

---

## Notes de lecture

Rubrique préparée par Denis Maurel

*Université François Rabelais Tours, LI (Laboratoire d'informatique)*

---

**Gilles COL. Construction du sens : un modèle instructionnel pour la sémantique. Peter Lang. 2017. 294 pages. ISBN : 978-3-034-32570-7.**

Lu par **François NEMO**

*Université d'Orléans – Laboratoire ligérien de linguistique*

*Longtemps considérées séparément, les questions de la construction du sens des énoncés à partir des éléments qui la composent et celles de la polysémie des unités sémantiques sont abordées chez Gilles Col comme deux faces d'une seule question. Parler d'énoncés plutôt que de phrases tient aussi à ce que les contraintes prosodiques sont intégrées à l'analyse comme des contraintes linguistiques à part entière, ce qui rend caduque la distinction en question. Le modèle proposé est un modèle à base d'instructions sémantiques et de mises en relation de ces instructions. Cette mise en relation se fait progressivement et reste modifiable par tout élément séquentiellement plus tardif. Ce processus dynamique produit de plus un profilage de l'interprétation de chaque instruction qui est à l'origine de la polysémie des unités concernées.*

Comme son titre l'indique très explicitement, le livre de Gilles Col présente la particularité d'aborder comme inséparables la question de la polysémie des unités linguistiques et celle de la construction du sens des énoncés et d'en proposer un traitement unifié dans la cadre d'une sémantique instructionnelle.

Si poser que « *l'omniprésence de la polysémie dans les langues force à considérer la compréhension des énoncés comme un processus dynamique, dans lequel le sens de chaque unité linguistique se détermine en même temps que se construit le sens global de l'énoncé* » peut sembler aller de soi, force est de constater que les deux questions ont jusqu'ici été très largement conçues comme séparées, les modèles dédiés à l'étude de la compositionnalité du sens ayant très largement ignoré la question de la polysémie alors que les théories de la polysémie n'ont que rarement explicité la façon dont elles envisageaient la construction du sens au niveau des énoncés.

Sur cette base, l'ouvrage interroge la pertinence de la notion d'instruction du point de vue terminologique (instruction vs procédure vs consigne), théorique (relation avec les espaces mentaux et les formes schématiques) ou encore en termes de modélisation de la construction du sens et des scènes verbales. Sa deuxième partie (p.109-186) propose une description instructionnelle d'unités spécifiques de l'anglais, en l'espèce les unités *over*, *by*, *until* et *on*, alors que la troisième partie

(p.187-271) décrit et illustre le traitement des instructions dans les énoncés et, en particulier, la question de l'ordre de traitement, sur la base d'un répertoire instructionnel des unités de l'anglais. Cette logique de construction de la scène verbale comme interprétation se voit illustrée graphiquement et formalisée autour de deux exemples « *He was standing by the window* » et « *the children have already finished their lunches* », du traitement de *over* et de l'exemple « *j'ai vu un tableau dans la pièce* ». Le modèle proposé inclut une description détaillée de la façon dont la prosodie intervient dans la construction du sens de l'énoncé.

La notion de polysémie donne lieu à une présentation centrée sur la notion de proxémie, illustrée par l'adjectif *blanc* et le verbe *appear*, au travers de l'étude de cliques. Celle-ci inclut toutes les expressions contenant *blanc* telles que *carte blanche*, *cousu de fil blanc*, *de but en blanc*, mais aussi les notions associées à *blanc* comme l'innocence (*blanc comme neige*). Elle illustre la façon dont les différents emplois de *appear* en viennent à se répartir sur trois axes, la description du faisceau de relations important moins que la compréhension de ce qui l'a produit, à savoir la façon dont *appear* est utilisé dans des constructions où son association avec *to* ou *that* va peser sur son interprétation. La polysémie induite par la compositionnalité tiendra alors à ce que « *une unité peut avoir un sens sous-spécifié et non ambigu : son sens plein sera dérivé du sens des mots connectés par elle* » (p. 120-121) et variera selon la nature de ceux-ci.

La notion de compositionnalité qui est introduite, à savoir une compositionnalité gestaltiste, concerne la relation de dépendance entre le sens d'un énoncé et celui des éléments qui le composent, mais tient compte du caractère descendant du calcul du sens de chacun de ceux-ci lié à la polysémie. On a donc un modèle de la compositionnalité en rupture complète avec le parallélisme sémantico-grammatical de la plupart des approches classiques, puisqu'il ne s'agit pas de combiner séquentiellement et syntagmatiquement des sens préétablis, mais bien de construire un système de contraintes à partir de celles imposées par chaque unité. Et ce, en amont de toute forme spécifique de satisfaction de chacune d'elles, à la manière dont les éléments d'une charade posent chacun une contrainte qui ne sera *in fine* satisfaite que lorsqu'il sera possible de satisfaire également à l'ensemble des contraintes.

La compositionnalité « gestaltiste » décrite dans l'ouvrage se distingue aussi des formes plus classiques de compositionnalité sémantique par le fait de ne pas être séquentielle, mais différée et aussi par le fait que les contraintes apportées par chaque unité (instructions) soient modifiées par la mise en relation avec d'autres contraintes.

Dans le premier cas, si l'introduction des contraintes sémantiques à satisfaire se fait séquentiellement par les instructions associées à chaque unité, la satisfaction de ces contraintes ne se fait pas, quant à elle, séquentiellement, et peut rester en partie suspendue jusqu'à l'introduction et la satisfaction de l'ensemble des contraintes. Il y a à la fois possibilité de différer l'interprétation finale et possibilité de construire dès le début des représentations extrêmement détaillées en les modifiant ensuite si besoin est. De ce fait, « *une fois évoqués, les éléments de la scène en construction continuent d'être déterminés, voire transformés lors de convocations ultérieures par d'autres unités* ». De même, l'ordre de traitement des instructions associées à

chaque unité n'est pas l'ordre de la phrase, les unités introduites pouvant convoquer des éléments non encore disponibles.

Dans le second cas, à la manière où une équation autonome, quand elle s'intègre à un système d'équations, voit ses formes de satisfaction modifiées par la mise en relation avec d'autres équations, la mise en relation de plusieurs instructions conduit à imposer des formes de satisfaction plus étroites et plus spécifiques de chacune d'elles. Ceci rend inévitable que, d'un emploi à l'autre, ses formes spécifiques soient divergentes et explique très directement l'existence d'une polysémie par profilage, qui n'est que le nom de la diversité des formes de satisfaction d'une contrainte unique quand celle-ci est associée à des contraintes différentes et dans différents emplois.

Dans le cadre des sémantiques instructionnelles, la position de Gilles Col est assimilable à une conception lexématique, là où d'autres approches ont pu développer des conceptions beaucoup plus morphémiques.

Cette différence est cependant moins théorique qu'empirique, puisque tout indique qu'il s'agit de réalités sémantiques complémentaires, les approches les plus morphémiques étudiant la contrainte (par exemple, une équation) avant et indépendamment de sa mise en relation avec d'autres contraintes (un système d'équations), les approches lexématiques prenant acte de cette mise en relation comme niveau sémantique stable.

Pour pratiquer dans mes propres travaux la première approche, sur des objets sémantiques encore plus transcategoriels, je ne peux que souligner ici l'intérêt de ce second point de vue et donc la coexistence de deux niveaux instructionnels.

Si les deux approches définissent identiquement la compositionnalité sémantique en termes de compositionnalité des contraintes et non, comme les approches compositionnelles classiques, en termes de compositionnalité des formes de satisfaction de ces contraintes, il ne faut pas opposer le fait d'étudier la façon dont un signe peut se trouver employé dans les contextes categoriels et constructionnels les plus variés (approche morphémique) et le fait de prendre acte que, dans chaque type d'emploi, un profilage spécifique des contraintes instructionnelles puisse se lexicaliser (approche lexématique).

Adopter une approche morphémique des instructions sémantiques implique d'adopter comme principe heuristique qu'une unité n'anticipe pas sa mise en relation avec d'autres unités, alors qu'adopter une approche lexématique des instructions sémantiques implique à l'inverse de prendre acte de cette mise en relation et des formes de stabilisation sémantique qu'elle implique.

Sachant l'échec massif des tentatives de modélisation de la compositionnalité en termes de combinaisons du sens des unités impliquées, toute modélisation, comme celle que propose Gilles Col, qui :

- sépare les contraintes à satisfaire des formes de satisfaction de ces contraintes ;
- conçoit chaque contrainte comme une instruction ;
- conçoit la compositionnalité comme l'intégration d'une contrainte à un système de contraintes ;

– conçoit la polysémie comme l’effet de la mise en relation d’une contrainte avec d’autres contraintes ;

– reconnaît le caractère pleinement linguistique des contraintes prosodiques ;  
constitue une avancée importante dans la bonne direction.

De la même façon que la syntaxe, après avoir longtemps tenté de construire une approche strictement catégorielle, a finalement été conduite à reconnaître la personnalité des unités syntaxiques, la modélisation du rapport entre syntaxe et sémantique, dans l’étude de la compositionnalité, passe par une pleine reconnaissance du caractère individuel des instructions associées à chaque unité sémantique, et par une meilleure compréhension de la façon dont ces instructions sont associées entre elles au niveau de l’énoncé, et ce, à toutes les étapes du processus dynamique de construction du sens.

---

**Manfred STEDE, Jodi SCHNEIDER. Argumentation Mining. Morgan & Claypool publishers. 2019. 175 pages. ISBN : 978-1-681-73461-3.**

Lu par **Aurélien MAX**

*Université Paris-Sud – Limsi*

---

*Les manières dont le TAL peut aborder des questions en lien avec l’argumentation sont nombreuses, et l’ouvrage Argumentation Mining se propose d’en rendre compte. L’ouvrage vise à apporter une vue d’ensemble représentative des travaux ayant abordé l’un des aspects en lien avec le traitement de l’argumentation au travers d’une structuration pertinente pour accompagner les chercheurs souhaitant découvrir ce domaine. C’est selon cet objectif que nous proposons notre regard sur l’intérêt de cet ouvrage.*

L’ouvrage présente une structuration relativement fine divisée en neuf chapitres. Le chapitre 1 présente une introduction autour des sujets touchant à l’argumentation. Les auteurs retiennent comme point de départ la définition synthétique de van Eemeren et Grootendorst (2004) : « *Argumentation is a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of propositions justifying or refuting the proposition expressed in the standpoint.* », en insistant sur le fait que parfois le point de vue (*standpoint*) peut être laissé implicite, ce qui constituera une difficulté particulière pour les traitements automatiques. De plus, en vue d’influencer le point de vue du lecteur, il peut être pertinent d’anticiper les positions que celui-ci est susceptible de tenir, et l’organisation de la constellation d’arguments, dans ce qui peut prendre la forme d’une structure argumentative, peut se montrer plus ou moins efficace (champ de la rhétorique).

Le chapitre 2 porte sur les relations entre langue et argumentation et couvre de nombreuses notions présentées de manière digeste et illustrées : la subjectivité (incluant les notions d’opinion et d’émotion), les liens entre opinions, points de vue et arguments, les liens entre arguments et actes de langage (et donc, la recherche sur la persuasion), les liens entre l’argumentation et la justification et l’explication (nécessité de convaincre ou non). Le chapitre se termine par des éléments de

rhétorique concernant la construction de phrases et l'organisation des textes. Les dimensions du langage couvertes montrent la grande complexité de l'analyse argumentative du langage.

Le chapitre 3 aborde la question des modèles structurels pour l'argumentation. Sont présentés les composants principaux des arguments : des thèses soutenues (*claims*) accompagnées d'éléments en soutien à ces thèses, et possiblement des attaques contre ces éléments ainsi que leurs réfutations. Un lien essentiel est établi avec la question de l'inférence : alors que l'inférence logique garantit la validité d'une conclusion lorsque les prémisses sont vraies, une argumentation vise à établir des conclusions pouvant être révisées à la lumière de nouvelles informations (*defeasible reasoning*). Les types de modèles argumentatifs sont repris de Bentahar, Moulin et Bélanger (2010) et peuvent être de trois types :

- monologiques (*monological models*), où les arguments sont des preuves à établir, et les thèses et prémisses sont assemblées en une structure interne pour chaque argument ;
- rhétoriques (*rhetorical models*), où les arguments sont liés à la perception de l'auditoire, et les arguments sont agencés dans une structure rhétorique ciblant la persuasion des arguments ;
- dialogiques (*dialogical models*), où les arguments sont considérés comme prenant part à un raisonnement révisable et sont agencés dans une macrostructure dialogique.

Le chapitre 3 se conclut avec une description détaillée des cartographies argumentatives (*argument maps*), outils visuels qui contribuent à appréhender des argumentations complexes, en abordant les questions de leurs applications, de l'outillage existant, ainsi que des possibilités de représentations informatiques (en particulier, le langage XML Argument Interchange Format (AIF)), de leur stockage et de leur requêtage.

Le chapitre 4 aborde la question des ressources annotées pour les travaux en lien avec l'argumentation. Un certain nombre de schémas d'annotation sont brièvement décrits, puis une liste de corpus (essentiellement sur l'anglais) est fournie.

La tâche de fouille d'arguments (*argument mining*) est décomposée par les auteurs en traitements qui font l'objet des chapitres 5 à 8 : l'identification de fragments de textes argumentatifs, leur segmentation en unités de discours argumentatives (*argumentative discourse units (ADU)*), l'identification d'une thèse principale (*central claim*), l'identification du rôle des ADU (chapitre 5), l'identification des relations entre ADU (chapitre 6), la construction d'une structure complète (chapitre 7), et enfin l'évaluation de la qualité d'une argumentation (chapitre 8).

Le chapitre 5, qui aborde la question du repérage de thèses exprimées dans les textes, considère tout d'abord la classification automatique de textes en textes argumentatifs ou non argumentatifs au niveau des documents, des phrases et des segments sous-phrastiques. Concrètement, l'expression d'une thèse peut s'étaler sur plusieurs phrases, jusqu'à ne correspondre qu'à un segment de phrases, ce qui fait

retenir par les auteurs comme définition pour les ADU : « *a span of text that plays a single role for the argument being analyzed, and is demarcated by neighboring text spans that play a different role, or none at all* ». En pratique, rares sont les travaux qui abordent directement le problème de la segmentation en ADU, le choix le plus répandu étant de reposer sur le niveau des phrases. La partie principale du chapitre porte sur l'identification de thèses exprimées, une problématique attirant de nombreux travaux. La description couvre tour à tour le cas de textes juridiques (les conclusions jouant le rôle de thèse), de textes éducatifs (instructions à faire ou ne pas faire quelque chose), de dissertations scolaires (opinion exprimée dont on cherche à convaincre le lecteur de la validité), de textes encyclopédiques (Wikipédia), et enfin de textes issus du Web et des réseaux sociaux, décrivant différentes situations concrètes dans lesquelles un auteur explicite ce qu'il veut que son lecteur croie.

Le chapitre 6 porte sur le repérage d'éléments venant à l'appui (*support*) ou au contraire en opposition (*objection*) avec une thèse. Le repérage d'éléments à l'appui peut se fonder sur des indicateurs de relations discursives (tels que certains travaux en ont recensé), ou bien faire l'objet d'une classification de phrases en isolation pour laquelle des mesures de similarité sémantique semblent être de bons indicateurs. Enfin, quelques travaux ont abordé le repérage joint de thèses et de leurs éléments à l'appui, notamment sous forme d'analyse de séquences. Les auteurs abordent également ici la question pertinente de l'identification de position (*stance*) dans laquelle il faut déterminer si une unité de texte est en faveur, contre ou sans lien avec une thèse, avec des exemples de tâches sur des données issues de réseaux sociaux, de débats sur Internet et de dissertations scolaires.

Le chapitre 7 aborde la question de la détermination de relations explicites entre éléments précédemment identifiés. Les auteurs décrivent tout d'abord des travaux de TAL visant à établir de telles relations, puis la question bien plus complexe de la construction de structures argumentatives, qui peuvent être hiérarchiques ou séquentielles. Les auteurs remarquent qu'aucun consensus n'existe vraiment sur le type de description approprié pour de telles structures, ce qui est reflété au travers des travaux décrits. On notera une grande diversité d'approches essayées, avec une importante densité de travaux récents.

Le chapitre 8 est consacré à trois questions d'approfondissement correspondant à des tâches moins abordées que les précédentes. La reconstruction d'enthymèmes correspondant à des prémisses et laissés implicites car supposés connus, est un problème complexe n'ayant à ce jour attiré que peu de travaux d'automatisation. L'identification de schémas argumentatifs, visant à spécifier un motif inférentiel sous-jacent permettant d'établir des relations d'appui ou d'attaque entre ADU, a, quant à elle, fait l'objet d'études d'annotation manuelle. Ces tâches d'annotation mettent toutefois en évidence un large rôle des jugements subjectifs, qui rendent complexe l'obtention d'annotations consensuelles, et peuvent expliquer le faible nombre de travaux ayant cherché à automatiser ces annotations. La troisième question traitée dans le chapitre est celle de l'évaluation de la qualité argumentative. Les auteurs insistent tout d'abord sur l'importance d'analyses discursives qui iraient au-delà des formes linguistiques de surface, et abordent la question de la vérification

de la cohérence argumentative en décrivant des travaux à l'interface du TAL et des solveurs argumentatifs (pour lesquels existe une compétition internationale se tenant tous les deux ans). Sur la question très vaste de l'estimation de qualité, les auteurs reprennent une décomposition logique, rhétorique et dialectique, qu'il est toutefois difficile de détacher d'un but particulier (tel que persuader un auditoire, résoudre un conflit, obtenir un consensus, recommander une action, etc.). Les travaux décrits couvrent trois approches : la focalisation selon une dimension de la qualité et un genre textuel, la comparaison de deux arguments selon certains aspects de la qualité et la séparation de l'évaluation de la qualité en des dimensions pouvant être abordées séparément. Enfin, les auteurs abordent le rôle controversé que peuvent jouer les émotions en lien avec l'argumentation, menant à des tâches visant par exemple à déterminer la possibilité d'une évolution des croyances d'un lecteur si des arguments particuliers lui étaient présentés.

Le chapitre 9 s'écarte, quant à lui, du titre de l'ouvrage puisqu'il aborde la question, jusqu'ici très peu traitée, de la génération automatique de textes argumentatifs. Une première partie décrit des travaux historiques en génération de textes à partir de données, puis une seconde partie traite de la génération à partir de textes existants, où le repérage et la synthèse d'arguments peuvent se rejoindre. Les auteurs notent que la génération de textes argumentatifs par réutilisation de textes existants est bien évidemment limitée par la performance de ce qui peut être déduit automatiquement des segments réutilisés, ce qui est donc l'objet des chapitres précédents.

Le dernier chapitre permet aux auteurs de faire trois contributions notables. Tout d'abord, ils proposent un noyau pour la tâche de fouille d'arguments qui reprend les principaux enseignements des chapitres précédents, offrant ainsi une vue synthétique originale et bienvenue à la fin de l'ouvrage. Puis ils décrivent des cas d'utilisation faisant suite à la fouille d'arguments, et introduisent brièvement une variété d'applications. Finalement, les auteurs livrent leur point de vue sur le domaine : le traitement des arguments en TAL est un domaine en pleine évolution dont les avancées à venir trouveront un rôle dans de nombreuses activités humaines. On retiendra particulièrement l'un des derniers messages : certains travaux de la discipline pourront être utilisés à des fins d'influence des individus, et les chercheurs ne peuvent se détourner de la question des aspects éthiques de l'application de leurs travaux.

Évaluer la durée de vie de l'ouvrage (qui ne recense pas moins de 300 références, dont un certain nombre de 2017 et de début 2018) est une prédiction difficile, les auteurs souhaitant eux-mêmes que l'ouvrage devienne rapidement obsolète. Il nous semble néanmoins que la couverture et la structure des sujets traités offrent une vue d'ensemble qui s'avérera très utile à quiconque voudrait comprendre les enjeux et difficultés de la question traitée, ainsi que les travaux abordés jusqu'au début de 2018. Si certains passages requièrent des connaissances récentes en TAL, les aspects langagiers et applicatifs intéresseront également un public plus large. On retiendra, plus globalement, une famille de traitements de TAL aux difficultés d'annotation et à la performance relativement limitée, ainsi que des travaux

principalement abordés sur l'anglais, mais qui devraient, selon la prédiction des auteurs, connaître de nombreuses évolutions dans les années à venir.

---

**Anders SØGAARD, Ivan VULIĆ, Sebastian RUDER, Manaal FARUQUI. Cross-Lingual Word Embeddings. Morgan & Claypool publishers. 2019. 120 pages. ISBN : 978-1-681-73572-6.**

Lu par **Caio Filippo CORRO**

*Université Paris-Sud – Limsi*

---

*Cet ouvrage résume la littérature sur l'apprentissage des plongements de mots cross-lingues. Il décrit à la fois l'utilisation de différentes sources de supervision (c'est-à-dire les différentes sources de données alignées) et les méthodes de projection des plongements de mots d'une langue à une autre. Un chapitre est également consacré aux méthodes non supervisées, c'est-à-dire qui ne requièrent pas de données alignées.*

L'apprentissage supervisé de réseaux de neurones requiert de grands jeux de données souvent non disponibles en dehors de quelques langues dites très dotées. Un enjeu majeur est donc de pouvoir construire des modèles multilingues pouvant faire profiter les langues peu dotées des ressources existantes pour d'autres. En traitement automatique des langues, l'entrée de ces réseaux est des plongements de mots, c'est-à-dire des représentations sous forme de vecteurs du vocabulaire. Il est donc intéressant d'explorer la construction de plongements de mots qui permettraient de transférer les réseaux de neurones d'une langue à une autre. Cet ouvrage propose un résumé de l'état de l'art sur la construction de ces plongements de mots *cross-lingues*, c'est-à-dire sur la construction de plongements de mots où plusieurs langues partagent le même espace vectoriel. Intuitivement, l'objectif souhaité est que le plongement du mot français « *chat* » soit similaire à celui du mot anglais « *cat* ». Cette direction de recherche ouvre trois grandes questions :

1. Comment construire ces espaces vectoriels ?
2. Comment les évaluer ?
3. Comment les utiliser en pratique ?

L'ouvrage se concentre principalement sur la première problématique, les deux autres étant traitées succinctement dans le chapitre 11.

La construction de ces plongements de mots *cross-lingues* se fait généralement en deux étapes. D'abord, des plongements de mots monolingues sont générés indépendamment pour chaque langue. Ensuite, une fonction de projection est apprise pour aligner une langue cible sur une langue source. On parle d'approches supervisées lorsque cette seconde étape s'appuie sur un dictionnaire de traduction, et, inversement, d'approches non supervisées si elles déduisent automatiquement un tel dictionnaire. Notons que les méthodes non supervisées peuvent bénéficier de dictionnaires créés *via* des heuristiques très simples, par exemple en définissant des entrées contenant les mots ayant la même graphie dans deux langues (« *week-end* »

en français et en anglais – évidemment cela peut introduire du bruit comme pour le mot « *lecture* » dans ces mêmes langues) ou les nombres.

Les auteurs proposent une typologie des méthodes de construction (chapitre 3) suivant deux axes : le type de données alignées utilisé pour la supervision et la méthode de régularisation pour l'apprentissage de la fonction de projection. Cette exposition est pertinente, car elle donne en deux tableaux une vue d'ensemble de la littérature. Il est intéressant de remarquer que le type d'alignement des données bilingues a souvent plus d'impact sur les performances pour les applications cibles que la méthode de régularisation.

Le chapitre 2 contient un très bref rappel des méthodes d'apprentissage des plongements de mots monolingues. Ces plongements de mots servent de base pour l'approche *cross-lingues*. Avant de s'attaquer au cœur du sujet, les auteurs proposent dans le chapitre 3 une méthode de classification des approches *cross-lingues* (voir plus haut). Ensuite, un historique du développement des méthodes *cross-lingues* utilisant le classement est exposé dans le chapitre 4. En trois chapitres, le lecteur a donc une vue d'ensemble des méthodes développées dans la littérature. Les chapitres suivants peuvent ensuite être lus indépendamment, en fonction des intérêts de chacun.

Les chapitres 5 à 7 se focalisent sur les différentes sources d'alignement utilisées pour apprendre la fonction de projection de façon supervisée en se concentrant sur les plongements de mots bilingues. Le chapitre 5 s'intéresse aux méthodes fondées sur l'alignement de mots, c'est-à-dire en utilisant des dictionnaires bilingues. Cette approche est probablement la plus populaire, car elle repose sur l'existence de données facilement accessibles pour beaucoup de langues, ressources existantes même pour des langues généralement peu dotées. Cependant, elle peut être limitée, car il est difficile de prendre en compte des phénomènes comme la compositionnalité ou la polysémie avec cette source de supervision. Le chapitre 6 se focalise donc sur l'utilisation de corpus de traduction de phrases : plutôt que d'apprendre explicitement une projection des plongements de mots d'une langue à une autre, cette projection est apprise implicitement afin qu'une phrase et sa traduction aient la même représentation vectorielle. Si cette approche est très intéressante, elle n'est cependant possible que pour les langues très dotées qui possèdent ce genre de corpus. Lorsque ces ressources ne sont pas disponibles pour une paire de langues, une approche plus générale décrite dans le chapitre 7 consiste à utiliser des documents alignés, en particulier provenant de Wikipédia.

Après s'être concentrés sur l'apprentissage de projections bilingues, les auteurs s'intéressent aux méthodes multilingues dans le chapitre 8. Notons que la motivation des méthodes multilingues n'est pas uniquement de pouvoir utiliser un unique modèle pour résoudre une tâche applicative dans beaucoup de langues, mais est aussi justifiée par le fait qu'un modèle bilingue peut être non adapté pour certains couples de langues. En effet, les langues naturelles diffèrent dans leurs constructions (voir les travaux effectués en typologie linguistique) et leurs ambiguïtés. En utilisant des modèles multilingues, il est alors possible d'avoir des alignements sur plusieurs langues en fonction des constructions partagées. Il existe deux sortes d'approches.

D'une part, les approches qui choisissent une langue pivot qui fait office de pont avec les autres, d'autre part, les approches qui mixent les langues entre elles.

Enfin, le chapitre 9 décrit les méthodes non supervisées, c'est-à-dire qui n'utilisent pas de données alignées. Ces méthodes commencent par induire un dictionnaire puis utilisent ce dictionnaire comme source d'alignement pour une méthode « supervisée ». Les différents travaux diffèrent sur la méthode de création et de raffinement des dictionnaires.

Cet ouvrage présente de façon claire et concise les travaux portant sur la construction de plongements de mots *cross-lingues*. On regrettera cependant l'absence totale de linguistique : il est surprenant de traiter du mot sans parler de morphologie. Les auteurs soulignent d'ailleurs eux-mêmes ce manquement, en conclusion, qui est dû au peu de travaux réalisés par la communauté sur ce sujet (chapitre 12).

---

**Alexandra Y. AIKHENVALD. The Art of Grammar. A Practical Guide. Oxford University Press. 2014. 408 pages. ISBN : 978-0-199-68321-5.**

Lu par **Liliane HODIEB**

*Inalco*

---

*La documentation linguistique est une discipline qui suscite de plus en plus d'intérêt et de recherches. Chaque langue est porteuse d'une vision unique du monde. Aussi, étant donné la vitesse à laquelle les langues tendent à disparaître, il devient urgent de s'y atteler. Documenter une langue consiste à fournir une description de la langue, accompagnée de textes – entre autres narratifs et épiques – d'extraits de conversations et d'un lexique. Pour l'auteur, décrire une langue est non seulement une entreprise scientifique, mais c'est aussi un art, dont les techniques doivent être maîtrisées pour que le travail produit soit satisfaisant. Ce sont ces techniques que l'auteur présente dans cet ouvrage, en mettant un point d'honneur sur l'importance et l'inévitabilité des aspects socioculturels dans la grammaire. En quinze chapitres, les différentes parties que doit comporter une grammaire de référence sont explicitées. La grammaire de référence, loin d'être une fin en soi, représente un tremplin pour la documentation des langues, notamment la production de documents pour l'éducation et l'alphabétisation, de dictionnaires, de lexiques, de vidéos et de ressources multimédias.*

Pour le linguiste, la grammaire de référence est la ressource la plus importante, car elle présente de façon détaillée le système linguistique, ses spécificités, son « génie ». Elle ne se contente pas de décrire le fonctionnement de la langue, mais elle explore également l'histoire, la géographie et l'organisation de la société dans laquelle la langue est parlée. Contrairement à une esquisse de grammaire qui fournit les éléments de base du système linguistique sans apporter de justifications ou d'explications, la grammaire de référence analyse les faits. De plus, elle justifie les choix effectués dans les analyses. Encore appelées grammaires « descriptives » ou « analytiques », les grammaires de référence constituent, en outre, le document sur lequel se fondent d'autres grammaires, notamment les grammaires historiques, prescriptives et pédagogiques. Elles sont par ailleurs la fondation de la linguistique

typologique qui se sert de ces grammaires pour faire des généralisations et des prédictions sur les langues. Les linguistes – « typologues » et « généraux » – sont par conséquent les premiers destinataires des grammaires de référence. Les anthropologues et les locuteurs de la langue en question font également partie de l'audience.

Le plan du livre reflète le plan d'une grammaire de référence. Après avoir présenté la langue et son contexte – les locuteurs, les langues proches sur le plan géographique et linguistique, la situation sociale, etc. – les éléments fondamentaux du système linguistique sont exposés. Il s'agit de la structure et du fonctionnement des sons (la phonétique et la phonologie), des catégories grammaticales, des relations grammaticales (la morphologie), des propositions, des syntagmes et des phrases (la syntaxe). Un mot sur l'usage de la langue en contexte est ensuite mentionné, notamment en ce qui concerne la structure de l'information. Enfin, il est question d'identifier les facteurs linguistiques et extralinguistiques faisant la particularité de la langue, afin de ressortir ses subtilités. L'auteur clôture en proposant des indications à suivre dans la construction de la grammaire, ainsi que dans la lecture d'une grammaire de référence.

Plusieurs principes essentiels à la bonne réalisation d'une grammaire sont définis. Au niveau de la forme, l'auteur insiste sur la clarté de la présentation. Il est préférable d'avoir plusieurs chapitres courts, ce qui est plus facile à lire, qu'un petit nombre de chapitres extrêmement longs. Au niveau du contenu, les informations doivent être présentées de manière explicite et dans une suite « logique », de sorte que l'on puisse lire le document comme on lit un roman ou une biographie. La structure de la langue doit donc se déployer au fur et à mesure que l'on progresse dans la lecture. Par ailleurs, une grammaire de référence étant une analyse, les différentes explications possibles pour un phénomène sont proposées, évaluées, et le choix d'une telle analyse ou d'une autre doit être justifié. Aussi, il est capital pour l'auteur de la grammaire de choisir des exemples appropriés qui illustrent chaque phénomène et appuient l'argumentation.

Le chapitre introductif présente de façon sommaire les objectifs de l'ouvrage et le thème abordé dans chacun des chapitres. La place cruciale du terrain dans l'écriture d'une grammaire est de surcroît soulignée. Le deuxième chapitre, qui correspond au premier chapitre d'une grammaire de référence, porte sur la langue et son contexte : sa localisation géographique, sa classification linguistique, ses traits grammaticaux majeurs, c'est-à-dire ceux qui font sa particularité et ses locuteurs. Il est également important de mentionner le phénomène du contact des langues, l'existence ou non de dialectes ainsi que l'attitude des locuteurs envers leur langue d'une part, et, d'autre part, envers les langues voisines dont les éventuels dialectes. L'organisation de la vie (les activités, les relations, le mode de vie, la religion, etc.) est une information de plus à relever. Le troisième chapitre définit les éléments fondamentaux d'une grammaire, lesquels sont développés dans les chapitres suivants. Les chapitres 4 à 13 constituent l'analyse de la langue proprement dite. Ainsi, sont examinées tour à tour : la phonologie, la morphologie, la syntaxe et la sémantique-pragmatique.

La phonologie est considérée comme la porte d'entrée de tout système linguistique. Il serait donc inconvenant de commencer l'analyse grammaticale d'une langue autrement. La phonologie segmentale, qui concerne les consonnes et les voyelles, est présentée en premier. Vient ensuite la phonotactique, qui est l'ensemble des règles qui régissent l'agencement des sons. La partie suprasegmentale s'intéresse à tout ce qui se passe au-delà des segments, notamment les phénomènes tonals et accentuels.

Toute langue possède des catégories grammaticales encore appelées dans la grammaire traditionnelle « parties du discours », divisées en deux groupes. D'un côté, les catégories ouvertes, dont les membres ne peuvent être listés de manière exhaustive et qui peuvent être créés à l'infini. On y retrouve notamment les noms et les verbes ; de l'autre côté, les catégories fermées ont un nombre limité et aucun élément nouveau ne peut y être ajouté. Parmi celles-ci se trouvent entre autres les articles, les pronoms interrogatifs et indéfinis, et les appositions, c'est-à-dire les prépositions et les postpositions. Deux critères essentiels permettent de définir la catégorie grammaticale d'un mot : sa structure morphologique et sa fonction syntaxique. Le critère sémantique peut également être appliqué dans certains cas.

Les adjectifs et les adverbes constituent des catégories grammaticales complexes. Ceci est dû au fait que leurs propriétés morphologiques et syntaxiques sont disparates, tantôt partagées avec les noms, tantôt avec les verbes et parfois même avec les deux pour ce qui concerne les adjectifs en particulier. La catégorie des adverbes, quant à elle, n'est pas seulement difficile à saisir, mais, selon l'auteur, c'est une notion héritée de la grammaire traditionnelle dont la réalité est illusoire.

Une grammaire de référence serait incomplète si elle ne soulevait pas la question des relations grammaticales. Ici la structure des propositions est examinée, sachant que deux grands types de structures propositionnelles ont été recensés à travers les langues. La première structure est intransitive, avec comme argument principal un sujet intransitif ou « S ». La seconde est transitive, avec au moins deux arguments principaux que sont le sujet transitif ou « A » et l'objet transitif ou « O ». Sur cette base, on distingue deux catégories de langues : les langues de type nominatif-accusatif où  $S = A$  et les langues de type absolutif-ergative où  $S = O$ . On constate donc que le cas joue un rôle primordial dans les relations grammaticales, de même que les appositions. Une analyse détaillée de ces relations doit pouvoir être proposée et justifiée dans la grammaire de référence.

Pour terminer l'analyse, une discussion sur la structure de l'information est de mise. Comment les structures grammaticales de la langue se comportent-elles effectivement dans un discours ? Cette question, indissociable de la pragmatique, renvoie aux notions de topique et de focus, dont les propriétés morphologiques et syntaxiques apportent des informations intéressantes sur le fonctionnement et l'usage de la langue en contexte.

D'une manière générale, cet ouvrage est un excellent guide pour tout linguiste souhaitant entreprendre la description d'une langue. L'abondance d'exemples clairs et provenant de diverses familles linguistiques – des langues indo-européennes (l'anglais, le français, l'espagnol), Niger-Congo (le babungu), des langues

aborigènes d'Australie (le dyirbal, le yawuru) sont citées en guise d'illustrations – rend la lecture à la fois passionnante et enrichissante. En définitive, il ressort de cela qu'écrire une grammaire est bel et bien un art, mais aussi un travail de bénédictin.