Notes de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Sylvain KAHANE, Kim GERDES. Syntaxe théorique et formelle. Volume 1 : Modélisation, unités, structures. *Language science press*. 2022. 627 pages. ISBN : 978-3-98554-037-2.

Lu par **Quentin FELTGEN**

Ghent University, Belgique

Avec cet ouvrage ambitieux, les deux auteurs offrent un modèle qui permet de décrire les mécanismes sous-jacents à la production des énoncés linguistiques, depuis la structuration du sens jusqu'à l'ordonnancement linéaire des unités linguistiques, l'une et l'autre étant articulés par une double structure syntaxique, profonde et de surface. Le contenu de l'ouvrage ne se limite cependant pas à l'exposé d'un modèle linguistique donné ; les auteurs décrivent également le cadre épistémologique plus général dans lequel cette théorie s'inscrit et peut être scientifiquement évaluée. En ce sens, ils situent la syntaxe des dépendances comme un modèle possible répondant à des problèmes de recherche bien définis, modèle auquel il existe donc des alternatives. Cette démarche invite à la fois à la réflexion et à la discussion, et d'autres théories linguistiques (comme la syntaxe X-barre ou la grammaire des constructions) sont convoquées à l'occasion de dialogues qui se veulent ouverts et féconds. Le propos est par ailleurs structuré en des sections courtes et bien agencées, entrecoupées d'encadrés qui permettent d'apporter des points de réflexion supplémentaires, ou un historique des notions discutées. Enfin, on saluera tout à la fois l'autonomie et la complémentarité des différents chapitres, qui permettront à cet ouvrage de servir efficacement d'outil de travail, même si l'âpreté de certaines définitions ainsi que leur foisonnement, l'étendue monumentale du sujet, et le caractère parfois limité des exercices, en font un instrument pédagogique certainement difficile d'accès, malgré la collection de manuels académiques dans laquelle il s'inscrit.

Le premier volume de cet ouvrage se compose de trois parties assez inégales, la troisième constituant les deux tiers de l'ouvrage. La première partie (chapitres 2 à 4) complète essentiellement l'introduction, la deuxième (chapitres 5 à 7) définit formellement les unités de la langue considérées comme pertinentes du point de vue de la syntaxe, et la troisième (chapitres 8 à 13) s'attache à la présentation et à la caractérisation des structures syntaxiques, objet principal de ce volume. S'inscrivant dans la collection « Textbooks in Language Science », l'ouvrage propose également, à l'issue de chaque chapitre, une batterie d'exercices avec leurs corrigés et un aperçu des références pertinentes, le texte étant lui-même enrichi de nombreux encadrés qui mettent en perspective le propos ou viennent offrir un complément (voire un

contrepoint) sur tel ou tel point abordé dans le corps principal. Si l'ouvrage ne se veut pas un manuel de syntaxe en dépendance, les chapitres 9, 10 et 12 constituent néanmoins un ensemble cohérent et très complet qui fournira une introduction tout à fait solide au sujet.

Le chapitre 1, qui fait office d'introduction, situe clairement les objectifs, la portée, et les orientations théoriques de l'ouvrage. Les deux auteurs indiquent ainsi s'inspirer, dans leur démarche, des travaux de Nicolas Bourbaki, un groupe de mathématiciens ayant entrepris de proposer une construction des mathématiques depuis leurs fondations, sans se référer au moindre prérequis et à la moindre notion préexistante. C'est donc l'ambition que se donne cet ouvrage : aboutir à un compte rendu des relations syntaxiques entre éléments de la langue, sans rien considérer comme acquis ou théoriquement établi. Par ailleurs, les auteurs précisent ne pas souhaiter présenter la syntaxe des dépendances comme un outil théorique déjà constitué dont il reviendrait à un manuel d'enseigner le maniement, mais comme un modèle scientifique visant à capturer certaines propriétés des énoncés de la langue, modèle dont on peut dès lors discuter la structure, la motivation, la portée, et la validité.

Le chapitre 2 présente quelques généralités sur la langue, en particulier la notion saussurienne de signe, la distinction langue et parole ou compétence et performance, ou encore les deux perspectives de la théorie linguistique, l'analyse (des énoncés produits) et la synthèse (les règles sous-jacentes à l'expression linguistique d'un sens) privilégiée par les auteurs. Le chapitre 3 explicite justement cette démarche et montre comment un sens peut être représenté schématiquement au moyen d'un graphe orienté, dont différents énoncés alternatifs mais équivalents du point de vue sémantique peuvent alors rendre compte. Cependant, cette perspective orientée vers la production ne sera vraiment développée qu'aux chapitres 12 et 13, et une place très large est laissée à l'analyse des énoncés.

Le chapitre 4 vient donner une assise épistémologique à l'ouvrage en proposant une discussion sur la notion de modèle, laquelle souffre en linguistique d'une certaine imprécision, si bien que la clarification apportée par les auteurs est bienvenue. En prenant pour exemple le modèle gravitationnel des marées en sciences physiques[1], les auteurs mettent en contraste modèle descriptif et modèle explicatif, en soulignant très justement qu'un modèle descriptif a lui-même capacité à prédire s'il est suffisamment développé. Les modèles de type « réseaux de neurones », au fort pouvoir prédictif mais sans valeur explicative, sont également évoqués, les auteurs, s'inscrivant en faux vis-à-vis de ce type, revendiquent de vouloir proposer un modèle proprement explicatif ; à cet égard, on regrettera seulement que la notion d'explication, si difficile en épistémologie, n'ait pas fait elle-même l'objet d'une clarification. Par ailleurs, la dimension prédictive dont se targuent les auteurs n'est en définitive guère sous-jacente dans le reste de l'ouvrage, l'ensemble du modèle

1 Il s'agit au demeurant d'un choix curieux, les marées n'ayant été comprises que tardivement par rapport, par exemple, aux trajectoires planétaires, le modèle newtonien ne parvenant pas à rendre compte des subtilités nombreuses du phénomène sans un effort théorique additionnel.

proposé paraissant essentiellement orienté vers des fins descriptives (ce qui est peut-être un mérite en soi).

La deuxième partie de l'ouvrage, consacrée aux unités de la langue, est introduite par le chapitre 5, qui reprend les principales notions du signe linguistique, et pose une troisième composante en plus du signifié et du signifiant, son syntactique, c'est-à-dire le potentiel combinatoire qui en caractérise l'usage. Dès lors, il se trouve trois manières d'appréhender les constituants minimaux de la langue : du point de vue du sens (sémantèmes), du point de vue de la forme (morphèmes) et du point de vue du syntactique (syntaxème). Cette notion de syntaxème se retrouve justement au cœur du chapitre 6, qui introduit un ensemble de concepts et de définitions fondamentales.

La décomposition propre est d'abord définie : un signe AB peut se décomposer proprement en deux signes A et B, si, par des rapports analogiques du type défini par Saussure, on peut aboutir à un signe A'B' ; par exemple *broyeur* peut se décomposer en *broy* + *eur* parce qu'il existe une analogie *broyeur* : *broyons* égale à *compresseur* : *compressons*. Le morphème est ainsi défini comme un signe qui ne peut pas être décomposé plus avant au sens de la décomposition propre. Cette combinaison propre devient libre lorsque la classe des éléments qui se combinent avec l'un forme une classe indépendante de l'autre. Ainsi, la combinaison de *bless-* et *-ure* n'est pas libre (quand bien même elle est propre) car la classe des éléments qui se combinent avec *-ure* est spécifique de ce signe (avec notamment des radicaux comme *struct-* qui ne correspondent à aucun verbe). Cette combinaison libre permet de définir une unité syntaxique comme un signe se combinant librement avec son environnement. Les unités syntaxiques minimales (qui ne peuvent donc pas être décomposées plus avant au sens de la décomposition libre) définissent alors les syntaxèmes.

Le chapitre 7 complète ces définitions en abordant cette fois les unités sémantiques, définies comme des signes résultant d'un ou plusieurs choix du locuteur. Si ce choix est en plus indivisible, c'est-à-dire qu'il ne peut plus être décomposé en choix successifs, l'unité ainsi délimitée est minimale : il s'agit d'un sémantème. Ce chapitre est également l'occasion de préciser la notion de signème, introduite au chapitre précédent : un signème est défini comme un faisceau de signes, c'est-à-dire un paradigme de formes (comme les différents morphèmes associés au radical du verbe *aller*) mis en relation avec un ensemble de sémantèmes (les différentes acceptions sémantiques). Pour les auteurs, les signèmes (qui sont des constructions théoriques) constituent les unités de la langue, les signes (qui sont des observables), les unités de la parole : la production d'un signe suppose la sélection par le locuteur d'une des acceptions sémantiques du signème, et l'énonciation de la forme appropriée compte tenu des contraintes contextuelles.

La présentation de la syntaxe commence réellement avec le chapitre 8, qui s'appuie sur les chapitres précédents pour préciser la définition des unités syntaxiques (toujours à l'aide de la notion de combinaison libre), qui peuvent être soit minimales (les syntaxèmes), soit une combinaison libre d'unités minimales (les syntagmes). La syntaxe est elle-même définie comme l'étude des combinaisons

libres de signes linguistiques. Le chapitre 9 présente, quant à lui, le concept de connexion syntaxique, qui sous-tend à lui seul la notion de structure syntaxique. Il y a connexion syntaxique lorsque deux éléments syntaxiques se combinent pour former un syntagme – c'est donc, là encore, la combinaison libre qui permet de définir la connexion. Plus précisément, les auteurs définissent la connexion de manière formelle comme une classe d'équivalence, c'est-à-dire un ensemble de combinaisons plus ou moins fines, mais toujours libres, opérant sur la même frontière.

Le chapitre 10 constitue, sans aucun doute, l'apport central de l'ouvrage, tant par son ampleur que par son importance. Il précise en effet la notion de tête et énonce un grand nombre de critères pratiques permettant d'identifier le gouverneur d'une unité syntaxique et, par là, de construire l'arbre de dépendance associé à l'énoncé. On regrettera cependant que la notion de dépendance ne fasse pas l'objet de la même construction formelle que celle, par exemple, de connexion : l'existence d'une hiérarchie entre les unités syntaxiques est immédiatement postulée, sans qu'il soit détaillé ce que signifie la relation de dépendance qui en résulte. Cela n'enlève rien au caractère opératoire de la notion, comme en témoignent les très riches discussions et les nombreux exemples d'applications de critères dont la variété permet de rendre l'analyse en dépendance efficace dans un large éventail de situations. On appréciera également la position très ouverte des auteurs, par exemple sur le choix du nom ou du déterminant comme tête : ceux-ci discutent les différentes possibilités en prenant soin de souligner que la décision finale ne constitue qu'un choix théorique susceptible de révision (ou même d'adaptation à des fins pédagogiques et de présentation).

Le chapitre 11 illustre à nouveau ce positionnement épistémologique qui vise à favoriser le dialogue entre les différentes approches en présentant les relations qu'entretiennent l'analyse en dépendance et l'analyse en constituants. Les limites et les avantages de chacune sont présentés avec détail, et l'exposé, quoiqu'il brasse de nombreux cadres distincts, reste toujours d'une parfaite clarté, supporté notamment par de nombreuses figures qui illustrent bien la fluidité existante entre les divers cadres de représentation. Tout comme le chapitre 10, le chapitre 11 est assorti d'un historique passionnant de ces notions et de ces représentations.

Le chapitre 12 est tout entier dédié à la mise en place d'un modèle topologique, permettant de conformer la structure syntaxique en dépendance à l'ordre linéaire de l'énoncé, dont les contraintes sont propres à chaque langue. Cette approche se veut donc générative dans le sens où ce modèle topologique fournit, à partir d'un arbre en dépendance, les règles permettant la production d'un énoncé linéaire obéissant aux exigences topologiques de la langue. Deux notions sont particulièrement discutées : la projectivité d'abord, réalisée lorsque les dépendances entre unités ne se coupent pas une fois celles-ci arrangées selon l'ordre linéaire, et dont les liens avec la théorie des graphes sont explicités de façon éclairante. La notion de gabarit fait ensuite l'objet de l'essentiel du chapitre : un gabarit est un ensemble de champs ordonnés associé à un constituant, chaque champ venant spécifier quels éléments peuvent l'occuper. Ainsi, pour chaque nœud de l'arbre de dépendance, un gabarit approprié

vient spécifier comment se situent ses dépendants par rapport à lui, suivant la nature de ceux-ci.

Le chapitre 13 se consacre à la structure syntaxique profonde, détaillant le passage d'une structure prédicative du sens à une structure syntaxique profonde, laquelle agence les sémantèmes dans une structure de dépendance, intermédiaire entre le plan du sens et l'arbre de dépendance. Là encore, une approche de type génératif est proposée en associant notamment à chaque unité lexicale un tableau de régime (ou une structure élémentaire d'arbre syntaxique), c'est-à-dire la spécification complète de ses arguments et de leur régime. En combinant les lexèmes suivant cette structure prédicative, on obtient l'arbre correspondant au sens initial. Les auteurs achèvent ainsi de construire l'édifice menant du sens jusqu'à l'énoncé linéairement ordonné.

Silviu PAUN, Ron ARTSTEIN, Massimo POESIO. Statistical Methods for Annotation Analysis. *Morgan & Claypool Publishers*. 2022. 198 pages. ISBN : 978-1-63639-253-0.

Lu par **Lydia-Mai HO-DAC**

Université de Toulouse Jean-Jaurès / CLLE UMR 5263

Cet ouvrage propose un état de l'art « appliqué » des méthodes développées pour évaluer la qualité d'une ressource annotée au niveau linguistique en vue de son utilisation pour le TAL. L'ouvrage se veut avant tout didactique, offrant des explications très claires sur les notions et les formules nécessaires pour maîtriser des méthodes généralement développées dans d'autres domaines que celui de la linguistique (par exemple, la médecine), et des cas d'usage dans le domaine de la linguistique. Les auteurs profitent de chaque explication et cas d'usage pour souligner les précautions à prendre lors de la mise en place de campagne d'annotation et l'importance de l'évaluation dans la constitution de ressources annotées. Trois familles de méthodes sont présentées, des plus simples aux plus complexes (et récentes) : mesures traditionnelles de l'accord entre annotateurs, modèles probabilistes sur l'accord, modèles probabilistes sur l'annotation. La diversité des cas d'usage proposés pour illustrer chaque méthode a été pensée pour couvrir la réalité de la diversité des annotations en linguistique : annotation par les foules vs *par quelques individus au long court, catégorisation binaire* vs *multicatégorisation, avec ou sans jeu d'étiquettes prédéfini, avec ou sans l'étape de délimitation des unités à annoter, annotation pour évaluer* vs *pour entraîner des modèles.*

Organisation générale de l'ouvrage

Après une introduction revenant sur l'activité d'annotation et la différence fondamentale pour les auteurs entre **fiabilité des annotations** (c'est-à-dire le fait que les annotations produites semblent cohérentes entre elles – reliability en anglais) et **validité des annotations** (c'est-à-dire le fait que les annotations soient correctes par rapport à une référence – *validity* en anglais), les auteurs listent certains éléments essentiels à maîtriser pour appréhender l'ouvrage (la notion de processus génératif et d'inférence, l'importance des méthodes existantes pour évaluer l'adéquation entre un modèle et des données, l'influence des préalables dans les modèles statistiques et

les moyens linguistiques à utiliser pour exprimer des probabilités). L'introduction s'achève avec un avertissement sur les connaissances statistiques nécessaires pour une bonne compréhension de l'ouvrage et pointe vers des suggestions de lecture à faire pour acquérir ces connaissances.

Le corps de l'ouvrage est organisé en deux parties qui répartissent les méthodes présentées selon qu'elles mesurent la fiabilité des annotations ou leur validité. Pour chaque famille plusieurs mesures sont présentées, détaillées et illustrées. Après la présentation théorique des concepts de base et des formules impliqués dans la méthode, celle-ci est appliquée à un cas d'usage dont les données sont systématiquement décrites et rendues accessibles sur un site associé (voir *infra*). Des encarts au fil du texte proposent en complément des définitions ou des rappels des notions élémentaires nécessaires pour bien appréhender la ou les mesures présentées (par exemple, l'analyse de contenu, les différentes lois de probabilité utilisées dans le chapitre en cours). Un résumé et un bilan closent la présentation de chaque famille de méthodes permettant ainsi de dégager les spécificités, les avantages et les inconvénients des différentes mesures.

Tout au long de l'ouvrage, des sections sont proposées pour alerter sur les biais possibles auxquels il faut s'attendre au moment de l'évaluation de la fiabilité ou de la validité des annotations.

Un site est proposé en complément de l'ouvrage pour en accompagner la lecture et donner accès aux données mentionnées dans les différents cas d'usage utilisés pour illustrer les différentes méthodes.

Fiabilité des annotations selon la mesure de l'accord entre annotateurs

Dans cette partie, les auteurs distinguent les méthodes utilisant des coefficients d'accord entre annotateurs (*coefficient of agreement*) des méthodes fondées sur des modèles probabilistes de l'accord (*probabilistic models of agreement*).

Selon la même progression que pour toutes les autres sous-parties, les coefficients d'accord entre annotateurs sont présentés du plus simple au plus complexe, en ce sens que les plus complexes permettent de prendre en compte une plus grande diversité de paramètres. La première mesure présentée est celle du pourcentage d'accord (c'est-à-dire le pourcentage d'items sur lesquels les annotateurs sont d'accord par rapport au nombre total d'items annotés) que les auteurs utilisent pour illustrer à quel point les mesures les plus simples ne sont pas adaptées à la réalité d'une grande partie des annotations produites en linguistique car, entre autres, les catégories à annoter sont rarement distribuées équitablement dans la langue (sans parler de la question de l'unité à annoter qui sera traitée plus loin). Le fait que certaines catégories sont plus fréquemment rencontrées dans la réalité de la langue va avoir une influence sur le comportement des annotateurs qui vont avoir tendance à annoter ces catégories plus facilement et donc plus souvent et/ou avec un plus grand accord entre annotateurs. Il est donc nécessaire de disposer de méthodes qui s'adaptent à la diversité des annotations pour ne pas biaiser les résultats de l'évaluation. Complexifiant au fur et à mesure le type d'annotation à évaluer, les auteurs progressent vers des mesures permettant de manipuler des

paramètres de plus en plus variés : nombre des annotateurs et nécessité de pondérer certaines annotations liées à des catégories plus fréquentes ou faciles à annoter que d'autres.

Les mesures détaillées sont les suivantes : pourcentage d'accord, mesures d'association, coefficient de corrélation, le *S* de Bennet *et al.* (1954), le π de Scott (1955), le κ de Cohen (1960), le κ de Fleiss (1971), l'α de Krippendorff (1980) et la variante pondérée du κ de Cohen (1968). Toutes ces mesures sont présentées de façon théorique puis appliquées à un même cas d'usage : la classification par deux ou plus d'annotateurs des actes de paroles selon le modèle DAMSL (*Dialog Act Markup in Several Layers*, Core *et al.* 1997). Ce cas d'usage commun permet de généraliser les différences principales entre les coefficients proposés.

Le chapitre 2 s'achève par la présentation des intérêts et des limites de ces méthodes mesurant le coefficient d'accord entre annotateurs. L'intérêt principal est de prendre en compte la part de hasard qui peut intervenir dans une tâche d'annotation. Cet intérêt est explicite dans les termes anglais utilisés pour les désigner : *Chance-corrected Agreement Coefficients*. Mais le hasard est loin d'être le biais le plus important à éviter lorsque l'on évalue un jeu d'annotations. Parmi les limites des coefficients présentés, les auteurs reviennent en détail sur (i) les difficultés pour gérer les « données manquantes » c'est-à-dire le fait que les items associés à des catégories plus difficiles à annoter ont tendance à être moins (bien) annotés ; (ii) la nécessité de prendre en compte les spécificités de certains annotateurs ou d'évaluer des annotations réalisées par les foules ; (iii) l'impossibilité d'évaluer correctement des annotations incluant une étape de délimitation des unités à annoter (à ce stade, seule la catégorisation d'unités présegmentées a été discutée) ; (iv) les problèmes liés à des modèles d'annotation impliquant des catégories présentant une très forte disparité de fréquence d'apparition ou une certaine difficulté à être identifiée.

Le chapitre 3 prolonge le chapitre 2 en illustrant comment les coefficients d'accord entre annotateurs sont utilisés pour évaluer les tâches propres au TAL : étiquetage morphosyntaxique, identification des actes de paroles, reconnaissance des entités nommées, détection de la subjectivité, segmentation thématique, annotation de la prosodie, annotation des phénomènes d'anaphore et de deixis discursive, résumé automatique, désambiguïsation lexicale. Face à cette diversité et cette complexité des annotations linguistiques, force est de constater que les coefficients présentés ne semblent pas les mieux adaptés pour évaluer la qualité d'une ressource annotée linguistiquement. Ce constat permet aux auteurs d'introduire les modèles probabilistes comme une solution préférée pour évaluer la fiabilité des annotations.

Fiabilité des annotations selon les modèles probabilistes

Le principe de base de l'évaluation de la fiabilité des annotations avec des modèles probabilistes consiste à distinguer deux étapes dans l'évaluation : dans un premier temps les items à annoter sont caractérisés selon leur probabilité à être difficile à annoter et, dans un deuxième temps, selon leur probabilité à faire consensus entre les annotateurs. La première étape part du constat général fait par

tout chercheur ayant mené une campagne d'annotation en linguistique : il existe des cas clairs mais aussi des cas limites beaucoup plus difficiles à juger.

Les modèles probabilistes permettant d'évaluer cette difficulté de jugement (et donc le risque d'annoter différemment les uns des autres) sont largement utilisés dans le domaine médical, ce qui explique que les cas d'usage utilisés dans ce chapitre sont issus de ce domaine. Trois premiers modèles sont présentés : les modèles de Aickins's α (1990), de Gwet's AC_1 (2008) et de Guggenmoos-Holzmann (1996). Le dernier diffère un peu des autres en ce sens qu'il cherche à évaluer l'instabilité plus que la difficulté des items (plus un jugement est instable plus il sera dur à faire).

Là encore, les auteurs soulignent les limites de ces modèles, et notamment le fait qu'ils ne permettent pas d'évaluer la validité (voire la véracité) des annotations. Par exemple, ils ne permettent pas d'analyser si les annotateurs font les mêmes erreurs d'annotation (et sont donc d'accord mais sur de mauvaises catégories ou délimitation d'unités). Ils ne permettent pas non plus d'analyser si les annotateurs sont plus enclins à annoter d'une certaine manière. De façon générale, ces modèles ne permettent pas de prendre en compte le biais lié aux aptitudes propres à chaque annotateur, aptitudes qui sont souvent surestimées par les modèles. De plus, ces modèles ne fournissent qu'un score d'accord, sans donner une idée de la qualité de l'annotation en termes de « vérité », ce qui est problématique si l'objectif est d'utiliser ces annotations pour entraîner ou évaluer des modèles.

La solution à ces limites réside, selon les auteurs, dans l'application de modèles de variables latentes (*Latent class model*), ce qu'ils démontrent en appliquant ce type de modèle à deux cas d'usage : annotations en médecine réalisées par un panel diversifié *vs* expert d'annotateurs (Uebersax *et al.* 1989), et classification de phrases en termes de contenu subjectif *vs* objectif où il est plus facile et fréquent d'annoter les items objectifs que les autres (Wiebe *et al.* 1999). Cette dernière partie permet d'annoncer la deuxième partie qui propose d'appliquer le modèle de variables latentes à l'évaluation de la validité plutôt que de la fiabilité.

Validité des annotations selon les modèles de variables latentes

La deuxième partie commence par souligner qu'à l'heure où de plus en plus d'annotations sont réalisées par des foules (*crowdsouring* ou myriadisation) ou des machines (avec les techniques d'*active learning*) plutôt que par des experts, le fait de concentrer l'évaluation sur l'accord entre annotateurs est à revoir ; le plus important étant davantage d'évaluer si les annotations sont correctes ou pas.

Après une introduction sur les modèles probabilistes utilisés pour évaluer la validité d'un jeu d'annotations, les auteurs soulignent un des intérêts majeurs du modèle de variables latentes : la capacité à modéliser à la fois le comportement des annotateurs et l'impact de la difficulté des items sur ce comportement. Concernant le comportement des annotateurs, l'intérêt est de considérer le fait que les annotateurs n'ont pas tous les mêmes capacités d'annotation et qu'ils ne sont pas toujours réguliers notamment à cause de l'influence des items précédemment annotés. Deuxième problème : ces variations de comportement sont augmentées par le degré

et le type de difficulté des items. En effet, certaines difficultés altèrent les aptitudes des annotateurs comme le fait qu'ils appartiennent à des catégories rarement rencontrées, qu'ils soient ambigus (certains items peuvent, selon l'interprétation de l'annotateur, appartenir à des catégories différentes), que les catégories du modèle soient difficiles à distinguer, ou encore que la tâche d'annotation soit lucrative, ou peu motivante ou encore très ludique (par exemple, les *games with purpose*).

Les modèles présentés dans cette partie sont les suivants : le modèle de David et Skene (1979) et celui de Carpenter (2008) pour modéliser le comportement des annotateurs, Whitehill *et al.* (2009) pour la difficulté d'items et trois modèles dont celui défendu par les auteurs pour tenter de prendre en compte ces deux aspects (Simpson *et al.* 2011, Felt *et al.* 2015 et les auteurs Paun *et al.*, 2018). Une dernière option est présentée, le modèle discriminatoire de Raykar *et al.* (2010). Celui-ci propose de caractériser dans un premier temps les items afin de nuancer l'évaluation des annotations par rapport à ces caractéristiques.

Le premier cas d'usage utilisé pour illustrer ces modèles et comparer leurs résultats est la tâche de reconnaissance de l'inférence textuelle où il est demandé aux annotateurs de dire si un argument peut être inféré depuis un premier argument, oui ou non. Les annotations utilisées pour ces illustrations sont issues du travail de Snow *et al.* (2018) qui a récolté des annotations *via* le service d'Amazon Mechanical Turk. Ce cas d'usage est un bon exemple pour illustrer les variations entre annotateurs et les degrés et types de difficulté qu'un annotateur peut rencontrer.

Deux autres cas d'usage permettent une description plus approfondie de l'application de certains des modèles présentés : l'analyse de séquences et l'annotation des anaphores, qui illustrent des cas d'annotation plus complexes car ne faisant pas appel à une catégorisation binaire comme la tâche de reconnaissance de l'inférence textuelle.

L'ouvrage s'achève avec un dernier chapitre qui prône l'intérêt d'utiliser les résultats des évaluations issus des différents modèles pour mieux définir le phénomène linguistique que l'on cherche à annoter et améliorer les schémas et les processus d'annotation afin d'obtenir des jeux de données dont la qualité assurera une certaine validité pour les systèmes d'apprentissage automatique.

L'ouvrage de Paun, Artstein et Poesio est complet et très dense. Chaque formule est expliquée, puis appliquée pas à pas aux cas d'usage qui ont le mérite d'être en lien avec la réalité des annotations linguistiques. Les auteurs montrent bien qu'évaluer un jeu d'annotations est complexe et que de nombreux biais sont présents dans les données, notamment ceux liés au comportement des annotateurs et aux difficultés des items en langage naturel. Enfin, les auteurs discutent régulièrement de l'importance d'évaluer pour s'assurer de la qualité des données sur lesquelles entraîner et évaluer des modèles.