La mathématisation des formalismes syntaxiques

Marcel Cori

La mathématisation des formalismes syntaxiques

Marcel Cori

Université Paris X et CNRS MoDyCo
mcori@u-paris10.fr

1. Introduction*

L'objet de cet article est de s'interroger sur la nature des formalismes syntaxiques contemporains, et plus exactement de savoir s'ils répondent au projet de mathématisation qui a été initié dans les années 1950.

Une des raisons essentielles pour lesquelles les connaissances sur les langues ont reçu une expression mathématique est la volonté d'effectuer des traitements informatiques de productions langagières. On sait (voir Cori et Léon, 2002) que dès l'apparition des premiers ordinateurs s'est posée la question de la traduction automatique, ce qui nécessitait de trouver des codages (numériques) des données linguistiques. Telle est l'origine reconnue du *Traitement automatique des langues* (*TAL*).

C'est à peu près à la même époque, mais avec une claire volonté d'indépendance vis-à-vis du TAL[1], que Chomsky lance son programme de recherches. Il s'agit d'avoir une démarche scientifique en linguistique, sur le modèle des sciences de la nature. La mathématisation est présentée comme une composante *sine qua non* de la démarche.

Les formalismes syntaxiques, ainsi qu'on les a appelés[2], s'inscrivent dans l'un ou l'autre de ces projets, si ce n'est dans les deux. Ces formalismes, qui se sont multipliés depuis un demi-siècle, ont permis de grandes avancées en linguistique, et ont donné lieu à des réalisations informatiques prometteuses. Néanmoins, il est légitime de s'interroger sur le lien entre ces formalismes et le projet de mathématisation.

Dans ce qui suit, nous commençons par situer la mathématisation par rapport à la perspective des recherches linguistiques (§ 2) et par rapport à celle du TAL (§ 3), en précisant quels sont les critères propres à chacune de ces perspectives. Nous essayons ensuite (§ 4) de déterminer quels sont les courants qui, au cours des cinquante dernières années, ont été les porteurs du projet de mathématisation. Enfin (§ 5), le formalisme des HPSG est examiné de plus près, en tant

* Je remercie très vivement Gabriel G. Bès, Sophie David, Françoise Kerleroux, Danielle Leeman et Jean-Marie Marandin pour leurs lectures critiques d'une première version de cet article.

[1] Ainsi, dans *Structures syntaxiques*, il n'est pas fait mention du traitement automatique. Près de vingt ans plus tard, Chomsky (1975 : 39) se justifiera en ces termes : « for machine translation and related entreprises, they seemed to me pointless as probably quite hopeless ».

[2] Voir par exemple Miller et Torris (1990). Pour un panorama de différents formalismes syntaxiques et une bibliographie plus complète, on peut citer également Abeillé (1993) et Ligozat (1994).

que point d'aboutissement actuel et provisoire de l'évolution d'un certain nombre de tendances des formalismes syntaxiques.

2. La mathématisation en linguistique

2.1. Des méthodes scientifiques[3]

Chomsky a sans conteste joué un rôle fondamental dans le projet de mathématisation en linguistique. Il s'en est justifié dans les termes qui suivent[4] (1957 : 5) : « Precisely constructed models for linguistic structure can play an important role, both negative and positive, in the process of discovery itself. By pushing a precise but inadequate formulation to an unacceptable conclusion, we can often expose the exact source of this inadequacy and, consequently, gain a deeper understanding of the linguistic data. More positively, a formalized theory may automatically provide solutions for many problems other than those for which it was explicitly designed. Obscure and intuition-bound notions can neither lead to absurd conclusions nor provide new and correct ones, and hence they fail to be useful in two important respects. I think that some of those linguists who have questioned the value of precise and technical development of linguistic theory have failed to recognize the productive potential in the method of rigorously stating a proposed theory and applying it strictly to linguistic material with no attempt to avoid unacceptable conclusions by *ad hoc* adjustments or loose formulation. »

Qu'apportent des « modèles précisément construits » ? La réponse, à notre sens, tient en trois mots : falsifiabilité, prédictivité, objectivité. Tout d'abord, il est impossible de réfuter des propositions énoncées de manière imprécise. Seules des propositions formulées rigoureusement, dont on peut sans équivoque tirer des conséquences, peuvent être réfutées. Or la connaissance scientifique n'avance que par la falsification successive des hypothèses posées[5].

Ensuite, une science doit pouvoir trouver des solutions qui s'appliquent à des faits qui n'ont pas été examinés ou à des problèmes qui n'ont pas été envisagés : c'est son caractère prédictif. Cela nécessite encore que puissent être tirées sans équivoque les conséquences des propositions énoncées.

Il apparaît ainsi essentiel qu'un modèle soit objectif[6]. Autrement dit, mises entre n'importe quelles mains, les mêmes hypothèses dans un même modèle doivent conduire aux mêmes conclusions, comme les mêmes calculs doivent conduire aux mêmes résultats quel que soit le calculateur. C'est dire qu'un modèle « précisément construit » ne peut être qu'un modèle mathématique. On peut se passer de l'auteur

---

[3] Pour une théorie générale des méthodes scientifiques, nous renvoyons à Popper (1959).

[4] Ce passage est cité notamment par Pollard et Sag (1994 : 7-8).

[5] « *Un système faisant partie de la science empirique doit pouvoir être réfuté par l'expérience* » (Popper, 1959 : 37, souligné par l'auteur).

[6] Selon Popper (1959 : 41, souligné par l'auteur) « *l'objectivité* des énoncés scientifiques réside dans le fait qu'ils peuvent être intersubjectivement *soumis à des tests* ».

d'une théorie mathématisée pour la tester expérimentalement, ou pour essayer de l'appliquer à tel ou tel fait. On peut transmettre la théorie à d'autres personnes, qui en deviennent des dépositaires aussi légitimes que les auteurs originels. La théorie n'est plus attachée au théoricien.

Cela ne signifie évidemment pas qu'en dehors de la modélisation (mathématique) il n'y a pas de pratique scientifique, et encore moins qu'avant Chomsky la linguistique n'était pas scientifique. Nous dirons, de manière très rapide, que la démarche scientifique consiste à observer, décrire, classer, généraliser, modéliser. L'observation est la toute première phase de la démarche scientifique, tandis que la modélisation en constitue la phase ultime, phase exclusive de la mathématisation.

2.2. Modèles et langages

Toute mathématisation passe par un langage symbolique, mais l'usage d'un langage symbolique n'est nullement gage de mathématisation : il faut que sur les symboles puissent être effectués des calculs. Et un modèle (mathématique), exprimé dans un langage donné, ne s'identifie pas à ce langage.

En effet, un même objet mathématique peut être exprimé selon différents langages. Prenons un exemple simple. L'addition de nombres entiers est une opération bien définie en mathématiques. Les propriétés de cette opération ne changent pas selon qu'on écrit *a + b*, *plus(a,b)* ou *+ a b*. Ce qui importe, c'est la sémantique de l'addition, qui est régie par les axiomes qui définissent l'opération.

Il en résulte que, si on confond modèles et langages, on risque de ne pas s'apercevoir que deux approches qui utilisent le même langage sont radicalement différentes, ou qu'il y a une rupture qui s'est produite au sein d'une même approche. Inversement, on peut être amené à penser que deux mathématisations n'ont rien de commun parce que les langages utilisés diffèrent.

Sachant que des modèles mathématiques censés représenter les connaissances linguistiques ont été élaborés les uns indépendamment des autres dans des cadres distincts, il a fallu, *a posteriori*, étudier en quoi ces modèles se différenciaient. Mais il n'est pas toujours immédiat de démontrer que deux objets mathématiques sont différents ou identiques. La notion même d'« identité » ou d'« équivalence » entre objets mathématiques ne va pas de soi. Il est par conséquent spécialement important, si on veut comparer des cadres théoriques, de définir ce qu'on entend par « équivalence ».

Nous renvoyons à Miller (1999) qui étudie le problème de la recherche d'équivalences entre formalismes syntaxiques. Il montre en quoi la notion d'équivalence faible[7] est insuffisante, et en quoi la notion d'équivalence forte[8] est au contraire trop stricte. Il définit en

[7] Deux formalismes *T* et *T'* sont faiblement équivalents si pour toute grammaire *G* de *T* il existe une grammaire *G'* de *T'* telle que l'ensemble des énoncés engendrés par *G'* soit identique à l'ensemble des énoncés engendrés par *G*, et inversement.

[8] Deux formalismes *T* et *T'* sont fortement équivalents si pour toute grammaire *G* de *T* il existe une grammaire *G'* de *T'* telle que l'ensemble des descriptions structurales

conséquence des fonctions ou des relations significatives, en vertu desquelles on détermine s'il y a équivalence ou non entre les descriptions structurales d'un même énoncé. Les relations significatives qu'il considère sont notamment la constituance, l'ordre linéaire, la dépendance.

On observera que les résultats de ce type d'étude sont plus fins qu'un simple jugement d'équivalence ou de non-équivalence. Entre des théories non équivalentes, on peut établir des relations, de deux ordres différents. En premier lieu, on peut constater une sorte d'inclusion entre formalismes. Par exemple, les automates finis sont inclus dans les grammaires syntagmatiques indépendantes du contexte. Mais, en second lieu, on pourra s'apercevoir que certaines dimensions, autrement dit certaines relations significatives prises en compte dans le cadre d'un certain formalisme, ne sont pas définissables dans un autre cadre. Par exemple, la notion de gouverneur (et de dépendants) que l'on trouve dans les grammaires de dépendance ne peut être définie dans les grammaires indépendantes du contexte (*Context-Free Grammars, CFG*)[9].

De la sorte, les grammaires syntagmatiques, les grammaires de dépendance et les grammaires catégorielles, qui sont d'origines presque totalement distinctes, deviennent comparables.

3. La problématique du TAL

3.1. Qu'est-ce que le TAL ?

On peut définir de manière très simplifiée le traitement automatique des langues comme étant constitué des méthodes et des programmes qui prennent pour données des productions langagières, quand ces méthodes et programmes tiennent compte des spécificités des langues humaines.

Ce n'est certainement pas un hasard si la mathématisation en linguistique a été concomitante du développement du TAL. Quand on informatise un problème, il est nécessaire d'être explicite, précis et objectif : les règles que l'on énonce devant entrer dans des processus automatisés, il n'est pas possible de rester dans le vague ou d'être ambigu. Une machine n'est pas susceptible d'interpréter en quoi que ce soit les informations qu'on lui communique. D'où la construction de systèmes de description rigoureux, ou le perfectionnement de systèmes existants pour les rendre plus proches des critères du TAL[10].

Dès les origines du TAL, cependant, deux logiques se sont opposées : une logique « scientifique », qui cherche à s'appuyer sur les recherches linguistiques et à faire progresser celles-ci, face à une logique que l'on peut qualifier d'« utilitariste », selon laquelle la fin justifie les moyens. L'opposition se poursuit aujourd'hui (cf. Bès

---

engendrées par *G'* soit identique à l'ensemble des descriptions structurales engendrées par *G*, et inversement.

[9] Cela sera en revanche possible si on ajoute la notion de tête aux CFG, comme cela se fait par exemple dans le cadre de la théorie X-barre.

[10] Par exemple les arbres de dépendance de Tesnière, dont on peut faire remonter les origines aux années 1930, ont reçu de nouvelles définitions dans les années 1960.

2002), la logique utilitariste se manifestant dans les travaux dits de TAL robuste, avec notamment les outils de désambiguïsation fondés sur les statistiques et les probabilités, ou les analyses partielles qui cherchent uniquement à délimiter certains constituants dans les phrases en ignorant les ambiguïtés. Ces travaux, même s'ils font appel à des outils mathématiques, sont très éloignés d'une mathématisation des connaissances linguistiques. Prenons les modèles probabilistes markoviens : ils peuvent servir aussi bien à prédire le prochain mot que va prononcer un locuteur que le temps qu'il fera demain. Quant aux automates finis, très utilisés par les systèmes d'analyse robuste, leur inadéquation en tant que modèle de la syntaxe a été démontrée par Chomsky (1957). Quoi qu'il en soit, l'objet des travaux de TAL robuste n'est pas de construire des généralisations sur les langues.

Or ces dernières années, en raison du développement des industries de la langue, le TAL a plutôt penché du côté de la logique utilitariste. Une conséquence semble être une moindre demande de formalismes syntaxiques qui soient mathématiquement rigoureux.

3.2. TAL et syntaxe

La syntaxe occupe une place à part dans le TAL, car centrale. Centrale, cela signifie que, si on décompose les traitements automatiques en des successions de sous-traitements, elle constitue un passage presque obligé, avec par exemple en amont des prétraitements qui permettent d'obtenir des découpages en unités de l'ordre du mot, et en aval des tâches spécifiques aux applications envisagées.

Historiquement, on notera que, après l'échec reconnu de la traduction automatique, les algorithmes d'analyse syntaxique sont devenus, pendant les années 1960, l'axe des recherches en traitement automatique. Cela en lien avec l'importance prise par la mathématisation de la syntaxe, importance provenant des travaux de Chomsky, mais également des recherches qui se sont développées indépendamment du programme de la grammaire générative.

3.3. L'opposition entre procédural et déclaratif

Depuis la fin des années 1970, il est admis qu'une réalisation informatique est d'autant meilleure que la séparation est bien marquée entre les algorithmes ou les programmes (*procéduraux*) et la représentation des données, effectuée de manière *déclarative*, et qu'un maximum des connaissances humaines sur lesquelles s'appuie la réalisation figurent dans les données (et par conséquent un minimum dans les programmes).

Nous allons illustrer cette remarque par un exemple simplifié à l'extrême de traitement automatique des langues. Supposons que l'on veuille trouver les étiquettes catégorielles d'une suite de mots dont on connaît les étiquettes lexicales, un fragment de programme dans lequel seraient intégrées les informations linguistiques pourrait se présenter comme suit[11]:

[11] Nous adoptons la syntaxe du langage Python.

(1)
```
if a[0]=='V':
    if categ(a[1:])=='SN': return 'SV'
    elif categ(a[1:])=='SP': return 'SV'
```

Cela signifie que l'on demande à l'ordinateur de vérifier si le premier mot de la suite est un verbe. Si oui, on lui demande de vérifier si tous les mots qui suivent forment un SN (ce qu'il effectue à l'aide d'instructions non présentées ici). Auquel cas l'ordinateur aura reconnu que la suite de mots forme un SV. Si tous les mots à partir du deuxième ne forment pas un SN, l'ordinateur vérifie s'ils forment un SP. Auquel cas, il reconnaît encore un SV.

Ces instructions intègrent, à l'évidence, des connaissances linguistiques. Connaissances qui pourraient être représentées par les règles d'une CFG sous la forme :

(2) a. SV → V SN
b. SV → V SP

Les règles de la CFG peuvent être prises comme les données d'un programme qui, alors, n'aura pas à intégrer de connaissances linguistiques spécifiques. On obtient en ce cas un traitement automatique dans lequel la représentation des connaissances est déclarative, alors que dans le premier cas elle était procédurale.

Le premier avantage d'un traitement automatique déclaratif est que celui-ci est plus général. Les programmes, écrits une fois pour toutes, ne sont pas à modifier quand on veut réviser une grammaire. Reprenons l'exemple ci-dessus. Si on veut admettre la possibilité d'un complément d'objet direct suivi d'un complément indirect dans le SV, il suffit d'ajouter une règle à la grammaire :

c. SV → V SN SP

On peut même envisager d'avoir des programmes qui ne soient pas à modifier si on change la langue sur lesquels ils s'appliquent.

Dans le cas du traitement procédural, il faut évidemment des programmes différents selon les langues traitées, et la mise à jour des informations linguistiques, qui revient à la correction d'un programme, est une opération techniquement complexe.

Un deuxième avantage du traitement déclaratif est qu'il permet une division du travail entre l'informaticien (qui n'a pas nécessairement à connaître la linguistique et les langues) et le linguiste en mesure de définir les grammaires et les lexiques (qui n'a pas à savoir programmer). Le linguiste exprime ses connaissances dans un ordre indifférent, et indépendamment de ce que le système informatique aura à en faire.

On se trouve alors placé dans la logique des *systèmes experts* : l'expert est un utilisateur privilégié qui communique des données au programme. Cet utilisateur est à opposer à l'utilisateur naïf qui, lui, fournit des données ne témoignant pas d'une quelconque science : par exemple un texte à traduire d'une langue dans une autre. En TAL l'expert est le linguiste, c'est-à-dire quelqu'un qui a des compétences en linguistique et/ou qui a des connaissances sur une langue donnée.

L'expert doit être capable d'exprimer ses connaissances dans un format imposé par le traitement automatique. C'est-à-dire connaître le formalisme dans lequel représenter ses données – en définissant ici le formalisme comme étant composé du modèle mathématique en vertu duquel les données sont structurées et du langage permettant de traduire le modèle. Le traitement automatique, si du moins il se place dans une optique déclarative, requiert donc bien la mathématisation des connaissances linguistiques.

Le formalisme constitue de la sorte une « langue commune », à la fois « comprise » par l'expert-linguiste et par la machine, une interface entre l'étude linguistique et le traitement automatique. Il s'ensuit que, même dans le contexte du traitement automatique, l'*expressivité* des formalismes, leur lisibilité, est importante. Il ne suffit pas d'avoir des modèles mathématiques rigoureusement définis et susceptibles d'être traités par des machines : encore faut-il que les êtres humains qui ont à intervenir soient susceptibles d'écrire des représentations dans le cadre du formalisme et de comprendre celles qui sont écrites sans trop de difficulté.

4. Les courants porteurs de mathématisation

Le projet de mathématisation en linguistique, énoncé par Chomsky, n'a pas été réellement suivi par son auteur. Ce sont plutôt des courants issus du chomskysme mais opposés à Chomsky qui ont tenté de le mener à bien, ainsi que des praticiens issus du TAL et des courants dont l'origine est indépendante du chomskysme.

4.1. Le paradoxe des grammaires syntagmatiques

Selon le programme de la grammaire générative, élaboré par Chomsky, il ne s'agit plus d'établir le répertoire de tous les énoncés attestés, mais de décrire les processus de construction ou de compréhension de tous les énoncés possibles. Grâce à la récursivité, une grammaire permet d'engendrer, à l'aide d'un nombre fini de règles (de réécriture), la description structurale d'un nombre infini d'énoncés. Plusieurs types de grammaires récursives sont définies, qui fondent la hiérarchie de Chomsky. Chomsky introduit ainsi les grammaires syntagmatiques (*Phrase Structure Grammars, PSG*) et, parmi ces grammaires, les grammaires indépendantes du contexte (*Context-Free Grammars, CFG*), censées modéliser l'analyse en constituants immédiats.

Chomsky se sert des propriétés des CFG afin d'invalider l'analyse en constituants immédiats. C'est dire que Chomsky se sert de la mathématisation à des fins de réfutation, et non dans la démarche positive de construction des grammaires transformationnelles qu'il promeut et qu'il va développer. Les transformations, en effet, ne donneront pas lieu à une réelle mathématisation, pas plus que les travaux ultérieurs successifs de l'école chomskyenne orthodoxe. Ceci a été noté par nombre de chercheurs, dont G. Pullum (1991 : 53) : « Government-binding syntax (or principles-and-parameters syntax ; who cares, it's only words) no longer makes any pretense at being formally intelligible. It is set to develop into a gentle, vague, cuddly

sort of linguistics that will sit very well with the opponents of generative grammar if they compromise just enough to learn a little easy descriptive vocabulary and some casually deployed and loosely understood labelled bracketing for which no one will be held accountable. » Chomsky est accusé d'avoir tourné le dos aux principes qu'il avait lui-même posés, d'avoir bâti un système non réfutable.

À l'inverse, les CFG connaîtront un destin que le rejet par leur propre concepteur comme inadéquates à la représentation des langues ne laissait pas présager. Ainsi, ces grammaires seront très largement exploitées par le TAL. Les algorithmes d'analyse syntaxique définis dans les années 1960 seront à base de CFG, et ce sont ces algorithmes qui seront adaptés aux formalismes développés par la suite. Formalismes dont la majorité partira des CFG ou des PSG. On peut dire que, dans la perspective de la mathématisation et du traitement automatique, les CFG se sont révélées bien plus fécondes que les grammaires transformationnelles.

4.2. La réforme

Très tôt des linguistes (Yngve, 1960 et Harman, 1963) ne se sont pas satisfaits des transformations et ont proposé de conserver les grammaires syntagmatiques en accroissant leur capacité descriptive. Pour eux, comme pour d'autres linguistes à leur suite, il fallait enrichir les grammaires syntagmatiques de telle sorte qu'elles permettent l'expression claire et distincte des principes organisateurs des langues. Les auteurs de ce type de démarche se sont présentés comme les plus fidèles continuateurs du programme initial de Chomsky, que celui-ci aurait trahi. C'est pourquoi nous les regroupons dans le courant de la réforme (Cori et Marandin, 2001).

Bresnan (1982), en définissant les LFG (*Lexical-Functional Grammar*), joue un rôle majeur dans l'affirmation de ce courant. Les arguments contre les grammaires transformationnelles, et donc en faveur des LFG, sont linguistiques et psycholinguistiques. Le formalisme est inspiré de travaux menés en TAL.

Mais, le modèle le plus achevé dans ce cadre est constitué par les GPSG (*Generalized Phrase Structure Grammars*, Gazdar *et al.*, 1985). Gazdar (1982 : 131) critique lui aussi les grammaires transformationnelles sur le plan de la rigueur mathématique, observant notamment qu'elles ne donnent pas lieu à des procédures de décision indépendantes de l'intuition des auteurs de grammaires.

L'innovation clef des GPSG est la définition des catégories syntaxiques comme des ensembles de spécifications de traits, où les traits sont des couples <attribut, valeur>. Cette définition des catégories autorise la prise en compte de catégories plus ou moins spécifiées, et par conséquent l'écriture de règles à plusieurs niveaux de généralité. Par ailleurs, des principes permettent d'énoncer des connaissances générales sur le langage.

On rattachera également au courant de la réforme les TAG (*Tree Adjoining Grammars*, Joshi, 1985) qui ont la particularité de redéfinir l'opération de composition des arbres. De même, les Grammaires d'arbres polychromes (Cori et Marandin, 1993, 1994) se rattachent au

courant « réformateur » en redéfinissant aussi la composition des arbres[12].

4.3. Les formalismes issus du TAL

Dans la période qui débute vers 1970, des dispositifs de TAL dont l'idée directrice était de s'affranchir des limites des CFG ont été bâtis. Ces dispositifs se sont posés comme des formalismes linguistiques ou ont inspiré des formalismes linguistiques. Les réseaux d'automates augmentés, ou ATN (Woods, 1970), constituent le premier de ces dispositifs, se présentant à la fois comme une implémentation des grammaires transformationnelles et comme un modèle alternatif. Certaines caractéristiques des ATN seront reprises par les formalismes ultérieurs, mais les ATN seront rejetés au début des années 1980 car ne satisfaisant pas aux exigences de déclarativité : en effet, dans les ATN les connaissances grammaticales sont mêlées aux opérations constitutives de la reconnaissance.

C'est au nom de la déclarativité que les DCG (*Definite Clause Grammars*, Pereira et Warren, 1980) seront défendues en tant que candidates à la succession des ATN. En fait, les DCG résultent du projet de traduire les CFG dans la logique du premier ordre afin d'effectuer l'analyse syntaxique des langues naturelles, projet qui a conduit à l'élaboration du langage de programmation Prolog (Colmerauer *et al.*, 1973). Les règles de grammaire (CFG) sont remplacées par des formules logiques (clauses de Horn), formules qui portent sur des termes logiques plutôt que sur des catégories. Ce qui permet de prendre en compte d'autres informations que l'étiquette catégorielle (des traits, comme le nombre, le genre, la personne, le temps, etc.) et de construire des interprétations des syntagmes. Ainsi, il est possible d'écrire une formule qui reprend la règle (2.a) en lui adjoignant des informations supplémentaires :

(3) $\text{verbe}(t_1,t_2,\text{temps}(x),\text{nombre}(y),\text{personne}(z),\ \text{transitif}(\text{oui}))$
$\wedge\ \text{sn}(t_2,t_3,\text{genre}(u),\text{nombre}(v),\text{personne}(w))$
$\Rightarrow\ \text{sv}(t_1,t_3,\text{temps}(x),\text{nombre}(y),\text{personne}(z))$

Cette formule indique que si on a un verbe entre les positions t_1 et t_2, et un SN entre t_2 et t_3, alors on a un SV entre t_1 et t_3, à condition que le verbe soit transitif. En ce cas, le temps, le nombre et la personne du verbe et du SV sont identiques, tandis que le SN possède un genre, un nombre et une personne qui sont indépendants.

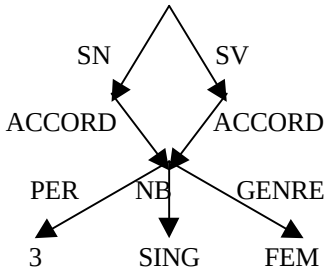
Les termes logiques ont été considérés trop rigides pour les représentations. En effet, tous les traits pertinents pour une catégorie doivent accompagner chaque occurrence de la catégorie dans chaque formule, dans un ordre qui est fixe. Par exemple, dans la formule (3) ci-dessus, l'indication du genre, du nombre et de la personne du SN est obligatoire, bien que sans utilité. C'est pourquoi on a introduit de nouveaux formalismes à base de structures de traits, qui permettent de

[12] Pour une présentation des Grammaires d'arbres polychromes, voir l'article d'Anne Lablanche dans ce numéro.

représenter de l'information partielle sur des objets. Les structures de traits sont représentées par des matrices, ou des graphes sans circuit quand elles autorisent le « partage de valeurs » (alors que les termes logiques sont représentés par des arbres). Le prototype de ces formalismes est PATR II (*Parse and Translate*, Shieber, 1986). PATR II, comme les DCG, s'appuie sur des CFG. Aux règles de la grammaire s'ajoutent des contraintes, et les représentations structurales construites sont des structures de traits (comme elles sont des termes pour les DCG). On trouvera en (4) la double représentation possible d'une structure de traits qui rend compte d'une contrainte de compatibilité d'accord accompagnant la règle `S → SN SV`.

(4)

$$\begin{bmatrix} \text{SN} & \begin{bmatrix} \text{ACCORD } \boxed{1} & \begin{bmatrix} \text{PERSONNE 3} \\ \text{NOMBRE SING} \\ \text{GENRE FEM} \end{bmatrix} \end{bmatrix} \\ \text{SV} & [\ \text{ACCORD } \boxed{1}\ ] \end{bmatrix}$$



Les grammaires d'unification fonctionnelle (*Functional Unification Grammars, FUG*) de Kay (1985) constituent un tournant, car elles font disparaître les grammaires en tant qu'objets distincts. Les règles sont incluses dans les structures de traits, et une grammaire n'est elle-même qu'une structure de traits obtenue par la disjonction de structures plus élémentaires, qui chacune correspond à une ou plusieurs règles. D'autres augmentations sont apportées aux structures de traits, par exemple l'utilisation de listes (écrites dans des parenthèses), la possibilité de représenter n'importe quelle suite de valeurs (à l'aide de points de suspension), etc. En (5) est donné un exemple simplifié à l'extrême de grammaire. La disjonction est marquée par des accolades. On envisage trois catégories possibles, S,

(5)

$$\left\{ \begin{array}{l} \left[\begin{array}{l} \text{(SUJET VERBE …)} \\ \text{CAT = S} \\ \text{SUJET = [CAT = SN]} \\ \left\{ \begin{array}{l} \text{[SCOMP = NON]} \\ \left[\begin{array}{l} \text{(… SCOMP)} \\ \text{SCOMP = [CAT = S]} \end{array} \right] \end{array} \right\} \end{array} \right] \\ \\ \text{[CAT = SN]} \\ \\ \text{[CAT = VERBE]} \end{array} \right\}$$

SN et Verbe, mais les catégories SN et Verbe ne sont pas décomposées. Les phrases (CAT = S) commencent par un sujet, de catégorie SN, suivi d'un verbe. Il y a ensuite deux possibilités (disjonction) : soit pas de complément phrastique, soit un complément phrastique, de catégorie S, qui termine la phrase.

Les formalismes issus du TAL et les formalismes de la réforme ayant des caractéristiques communes (des structures de traits sont utilisées dans les GPSG ou les LFG, le mécanisme de l'unification a été étendu des formules logiques aux structures de traits), cela a permis d'en faire une présentation unifiée sous la dénomination de grammaires d'unification (Shieber, 1986), puis de grammaires de contraintes (Shieber, 1992).

Néanmoins, derrière cette apparente unité il faut observer de sérieuses différences en ce qui concerne les modèles mathématiques sous-jacents et l'usage qui en est fait. Les structures de traits ne sont ainsi que marginales dans les GPSG et elles restent arborescentes. Quant à l'augmentation la plus importante introduite dans les FUG, c'est-à-dire la disjonction, elle confère une puissance bien plus grande au modèle qu'à celui par exemple sous-jacent à PATR II. Les modèles évoluent ainsi vers une puissance de plus en plus grande, et les HPSG seront dans une certaine mesure un aboutissement de cette évolution.

4.4. Les courants non chomskyens

Il ne faut pas oublier de mentionner les courants qui, indépendamment du chomskysme, ont développé des modèles mathématiques du langage, modèles qui ont évolué jusqu'à aujourd'hui. Citons en premier lieu les grammaires de dépendance, dont on peut fixer l'origine à Tesnière (1959), mais qui ont connu des variantes à la mathématisation plus rigoureuse, sous l'influence du traitement automatique. En second lieu, les grammaires catégorielles, nées essentiellement de préoccupations de logiciens[13], ont évolué en liaison avec les formalismes d'origine chomskyenne. Inversement, les formalismes d'origine chomskyenne ont adopté certaines solutions issues des grammaires catégorielles.

5. Le formalisme des HPSG est-il mathématisé ?

Il est intéressant de s'attarder un peu plus longuement sur les HPSG qui constituent un formalisme sans aucun doute actuel, résultat d'une évolution qui s'est produite au cours des dernières décennies. Les HPSG ont essayé de s'approprier et d'unifier les héritages des différents formalismes qui ont précédé (GPSG, FUG, grammaires catégorielles et même grammaires transformationnelles)[14]. Ainsi, ce formalisme peut être vu comme un héritier tout à la fois des courants issus du TAL et des courants de la réforme. Pollard et Sag (1994) s'inscrivent d'ailleurs explicitement dans la lignée de Chomsky (1957) et de son projet de mathématisation.

---

[13] Voir l'article de Béatrice Godart-Wendling dans ce numéro.

[14] En ce sens on peut qualifier ce formalisme de « syncrétique ».

5.1. Les principales caractéristiques des HPSG

Ce qui s'observe, tout d'abord, c'est la volonté de donner des descriptions uniformes des différentes dimensions du langage. L'uniformité de la modélisation se manifeste de deux façons. D'une part, le modèle de toute unité est construit sur le même patron quelle que soit sa taille : un mot (c'est-à-dire une unité du lexique) est représenté de la même manière qu'un syntagme ou qu'une phrase, voire un discours, tous ces objets étant des signes. D'autre part, les propriétés de ces entités, quelle que soit leur nature, sont exprimées dans le même format. Les grammaires intègrent ce qui relève du signifiant, du signifié et de l'usage des expressions linguistiques. Cette multidimensionnalité des objets linguistiques permet de ranger les HPSG dans le cadre des « grammaires de construction »[15].

Comme dans les FUG, les règles de grammaire et les grammaires sont des structures de traits. Mais, l'implication entre structures de traits devenant une opération autorisée, les principes généraux sont aussi des structures de traits. On trouvera ci-dessous une formulation du principe des traits de tête, tirée de (Pollard et Sag, 1987)[16] :

(6)

$$[\mathrm{DTRS}_{\text{structure à tête}}\ [\ ]] \Rightarrow \begin{bmatrix} \mathrm{SYN|LOC|HEAD}\ \boxed{1} \\ \\ \mathrm{DTRS|HEAD\text{-}DTR|SYN|LOC|HEAD}\ \boxed{1} \end{bmatrix}$$

À gauche et à droite du signe d'implication se trouve une structure de traits, et on admet que l'ensemble forme également une structure de traits. Il est dit que si un syntagme a parmi ses constituants (*daughters*, DTRS) une structure à tête (reconnaissable par son type), alors le syntagme et son constituant tête (HEAD-DTR) partagent les mêmes traits de tête. Ces traits de tête se trouvent, selon la version du formalisme considérée ici, dans les traits syntaxiques (SYN) et, parmi ces traits, dans les traits locaux (LOC).

Le modèle mathématique qui autorise de telles représentations est nécessairement très peu contraint. Cela ne peut être autrement si on veut, par un formalisme unique, représenter des dimensions distinctes et hétérogènes. Le formalisme des HPSG s'inspire ainsi largement des systèmes informatiques de représentation des connaissances, définis pour représenter n'importe quelle forme de connaissance. Plus précisément les HPSG se fondent sur des structures de traits typées[17].

5.2. Modèles et grammaires

Un modèle constitue le cadre mathématique dans lequel il est possible d'exprimer l'analyse des faits que l'on observe. Plus

[15] Voir l'article de Yannick Mathieu dans ce numéro.

[16] Pour une description plus détaillée des HPSG, voir l'article de Marianne Desmets *et al.* dans ce numéro.

[17] Pollard et Sag (1994 : 8-9), bien que se prononçant en faveur de la mathématisation, se refusent à donner une véritable définition mathématique des objets qu'ils manipulent. Ils renvoient à différents modèles, entre lesquels ils ne choisissent pas vraiment.

précisément, définir un modèle, cela consiste à se donner une ou plusieurs classes d'objets mathématiques, les occurrences d'objets permettant de donner des représentations du monde réel. Par exemple, un modèle peut déterminer la forme que doivent prendre les différentes grammaires censées décrire chacune une langue différente. Ce n'est d'ailleurs pas nécessairement aux mêmes personnes qu'il revient de définir un modèle et de construire des grammaires[18]. Cependant, le plus souvent, les concepteurs des modèles sont aussi les auteurs des grammaires, ce qui entraîne de la confusion sur ce qu'est un formalisme ou ce qu'est une théorie linguistique.

Il y a aussi de la confusion quant à l'usage du terme *grammaire*. Ainsi, les auteurs des HPSG, à la suite des auteurs des GPSG ou des LFG, emploient ce terme au singulier, identifiant de la sorte grammaire et formalisme. Mais, habituellement, on considère qu'un formalisme autorise une classe de grammaires possibles qui peuvent décrire des langues différentes[19]. Il s'ensuit des formulations un peu étonnantes, comme « une grammaire HPSG », où le terme grammaire prend deux sens différents. Nous préférons quant à nous, sans doute un peu abusivement en regard de la volonté des auteurs, systématiquement employer le pluriel pour désigner le formalisme quand dans l'intitulé de celui-ci se trouve le terme grammaire.

Un formalisme est pour nous, rappelons-le, constitué d'un modèle et du langage dans lequel est exprimé le modèle. Selon qu'un modèle est plus ou moins contraint, c'est-à-dire selon que la classe d'objets qu'il autorise est plus ou moins étroite, il y a un déplacement dans la prise de décision. Un modèle contraint minimise le nombre de décisions à prendre dans l'écriture des grammaires. On peut dire en ce sens que choisir un modèle contraint, c'est affirmer une position sur la structure des langues humaines. Avec le risque, si le modèle est trop contraint, de rendre impossible la représentation de certains phénomènes linguistiques.

À l'inverse, choisir un modèle très peu contraint, comme dans le cas des HPSG, cela s'apparente à ne pas définir de modèle du tout. L'auteur de grammaires dispose d'une très grande liberté, mais il se trouve un peu dans la même situation que s'il travaillait dans un cadre non mathématisé. Les décisions qu'il doit prendre ne se fondent sur aucun critère explicite, ou en tout cas aucun critère propre au modèle. Au mieux, elles peuvent être guidées par des commentaires fournis par les concepteurs des modèles, commentaires évidemment non mathématisés !

---

[18] Le partage du travail défini à propos des systèmes experts se trouve ici affiné. On peut en effet imaginer l'existence de deux types d'experts : d'une part un linguiste « théoricien » qui confectionne le modèle, et d'autre part des linguistes « praticiens » qui élaborent des grammaires décrivant des langues ou des parties de langues.

[19] Pollard et Sag (1987 : 147) définissent des *théories* qui décrivent une langue donnée : une théorie pour l'anglais, une théorie pour le français, etc.

5.3. Qu'est-ce qu'une théorie linguistique ?

Un modèle trop peu contraint n'est pas susceptible de définir une théorie linguistique, puisque à l'intérieur d'un tel modèle toutes les options restent possibles. On pourrait par conséquent donner une nouvelle définition de la théorie linguistique : elle ne résiderait plus dans le choix de l'outil mathématique de représentation, et donc dans le choix des grammaires *possibles*, mais dans les grammaires *effectivement choisies*. Une théorie serait ce que Bès (2002 : 64-65) appelle un « système d'hypothèses ». La théorie ne s'identifie pas au formalisme mais aux hypothèses qui sont écrites dans le formalisme. Et il est vrai que nombre de travaux fondateurs d'une théorie linguistique, en même temps qu'ils introduisent le formalisme qu'ils font leur, indiquent la manière de l'utiliser. Ainsi, Pollard et Sag (1987) décrivent les structures de traits qui sont à la base des HPSG et, dans le même ouvrage, énoncent des principes comme le principe des traits de tête ou le principe de sous-catégorisation, des règles de dominance immédiate, etc.

Il reste cependant en ce cas parfois très difficile de cerner les théories. Si, pour une raison ou pour une autre[20], on modifie le principe de sous-catégorisation des HPSG, tout en conservant les structures de traits typées, sommes-nous toujours dans les HPSG ? Plus globalement, si une théorie est définie par le choix des grammaires à l'intérieur d'un formalisme – très libéral –, y a-t-il des principes, argumentés scientifiquement, clairement énoncés et transmissibles, qui fondent ces choix ? Dit autrement, quel est exactement le système d'hypothèses des HPSG ?

La conséquence, sinon, n'est-elle pas l'identification d'une théorie linguistique aux auteurs de cette théorie et à leurs disciples, autrement dit à une « école », avec ses leaders dotés du pouvoir d'excommunier ? Ou bien l'identification de la théorie au langage de représentation ? La conscience d'appartenir à un même groupe passe en effet par des signes de reconnaissance, dont les notations – qui constituent le langage de représentation – ne sont pas les moins importants.

6. Conclusion

Deux mouvements convergents semblent conduire à un recul de la mathématisation des formalismes syntaxiques : le développement des méthodes de TAL robuste au détriment des méthodes fondées sur la linguistique et la recherche d'un traitement unifié de toutes les dimensions du langage qui entraîne la puissance grandissante des formalismes.

Les arguments en faveur de la mathématisation ne sont toutefois nullement remis en cause. Ce que l'on peut déceler, plutôt, c'est quelque chose comme un abandon à leurs « penchants naturels » des spécialistes du traitement automatique d'un côté et des linguistes non informaticiens de l'autre. Pour les premiers, il s'agit d'aller au plus

[20] Par exemple, comme Reape (1994) pour un traitement de certains phénomènes spécifiques de l'allemand.

vite afin de résoudre des problèmes pratiques, tandis que pour les seconds la mathématisation apparaît quelquefois comme une contrainte inutile, qui les empêcherait de décrire correctement et complètement les phénomènes qu'ils observent.

Il semble donc nécessaire, tout particulièrement dans la période actuelle, de rappeler les enjeux de la mathématisation.

Références

Abeillé A., 1993, *Les nouvelles syntaxes : Grammaires d'unification et analyse du français*, Paris : Armand Colin.

Bès G. G., 2002, La linguistique entre science et ingénierie, *TAL*, 43-3, 57-81.

Bresnan J., 1982, *The mental representations of grammatical relations*, Cambridge : MIT Press.

Chomsky N., 1957, *Syntactic structures*, La Haye : Mouton.

Chomsky N., 1975, Introduction 1973, *The logical structure of linguistic theory*, Chicago : The University of Chicago Press.

Colmerauer A., Kanoui H., Roussel P. et Pasero R., 1973, *Un Système de Communication Homme-Machine en Français*, Groupe de Recherche en Intelligence Artificielle, Université d'Aix-Marseille.

Cori M. et Léon J., 2002, La constitution du TAL, Étude historique des dénominations et des concepts, *TAL*, 43-3, 21-55.

Cori M. et Marandin J.-M., 1993, Grammaires d'arbres polychromes, *TAL*, 34-1, 101-132.

Cori M. et Marandin J.-M., 1994, Polychrome tree grammars (PTGs) : a formal presentation, [C. Martin-Vide, ed.] *Current issues in Mathematical Linguistics*, 141-149, Amsterdam : North-Holland.

Cori M. et Marandin J.-M., 2001, La linguistique au contact de l'informatique : de la construction des grammaires aux grammaires de construction, *Histoire Épistémologie Langage* 23-1, 49-79.

Gazdar G., 1982, Phrase structure grammar, [Jacobson P. et Pullum G., eds] *The Nature of Syntactic Representation*, 131-186, Dordrecht : D. Reidel Publishing Company.

Gazdar G., Klein E., Pullum G. et Sag I., 1985, *Generalized Phrase Structure Grammar*, Oxford : Basil Blackwell.

Harman G.H., 1963, Generative Grammars without Transformation Rules : A Defense of Phrase Structure. *Language* 39, 597-616.

Joshi A.K., 1985, Tree adjoining grammars : How much context-sensitivity is required to provide reasonable structural descriptions, [D.R. Dowty, L. Karttunen, A.M. Zwicky, eds] *Natural language parsing*, Cambridge University Press, 206-250.

Kay M., 1985, Parsing in functional unification grammar, [D.R. Dowty, L. Karttunen et A.M. Zwicky eds.] *Natural language parsing*, Cambridge University Press, 251-278.

Ligozat G., 1994, *Représentation des connaissances et linguistique*, Paris : Armand Colin.

Miller P. H., 1999, *Strong Generative Capacity : The Semantics of Linguistic Formalism*, Stanford : CLSI Publications.

Miller P.H. et Torris T., 1990, *Formalismes syntaxiques pour le traitement automatique du langage naturel*, Paris : Hermes.

Pereira F. et Warren D., 1980, Definite Clause Grammars for Language Analysis - A Survey of the Formalism and a Comparison with

Augmented Transition Networks, *Artificial Intelligence*, 13-3, 231-278.

Pollard C. et Sag I. A., 1987, *Information-Based Syntax and Semantics, Vol. 1: Fundamentals*, Stanford : CLSI Lecture Notes Series.

Pollard C. et Sag I. A., 1994, *Head-Driven Phrase Structure Grammar*, Chicago : University of Chicago Press.

Popper K. R., 1959, *The logic of scientific discovery*, traduction française, 1973, Paris : Payot.

Pullum G. K., 1991, Formal linguistics meets the Boolum, in *The great eskimo vocabulary hoax*, The University of Chicago Press, 47-55.

Reape M., 1994, Domain Union and Word Order Variation in German, [J. Nerbonne, K. Netter et C. Pollard, eds] *German in Head-Driven Phrase Structure Grammar*, Stanford : Lecture Note Series, CSLI, 151-197.

Shieber S. M., 1986, *An introduction to unification-based approaches to grammar*, Stanford : CSLI.

Shieber S. M., 1992, *Constraint-Based Grammar Formalisms*, The MIT Press.

Tesnière L., 1959, *Éléments de syntaxe structural*e, Paris : Klincksieck.

Woods W.A., 1970, Transition Network Grammars for Natural Language Analysis, *Communications of the ACM*, 13 : 10, 591-606.

Yngve V., 1960, A model and an Hypothesis For Language Structure. *Proceedings of the American Philosophical Society*, 104, 444-466.