not itself constitute a theory of electromagnetism. Indeed, I have seen theories of economics expressed using

> vector analysis. Rather, vector analysis is merely a vehicle with which Maxwell's theory can be expressed.

Un *SH* est composé de formules qui s'expriment dans un langage avec un système de notation comprenant un vocabulaire de symboles utilisés dans le cadre d'une syntaxe explicite. Ces formules, par le biais des *expressions déduites*, vont être mises en rapport avec les objets du domaine de l'observation au moyen du test de corroboration (cf. §3.3 suivant). Nous supposons que notre système d'hypothèses est informatisé ou informatisable, que les expressions déduites sont obtenues par une machine algorithmique et qu'elles ont la forme <*expr, Inf*>, où *expr* est toujours une suite de codes, et que *Inf* dénote les informations associées par la machine à *expr* à partir de *SH*.

Si l'on peut calculer sur la ou les formules en utilisant un formalisme de calcul – dans la citation précédente c'était l'algèbre vectorielle – nous avons une formule dite ici formalisée. Nous avons en revanche une formule non formalisée lorsque nous disposons d'un système de notation avec un vocabulaire et une syntaxe qui permettent de compacter l'information que nous voulons exprimer, mais qui ne permette pas, à partir de plusieurs formules de ce type, de calculer formellement d'autres formules. Une théorie formalisée est un ensemble de formules formalisées dont on peut déduire, grâce au formalisme de calcul, les conséquences ; dans notre cas ce sont les expressions déduites. Le rôle du programme informatique se limite ici à exécuter le calcul, étant donné un ensemble de formules formalisées, pour obtenir les expressions déduites. La machine n'invente rien. En général, une théorie formalisée est complètement indépendante de la machine qui la calcule.

Mais on peut avoir un *SH* – ensemble de formules exprimées dans un langage explicite – qui ne soit pas complètement formalisé et qui, dans un cas limite, ne le soit pas du tout. Les expressions déduites seront d'autant plus sujettes à controverse. Ne disposant pas d'un formalisme de calcul, ou dans la mesure où le formalisme de calcul est défaillant, c'est l'obtention effective des expressions déduites qui, elle-même, va parasiter la spécification de celles-ci.

Un programme informatique est, certes, toujours constitué de formules. Est-ce qu'on peut spécifier ce que fait ce programme et prouver qu'il fait ce que les spécifications exigent qu'il fasse ? Si oui, on dira que le programme calcule effectivement un *SH* formel. Si le programme obtient des résultats associant entrées et sorties, mais que l'on ne peut connaître les résultats qu'il obtient qu'en l'exécutant, ou, dans le meilleur des cas, en examinant son code et en supposant intuitivement ce qu'il devrait produire à partir du code et des commentaires que le programmeur aura bien voulu lui associer, on dira qu'on a un *SH* non formalisé mis en machine. Nous admettons donc que l'on peut avoir un *SH* non formalisé, ou non entièrement formalisé, et qu'il soit mis en machine. Dans ce cas, c'est l'algorithme de calcul effectif des expressions déduites qui va, largement, déterminer celles-ci.

Nous avons reconnu trois objets différents caractérisant un *SH* : les formules exprimées dans la syntaxe d'un formalisme de calcul ; le formalisme de calcul lui-même ; la machine algorithmique qui va spécifier effectivement les paires <*expr, Inf*>. Dans la mesure où le formalisme de calcul se distingue nettement de la machine algorith-

mique, nous dirons que nous avons un *système de calcul formel* ; si le formalisme de calcul et la machine algorithmique fusionnent, nous avons un *système de calcul algorithmique*. Un *SH* et un système de calcul sont des objets différents. Un système de calcul définit la syntaxe des formules d'un *SH* et, en principe, leur sémantique. S'il est formel, la machine algorithmique qui lui est associée permet de calculer effectivement les expressions déduites, mais sans que son intervention n'altère le résultat final ; s'il est algorithmique, c'est le couple <*formalisme de calcul, machine algorithmique*>, fusionné, qui va calculer effectivement les expressions déduites. Nous dirons qu'un système de calcul (formel ou algorithmique) est *sous-jacent* à un *SH* (respectivement formel ou algorithmique).

Comme exemple de calcul algorithmique dans notre domaine, on ne peut que penser aux ATN (*Augmented Transitions Networks*) : écrire un ATN est quasiment écrire un programme, un ATN étant un formalisme procédural, qui favorise une exploration de l'entrée de gauche à droite et une stratégie du haut vers le bas, cf. (Gazdar et Mellish, 1989, p. 96). Avec un ATN on peut dire beaucoup, mais on ne sait pas si ce qui est dit vient de la caractérisation que l'on veut faire des expressions observées, ou bien de la manière de calculer les expressions déduites. Les systèmes non formels d'hypothèses ont le gros inconvénient de nous empêcher de comprendre notre objet d'étude ; ils ne sont pas perspicaces, cf. (Gazdar et Mellish, 1989, p. 99).

C'est pourquoi on a voulu des grammaires (dites parfois particulières) en tant que *SH* formels, et des modèles de grammaire, en tant que calculs formels sous-jacents aux premières. On a voulu obtenir des *SH* déclaratifs où les formules exprimant les descriptions linguistiques ne devraient, tout au moins en principe, contenir aucune exigence procédurale. Ces grammaires devraient pouvoir être associées à une machine algorithmique qui va obtenir les paires <*expr, Inf*> des expressions déduites.

Pour caractériser l'éventail de nos *SH* par rapport au type de formalisme utilisé, nous pouvons donc fixer deux pôles avec des situations intermédiaires possibles. L'un est celui des théories formalisées, qui ne doivent rien dans la spécification des expressions déduites à l'outil informatique qui les obtient effectivement. L'autre, dans lequel la spécification des expressions déduites résulte de deux facteurs : les formules en tant qu'hypothèses visant à caractériser le domaine observationnel *via* les expressions déduites, d'une part ; les opérations effectives permettant d'obtenir les expressions déduites, d'autre part, les deux facteurs étant tellement imbriqués qu'il est impossible de distinguer ce qui vient de chacun d'eux. Les objets du premier pôle sont des *SH formels* avec des calculs formels sous-jacents ; les objets du deuxième pôle sont des *SH algorithmiques* avec des calculs algorithmiques sous-jacents. Dans les deux cas, nous supposons qu'il s'agit de *SH* informatisés ou informatisables. Dans les deux cas, un *SH* se comporte comme une fonction telle que *SH(expr)* = <*expr, Inf*>.

Le deuxième élément de notre sous-grille d'analyse des *SH* – extension et portabilité linguistiques – est relié aux deux constats suivants : il n'y a pas une seule langue naturelle, mais une pluralité ; dans chaque langue naturelle, il y a une gamme de structures de complexité différente à décrire et à exprimer dans un *SH*.

La notion d'extension linguistique peut se résumer ainsi. Soit un *SH* et son calcul sous-jacent, le *SH* spécifiant les *Inf* associées à un sous-ensemble d'expressions. Est-ce que, en conservant le même calcul sous-jacent, on peut étendre *SH* dans *SH'* de telle manière que les expressions auxquelles *SH* associe des informations soit un sous-ensemble d'expressions du même type que celles décrites par *SH'* ? Si c'est possible et que les modifications à introduire consistent à ajouter des formules sans altérer ou très peu les formules de *SH*, on dira que *SH* et son calcul sous-jacent satisfont au critère d'extensionnalité linguistique. Supposons par exemple une grammaire qui peut donner une description adéquate à *Jean a regardé la lune et a photographié l'éclipse*. Si nous pouvons en obtenir une autre grammaire, satisfaisant au même calcul sous-jacent que la première, avec laquelle on peut décrire aussi *Jean a regardé la lune et photographié l'éclipse*, la première et son calcul sous-jacent satisfont à l'extensionnalité linguistique.

La notion de portabilité linguistique concerne la possibilité d'utiliser un même calcul, soit formel soit algorithmique, pour traiter des langues différentes ou des aspects différents d'une même langue. Les arguments pour justifier les premières formulations de la grammaire générative transformationnelle étaient largement fondés sur la non-portabilité des grammaires hors contexte pour traiter une gamme de constructions.

Le troisième élément de la sous-grille d'analyse du domaine de l'hypothèse prétend discriminer entre des *SH* à visée descriptive et des *SH* à visée explicative.

Étant donné une *expr*, on aura <*expr, Obs*> obtenu par Ob, et <*expr, Inf*> obtenu par *SH*. Supposons un ensemble d'expressions *E* telles que, pour toute *expr* dans *E*, Ob lui associe *Obs* et *SH* lui associe *Inf*, *Inf* étant considérée adéquate par rapport à *Obs* (cf. §3.3). On note E_{Obs} l'ensemble obtenu par Ob, et E_{Inf} l'ensemble obtenu par *SH*. De là on peut obtenir la notion de *SH justifié par rapport à un ensemble d'expressions* et, à partir de celle-ci, la notion de *SH projeté par rapport à un autre SH'*.

Soient E_{Obs} obtenu par Ob et E_{Inf} obtenu par *SH*. Un *SH* est un ensemble de formules. *SH est justifié par rapport à* E_{Obs} si toute formule de *SH* est utilisée dans la spécification d'au moins une paire <*expr, Inf*> dans E_{Inf}. *SH' (SH'* $\neq$ *SH) est projeté par rapport à SH*, s'il existe un principe *P* permettant d'obtenir *SH'* à partir de *SH*, et ce de telle manière que *SH'* reste adéquat par rapport à E_{Obs} et, aussi, par rapport à d'autres <*expr, Obs*> qui ne sont pas dans E_{Obs} ; par ailleurs *SH'* n'est pas justifié par rapport à E_{Obs}, et *SH* et *SH'* satisfont le même calcul sous-jacent.

Comme *SH'* n'est pas justifié par rapport à E_{Obs}, toute formule de *SH'* n'est pas utilisée pour obtenir une paire <*expr, Inf*> de E_{Inf}, mais *SH'* « peut faire » tout ce que *SH* fait et encore plus. Les formules de *SH'* qui ne sont pas justifiées par rapport à E_{Obs} sont exigées non par l'extension des *expr* observées par Ob mais par le principe *P*. Un *SH'* projeté est descriptivement plus riche que le *SH* correspondant dans ce sens qu'il va décrire de manière adéquate des *expr* qui n'étaient pas parmi celles primitivement observées pour décrire *SH*. En d'autres termes, un *SH* projeté permet de décrire ce qui n'est pas encore observé. Reprenons l'exemple précédent de la coordination. Si on disposait d'un principe *P* tel que, appliqué à un *SH* permettant de décrire *Jean a regardé la lune et a photographié l'éclipse*, on pouvait obtenir *SH'* et avec *SH'* décrire

Jean a regardé la lune et photographié l'éclipse, *SH'* serait projeté. Ce sont les *SH'* projetés qui introduisent la notion de *prédiction explicative* [13]. Notons que si *SH'* est projeté à partir de *SH*, l'extensionnalité linguistique de *SH* est, par là-même, satisfaite, mais le contraire n'est pas vrai. Un *SH'* est projeté en fonction d'un principe *P*, dont la formulation est conditionnée par le calcul (formel ou algorithmique) sous-jacent à l'argument de *P*. C'est donc *P*, conjointement avec le calcul sous-jacent à *SH*, et *SH'* qui explicitent la notion de prédiction explicative.

Remarquons qu'aucun positionnement par rapport à l'un des trois éléments de la grille n'exclut un positionnement par rapport aux deux autres. Ainsi, par exemple, un *SH* algorithmique peut ou non être portable, tout comme un *SH* formel.

3.3. *Le domaine du test de corroboration*

Nous avons supposé que Ob et *SH* agissent comme des fonctions qui prennent comme argument une *expr* de manière à spécifier E_{Obs} et E_{Inf}. Le test de corroboration peut s'exprimer par une autre fonction, `test` avec deux arguments :

$$test(< expr_i, Obs >, < expr_j, Inf >) = résultat$$

où <*expr*$_i$ = *expr*$_j$> et où $résultat$ peut prendre, dans un cas idéal, deux valeurs `adéquat`, `inadéquat`. Spécifier `test` suppose spécifier selon quelles conditions sur Ob et sur *SH* on aura une évaluation déterminée pour $résultat$. De manière plus réaliste, il faut se préparer aussi à avoir des situations scalaires entre `adéquat` et `inadéquat`, et prévoir des valeurs du type `indéterminé`. Mais de toute manière, pour évaluer `test`, il faut se donner des critères explicites pour comparer *Obs* et *Inf*, chacun de ces objets étant exprimé, rappelons-le, dans un langage adéquat.

Selon cette manière de voir les choses, il n'existe pas de « continuité » entre <*expr*$_i$, *Obs*> et <*expr*$_j$, *Inf*> [14]. Ces couples sont des objets différents au niveau de *Obs* et de *Inf*, bien que l'on pose l'identité <*expr*$_i$ = *expr*$_j$>. Celle-ci est possible grâce aux choix effectués sur le domaine observationnel : il s'agit d'identité au niveau des codes sur support magnétique. Si l'on travaillait en reconnaissance de la parole, la simplification adoptée serait strictement impossible. `test` met en rapport un *SH*, *via* les expressions déduites, avec E_{Obs}. De manière indirecte, le calcul sous-jacent à *SH* est ainsi, lui aussi, mis en rapport avec E_{Obs}. Et cela, que les *SH* et leurs calculs sous-jacents soient formels ou algorithmiques.

Les résultats de `test` doivent être interprétés selon les choix méthodologiques de la section 2. Selon notre grille, tout ce qu'il y a dans `test` est nécessaire pour corroborer, mais rien d'autre que `test` n'intervient dans l'obtention des résultats de la corroboration. Avec cette affirmation, nous écartons explicitement de la grille tout critère platonicien qui poserait la question de la naturalité (légitimité, etc.) des « notions linguistiques » utilisées dans la formulation des *SH*. Par ailleurs c'est le *SH* dans

13. Cf. les antécédents de cette notion dans (Chomsky, 1964) et (Peters, 1972).
14. Cf. une opinion différente dans (Auroux, 1998, p. 62-63).

son ensemble que `test` met en rapport avec <*expr*$_i$, *Obs*>, même s'il le fait *via* les expressions déduites ; nous suivons en cela Bunge (1969, p. 420-421 et p. 304-307). Nous écartons aussi de la grille tout appel à la « naturalité », « élégance » ou « commodité » de la syntaxe utilisée pour écrire les formules d'un *SH*, et tout critère concernant l'efficacité du traitement informatique.

4. L'approche texto-algorithmique

L'approche texto-algorithmique (cf. dans la section 1 les mots-clés qui la caractérisent) a fait l'objet de peu de discussions méthodologiques approfondies. Pourtant, la simple observation des publications de référence (cf. les conférences COLING) et le réalisme épistémologique obligent à s'interroger sur cette manière de voir les choses. À partir de la grille, nous proposons, dans le paragraphe 4.1, une caractérisation générale du domaine dans l'optique « fondée sur des règles » [15] ; dans le paragraphe 4.2 nous analysons des travaux spécifiques.

4.1. *Caractérisation générale*

Dans le domaine de l'observation de l'approche texto-algorithmique, Ob définit des classes de textes C^t en fonction des sources qui les ont produits et qui vont continuer à les produire (par exemple tel journal) et de leur domaine dénotationnel caractérisé en termes très larges (par exemple textes portant sur la politique étrangère). Les observations requises dans *Obs* de <*expr, Obs*> portent sur la segmentation, sur la morphologie et, en syntaxe, sur des descriptions parcellaires (les « chunks » ou syntagmes noyau). L'observation des dépendances entre constituants de l'énoncé en est à ses débuts et la sémantique vericonditionnelle strictement absente : Ob a ainsi rarement l'occasion d'accéder à la dénotation des expressions, à quelques exceptions près comme l'anaphore.

Dans le domaine de l'hypothèse, les *SH* sont des analyseurs non réversibles pour la génération. La description linguistique qui y est utilisée est partiellement fusionnée avec la machine algorithmique, et cela à un niveau qui n'est pas toujours facile à discriminer. En revanche, des calculs sous-jacents partiellement algorithmiques accèdent à la portabilité linguistique, et on exige la robustesse des *SH* : à tout énoncé dans son entrée, un *SH* doit associer une structure de sortie.

Les tests de corroboration sont systématiquement utilisés et ils ont été institués : cf. les conférences MUCS et TREC et, en France, GRACE et SENSEVAL. L'interprétation des résultats des tests a requis l'utilisation de concepts spécifiques, comme ceux de couverture et de précision.

La statistique est utilisée pour détecter les conditions d'utilisation les plus fréquentes d'une structure S dans un corpus d'étude (t_i dans C_t) et pour guider la des-

15. On ne traite donc pas l'approche statistique fondée sur les chaînes de Markov.

cription du *SH* censé rendre compte de tout t dans C_t. Les résultats obtenus montrent que, effectivement, à condition de se donner des C^t homogènes – et sur ce point on peut évaluer la « perspicacité » de Ob[16] –, la monotonicité des résultats statistiques est importante. Et elle aide à obtenir des produits performants.

L'approche texto-algorithmique, avec une expérience enrichie par un presque demi-siècle qui a vu passer, parmi d'autres, la grammaire générative chomskienne, les grammaires à traits et la sémantique vériconditionnelle, est revenue, pour traiter le langage humain, à la situation d'avant 1957, d'où Chomsky était venu la sortir. L'observation dans l'approche texto-algorithmique porte sur ce qu'il y a dans les énoncés de plus immédiatement accessible, la sémantique n'étant pas ou à peine abordée. L'utilisation des corpus d'études et l'induction de descriptions à partir d'eux revient à ce que Chomsky avait déclaré impossible à faire.

4.2. *Analyse de travaux de l'approche texto-algorithmique*

Une forte opacité dans les trois domaines de la grille caractérise les travaux de l'approche texto-algorithmique. Nous ciblons l'analyse sur deux travaux – (Karlsson, 1990) et (Aït-Mokhtar *et al.*, 2002) – produits à une douzaine d'années d'intervalle.

Dans le domaine de l'observation, au-delà de l'utilisation de C^t, (textes anglais dans (Karlsson, 1990) et français dans (Aït-Mokhtar *et al.*, 2002)), le rôle d'Ob pour exprimer *Obs* ne peut être inféré qu'à partir de la présentation des résultats quantitatifs du test de corroboration, non explicité en tant que tel. On a ainsi une idée du type de relation syntaxique que chaque sytème a voulu faire observer par Ob, mais il n'est pas possible de se comporter en Ob pour produire <*expr, Obs*> sur les énoncés d'un txt de C^t : ni C^t ni la manière attendue d'Ob pour observer les expressions de txt dans C^t ne sont suffisamment caractérisés.

Dans le domaine de l'hypothèse, (Karlsson, 1990) et (Aït-Mokhtar *et al.*, 2002) ont implanté un système portable de calcul algorithmique d'analyse et défini (ou esquissé) la syntaxe des règles dans lesquelles l'information linguistique pertinente est exprimée. Dans les deux travaux, on utilise des modules ordonnés, dont les fonctionnalités se recoupent largement. Ces modules correspondent aux étapes d'une analyse du bas vers le haut : (1) segmentation et analyse morphologique, (2) désambiguïsation des catégories morphologiques, (3) analyse syntaxique locale, (4) analyse de relations syntaxiques fonctionnelles. (Aït-Mokhtar *et al.*, 2002) peut être compris comme l'option symétrique à (Karlsson, 1990), dans ce sens que le système implanté n'élimine pas des possibilités mais construit le résultat final au fur et à mesure de l'analyse.

La syntaxe des règles est assez clairement explicitée dans (Karlsson, 1990). Un même type de règles – appelées « contraintes » – est utilisé pour la désambiguïsation morphologique contextuelle, pour l'obtention des limites des phrases et pour l'élimination des relations syntaxiques potentielles assignées aux expressions d'une catégorie

16. Les textes choisis suivront des traditions rhétoriques stables.

déterminée par un autre type de règles (les règles d'association morphosyntaxique). L'ordre d'explicitation des règles est pertinent pour les règles d'association morphosyntaxique et nous conjecturons qu'il l'est aussi pour les contraintes. On ne connaît ni l'algorithme utilisé ni l'information linguistique exprimée dans les règles.

La syntaxe des règles pour l'analyse syntaxique locale (les « *chunking rules* ») n'est pas donnée dans (Aït-Mokhtar *et al.*, 2002). On sait qu'elles s'organisent en couches (« *layers* »), que les couches sont ordonnées et que les règles sont ordonnées dans chaque couche. Deux types de règles sont évoquées : les règles ID/LP de GPSG et les « *sequence rules* », dont la différence expressive par rapport aux « *LP statements* » de GPSG n'est pas explicitée. Les « *chunking rules* » ont « *less generative power than CFG rules* » car elles ne sont pas récursives, mais en même temps elles « *may include constraints on the context where they apply* », ce qui implique qu'elles ont plus de pouvoir expressif que les règles hors contexte. La syntaxe des règles de dépendance, conceptuellement analogues aux contraintes de (Karlsson, 1990), est mieux définie, sans qu'il soit possible de connaître leurs limites expressives.

Avec le peu d'informations fournies dans les travaux analysés, nous ne pouvons faire que des conjectures sur les *SH* utilisés. Ceux-ci sont des programmes implantés en machine avec la capacité expressive d'une grammaire de type 0. Mais ils doivent avoir été construits selon des contraintes d'un calcul sous-jacent algorithmique. Ce sont des machines effectives, qui bénéficient donc de l'expertise informatique de ceux qui les ont implantées. Le calcul de leurs extensions linguistiques paraît impossible, tout comme la possibilité de leur associer des principes explicatifs.

L'ordre de déclaration des règles, des modules et des couches est *pertinent* : un ordre différent entraînerait des résultats différents au niveau des couples <*expr, Inf*>, ce qui donne à la formulation des règles une « coloration algorithmique ». L'ordre pertinent exprime l'incrémentalité invoquée dans (Aït-Mokhtar *et al.*, 2002), qui est, de plus, considérée comme la voie d'accès à la robustesse de l'analyse. La question que l'on peut se poser est : est-ce que l'ordre pertinent requis dans la formulation des règles n'est qu'une incommodité pour celui qui les formule, auquel cas notre grille ne s'en préoccupe pas, ou bien l'ordre pertinent est-il constitutif de l'analyse robuste, c'est-à-dire qu'il serait une caractéristique dont celle-ci ne pourrait pas se passer [17] ?

Les différents types d'opacité des deux travaux analysés dans les domaines de l'observation et de l'hypothèse excluent toute possibilité d'explicitation du test de corroboration. Les deux travaux ne présentent que des résultats globaux. On ne saura pas ainsi quelles motivations, soit linguistiques soit algorithmiques, ont conduit au pourcentage, faible, d'erreurs. Par exemple : est-ce qu'il n'y a pas eu de problème avec les sujets inversés parce que le *SH* a su les traiter ou parce qu'il n'y avait pas de sujets inversés dans les textes traités ? Cette impossibilité se conjugue avec une autre. On peut conjecturer que les *SH* analysés, qui doivent satisfaire à l'exigence de robustesse, acceptent des énoncés mal formés, mais les limites dans lesquelles ils le font sont impossibles à évaluer.

---

17. Nous conjecturons que, si l'on ne tient pas compte de l'efficacité de calcul, on devrait pouvoir accéder à la robustesse de l'analyse sans utiliser nécessairement un ordre pertinent, mais en exploitant au maximum les conditions dans la formulation des règles.

5. L'approche symbolique

La grille d'analyse portera sur la *Grammaire de Montague*, limitant strictement la dénotation de cette expression à UG – *Universal Grammar* – (Montague, 1974, p. 222-246) et à PTQ – *The Proper Treatment of Quantification in Ordinary English* – (Montague, 1974, p. 247-270) [18]. La Grammaire de Montague est une contribution qui a été à la base de l'émergence de l'approche symbolique. Elle se situe intellectuellement aux antipodes des travaux issus de l'approche texto-algorithmique. Elle a enrichi la connaissance des langues naturelles et a ouvert des perspectives, tout en portant en elle les limites de son développement et du développement de l'approche symbolique. Elle sera analysée avec un certain détail mais non exhaustivement. Notre analyse portera sur la syntaxe des expressions et sa traduction en formules de la logique intensionnelle (désormais LI), spécifiée par Montague, et qui, en tant que telle, reste en dehors de notre analyse et sur les effets de l'interprétation de ces formules dans l'univers d'interprétation (désormais *U*).

Dans les deux travaux, le domaine de l'observation est peu ou mal explicité. Nous nous limitons à remarquer que, sur le plan syntaxique, l'observation exclut des constituants au pluriel (ce qui simplifie le type de quantificateurs requis en sémantique), et, sur le plan de la dénotation, le rôle dévolu à l'Ob dépasse de loin les capacités d'un être humain. En effet, les observations portent sur l'association d'une expression à un univers U^i, étant donné $U^1 \ldots U^n$. Elles s'expriment par une valeur (*Va*) de vérité : *V* ou *F*. Pour une même expression, Ob est susceptible de formuler un ensemble infini d'observations pour couvrir tous les univers possibles. Le résultat de l'observation de *expr* par Ob sera ainsi noté $<expr, <U^i, Va>>_{Obs}$. Dans ce cadre, les fonctions utilisées pour caractériser l'univers de dénotation ne peuvent pas être effectivement évaluées. Cette « abstraction » par rapport à l'interprétation du formalisme dans un réel effectif relativise fortement l'argumentation proposée pour justifier ce type de sémantique [19].

5.1. *L'analyse de UG*

L'idée suivante, que l'on appellera *l'idée de la non-différence*, est l'élément de base du projet sémiotique de Montague :

> There is in my opinion no important theoretical difference between natural languages and the artificial languages of logicians; indeed, I consider it possible to comprehend the syntax and the semantics of both kinds of languages within a single natural and mathematically precise theory (Montague, 1974, UG p. 222).

Le problème que Montague se donne est celui d'associer une expression à sa dénotation pour lui associer *Va*. Cette idée peut – mais ne doit pas nécessairement – être comprise comme une hypothèse empirique sur les langues naturelles. Chambreuil semble la considérer comme telle, cf. (Chambreuil, 1998). Pour notre part,

18. Nous ne suivons ni l'injonction de (Partee et Hendriks, 1997) en sens contraire ni une pratique courante : le concept élargi dans des limites incontrôlables est impossible à évaluer.
19. Argument de Davidson repris par (Partee et Hendriks, 1997) et (Kamp et Reyle, 1993).

nous croyons que, telle qu'elle se présente instanciée dans la Grammaire de Montague, elle ne peut être utilisée que comme un principe méthodologique. L'idée de la non-différence, résumée à partir de UG, se concrétise ainsi : le calcul des conditions et des valeurs de vérité des langues naturelles (désormais *expr-LN* et *LN*) dans *U* peut se passer des représentations sémantiques intermédiaires (désormais *rsém*), tout comme on donne les valeurs de vérité aux expressions d'un langage logique [20].

Selon cette manière de voir les choses, il n'y a plus de signe saussurien, avec un signifiant associé à un signifié, il n'y a plus de forme logique chomskienne et les DRS à la Kamp ne sont pas nécessaires. Signifié, forme logique et DRS disparaissent en tant qu'objets nécessaires à l'association d'une expression d'une langue naturelle à sa dénotation. LI est un intermédiaire utile mais dont on peut se passer.

L'interprétation induite est le mécanisme pour associer *expr-LN* à *U*, en utilisant les formules de LI par commodité de calcul. Elle résulte basiquement de la composition de deux homomorphismes algébriques : l'un, ici noté ht (homomorphisme de traduction), et l'autre, ici noté ha (homomorphisme d'assignation de signification). On doit supposer une relation R permettant d'obtenir, à partir de l'ensemble d'expressions ambiguës d'une langue naturelle (*LNa*), l'ensemble *LN* ¬ *a* (ensemble d'expressions non ambiguës). On suppose enfin trois algèbres : *Alg-LN* ¬ *a*, permettant de décrire syntaxiquement *LN* ¬ *a* ; *Alg-LI*, décrivant syntaxiquement les expressions de *LI* ; Ad^{LI} (algèbre dérivée de *LI*), telle que *Op(Alg-LI)* = Ad^{LI}, *Op* étant un ensemble d'opérations polynomiales sur *Alg-LI*. Selon les conjectures formelles sous la forme de « Remarks » de UG, on a *ht(Alg-LN*¬ *a)* = AD^{LI} et *ha*(Ad^{LI} = *A'*, *A'* étant l'interprétation dans *U* de Ad^{LI} et, grâce à *ht*, de *Alg-LN* ¬ *a*.

L'interprétation induite avec ses composantes (*R, Alg-LN* ¬ *a, Alg-LI,* AD^{LI}, *ht, ha*) est le cadre général sous-jacent à des *SH*, *Alg-LN* ¬ *a* étant le calcul formel sous-jacent aux *SH*. *Alg-LN* ¬ *a* serait ainsi le « véhicule d'expression » formel de l'expression de la syntaxe (cf. §3.2). L'instanciation du cadre suppose que *Alg-LN* ¬ *a* soit spécifiée, que *R* soit spécifié et que les contraintes de la définition de *LN* ¬ *a* soient respectées. Si l'on avait un cadre ainsi instancié, on obtiendrait l'association d'une expression à *U* et à *Va* (grâce, en dernière instance, à *ha* qui va associer les formules de LI à *U*, les formules de LI ayant été obtenues par *ht*). On obtiendrait donc une expression déduite que nous notons <*expr,* <U^i, *Va*>>$_{Inf}$ et le test de corroboration pourrait fonctionner normalement :

`test(<expr,<U`^{i}`,Va`^{i}`>>`$_{Obs}$`, <expr,<U`^{i}`,Va`^{j}`>>`$_{Inf}$`) = adéquat` ssi $\mathrm{Va}_i = \mathrm{Va}_j$

L'idée de la non-différence, exprimée par l'interprétation induite et ainsi interprétée comme hypothèse empirique, serait corroborée dans la mesure où `test` donnerait `adéquat` dans un domaine d'observation bien défini. Au-delà du fait que, un tiers de siècle après avoir été envisagées, ni la preuve formelle de l'interprétation induite n'a été faite ni la « *single natural and mathematically precise theory* », qui devrait

20. Notons qu'il s'agit d'*expressions* au sens strict : à l'instar des objets de la logique, des objets graphiques nus, non associés à une structure de traits quelconque.

comprendre syntaxe et sémantique des langages des logiciens et des langues naturelles, n'a été formulée, deux raisons majeures empêchent de considérer l'idée de la non-différence comme une hypothèse empirique.

La première est qu'il n'y a pas de restrictions dans UG ni sur *R*, ni sur les opérations admissibles dans *Alg-LN* $\neg$ *a* ; de ce fait, celle-ci peut être justifiée sans restriction aucune, y compris sans l'exigence de sa calculabilité. De ce fait, une syntaxe pour *SH* supportant *ht* ne peut pas ne pas être trouvée, ce qui est strictement contraire à toute idée de corroboration d'une science de l'empirie. Elle devient ainsi une exigence méthodologique, cf. (Janssen, 1997, p. 419, 457) ; elle vaut pour ceux qui croient encore qu'il y a de bonnes recettes pour obtenir de bonnes hypothèses empiriques [21].

La deuxième raison pour ne pas considérer l'idée de la non-différence comme hypothèse empirique est qu'elle est, sur le plan formel, handicapée par le fait qu'une expression élémentaire dans *Alg-LN* $\neg$ *a* (c'est-à-dire une *expr* ne résultant pas de l'application d'une opération) peut appartenir, dans un langage non ambigu (tel que celui qui est défini dans UG, cf. (Chambreuil *et al.*, 1998)), à des ensembles indexés par des catégories différentes. Par exemple l'expression *contre* peut, dans un langage non ambigu tel que défini dans UG, appartenir à trois ensembles d'expressions différemment indexés (comme verbe, préposition et adverbe). Or *ht*, qui est une fonction, a comme domaine l'union des ensembles des expressions bien formées ; appliquée à une expression, elle ne peut pas obtenir dans son évaluation plus d'une entité. Ainsi `ht(contre)=résultat` aura en `résultat` sa traduction soit comme verbe, soit comme préposition, soit comme adverbe, ce qui, d'une part, fait perdre l'ambiguïté et, d'autre part, donne de mauvais résultats dans l'interprétation ultérieure par *ha* [22].

5.2. *L'analyse de PTQ*

Montague dans PTQ ne spécifie pas les expression d'un *LN* $\neg$ *a* de l'anglais à partir de la définition donnée dans UG et ne définit pas une Base de traduction (c'est-à-dire que *ht* du paragraphe précédent n'est pas défini comme dans UG). En revanche, PTQ introduit une définition récursive des indices catégoriels en syntaxe et une fonction de traduction de ces indices sur les types de *LI*, absentes de UG.

Le pouvoir expressif de la syntaxe de PTQ n'a pas été caractérisé. PTQ utilise des *opérations* qui sont associées à des *règles*. On répertorie 16 opérations (notées F_0 à F_{15}) et 24 règles (notées V_0 à V_{23}). Chaque opération est associée à une ou plusieurs règles, ce qui donne 24 couples <*règle, opération*>. Par exemple, l'opération F_6 concatène deux expressions : $F_6(\alpha, \beta) = \alpha\beta$. Elle est associée à quatre règles dans quatre couples différents. Chaque règle va spécifier la catégorie de α, de β et de l'expression résultante $\alpha\beta$. Ainsi le couple <V_6, F_6> dit que si α est une préposition

21. Ce qui est le cas de (Janssen, 1997) et de (Nazarenko, 1998).
22. Nous ne croyons pas que la solution formelle (adoptée dans (Janssen, 1997) et dans (Partee et Hendriks, 1997)), consistant à considérer que `résultat` est un ensemble avec trois traductions différentes, soit la bonne : on ne fait que transférer le problème à *ha*.

(indice catégoriel *IAV/T*) et β un syntagme nominal (indice catégoriel T), alors $\alpha\beta$ sera un adverbe (indice catégoriel *IV/IV*). L'opération F_7 définit une opération de permutation : $F_7(\alpha, \beta) = \beta\alpha$. Elle est utilisée dans le seul couple <V_{10}, F_7> pour permuter un verbe intransitif précédé d'un adverbe et obtenir un verbe intransitif suivi d'un adverbe (on passe ainsi de *slowly runs* à *runs slowly*).

Le couple <V_6, F_6> peut être exprimé par une grammaire 2 dans la hiérarchie chomskienne ou par l'application fonctionnelle des grammaires catégorielles, ce qui n'est pas le cas du couple <V_{10}, F_7>. Cet exemple montre déjà que l'affirmation selon laquelle Montague utiliserait en PTQ une grammaire catégorielle est fausse. L'analyse détaillée des 24 couples de PTQ (cf. (Bès, 2001, §4.1.3)) montre que certains couples ont un pouvoir expressif supérieur aux grammaires 2 (c'est le cas de <V_{10}, F_7>). D'autres correspondent au type 2, mais ils ne peuvent pas être exprimés par l'application fonctionnelle opérant sur les indices catégoriels des expressions élémentaires ou obtenus par application fonctionnelle sur ceux-ci (ce sont les couples sur la coordination, car il n'y a pas d'expressions élémentaires d'indice t et les expressions de cet indice ne s'obtiennent pas toutes par application fonctionnelle). D'autres couples encore sont la notation d'un ensemble infini d'objets qu'on nomme « règles », mais qui incorporent des *manipulations* : celles-ci sont des suggestions algorithmiques de recherche et de remplacement d'éléments dans une liste. Ces couples-ci devraient permettre d'analyser les relatives et la pronominalisation. De plus, les couples pour la pronominalisation incorporent des effacements qui exigeraient une grammaire de type 0.

PTQ utilise par ailleurs des *arbres dérivationnels*, objets mal définis et illustrés par un exemple dans (Montague, 1974, PTQ p. 154), qui sont essentiellement une procédure et qui ont été interprétés de différentes manières[23].

La syntaxe de PTQ, analysée en tant que *SH* avec son vocabulaire *V*, ses 24 couples et les arbres dérivationnels n'est associable à aucun calcul formel sous-jacent. La syntaxe de PTQ est un *SH* entre pseudo-formel et algorithmique, dont les expressions déduites ne peuvent être calculées qu'« à la main ». Si on les calcule ainsi et qu'on applique le test de corroboration, on trouve des résultats inadéquats, par exemple : mauvaise interprétation des pronoms (*him* dans *He likes him* peut être interprété comme coréférent avec *he*) ; mauvaise coordination des enchâssées, car *Peter asserts-that John likes Suzanne and Roberts walks* peut être interprété avec une coordination de complétives : *Peter asserts-that (John likes Suzanne and Roberts walks)* ; mauvaise interprétation des adverbes, car *Peter asserts-that voluntarily John does not love Mary* sera interprété avec *voluntarily* spécifiant *asserts-that* et non l'enchâssée. Enfin, par rapport à la traduction sur LI et l'interprétation des formules de LI dans la dénotation, il faut remarquer un problème majeur : s'il est vrai que PTQ peut dans certains cas exprimer la coréférence, le formalisme utilisé est incapable d'exprimer l'exigence de non-coréférence. Ainsi *Peter loves Mary* pourrait avoir la même dénotation dans un univers que *Peter loves himself*[24].

---

23. La différence concerne l'indexation des expressions : cf. (Montague, 1974, PTQ, p. 254-255) et (Chambreuil, 1989, p. 57-58).
24. Deux noms propres différents par leur genre sont associables à une même dénotation.

La grammaire de Montague n'a été analysée ici que par rapport à la syntaxe et à sa traduction sur LI, et avec application stricte de la grille de la section 3. Il est évident qu'il est impossible d'obtenir par traitement automatique toutes les expressions déduites. On sait, dans le domaine du TALN, que traiter des relatives, de la coordination et de l'anaphore relève de l'exploit à accomplir. Or, pour traiter ces structures, des manipulations sont requises dans les couples respectifs, et ces manipulations sont intrinsèquement mal conçues : elles ont été imaginées dans l'espoir qu'une exploration linéaire plate de la suite à analyser, sans tenir compte des parenthétisations qui devraient être associées à cette suite, pouvait aboutir à la bonne détection de l'élément à remplacer. C'est sur ces points concrets, identifiables à condition de se donner une définition rigoureuse de la *Grammaire de Montague*, que l'on peut mesurer la distance entre les idées directrices de Montague et leur application à l'étude des langues naturelles.

6. Éléments d'un bilan et perspectives

Le réalisme épistémologique exige que nous nous intéressions à ce qui est effectivement fait, au-delà de ce qu'on proclame ou propose. Notre grille devrait aider à appréhender ce faire effectif. Elle est volontairement pointue et les contraintes introduites dans le domaine de l'observation laissent de côté des phénomènes qui relèvent du langage humain. Mais le domaine sur lequel elle porte devrait pouvoir être mis en relation avec l'étude des comportements langagiers et nous ne voyons pas comment ceux-ci peuvent être étudiés sans cette mise en relation[25]. Notons cependant que la grille n'introduit aucune contrainte sur la conception possible de *Obs* au-delà des exigences d'opérationnalité et d'intersubjectivité. La notion de *Obs* est suffisamment flexible pour absorber ce que l'on considère souvent sous la rubrique de « adéquation descriptive forte ». Par ailleurs, d'autres facteurs possibles (temps de calcul, gestion de la formulation des règles...) peuvent être pris en considération pour choisir une solution. Mais le point de départ adopté est que la connaissance de l'empirie ne peut se construire qu'autour du test de corroboration, ce qui implique de confronter hypothèses et observations. C'est sur ce noyau central que porte la grille. Son application dans le cadre du réalisme épistémologique conduit à un certain pointillisme : on ne corrobore pas les grandes idées mais ce que l'on a effectivement fait, c'est-à-dire des résultats concrets et parcellaires.

L'application de la grille devrait être utile pour discriminer entre des *SH* susceptibles d'être soumis au test de corroboration et des exigences méthodologiques requises qui ne le sont pas. Pour interroger l'idée de la non-différence de Montague (cf. §5.2), il faut opérer avec la notion stricte de compositionnalité – la composition de deux homomorphismes algébriques – et non avec la notion interprétable à la demande de « compositionnalité ». L'idée de la non-différence, interprétée comme hypothèse empirique, si elle avait été corroborée, aurait été un pas gigantesque pour comprendre les langues naturelles. Elle aurait fait sauter une barrière entre objets

25. La grille n'est pas applicable dans l'état actuel au langage oral non transcrit.

considérés comme appartenant à des classes irréductiblement différentes. Mais elle n'est pas corroborée, ce qui ne veut nullement dire que l'on a démontré le contraire. Pour l'instant, si l'on veut une sémantique véric onditionnelle, au moins partiellement, on ne voit que la possibilité de conserver un signe classique saussurien à double face.

L'examen de Montague permet de détecter un autre type de limite. Nous avons remarqué l'impossibilité d'évaluer les fonctions qui permettent de caractériser U dans la Grammaire de Montague (cf. §5). Nous ajoutons que *U,* chez Montague, et en général dans la sémantique véric onditionnelle, n'est pas constitué par les objets du monde tels qu'on peut les percevoir, mais par des symboles d'un langage : dans la formule *<expr,* $<U^i$*, Va>>*$_{Inf}$, $<U^i$*, Va>* est aussi du langage. Si par ailleurs nous acceptons que Ob doit exprimer ses observations dans un langage adéquat (cf.§3.1), `test` corrobore un langage par rapport à un autre langage. Comme nous exprimons avec une portion de langue naturelle les conventions pour décrire le langage avec lequel nous exprimons les observations, de même que les conventions pour décrire le langage qui exprime *U,* nous devons conclure qu'il y aura toujours un résidu de la langue naturelle qui pose problème : c'est la portion qui, en même temps, est langage-objet et métalangage.

Les travaux qui privilégient le formalisme de calcul (différents modèles de grammaire, utilisation du calcul de Lambek, de la logique linéaire, sémantique véric onditionnelle) peuvent être mieux compris comme des tentatives de formulation d'un véhicule d'expression d'une théorie (cf. §3.2) que comme des théories de l'empirie. Ces formalismes devraient être évalués en relation avec le type d'observation – la spécification de *Obs* dans la grille – dont les formules qui s'expriment par le biais de ces formalismes sont censées être corroborées. Or, sur ce point, il faut rappeler que la hiérarchie des grammaires, souvent évoquée pour caractériser ces formalismes, ne porte que sur des grammaires à l'adéquation dite « observationnelle », autrement dit à des grammaires dont le domaine d'observation doit se limiter à des *expr* bien formées, observées à partir de l'utilisation de `bf`, sans aucune autre caractérisation possible de *Obs*[26].

Dans les travaux privilégiant le formalisme de calcul, peu d'efforts sont consacrés au test de corroboration[27]. Le lien avec le domaine de l'observation est limité à la participation de Ob opérant sur des exemples ciblés. On a ainsi un domaine d'obser-

---

26. Cf. les remarques très explicites dans ce sens dans (Chomsky, 1963, p. 325-326 et p. 357). L'étude de l'équivalence entre calculs formels sous-jacents aux grammaires a été située dès le départ, malgré quelques essais d'extension à l'équivalence forte (cf. (Chomsky, 1963, p. 395-401)), dans les limites de l'équivalence faible, portant donc sur des langages en tant qu'ensembles de *expr* ; cf. la conjecture formelle de (Chomsky, 1964) sur l'équivalence faible des grammaires catégorielles bidirectionnelles, le calcul de Lambek et les grammaires hors contexte, aujourd'hui formellement prouvée. Nous suivons ici (Legeret-Tessiot, 1995) qui présente un panorama clair des relations formelles d'équivalence faible entre grammaires hors contexte, différentes versions du calcul de Lambek et des grammaires catégorielles.

27. Une exception notable est (Kamp et Reyle, 1993), travail qui a soulevé des questions importantes, aussi bien sur des points précis du domaine de l'observation que sur les concepts qu'on utilise pour observer et qui s'appliquent « *unreservedly or not at all* » (Kamp et Reyle,

vation peu stable (cf. note 10) et les exemples ciblés suggèrent des généralisations d'observation avec des limites peu claires. Enfin, si l'on regarde de près des descriptions effectives, et cela sans sortir du domaine d'observation visé par ces descriptions, les résultats sont souvent troublants[28].

L'utilisation de la grille sur des exemples précis dans deux approches différentes suggère des généralisations sur ceux-ci. Dans l'approche symbolique, l'attention porte sur le formalisme en tant que véhicule d'expression d'une théorie, alors que dans l'approche texto-algorithmique on vise plutôt un texte « tout venant ». Dans le premier, on utilise des exemples ciblés, sans vérifier si ces exemples peuvent être effectivement traités par le formalisme invoqué et sans vérifier leur représentativité. Dans le deuxième, on peine à différencier ce qui relève de l'objet d'étude et ce qui relève de la machine utilisée, soit pour s'en servir (applications industrielles) soit pour l'étudier. Puisque avantages et inconvénients des deux approches dans le domaine de l'observation et de l'hypothèse semblent quasiment symétriques, pourquoi ne pas les considérer comme complémentaires ?

L'utilisation d'exemples ciblés a obscurci un problème crucial : le très grand nombre d'analyses possibles d'un énoncé. Dans les exemples ciblés, on raisonne sur des expressions élémentaires associées à une seule catégorie morphosyntaxique. Or, grâce aux travaux de l'approche texto-algorithmique, on a aujourd'hui une vision réaliste, même si elle est macroscopique, des difficultés introduites par les ambiguïtés de catégorisation morphosyntaxique, qui se conjuguent avec les ambiguïtés de segmentation et de rattachement. C'est peut-être le principal apport de cette approche en termes de connaissance de l'objet d'étude. Avec les formalismes de calcul existants, on ne prévoit pas aujourd'hui, étant donné un lexique et des règles, les conditions contextuelles qui déclenchent la multiplicité d'analyses (les analyses parasites de (Chanod, 1993)). Le problème à traiter a été bien diagnostiqué il y a une dizaine d'années : « le calcul d'une construction particulière est distribué non pas sur une règle spécifique, mais sur l'ensemble des règles de l'analyseur » remarque Chanod (1993, p. 239), et sur les entrées du lexique, ajouterions-nous. Les formalismes de calcul existants seraient-ils

---

1993, p. 226), mais le développement de la DRT est plus orienté vers le formalisme que vers l'observation, cf. (Amsili et Bras, 1998).

28. Nous nous limitons à donner deux exemples, le premier emprunté à (Bès et Baschung, 1985) qui en signale d'autres. Dans la description de l'anglais selon GPSG (Gazdar *et al.*, 1985) d'après les définitions de CAP (*Control Agreement Principle*), des traits Control, des FCR (*Feature Co-occurence Restrictions*) 12 et 13, et des FSD (*Feature Specification Default*) 4 et 10, on obtient comme spécification de l'accord verbal qu'un verbe s'accorde avec son complément d'objet, de telle manière que *I wants the book* est spécifié par la grammaire proposée. Selon (Kamp et Reyle, 1993) *Jacques aime un livre que Pierre regarde* et *Pierre regarde un livre que Jacques aime* devraient être associés à une même DRS, alors que les relations anaphoriques avec une continuation possible par *Il est content* sont très différentes. Ces exemples illustrent un questionnement de fond sur les calculs formels sous-jacents qui ont vocation à être utilisés commme des véhicules d'expression d'une théorie sur l'empirie : est-ce que les erreurs que l'on trouve dans les descriptions effectives relèvent d'une inadvertance ponctuelle ou d'une carence dans le pouvoir expressif de leur formalisme de calcul ?

capables d'être utilisés pour cette tâche ? La question reste ouverte et elle est suffisamment bien posée. On ne voit pas pourquoi les deux approches ne pourraient pas collaborer pour avancer vers une solution.

Nous ne pouvons que souscrire aux idées de (Chanod, 1993, p. 226) lorsqu'il propose de « soumettre l'analyseur à des expériences cruciales, pour reprendre la terminologie de K. Popper [...] les tests d'analyseurs [doivent] être conçus comme des expériences de réfutation, destinées à établir les limites des hypothèses explicites ou implicites de la description sous-jacente à l'analyseur ». Or, s'il y a un enseignement que l'on peut tirer de Popper, c'est que les hypothèses doivent être explicites, alors qu'on observe que ce qui est communiqué à propos de ce que l'on fait dans l'approche texto-algorithmique est entièrement opaque.

La linguistique est très loin aujourd'hui de s'être constituée en tant que science de l'empirie portant sur les langues naturelles selon le paradigme de la physique newtonienne. On n'a pas (encore ?) réussi à formuler des principes spécifiant des projections explicatives qui tiennent la route sur des domaines observationnels quelque peu étendus, c'est-à-dire en appliquant sur les projections obtenues le test de corroboration. Mais nous admettons que le paradigme newtonien n'est pas le seul possible (Auroux, 1998, p. 9, 47) et que l'induction peut être utile comme stratégie de recherche.

Les travaux du domaine texto-algorithmique ont ouvert la voie à un processus inductif de construction de connaissances. Il serait vraisemblablement possible de tenter la formulation de généralisations inductives sur le type d'information linguistique qui est effectivement manipulée, ce qui exigerait de discriminer dans les résultats obtenus les effets provoqués par le type d'information et par les choix algorithmiques.

Deux conditions sont nécessaires (mais non suffisantes) pour avancer dans cette voie : une application sincère du test de corroboration et l'accès à l'information linguistique sous-jacente aux réussites et échecs du test. Nous ne voyons rien dans notre objet d'étude qui soit susceptible de nous empêcher d'avancer dans cette direction. Et nous laissons à d'autres travaux relevant de la sociologie, de l'économie ou de l'histoire le soin de s'interroger sur les obstacles sociologiques, politiques, économiques ou psychologiques qui peuvent empêcher de poursuivre dans cette direction.

Pour la première soumission de ce travail, je remercie Jean-Pierre Chanod, qui, par ses commentaires, a signalé la possibilité et l'intérêt d'avancer dans une voie inductive de construction de la connaissance à partir des travaux dans le domaine texto-algorithmique. La première soumission a également profité des remarques de Caroline Hagège. La deuxième bénéficie des commentaires, questionnements et suggestions des relecteurs de la première, que je remercie. Elle a bénéficié aussi des commentaires de Daniel Guillot et de Cassian Braconnier. Enfin, pour la préparation des deux soumissions, les discussions avec François Trouilleux et ses critiques constructives, pour la plupart incorporées, ont été une aide déterminante. Dans tous les cas, la responsabilité du travail est entièrement mienne.

7. Bibliographie

AÏT-MOKHTAR S., CHANOD J.-P., ROUX C., « Robustness beyond Shallowness : Incremental Deep Parsing », *Robust Methods for the Analysis of Natural Language Data* ; a special issue of the *Natural Language Engineering Journal* (NLE), 8, 2/3, p. 121-144, 2002.

AMSILI P., BRAS M., « DRT et compositionnalité », in (Nazarenko, 1998, p. 131-160).

AUROUX S., *La raison, le langage et les normes*, Paris, PUF, 1998.

BACH E., « Linguistique structurale et philosophie des sciences », *Diogène*, LI, 1965, p. 117-136.

BÈS G. G., Empiricité en linguistique et grammaire de Montague : la sémantique en 5P et la compositionnalité, GRIL, Université Blaise-Pascal, 2001.

BÈS G. G., BASCHUNG K., « Feasibility of a GPSG French Grammar », Rapport de recherche du projet ESPRIT 393 ACORD, GRIL, Université Blaise-Pascal, 1985.

BLOOMFIELD L., « A set of postulates for the science of language », in JOOS M. [ed.], *Readings in Linguistics*, New York, American Council of Learned Societies, 1957, p. 26-31.

BLOCH B., « A set of postulates for phonemic analysis », *Language*, vol. 24, 1948, p. 3-46.

BRESNAN J., KAPLAN R. M., *« Grammars as Mental Representations of Language »,* BRESNAN J. [ed.], *The Mental Representation of Grammatical Relations*, Cambridge, Mass., The MIT Press, 1982, p. xvii-liii.

BUNGE M., *La investigación científica ; Su estrategia y su filosofía*, Barcelona, Ariel, 1969.

CHAMBREUIL M., *Grammaire de Montague. Langage, traduction, interprétation*, Clermont-Ferrand, Adosa, 1989.

CHAMBREUIL M. [dir.] *Sémantiques*, Paris, Hermès, 1998.

CHAMBREUIL M., BEN GHARBIA A., GAMALLO OTERO P. « Variations sur la compositionnalité Montaguienne », in (Nazarenko, 1998, p. 35-65).

CHANGEUX J.-P., *L'Homme de vérité*, Paris, Odile Jacob, 2002.

CHANOD J.-P., « Problèmes de robustesse en analyse syntaxique », *Colloque "Informatique & Langue naturelle" I.L.N.'93*, Nantes, 1993, p. 223-243.

CHANOD J.-P., « TALN et industrie », *TAL*, vol. 38, n° 1, 1997, p. 135-144.

CHEVALIER J.-C. « Le jeu des exemples dans la théorie grammaticale ; étude historique », in CHEVALIER, J.-C. [dir.], *Grammaire transformationnelle, syntaxe et lexique*, Lille, Presses Universitaires de Lille, 1976, p. 235-263.

CHOMSKY N., *Syntactic Structures,* La Haye, Mouton, 1957.

CHOMSKY N., « Formal Properties of Grammars », in (Luce *et al.,* 1963, p. 323-418).

CHOMSKY N. « Current Issues in Linguistic Theory », FODOR, J.J. & KATZ, J.J. [ed.] *The structure of Language ; Readings in the Philosophy of Language*, Englewood Cliffs, Prentice-Hall, 1964, p. 50-118 [Publié en première version en 1962].

CHOMSKY N., *Aspects of the Theory of Syntax,* Cambridge, Mass, The MIT Press, 1965.

CHOMSKY N., MILLER G. A., « Introduction to the Formal Analysis of Natural Languages », in (Luce *et al.,* 1963, p. 269-321).

CHOMSKY N., HALLE M., *The Sound Pattern of English*, New York, Harper & Row, 1968.

CORI M., MARANDIN J.-M., « La linguistique au contact de l'informatique : de la construction des grammaires aux grammaires de construction », *HEL*, vol. 23, n° 1, 2001, p. 49-79.

DESCLÉS J.-P., « Énoncés et énonçables », *Lingua e Stile*, vol. XIII, n° 2, 1978, p. 175-223.

GAZDAR G., KLEIN E., PULLUM G., SAG I., *Generalized Phrase Structure Grammar*, Oxford, Blackwell, 1985.

GAZDAR G., MELLISH C., *Natural Language Processing In Prolog*, Wokingham, UK, Addison-Wesley, 1989.

GRANGER G.-G., *La vérification*, Paris, Odile Jacob, 1992.

HABERT B., NAZARENKO A., SALEM A., *Les linguistiques de corpus*, Paris, Armand Colin, 1997.

ITKONEN E., *Grammatical Theory and Metascience*, Amsterdam, Benjamins, 1978.

JANSSEN T.M.V., « Compositionality », in (Van Benthem et Ter Meulen, 1997, p. 417-473).

KAMP H., REYLE U., *From Discourse to Logic*, Dordrecht, Kluwer, 1993.

KARLSSON F., « Constraint grammar as a framework for parsing running text », *Papers presented to the 13th International Conference on Computational Linguistics*, vol. 3, Helsinki 1990, p. 168-173.

KATZ J. J., *Language and Other Abstract Objects*, Rowman and Littlefield, Totawa, New Jersey, 1981.

KATZ J. J., POSTAL P. M., « Realism *vs.* Conceptualism in Linguistics », *Linguistics and Philosophy*, vol. 14, 1991, p. 515-554.

KING P. J., « Theory and formalism in LFG : too much inertia ? », *Lexical-Functionnal Grammar List*, 1999 (lfg@listserv.linguistlist.org).

LEGERET-TESSIOT M.-A., Algèbres de démonstrations et grammaires catégorielles, Thèse de doctorat, Université Blaise-Pascal, 1995.

LUCE R. D., BUSH R. R., GALANTER E., [ed.] *Handbook of Mathematical Psychology*, New York, John Wiley and Sons, 1963.

MARANDIN J.-M., « Analyseurs syntaxiques, équivoques et problèmes », *TAL*, vol. 34, n° 1, 1993, p. 5-33.

MARTIN R. [dir.], *La notion de recevabilité en linguistique*, Paris, Klincksieck, 1978.

MILLER G. A., CHOMSKY N., « Finitary Models of Language Users », in (Luce, 1963, p. 419-491).

MILNER J.-C., *Introduction à une science du langage*, Paris, Le Seuil, 1989.

THOMASON R. H. [ed.], *Formal Philosophy. Selected Papers of Richard Montague*, New Haven, Yale University Press, 1974.

NAZARENKO A. [dir.], « Compositionnalité », *TAL,* vol. 39, n° 1, 1998.

NAZARENKO A., « Présentation », in (Nazarenko, 1998, p. 3-7).

PARTEE B. H., HENDRIKS H.L.W., « Montague Grammar », in (Van Benthem et Ter Meulen, 1997, p. 5-91).

PETERS S., « The Projection Problem : How is a Grammar to Be Selected ? », in PETERS S. [ed.], *Goals of Linguistic Theory*, Englewood Cliffs, Prentice-Hall, 1972, p. 171-188.

SOKAL A., BRICMONT J., *Impostures intellectuelles*, Paris, Odile Jacob, 1997.

VAN BENTHEM J, TER MEULEN A. [ed.], *Handbook of Logic and Language*, Amsterdam, Elsevier, 1997.