

[Version préliminaire de la contribution parue dans *Le Modèle*, Amadis 9, Presses de l'université de Bretagne Occidentale, 2011, pp. 331-340.]

Quelques remarques sur la notion de modèle en linguistique

Catherine FUCHS

LATTICE (CNRS/ENS)

Après avoir rappelé ce que l'on entend par 'modèle' en science (§ 1.), j'esquisserai un bref survol historique concernant le recours à la notion de modèle en linguistique (§ 2.), avant de proposer quelques pistes de réflexion autour des enjeux de la modélisation (§ 3.), puis terminerai par la question du statut scientifique de la linguistique (§ 4.).

1. Définition

En science, un modèle est une représentation simplifiée d'un processus ou d'un système, qui permet d'en reproduire ou d'en simuler le fonctionnement dans ses propriétés jugées essentielles. Modéliser, c'est donc abstraire pour pouvoir calculer, expérimenter et prédire.

La modélisation ne peut se faire que sur la base d'une théorie préalable. Le modèle ne se confond pas avec la théorie, qui a permis de dégager les propriétés du système.

En logique, un modèle doit conférer à toute formule de la théorie la valeur 'vraie' (cf. la théorie des modèles d'Alfred Tarski).

2. La notion de modèle en linguistique

En linguistique, le recours à la notion de modèle est assez divers et fluctuant. Le modèle est conçu, tantôt dans un sens technique (correspondant à l'anglais 'model'), tantôt dans un sens non technique (correspondant plutôt à l'anglais 'theory').

2.1. Le modèle au sens technique

L'usage du terme 'modèle' au sens technique revient aux linguistiques dites computationnelles qui, sur la base d'une théorie donnée, recourent à un formalisme de représentation (de nature généralement algébrique) correspondant à une certaine puissance de calcul, en vue d'une implémentation informatique.

Cette tradition remonte à la fin des années 1950 – début des années 1960, qui marque l'émergence des recherches en matière de traitement automatique des langues, et tout particulièrement en vue de la traduction automatique. C'est l'époque où, en France, Gross et Lentin (qui publieront en 1967 leur ouvrage *Notions sur les grammaires formelles*), à la suite de Schützenberger, se penchent sur les liens entre théorie des grammaires formelles et théorie linguistique.

A la même époque, outre-Atlantique, Harris s'efforce de mathématiser les structures de la langue (d'où son ouvrage de 1968 *Mathematical Structures of Language*), et son élève Chomsky s'interroge sur l'adéquation de divers types de grammaires formelles pour modéliser les structures syntaxiques. Réfléchissant sur la parenté entre théorie des grammaires et théorie des automates, Chomsky évalue divers types de grammaires formelles, dans un article fondateur de 1956 ("Three models for the description of language"), où il distingue :

- (a) les grammaires de 'type 3', ou grammaires régulières, à nombre fini d'états (chaînes de Markov) qui permettent de traiter la contiguïté syntagmatique, mais pas les constituants discontinus, les renvois croisés, ni la hiérarchisation des constituants ;
- (b) les grammaires de 'type 2', ou grammaires de constituants (càd. munies d'un axiome, d'un vocabulaire, et de règles de réécriture dont certaines sont récursives) non contextuelles ('context-free') qui sont trop puissantes, car elles engendrent des phrases non grammaticales ;
- (c) les grammaires de 'type 1', ou grammaires de constituants contextuelles ('context-sensitive') qui sont adéquates, mais ne peuvent pas représenter les liens de paraphrase entre structures ni les ambiguïtés de structures (ce qui justifiera l'introduction de la notion de 'transformation') ;
- (d) les grammaires de 'type 0', ou grammaires sans contrainte sur règles de réécriture, qui sont reconnues par une machine de Turing générale, mais sont trop peu structurées pour pouvoir représenter les structures syntaxiques.

En 1959 ("On certain formal properties of grammars"), Chomsky s'intéresse, toujours dans le même esprit, aux propriétés mathématiques de diverses classes de grammaires formelles.

On rappellera que le n° 9 de la revue *Langages*, paru en 1968 sous le titre "Les modèles en linguistique", réunissait des contributions de Gross, Bar-Hillel, Harris, Chomsky et Schützenberger.

Par la suite, on le sait, l'évolution de la théorie chomskienne, qui s'est éloignée de cet objectif initial de computation effective, a conduit au développement d'autres types de grammaires, jugées plus faciles à implémenter. Citons, entre autres, la 'grammaire lexicale fonctionnelle' (LFG) de Bresnan, la 'grammaire syntagmatique généralisée' (GPSG) de Gazdar,

la ‘grammaire syntagmatique guidée par les têtes’ (HPSG) de Pollard et Sag, et la ‘grammaire d’arbres adjoints’ (TAG) de Joshi : voir Abeillé (1993).

Les approches formelles de la langue ont également conduit à élaborer, dès le début des années 1970, des modèles sémantiques conçus à partir de la logique : voir par exemple Keenan (ed., 1975). L’un des pionniers dans ce domaine est le logicien Richard Montague, dont la sémantique intensionnelle constitue une référence essentielle en matière de sémantique formelle. La démarche se fonde sur l’idée que les langages artificiels des logiciens et les langues naturelles relèvent d’une même théorie mathématique et que l’étude de la syntaxe (en l’occurrence l’étude des expressions déclaratives) permet d’accéder ensuite à celle de la sémantique (conçue comme devant rendre compte des notions de vérité et de conséquence). La sémantique intensionnelle qu’il élabore consiste donc à ‘interpréter’ les expressions de la langue, de façon homologue à l’interprétation d’une logique. Cette démarche, dont il établit les principes généraux, est illustrée sur certains problèmes particuliers comme le traitement de la quantification, celui de la coordination, ou encore celui du verbe *être* ou des verbes dits ‘d’attitude propositionnelle’ (*penser, croire, etc.*) en anglais.

Dans ce sens technique, modéliser les structures syntaxiques ou sémantiques de la langue, c’est donc se donner un formalisme qui permet de représenter de façon opératoire les connaissances sur ces structures, connaissances que la théorie linguistique a permis de dégager et de formuler. C’est bien dans ce sens que, par exemple, Sabah utilise en 1988 le terme ‘modèle’ dans le premier volume (intitulé *Représentation des connaissances*) de son ouvrage *L’intelligence artificielle et le langage*. Dans la première partie de ce volume, en effet, il présente successivement, sous la dénomination générale de ‘modèles linguistiques’ : les grammaires formelles, les grammaires transformationnelles, les grammaires de cas, les grammaires systémiques, la grammaire LFG, les grammaires du LADL de Gross et le modèle ‘Sens-Texte’ de Mel’chuk.

2.2. Le modèle au sens non technique

Par différence avec l’usage strict qui vient d’être rappelé, on constate que certains linguistes utilisent parfois le terme ‘modèle’ en un sens plus lâche. Dans un tel emploi, ‘modèle’ équivaut tantôt à ‘méthode d’analyse’ (par exemple, lorsque l’on parle du ‘modèle distributionnel’), tantôt à ‘théorie’ (par exemple, s’agissant du ‘modèle psychomécanique’, qui ne met en œuvre aucun formalisme au sens plein de ce terme).

La citation suivante, extraite d’un article de 2004 de Cadiot & al.(p. 17), est tout à fait révélatrice d’une telle assimilation entre ‘modèle’ et ‘théorie’ : “La section suivante s’attache à

discuter un ‘modèle’ perceptif et pratique (...). Soulignons que *sous le nom de ‘modèle’, nous renvoyons en réalité au choix d’une perspective théorique* sur l’expérience perceptive — opération décisive pour toute linguistique qui penserait trouver là un domaine d’application privilégié, *a fortiori* un de ses fondements.” [mon soulignement, C.F.].

Dans un sens que l’on pourrait qualifier de ‘semi-technique’, on voit aussi le terme ‘modèle’ utilisé parfois pour référer à l’architecture d’ensemble des niveaux de l’analyse linguistique, telle qu’elle peut être utilisée dans une perspective psycholinguistique (de production ou de compréhension) et qui, à ce titre, suppose un minimum de dispositif formel assurant l’articulation entre niveaux. C’est ainsi, par exemple, que dans l’ouvrage intitulé *Production du langage*, paru en 2002 sous la responsabilité de Fayol, sont déclinés dans la première partie, sous la dénomination générale de ‘modèles et composantes de la production verbale’ : les ‘modèles de la production de la parole’, les ‘modèles de rédaction de texte’, et les ‘modèles linguistiques de production’ (grammaires génératives, fonctionnalisme et grammaires cognitives, théorie des opérations prédicatives et énonciatives).

3. Les enjeux de la modélisation

La recherche d’un modèle se définit, on l’a vu, comme la recherche d’un formalisme permettant de représenter les connaissances qu’une théorie donnée permet de spécifier sur un certain objet, en vue de reproduire ou de simuler (informatiquement) les propriétés essentielles de cet objet. En conséquence, l’activité de modélisation engage cinq grandes questions : celle de l’objet à modéliser, celle du niveau de représentation, celle du type de formalisme, celle de la démarche de construction du modèle, et enfin celle du lien à l’informatique.

3.1. L’objet à modéliser

Quelle est la nature de l’objet qu’un linguiste cherche à modéliser ? Cette question peut se décliner de diverses manières, en fonction des objectifs qu’il poursuit, et de la façon dont la théorie linguistique à laquelle il se réfère s’insère dans la problématique plus large du langage et de son lien possible à la cognition (cf. Fuchs ed. 2004).

S’agissant de l’architecture ‘structurelle’ des connaissances sur la langue : peut-on modéliser uniformément l’entier du système de la langue ? ou bien, face à la complexité et à l’hétérogénéité de cet objet, faut-il envisager une pluralité de modèles locaux correspondant à autant de sous-systèmes (càd. des ‘modules’ requérant éventuellement des types de modélisation distincts) ?

S'agissant de l'architecture 'fonctionnelle' des connaissances sur la langue : l'objet construit par le linguiste est-il compatible avec ce que l'on sait de la mise en œuvre effective des connaissances dans la production et/ou la compréhension de textes ? (en particulier, si modules il y a, ceux-ci sont-ils encapsulés ? sont-ils traités séquentiellement ou en parallèle ? interagissent-ils ?) ; par ailleurs, l'objet 'discours' est-il homologue à l'objet 'phrase' ?

Enfin, s'agissant de l'architecture 'neuronale' des connaissances sur la langue : l'objet construit par le linguiste est-il compatible avec ce que l'on sait du traitement des connaissances par le cerveau humain ?

3.2. Le niveau de représentation

La 'représentation' des connaissances linguistiques qui se trouve engagée dans le processus de modélisation se situe nécessairement à un niveau 'métalinguistique'. Or il importe de rappeler que le terme de 'représentation' est, en soi, équivoque, puisqu'il peut renvoyer également aux représentations 'mentales' ou aux représentations 'linguistiques' (dites encore 'textuelles'). La nécessité de distinguer ces trois niveaux de représentation a été très clairement indiquée par Culioli (1990, pp. 21-24) :

- (a) Le niveau 1 est celui des 'représentations mentales' (cognitives), que l'homme construit à partir de ses relations au monde, aux objets, à autrui, de son appartenance à une culture, etc. Aux représentations de ce niveau, le linguiste ne peut avoir d'accès direct.
- (b) Le niveau 2 est celui des 'représentations linguistiques' (les textes), qui sont la trace des représentations du niveau 1 : ce sont donc des représentations au second degré, mais il n'y a pas de relation terme à terme entre les représentations de ces deux types.
- (c) Le niveau 3 est celui des 'représentations métalinguistiques' construites de façon explicite par le linguiste ; c'est évidemment à ce niveau que se situent les représentations à l'œuvre dans l'activité de modélisation.

Autrement dit, « il nous faut un système de représentation qui supporte la généralisation, qui soit robuste, et qui soit dans une relation d'extériorité par rapport à son objet (...) *il nous faut construire un système de représentation qui porte sur ce système de représentation qu'est la langue.* » (Culioli, 1990, p. 23) [mon soulignement, C.F.].

Mutatis mutandis, on retrouve l'équivalent de cette tripartition chez les tenants des grammaires cognitives, selon lesquels les représentations 'conceptuelles' universelles (cf. le niveau 1 *supra*) sont à distinguer des représentations 'sémantiques' variables selon les langues (cf. le niveau 2 *supra*), ainsi que des représentations 'métalinguistiques' recourant à des outils

généraux qui permettent d'appréhender l'invariance derrière les variations (cf. le niveau 3 *supra*).

3.3. Le type de formalisme

En linguistique formelle, on assimile généralement la modélisation à la transcription des structures linguistiques en formules empruntées à un métalangage logico-algébrique.

En linguistique cognitive au contraire, où l'on attend en outre de la modélisation qu'elle soit cognitivement pertinente (càd. qu'elle s'articule de façon plausible avec l'architecture fonctionnelle de l'esprit et/ou l'architecture neuronale du cerveau), les grammaires cognitives se démarquent du métalangage logique pour proposer des représentations iconiques inspirées, de façon d'ailleurs assez lâche, de la topologie. Mais l'aspect véritablement computationnel (au sens technique du terme) demeure chez elles quasi-inexistant.

La question de savoir où trouver les métalangages adéquats reste une question ouverte : faut-il les chercher dans des langages formels existants ? ou bien faut-il essayer d'inventer des systèmes de représentation nouveaux ? ou bien encore, est-il possible de s'en tenir à une métalangue 'naturelle' non formelle (à l'instar de Harris pour la syntaxe, ou bien de Wierzbicka pour la sémantique) ?

Reconnaissons, à tout le moins, avec Culioli (1999, p. 29) que « (...) la difficulté centrale de la formalisation en linguistique ne réside ni dans la formalisation de systèmes algébriques syntaxiques, ni dans l'étude distributionnelle des combinaisons de mots-objets en correspondance ponctuelle avec la réalité extra-linguistique, mais dans le domaine intermédiaire, spécifique des langues naturelles, où il nous faut découvrir sur quels êtres travailler, construire des types de logiques inconnus à ce jour et qui ne fonctionnent sans doute pas de façon homogène, doser la force des concepts, "ces instruments d'effraction", que nous proposent les mathématiques et les adapter à nos fins. Ainsi, nous ignorons les structures mathématiques qui se révéleront adéquates et fécondes : (...) nous aurons vraisemblablement à les inventer avec l'aide du mathématicien, puis, encore plus vraisemblablement, à les "bricoler", du moins dans une première étape. Qu'il soit bien compris que l'on n'importe pas des techniques logico-mathématiques pour les plaquer sur un objet quelconque. (...) Il n'est pas question de puiser dans un stock d'outils, mais de prendre son bien où on le trouve (combinatoire et algèbre, topologie, etc.). »

3.4. La démarche de construction du modèle

Ici encore, une diversité d'options s'ouvre au linguiste soucieux de modélisation : faut-il privilégier, à l'instar des grammaires formelles, la démarche hypothético-déductive ? ou bien préférer, comme le font par exemple certains typologues, une forme d'induction (qui pourrait éventuellement être assistée par ordinateur, sous forme d'un calcul 'bottom-up') ? ou bien encore favoriser, ainsi que le préconisent des chercheurs comme Seiler ou Desclés, une démarche abductive ?

En tout état de cause, l'exercice quotidien du linguiste réside dans les constants allers et retours entre observation et théorisation, ainsi que dans l'effort d'abstraction conduisant à la généralisation : « N'oubliez pas qu'en linguistique nous n'avons pas de ces concepts élaborés au cours des siècles qui résument toute une activité d'idéalisation, par laquelle on se dégage de l'intuition locale ; nous n'avons pas eu une période galiléenne pour, à un moment donné, mathématiser les concepts ; nous n'avons pas de ces instruments et dispositifs expérimentaux qui, sans que nous en ayons conscience, sont presque des concepts ramassés dans un objet technique (un télescope, par exemple, qui résume tout un ensemble de recherches préliminaires). Le linguiste est obligé de travailler de façon rudimentaire : produire des observations, travailler sur des valuations (*c'est la même chose ; c'est différent ; c'est la même chose à telle modulation près ; c'est acceptable ; c'est inacceptable*) ; théoriser pour pouvoir représenter ; retourner aux observations, dans ce va-et-vient indispensable entre l'observation et la théorisation. Mais cela suppose, on le voit, que l'on ne se contente pas de représenter. Il faut que cette métalangue de représentation soit une langue de calcul. Calculer, c'est opérer en dehors de mes interventions subjectives et de mes courts-circuits intuitifs. Calculer, c'est aussi pouvoir s'engager dans ce passage du local au régional, voire au global (...). C'est donc pouvoir décomposer les procédures de généralisation par lesquelles on passe d'une classe de phénomènes à une autre classe, d'une langue à une autre, n'abandonnant jamais la variation empirique dans notre recherche de l'invariance. Car c'est bien d'invariants qu'il s'agit et non pas de grammaire universelle (...) » (Culioli, 1990, p. 23).

3.5. Le lien à l'informatique

Dans le cadre de la modélisation en linguistique, le rapport à l'ordinateur est sujet à des pratiques fort différentes selon les cas.

Pour certains, l'informatique permet simplement de gérer de grandes quantités de données qu'il ne serait pas possible de manipuler manuellement. Dans ce cas, l'ordinateur n'est qu'un outil d'aide : une telle pratique empirique assistée par ordinateur ne relève pas

véritablement de la modélisation au sens strict — sauf à espérer voir émerger, *sponte sua*, certaines propriétés de l'objet.

Pour d'autres, au contraire, l'implémentation informatique d'un modèle mathématique doit permettre, non seulement de tester des hypothèses théoriques, mais aussi de mettre au jour à partir du modèle certaines propriétés que l'on pourrait ensuite tester expérimentalement — donc de prédire certains faits de langue nouveaux : c'est ici le va-et-vient entre modèle et hypothèses qui est privilégié.

Entre ces deux extrêmes, nombre de chercheurs recourent à l'informatique dans le cadre d'activités de traitement automatique de la langue et d'ingénierie linguistique, où l'implémentation permet avant tout de contrôler le traitement de données langagières.

4. Le statut scientifique de la linguistique

Au stade actuel de développement de la linguistique, la multiplicité des théories et des (pseudo-)modèles — bien souvent incomplets et non implémentés — pose problème, car ces théories et modèles concurrents sont, de fait, incomparables entre eux et non cumulatifs. Pour citer à nouveau Culioli (1999, p. 21) : « Les modèles sont-ils équivalents, compatibles ? Les représentations sont-elles isomorphes ? Tel mode de représentation est-il opératoire, càd, sait-on l'utiliser pour calculer ? Ici devrait se greffer une théorie de l'approximation qui permettrait d'évaluer la force et la régionalité d'un modèle. (...) Formaliser devrait amener à reconnaître qu'aucun modèle n'est exhaustif et à en tirer les conséquences scientifiques ».

Il serait grand temps, en effet, que la linguistique procède, comme le font les spécialistes de traitement automatique, à une mise à plat et à une évaluation comparative de ces (pseudo-)modèles. Mais il est difficile de comparer des modèles qui ne s'intéressent pas aux mêmes phénomènes, sauf s'il s'agit de modèles proches.

Pour conclure, une question plus fondamentale encore me semble se poser, en amont de ces divers types d'invocation de la notion de modèle. Cette question est la suivante : la linguistique actuelle a-t-elle atteint un statut scientifique suffisant pour que la modélisation soit une question pertinente ? En d'autres termes, la linguistique se trouve-t-elle encore à un stade pré-galiléen (cf. la citation de Culioli *supra* : « nous n'avons pas eu une période galiléenne pour, à un moment donné, mathématiser les concepts »), ou à un stade seulement « localement galiléen » (comme le prétend Chomsky) ? Ou, pour reprendre la formule que Lazard emprunte, dans son ouvrage de 2006, à Granger (1987) : la linguistique n'est-elle encore qu'une « proto-science » ? Si tel est le cas, que doit-on en conclure : que nous avons assez, trop ou trop peu de

données ? que nous avons assez, trop ou trop peu de théories ? ou bien que c'est le statut même de la modélisation en linguistique qu'il s'agit d'éclaircir ?

En un mot, la linguistique aurait certainement besoin de plus d'épistémologie.

Références bibliographiques

- Abeillé, Anne (1993) : *Les nouvelles syntaxes : grammaires d'unification et analyse du français*, Paris, Colin.
- Chomsky, Noam (1956) : "Three models for the description of language", *I.R.E Transactions on Information Theory*, II : 2,113-114. trad.fr. 1968 : "Trois modèles de description du langage", *Langages*, 9, 51-76.
- Chomsky, Noam (1959) : "On certain formal properties of grammars", *Information and Control*, 2, 137-167.
- Cadiot, Pierre, Franck Lebas & Yves-Marie Visetti (2004) : "Verbes de mouvement, espace et dynamiques de constitution", *Histoire, Epistémologie, Langage*, 26 : I, 7-42.
- Culioli, Antoine (1990) : "La linguistique, de l'empirique au formel", *Pour une linguistique de l'énonciation*, tome 1, Paris, Ophrys, 9-46 ; trad. angl. "Three levels of representation", *Cognition and Representation in Linguistic Theory*, Amsterdam, Benjamins, 21-31.
- Culioli, Antoine (1999) : "La formalisation en linguistique", *Pour une linguistique de l'énonciation*, tome 2, Paris, Ophrys, 17-29.
- Fayol, Michel (ed.) (2002) : *Production du langage*, Paris, Hermès.
- Fuchs, Catherine (ed.) (2004) : *La linguistique cognitive*, Paris, Ophrys.
- Granger, Gilles Gaston (1987) : *Leçon inaugurale faite le 7 mars 1987*, Paris : Collège de France.
- Gross, Maurice (ed.) (1968) : "Les modèles en linguistique", *Langages*, 9.
- Gross, Maurice & André Lentin (1967) : *Notions sur les grammaires formelles*, Paris, Gauthier-Villars.
- Harris, Zellig (1968) : *Mathematical Structures of Language*, New-York, Wiley. trad.fr. 1971 : *Structures mathématiques du langage*, Paris : Dunod.
- Keenan, Edward (ed.) (1975) : *Formal Semantics of Natural Language*, Cambridge : Cambridge University Press.
- Lazard, Gilbert (2006) : *La quête des invariants inter-langues*, Paris, Champion.
- Sabah, Gérard (1988) : *L'intelligence artificielle et le langage*, tome 1 ("Représentation des connaissances"), Paris, Hermès.