Notes de lecture

Rubrique préparée par Denis Maurel

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

Roland SCHÄFER, Felix BILDHAUER. Web Corpus Construction. *Morgan & Claypool publishers*. 2013. 129 pages. ISBN 978-1-6084-5983-4.

Lu par **Noémie COLIN**

Chef de projet en informatique et linguistique

L'ouvrage traite de la tâche de construction de corpus à partir de données Web. Il fournit des exemples de création de corpus venant de données récupérées sur le Net et également des outils existants qui ont déjà été utilisés pour la construction de corpus. Les principales étapes de cette tâche y sont présentées avec des volumes conséquents. Dans cet ouvrage, les auteurs détaillent l'indexation du Web et les filtrages de ces données (tels que la suppression des doublons). Cet ouvrage couvre également le traitement linguistique de ces données, en particulier par rapport aux bruits provenant spécifiquement des données Web. Finalement, les auteurs expliquent comment les corpus Web peuvent être évalués et comparés à d'autres corpus.

L'ouvrage porte sur la collecte de données Web pour la construction de corpus. Ceux-ci sont utiles pour différents usages. C'est pourquoi il est difficile d'énoncer une méthode générale de construction de corpus Web. Les auteurs ont choisi ici de se placer du point de vue de la linguistique empirique. Plusieurs outils, disponibles également sur le Web, sont énoncés au fur et à mesure de l'œuvre en fonction de la tâche concernée. Il est souligné que l'utilisation d'un *crawler* (robot d'indexation) est une tâche qui requiert des compétences de programmation avancées. Avant toutes choses, il est donc crucial de savoir qu'en fonction des décisions prises sur certains filtrages ou nettoyages, les données originales vont être altérées. En effet, toutes les étapes de cette procédure seront automatisées autant que possible (ce qui est particulièrement évident étant donné que l'on souhaite créer des corpus volumineux), ce qui explique que le contenu et la qualité du corpus final seront influencés par ces choix appliqués automatiquement. Ce livre a été divisé en quatre chapitres principaux portant sur une étape importante de la construction de corpus Web. Ces étapes sont évoquées dans l'ordre habituel de la pratique de construction de corpus Web.

Le chapitre 1 permet d'introduire le sujet de la construction de corpus Web à travers des exemples, et énonce les intérêts de cette tâche. Le chapitre 2 s'intéresse à la théorie et à la technologie de la collecte de données (« *crawling* »). Le chapitre 3

explique le nettoyage non linguistique à appliquer aux données collectées. Le chapitre 4 détaille les problèmes issus du traitement linguistique de ces données. Finalement, le chapitre 5 introduit les méthodes d'évaluation de la qualité du corpus compilé.

Le chapitre 1 est une introduction aux corpus Web. La première partie de ce chapitre est composée d'une liste de corpus téléchargeables gratuitement sur Internet. Les auteurs nous avertissent, dans la deuxième partie, des dangers et des désavantages liés à la collecte de données sur le Web élaborée à l'aide de moteurs de recherche utilisés pour répondre aux limites des ressources disponibles gratuitement. En revanche, cela permet de nous énoncer les avantages de la collecte de données Web pour la construction de corpus. Les auteurs illustrent ces avantages grâce à deux projets : *the WaCky initiative*[1]et *the Leipzig Corpora Collection*[2].

La collecte de données est la partie la plus simple dans la construction de corpus Web, il s'agit du sujet du chapitre 2. Elle peut être décomposée en deux parties : tout d'abord, la récupération de plusieurs URL, et finalement l'utilisation d'un outil pour indexer le Web. Un *crawler* (robot d'indexation ou encore collecteur) est un logiciel qui explore automatiquement le Web. Il permet de télécharger récursivement des pages Web à partir de liens extraits d'une ou plusieurs autres pages pivots connues. Il ne faut pas oublier que les données collectées sont un extrait d'un ensemble volumineux de documents Web. Il est donc important, à cette étape, de prendre un échantillon pour déterminer la nature du corpus final obtenu. Afin de bien comprendre la stratégie d'indexation à adopter en fonction du corpus que l'on souhaite obtenir, une analyse de la structure du Web et des différentes stratégies de collecte de données sont présentées en détail : certaines stratégies permettent d'obtenir des corpus très volumineux (comme les collecteurs qui utilisent l'algorithme de parcours en largeur), et d'autres de plus petits corpus (grâce aux systèmes d'échantillonnage en marche aléatoire). À partir de ces données brutes collectées, les chapitres 3 et 4 permettent de comprendre quelles étapes de traitements des données doivent être suivies.

Dans le chapitre 3, les auteurs décrivent les différentes techniques de filtrage à employer pour nettoyer les données brutes collectées sur le Web, telles que l'emploi des balises, la détection des langues, les doublons... Ils insistent sur le fait que tous ces filtres permettent effectivement d'obtenir au final des données propres, mais qui sont également altérées par tous ces filtrages. Pour chaque type de nettoyage appliqué, il faut faire un choix de traitement, comme par exemple si l'on applique un changement au niveau du mot, du segment, de la phrase ou du paragraphe, car, selon la décision prise, cela affectera les données d'une façon différente.

---

1 http://wacky.sslmit.unibo.it

2 http://corpora.uni-leipzig.de/

Le chapitre 4 porte sur le traitement linguistique et la normalisation des données Web (tokenisation, détection de phrases, étiquetage morphosyntaxique, lemmatisation). La première section de ce chapitre est destinée en particulier aux informaticiens non linguistes, car elle définit et explique les bases des différentes étapes du traitement linguistique. Dans la deuxième section, il est montré que cette partie du travail est la plus compliquée car il y a beaucoup de « bruit » sur ce type de données. Il n'y a pas de méthode générale qui puisse être appliquée pour tous les cas dans le nettoyage des données Web, mais certaines étapes ont été un peu plus approfondies afin d'améliorer les résultats obtenus au niveau qualitatif. Le plus difficile, et ce à quoi les auteurs ont tenté de répondre, est d'identifier les différences entre le « bruit » et les « données ». En fonction de la définition que l'on a du « bruit », on considère que certaines étapes, comme la normalisation, peuvent introduire de nouveaux « bruits ». C'est ce qui nous pousse à lire le chapitre 5 qui porte sur le contrôle qualité et qui nous permet de comprendre comment on peut évaluer la qualité des données collectées et filtrées linguistiquement.

Le chapitre 5 s'intéresse donc au contrôle qualité et à l'évaluation des données Web collectées, sachant, qu'une fois de plus, il faut considérer que la « qualité » évoquée ici est une notion relative. Différentes méthodes sont donc expliquées et montrées sur des corpus, comme celui de DECOW2012, par exemple, selon certains critères :

- à l'intérieur du corpus même, distributions des mots en fonction de la longueur des phrases, par exemple ;
- sur l'aspect du corpus, comme la fréquence des mots (ou mots-clés) par rapport à d'autres corpus. Cela peut être très utile pour comparer ce type de textes avec des corpus de référence, des corpus de spécialités ou tout autre corpus générique.

Toutes ces techniques d'évaluation permettent de mettre en valeur et de vérifier les avantages de l'indexation et des différentes procédures de traitement linguistique.

Ce livre doit être considéré comme un tutoriel pour toutes les personnes (ayant de préférence des compétences en programmation et/ou en linguistique informatique) qui souhaitent créer un corpus à partir de données collectées sur le Web.

Gaston GROSS. Manuel d'analyse linguistique. *Presses universitaires du Septentrion*. 2012. 369 pages. ISBN 978-2-7574-0397-6.

Lu par **Pascal Vaillant**

Université Paris-13, LIM&Bio

Par sa présentation, claire et didactique, sans prérequis conceptuels ou terminologiques, ce livre se positionne comme un ouvrage de référence. Il n'est pas lié à un formalisme ou à une technologie et s'ancre dans une lignée d'ouvrages majeurs de description empirique de la langue, fondée sur des critères distributionnels. Il mérite donc, à notre avis, de devenir une référence durable présente dans la bibliothèque de tout spécialiste de modélisation du lexique et de la syntaxe.

Le *Manuel d'analyse linguistique* de Gaston Gross est le legs de trente ans d'expérience de description grammaticale du français. Il fournit une synthèse précieuse du savoir-faire accumulé dans ce domaine, et il est probablement, en termes de couverture du lexique, sans équivalent dans l'espace francophone. C'est sur ce savoir-faire que s'est construit, à l'université Paris 13, une équipe qui exporte ses compétences en lexicographie numérique auprès d'entreprises et de partenaires de plusieurs pays. Gaston Gross, aujourd'hui professeur émérite, publie ici dans un volume unique des principes et des méthodes de description dont il avait auparavant rendu compte de sous-ensembles plus spécifiques.

L'ouvrage n'exige pas de prérequis conceptuels autres que ceux du vocabulaire de la grammaire française classique telle qu'elle est enseignée au collège. Il peut donc figurer sur la liste d'achats utiles de tous ceux qui vont être amenés à pratiquer de la description lexicale dans une perspective de traitement automatique – qu'ils soient étudiants ou professionnels, et quelle que soit leur formation de base (informatique ou linguistique).

Un point de vue lexicographique centré sur les emplois

Ce que propose Gaston Gross dans ce *Manuel d'analyse linguistique* est, en résumé, une méthode pour constituer des dictionnaires électroniques. Au cœur de la problématique du traitement des langues, le dictionnaire doit contenir suffisamment d'information pour décrire la structure d'un texte, avec un niveau de profondeur adéquat à la tâche à effectuer. Il doit permettre, par exemple, à un système automatique de traduire le français « *voulez-vous prendre un café ?* » en allemand par la phrase « *möchten Sie einen Kaffee trinken ?* », mais de ne pas traduire « *voulez-vous prendre un bain ?* » par « *möchten Sie ein Bad trinken ?* ».

L'approche suivie par Gaston Gross est fondée sur le postulat que l'on ne peut séparer lexique, syntaxe et sémantique. En lecteur assidu de Zellig Harris, il tire les conséquences de l'idée que l'usage d'un mot est défini par les distributions de ses contextes : « les textes ne sont pas constitués de mots interchangeables comme le

seraient les briques d'un mur, mais ils forment entre eux des structures élémentaires, qui sont à la base de tout discours. Ces structures élémentaires, nous les appelons des *schémas prédicatifs* […] Dès qu'un schéma prédicatif a été reconnu à un endroit du texte, on se trouve devant une sélection de mots qui ensemble forment une unité, excluant toutes les autres lectures possibles de ces mots, s'ils s'inséraient dans d'autres schémas » (p. 12). Les propriétés sémantiques et syntaxiques d'un mot sont donc indissociables, et « décrire une langue, c'est faire le recensement organisé de l'ensemble des emplois qu'elle comporte » (p. 8). Ce livre propose « un traitement intégré du lexique, de la syntaxe et de la sémantique » (p. 23).

Dans cette perspective, la syntaxe et la sémantique sont intégrées au lexique. Ce principe peut paraître évident à une génération d'étudiants et de chercheurs en TAL élevés aux grammaires à adjonction d'arbres (TAG), mais il faut souligner que Gaston Gross a raison depuis longtemps ; qu'il a longtemps eu raison contre la pensée dominante en linguistique formelle, enfin, que le fait que son approche soit fondée sur des principes maintenant largement reconnus comme valables par la communauté des chercheurs, n'indique pas que ceux-ci aient achevé de percoler jusqu'aux applications réellement utilisées dans le monde extérieur.

Le chapitre 2 montre, sans faire le choix d'un formalisme de représentation particulier, l'importance de la notion d'emploi de prédicat. Celle-ci définit des catégories qui permettent de rassembler commodément des faisceaux de traits linguistiques empiriquement corrélés et relevant de différents « niveaux » de description dans la partition traditionnelle du système des langues en lexique, morphologie, syntaxe et sémantique : le choix de l'un des sens d'un verbe, la structure de la phrase déployée autour de lui, le nombre de ses compléments, les prépositions et flexions de chacun d'entre eux, le degré de figement de certains, les transformations activables ou au contraire inhibées. C'est en ces termes que Gaston Gross constate le fait que le sens d'une unité linguistique dépend de son contexte d'emploi (« le sens n'est pas premier ni "isolable" mais […] il est en connexion avec bien d'autres propriétés d'une structure phrastique », p. 43).

L'un des fondements de la méthode de Gaston Gross est la notion de « classes d'objets » ; elle constitue la matière du chapitre le plus important de l'ouvrage (le chapitre 4). En résumé, pour éviter de produire « *ich habe einen feindlichen Soldaten abgerissen* », il faut identifier les paradigmes de compléments d'objet correspondant aux différents emplois du verbe *abattre*, et noter que ce verbe se traduit par *abreißen* s'il s'agit de démolir une construction, par *fällen* s'il s'agit d'un arbre, et par *erschießen* s'il s'agit d'un homme (p. 35).

Une description exhaustive du français

Au-delà de ces principes fondamentaux, les qualités remarquables de l'ouvrage de Gaston Gross sont sa finesse et son exhaustivité. Il met en garde dès l'abord contre une vision naïve de la structure du lexique, qui permettrait une description à peu de frais du fonctionnement de la langue sous forme de règles simples, en posant des équations entre catégories morphosyntaxiques et sémantiques. Il établit ainsi

l'existence de prédicats nominaux et de prédicats adjectivaux, dont l'étude détaillée fait l'objet de plusieurs chapitres (chapitres 5 et 6 respectivement, le chapitre 8 étant consacré à l'étude spécifique de leurs verbes supports). S'il n'entre pas dans l'objectif de cet ouvrage de donner des descriptions extensives de certains cas d'usage[3], on y trouve en revanche un large panorama des catégories de phénomènes décrits et des questions qui surgissent lors de leur examen.

De même, de nombreuses pages sont consacrées, avec à chaque fois un souci de complétude, à l'analyse du fonctionnement des prépositions, des déterminants, des adverbes, des différents types de prédicats (événements, actions, états) et de l'expression des subordonnées et compléments circonstanciels. L'auteur développe notamment une réflexion très intéressante sur la base de l'intuition que la différence fondamentale entre compléments et circonstants peut être définie comme suit : le complément est l'argument d'un prédicat de premier ordre, alors que le circonstant est un prédicat de second ordre, dont les arguments sont eux-mêmes des prédicats (p. 52-53). Cette vision est très éclairante pour ceux qui se confrontent à la problématique de la correspondance entre une structure syntaxique et une structure sémantique.

Gaston Gross consacre enfin son plus long chapitre (chapitre 10) aux phénomènes de figement. L'auteur note très justement qu'on a affaire, avec le figement, à « un fait massif, qui doit être considérer comme une propriété définitionnelle des langues naturelles et qui a totalement échappé à la tradition grammaticale » (p. 27).

C'est également lorsque l'on confronte le modèle de Gaston Gross au cas des figements que l'on en découvre peut-être, nous semble-t-il, les limites – liées à ce qu'il implique de tracer des frontières entre catégories, sans pouvoir exprimer la liberté qu'a la langue de faire varier de façon continue la contrainte de restriction d'emploi liée au contexte. Malgré sa perception très fine des variations d'usage, Gaston Gross finit en effet toujours, dans l'optique de la constitution de dictionnaires électroniques, par résoudre les questions de description des emplois par le découpage en catégories étanches. C'est sans doute un mal nécessaire, mais cela impose une simplification parfois arbitraire aux possibilités d'usage de la langue.

Linguistique ou traitement automatique des langues ?

La question la plus fondamentale que l'on se pose, en fin de compte, à la lecture du *Manuel d'analyse linguistique* de Gaston Gross, est de savoir s'il s'agit d'un livre de linguistique ou d'un livre de TAL. On a le sentiment que l'auteur voudrait qu'il soit avant tout un livre de linguistique : cela transparaît dans le titre ou dans la

3 Au contraire de l'approche de la *Lexicographie explicative et combinatoire* de l'école Sens-Texte.

première phrase du texte présenté au lecteur en quatrième de couverture (ce livre est une présentation « […] d'une méthode d'analyse des mécanismes qui sont à la base du fonctionnement du français »). Et, en effet, l'ouvrage est intégralement consacré à la description de phénomènes de syntaxe ou de sémantique, sans s'inféoder à une quelconque école de représentation formelle.

Cependant, considérer comme un livre de pure description de la langue, le *Manuel d'analyse linguistique*, pourrait prêter le flanc à un certain nombre de critiques, comme la faible justification du nombre des hyperclasses ou du niveau de spécificité choisi dans le découpage des emplois des unités, la vision statique qu'il impose des paradigmes linguistiques, l'incapacité à représenter la gradualité des phénomènes de figement (alors même que celle-ci est affirmée en préambule, p. 202), ou la difficulté à gérer les recouvrements de classes sans faire appel à l'encombrant mécanisme de la métaphore.

En réalité, l'ouvrage de Gaston Gross est bien un livre qui traite essentiellement de phénomènes linguistiques, mais l'application informatique, qu'il ne perd jamais de vue, est le fondement qui lui impose ses contraintes de faisabilité (contraintes de finitude et de complétude, dans l'état de l'art actuel des formalismes de représentation de l'information lexicale et syntaxique). La formulation la plus honnête de ce qu'est fondamentalement cet ouvrage est donc celle qu'a choisie l'auteur dans la première phrase de son introduction : une « méthode d'analyse de la langue destinée au traitement automatique, vu du côté de la linguistique » (p. 7).

Conclusion

Le *Manuel d'analyse linguistique* est une contribution majeure aux travaux de description du français. Il apporte aux spécialistes de lexicographie électronique des méthodes de modélisation de la langue où la syntaxe est intégrée à la description du vocabulaire, et qui ne laisse dans l'ombre aucune partie importante du système de la langue française.

Les limites de l'approche de Gaston Gross sont d'une part qu'il est fortement centré sur le français, et que l'auteur ne donne pas toujours les clés pour pouvoir généraliser ses méthodes à d'autres langues, et, d'autre part, qu'il reflète un état de la pensée linguistique qui donne une vision statique du lexique. Ces limites ne sont pas un défaut spécifique au travail de Gaston Gross, mais reflètent un état de l'art en description linguistique formelle, et elles n'enlèvent rien au caractère monumental de ce qu'il faut appeler une somme dans le domaine de la description et de la modélisation de la langue. Il nous semble qu'il ne sera pas raisonnable, à l'avenir, de travailler sur un projet de linguistique informatique impliquant une description de la langue française sans se référer à l'ouvrage de Gaston Gross.

Noël NGUYEN, Martine ADDA-DECKER. Méthodes et outils pour l'analyse phonétique des grands corpus oraux. *Hermès-Lavoisier*. 2013. 320 pages. ISBN 978-2-7462-4530-3.

Lu par **Olivier CROUZET**

Université de Nantes, Laboratoire de Linguistique (EA3827)

Cet ouvrage collectif dirigé par Noël Nguyen et Martine Adda-Decker propose une description des fondements théoriques et méthodologiques de l'analyse phonétique de grands corpus oraux à travers les techniques issues de la reconnaissance automatique de la parole et fournit un aperçu de l'état de l'art dans ce domaine. Cet ouvrage est découpé en trois parties que l'on pourrait classifier ainsi : fondements théoriques et méthodologiques des analyses et annotations phonétiques (chapitres 1, 2, et 6) ; méthodes et problèmes de la mise en œuvre de segmentations automatiques de grands corpus (chapitres 3 et 4) ; illustrations par des études ciblées de phénomènes phonétiques et phonologiques (chapitres 5 et 7).

Le chapitre 1 propose une introduction aux principaux concepts de physique acoustique utiles à l'investigation phonétique (physique acoustique et modèle source-filtre, transformée de Fourier, codage par prédiction linéaire), puis offre une présentation des principes et problèmes liés à l'étude des relations articulatoire et acoustique suivie d'une description des connaissances actuelles sur les propriétés acoustiques des différentes macroclasses sonores du français. Le chapitre se termine par une approche des variations et des modifications plus ou moins marquées qui peuvent apparaître entre les propriétés « prototypiques » des classes phonétiques et leur réalisation effective dans les différentes conditions de production de la parole continue (coarticulation, réductions, modifications massives...).

Le chapitre 2 introduit les notions associées à la digitalisation des signaux et à leur analyse spectro-temporelle pour se concentrer ensuite sur la définition des concepts impliqués dans les procédures de segmentation et étiquetage phonétique. Après avoir discuté les connaissances et problèmes concernant la segmentation et l'étiquetage manuels de la « parole prototypique » et les conditions de leur extension à de la parole « non préparée », les auteurs rappellent le caractère purement conventionnel de la segmentation de la parole et les limites que ceci impose aux outils de segmentation automatique. Ils offrent ensuite une présentation critique des potentialités respectives offertes par les méthodes automatiques appliquées sur de grands corpus comparées à la mise en œuvre de « corrections manuelles », celles-ci étant nécessairement appliquées à des « petits » corpus. Cette analyse permet d'identifier quels problèmes théoriques peuvent se prêter à de telles approches dans l'état actuel des recherches.

Le chapitre 3 présente les principes de développement d'un corpus pour l'étude des formes sonores. Les aspects liés aux questions techniques et éthiques constituent la première partie du chapitre. Les auteurs présentent ensuite le développement

d'une base de données en illustrant leur propos par l'analyse des problèmes de choix des conventions de notation (mode de transcription, conventions de codage de phénomènes phonologiques, notation des tours de parole et des interventions ponctuelles des locuteurs...) puis abordent la question de la mise à disposition d'outils (site Web et interfaces). Le chapitre se termine par une description du développement en cours d'une plate-forme d'interrogation du corpus PFC (phonologie du français contemporain).

Le chapitre 4 présente, après un aperçu historique du développement des corpus oraux, les principes sous-jacents aux systèmes de transcription automatique de la parole. Les techniques de « modélisation statistique » de la parole sont explicitées dans le but de pouvoir décrire les composants d'un tel système : modèles acoustiques, modèles de langage et dictionnaires de prononciation. Les systèmes de transcription (actuels) sont vus comme des décodeurs dans lesquels le modèle de langage est inutile en raison de la mise à disposition de la séquence de mots à transcrire. Ce chapitre permet de bien comprendre les différents « étages » des systèmes de transcription automatique et fournit une base de réflexion précieuse sur le rôle que jouent ces différents sous-systèmes dans les analyses de grands corpus. S'ensuit une discussion sur la loi de Zipf qui décrit les relations s'établissant entre la fréquence des mots et leur rang de fréquence ainsi que les limites qu'elle impose pour l'analyse et la représentativité des données. Le chapitre se termine par quelques illustrations issues de travaux conduits dans des buts d'évaluation de ces procédures.

Le chapitre 5 fait partie des deux chapitres qui illustrent des travaux d'analyse linguistique issus des techniques discutées dans l'ouvrage. Sur la base des résultats venant d'une étude perceptive de six variantes dialectales de la langue française, les auteurs présentent deux ensembles de travaux portant respectivement sur des analyses acoustiques (mesures de formants) et phonologiques (variantes de prononciation). Ces études, réalisées sur un grand corpus segmenté automatiquement, illustrent parfaitement les principes discutés dans le chapitre 4. Plusieurs types d'analyses statistiques sont conduits sur ces données dans la perspective de rendre compte des potentialités offertes par l'analyse automatique de grands corpus pour l'avancement des connaissances linguistiques.

Le chapitre 6 décrit une sélection des différentes techniques disponibles pour mesurer les propriétés spectrales des voyelles en relation avec les échelles de mesure de la fréquence et/ou leurs différentes sources de variation. On y trouve d'abord une présentation des techniques de normalisation intrinsèques aux voyelles (transformations psychoacoustiques, estimation du second formant effectif –F_2', ratios / différences de fréquences). La seconde partie est dédiée à la présentation de techniques de normalisation extrinsèques aux voyelles (expression des fréquences de formants en fonction des mesures réalisées sur d'autres observations). On notera que la formule fournie pour l'estimation acoustique du second formant effectif (F_2', p. 246) comporte des omissions et des imprécisions qui la rendent inapplicable en l'état. La référence à Mantakas, Schwartz & Escudier (1986) est par ailleurs erronée puisque la formule indiquée est en réalité une simplification indépendante de

l'énergie des formants *dérivée* de Mantakas *et al.* (1986), mais présentée dans Schwartz, Boë, Vallée & Escudier (1997), qu'il conviendra de consulter pour obtenir une formule adéquate de l'estimation de F_2'.

Le chapitre 7 propose une vision englobant l'ensemble des questions abordées dans l'ouvrage pour mettre en perspective les méthodes décrites avec les potentialités offertes en termes de recueil de données pour l'analyse phonétique. Ce chapitre présente assez succinctement une série d'études ayant chacune cherché à évaluer l'impact de ces méthodes automatiques sur les analyses des voyelles. Les investigations présentées portent sur la localisation des frontières, les durées, les mesures spectrales, les variantes ciblées de prononciation.

Si cet ouvrage ne peut être vu comme un « manuel » pratique pour la mise en œuvre des méthodes décrites, il constitue une source précieuse d'information sur ce type d'approches. Il permettra à chacun (1) d'estimer les potentialités que peuvent ou non représenter ces approches pour les questions qui l'intéressent et (2) d'acquérir les bases théoriques et méthodologiques nécessaires pour préparer la mise en œuvre de ces techniques. Globalement, la lecture de cet ouvrage est très enrichissante. Elle fournit une approche critique des méthodes employées, même si émergent parfois des formulations discutables concernant l'interprétation statistique des données issues de ces outils (p. 217, par exemple) ou quelques omissions qui pourraient être gênantes pour certains lecteurs (p. 246). Il aurait aussi été appréciable que certaines limites ou solutions suggérées soient discutées plus en détail (approches et techniques spécifiques permettant de compenser les déformations liées à la loi de Zipf, p. 183 ; problèmes posés par les bases MySQL en termes de compatibilité avec d'autres formats, p. 143-144). Malgré les quelques défauts mentionnés, cet ouvrage constitue un apport scientifique considérable au développement des recherches en linguistique et contribuera à susciter le désir de mettre en œuvre ce type d'approches au sein de la communauté.

Ido DAGAN, Dan ROTH, Mark SAMMONS, Fabio Massimo ZANZOTTO. Recognizing Textual Entailment: Models and Applications. *Morgan & Claypool publishers*. 2013. 200 pages. ISBN 978-1-5982-9834-5.

Lu par **Natalia Grabar**

UMR 8163 Savoirs, Textes, Langage STL Université de Lille 3

---

*Le présent ouvrage, dédié à l'inférence textuelle (*textual entailment *ou* textual inference *en anglais), se propose de dresser le bilan de différents travaux menés dans ce domaine de recherche depuis plusieurs années. Nous cernons d'abord la notion de l'inférence textuelle et présentons ensuite le contenu de l'ouvrage.*

L'inférence textuelle consiste à établir une relation entre deux segments textuels, appelés Texte et Hypothèse. L'inférence textuelle est une relation directionnelle,

dans laquelle la vérité de l'Hypothèse peut être inférée à partir du sens du Texte, ou, en d'autres mots, il est possible de vérifier si l'Hypothèse est subsumée par le Texte (si, étant donné le Texte, l'Hypothèse est vraie, ou encore si le sens de l'Hypothèse est inclus dans le sens du Texte). Par exemple, le Texte « *The drugs that slow down or halt Alzheimer's disease work best the earlier you administer them* » permet d'inférer que l'Hypothèse *Alzheimer's disease is treated by drugs* est vraie ; en revanche, l'Hypothèse *Alzheimer's disease is cured by drugs* ne peut pas être inférée à partir de ce Texte.

La détection automatique des inférences textuelles s'avère utile pour différentes applications du TAL. Par exemple, dans les systèmes de questions/réponses, l'inférence textuelle permet de valider ou de réordonner les réponses candidates retrouvées ; en extraction de relations ou d'information de manière générale, l'inférence textuelle permet d'extraire et de relier des candidats supplémentaires grâce à l'établissement de liens indirects calculés avec l'inférence textuelle ; en résumé automatique, l'inférence textuelle peut servir pour mesurer la similarité sémantique entre deux phrases, mais aussi pour vérifier que le résumé généré est en effet subsumé par le texte traité. Par ailleurs, l'inférence textuelle peut également être utilisée dans les tâches d'évaluation pour établir des relations indirectes entre les données de référence et les résultats générés automatiquement.

La complexité de la détection d'inférences textuelles implique l'exploitation et la combinaison de plusieurs fonctionnalités et modules, comme par exemple le prétraitement du Texte et de l'Hypothèse, leur enrichissement avec des ressources dédiées, la génération d'alignements, la sélection d'alignements générés, la classification et la décision finale sur le statut de l'Hypothèse (inférée/vraie, non inférée/fausse, et parfois aussi contradictoire ou inconnue).

L'ouvrage *Recognizing Textual Entailment: Models and Applications* propose une revue des travaux menés autour de l'inférence textuelle depuis plusieurs années. La présentation de travaux est d'autant plus exhaustive et pertinente que l'un des auteurs a été le fondateur et promoteur de ce domaine de recherche. En collaboration avec d'autres auteurs, il a su structurer le domaine, en particulier grâce à l'organisation de compétitions de TAL dédiées à la reconnaissance de l'inférence textuelle (*Recognizing Textual Entailment* ou RTE).

L'ouvrage est structuré en six chapitres, qui peuvent être lus en continu ou de manière indépendante. L'ouvrage est également accompagné d'une annexe et d'une bibliographie représentative et exhaustive, qui viennent appuyer les propos des auteurs.

Le chapitre 1 décrit le contexte dans lequel se situe la recherche sur l'inférence textuelle. L'accent est mis en particulier sur les raisons et les besoins qui la motivent, en indiquant les utilisations possibles de l'inférence textuelle dans les applications de TAL (extraction de relations et d'information, questions/réponses, évaluation des réponses, résumé automatique, résumé automatique multidocument, traduction automatique, évaluation de la traduction automatique, évaluation des

analyseurs syntaxiques...). Les auteurs proposent également une description des compétitions de RTE établies depuis 2004 et de leur évolution. Cette évolution reflète une attention particulière consacrée, par exemple, à :

– la création des données de référence, avec le souci d'obtenir un meilleur consensus entre les annotateurs et de couvrir des phénomènes diversifiés ;

– la représentativité de l'inférence textuelle dans les données réelles (alors que, dans les premières compétitions les inférences textuelles apparaissaient dans 50 % des données, dans les compétitions suivantes elles ont désormais une distribution moindre et plus naturelle) ;

– l'évolution des mesures d'évaluation ;

– la dépendance des inférences textuelles au contexte ou au discours de manière plus générale ;

– l'étude de l'impact de différents modules et ressources utilisés par les systèmes, en particulier grâce aux *ablation tests*, qui correspondent aux tests supplémentaires et pour lesquels il est demandé aux participants de supprimer une ressource ou un module donné de leur système afin de mieux analyser son impact sur la tâche de RTE.

Le chapitre 2 débute par la présentation d'un modèle intuitif d'inférence textuelle, qui repose sur une simple subsomption lexicale entre le Texte et l'Hypothèse. Ce modèle est impuissant face à plusieurs situations, comme le changement de dépendances syntaxiques, la synonymie ou la négation. Face à ces limites, les auteurs introduisent quelques modules communément utilisés pour l'inférence textuelle, comme par exemple, l'analyse syntaxique et l'établissement des dépendances entre les constituants, l'étiquetage des rôles sémantiques des constituants, le calcul des similarités lexicales ou syntaxiques entre le Texte et l'Hypothèse, l'alignement des segments de texte, l'application des formalismes logiques (logique de proposition, logique du premier ordre, démonstration de théorèmes, clauses de Horn...), les règles de transformation des arbres syntaxiques, l'utilisation de ressources (PropBank, FrameNet, WordNet, VerbNet, VerbOcean, DIRT) ou de bases de connaissances dédiées. Finalement, les auteurs introduisent l'architecture typique d'un système de RTE, qui comprend les étapes indiquées plus haut : le prétraitement du Texte et de l'Hypothèse, leur enrichissement avec des ressources dédiées, la génération d'alignements, la sélection d'alignements générés, ainsi que la classification et la décision finale sur le statut de l'Hypothèse.

Dans le chapitre 3, les auteurs montrent la nécessité d'utiliser l'apprentissage automatique pour l'établissement de la décision finale sur le statut de l'Hypothèse. Les raisons principales sont que plusieurs facteurs peuvent contribuer à cette décision et, de plus, ces facteurs peuvent cacher des imperfections et des incertitudes héritées des étapes précédentes (comme les erreurs produites par les outils automatiques), des insuffisances des ressources disponibles ou encore de la complexité de l'espace de décision. Les auteurs introduisent ainsi des exemples de descripteurs générés par les étapes précédentes (par exemple, les descripteurs

lexicaux ou syntaxiques, la polarité, les structures passives) et pouvant être exploités pour la prise de la décision. De même, la fonction d'apprentissage peut être utilisée pour effectuer d'autres étapes prises en charge par le système (calcul de la similarité, l'alignement...).

Le chapitre 4 présente quelques systèmes individuels. Il s'agit essentiellement de systèmes propriétaires d'équipes de recherche travaillant sur le sujet, mais notons qu'un système (EDITS) est utilisable librement. Le fonctionnement de ces systèmes (méthodologies et ressources exploitées) est décrit en faisant le lien avec les modèles et modules introduits dans les chapitres 2 et 3. Par ailleurs, les performances obtenues par ces systèmes lors des compétitions de RTE sont présentées. Ces performances ne permettent toutefois pas de positionner l'efficacité des méthodologies et des ressources les unes par rapport aux autres, essentiellement à cause de la très grande variété dans la conception des systèmes, ainsi que des utilisations fort différentes qui peuvent être faites d'une ressource ou d'une méthodologie données. En revanche, l'influence de certains modules et de certaines ressources sur le fonctionnement d'un système donné est observable grâce aux *ablation tests*. Les auteurs font également le constat qu'il est impossible de savoir quelles sont les meilleures approches pour la tâche d'inférence textuelle actuellement.

Le chapitre 5 se focalise sur l'importance des ressources et des connaissances adaptées pour le fonctionnement des systèmes de RTE. Il apparaît, en effet, que les ressources actuellement disponibles sont insuffisantes, ce qui se trouve être très souvent la cause principale des faiblesses des systèmes automatiques. Afin de motiver les travaux de ce type, les auteurs proposent une présentation des principales méthodes orientées sur l'acquisition de ressources sémantiques et des connaissances pour l'inférence textuelle. L'inférence textuelle peut, en effet, introduire des contraintes spécifiques vis-à-vis des ressources existantes, comme par exemple l'orientation des relations (l'inférence textuelle est une relation orientée, alors que la relation de synonymie n'est pas orientée), ou encore leur exploitation et utilité en fonction des contextes et des types de documents. Une des conséquences est que les ressources souhaitables pour la tâche de RTE doivent souvent être spécifiquement conçues ou adaptées. Lors de la présentation des méthodes, les auteurs font la distinction entre l'exploitation ou l'adaptation de ressources construites manuellement (ressources de la famille WordNet, FrameNet, différents dictionnaires existants ou les Wikipedias) et l'exploitation de corpus avec des méthodes distributionnelles ou par cooccurrence. Notons que les corpus exploités peuvent être monolingues ou bien avoir des caractéristiques plus spécifiques (monolingues parallèles, monolingues comparables ou bilingues parallèles). Comme il a été noté précédemment, une difficulté supplémentaire est liée au fait que les connaissances et les ressources n'ont pas de validité absolue : la possibilité de leur utilisation adéquate dépend des contextes au sens large (domaine, corpus, contexte donné...).

Le chapitre 6 propose une conclusion et souligne encore les directions de futurs travaux de recherche nécessaires dans le domaine de RTE. Il s'agit par exemple de modules de prétraitement plus efficaces et de méthodes plus performantes et exhaustives pour l'acquisition de connaissances et de ressources dédiées. Les auteurs proposent aussi plusieurs pistes pour l'évolution de l'envergure des compétitions de RTE. Ainsi, la disponibilité de plates-formes ouvertes ou bien de systèmes interopérables permettrait de mieux structurer et motiver les travaux sur l'inférence textuelle et de dynamiser l'évolution de ce domaine de recherche.

De manière générale, il s'agit d'un ouvrage très précis et didactique, qui peut satisfaire plusieurs types de lecteurs. Les chercheurs qui voudraient se lancer dans le développement d'un système pour la détection de l'inférence textuelle peuvent trouver une description de l'état de l'art actuel du domaine, de même que la description des systèmes participant dans les compétitions de RTE, ainsi que des modules et des ressources utilisés habituellement. Les étudiants qui voudraient découvrir ce domaine de recherche peuvent exploiter cet ouvrage et en trouver une description précise. Les chercheurs qui voudraient contribuer à l'évolution des capacités du domaine peuvent trouver les directions de recherche futures, dont la nécessité est observée par les systèmes et les méthodes actuels. Les enseignants, qui voudraient monter un cours sur l'inférence textuelle, en particulier parce qu'un tel cours permettrait de donner une visée assez globale et imbriquée des travaux de TAL, peuvent utiliser cet ouvrage et y trouveront les éléments nécessaires et suffisants.

Frédéric LANDRAGIN. Dialogue homme-machine : conception et enjeux. *Hermès-Lavoisier*. 2013. 210 pages. ISBN 978-2-7462-4522-8.

Lu par **Laurie ACENSIO**

Altra Consulting et Lip6 (Laboratoire informatique, Paris 6)

Le domaine du dialogue homme-machine (DHM) vise à permettre une meilleure interaction entre l'homme et la machine en se fondant, en autres, sur le dialogue en langage naturel. Ce travail de recherche universitaire aborde les théories, méthodes et techniques à chaque étape de la conception d'un programme informatique capable de comprendre et de produire de la parole. En effet, la réalisation d'un système de DHM est complexe et soulève de nombreuses problématiques dont principalement celles liées à l'interprétation de la dynamique du langage en entrée, les traitements et raisonnements internes au système et la gestion des messages en sortie avec la génération automatique et la présentation d'information multimédia. Dans son ouvrage, l'auteur s'appuie sur des exemples de dialogues liés au système de réservation de billets de train prévu dans le cadre du projet européen Ozone (Issarny et al.*, 2005).*

L'ouvrage se décompose en trois grandes parties : les repères historiques et méthodologiques, les traitements des entrées, et enfin le comportement du système et son évaluation. La première partie présente un bref historique des expérimentations passées (système Eliza, Parry, Shrdlu, Trains) jusqu'à l'explosion des projets et des techniques à partir des années 1990. Les travaux scientifiques actuels mettent en avant l'amélioration des DHM *via* des techniques informatiques d'apprentissage automatique pour le traitement de gros corpus, l'essor des efforts de standardisation (W3C, ISO, TEI, DAMSL…), la multiplication des modalités de communication avec les modèles du dialogue multimodal et l'ouverture avec d'autres domaines scientifiques récents comme la robotique et le SQR. En effet, le domaine du DHM étant interdisciplinaire, de nombreuses thématiques sont rapidement abordées dans les deux premiers chapitres dont l'intelligence artificielle, l'interaction homme-machine et les agents conversationnels animés.

Cependant, ce sont avec les analyses du TAL (lexicales, syntaxiques, sémantiques et pragmatiques) et l'étude des dialogues avec les approches de l'analyse conversationnelle que le domaine du DHM entretient des liens privilégiés. Le troisième chapitre décrit les étapes de réalisation d'un système de DHM *via* des implémentations symboliques et/ou statistiques tandis que le quatrième chapitre aborde la question des architectures logicielles et de ses enjeux cruciaux tels que la réutilisabilité et la conception de modèle générique.

L'une des facettes essentielles d'un système de DHM est sa capacité à identifier les actes de langage en entrée, ce qui est largement abordé dans la deuxième partie.

Le cinquième chapitre est centré particulièrement sur le traitement des énoncés en entrée et leurs propriétés prosodiques, lexicales, syntaxiques et sémantiques. L'auteur insiste particulièrement sur la reconstruction du sens explicite et implicite de l'énoncé afin que celui-ci coïncide au mieux avec les intentions de l'utilisateur. Les principales caractéristiques du langage naturel sont abordées dont la polysémie, la métonymie, la métaphore, la sémantique verbale, l'implicite, l'ambiguïté et la structure informationnelle. Les techniques actuelles trouvent leurs limites notamment dans l'identification des ambiguïtés, la gestion des mots inconnus, l'identification d'un usage non littéral de la langue (ironie, sarcasme, exagération). Pour résoudre ces problèmes, les systèmes de DHM doivent abandonner le défi d'une compréhension complète du langage et privilégier plutôt une analyse syntaxique locale ou partielle qu'une analyse syntaxique globale. Il en est de même pour l'analyse sémantique qui peut rester incomplète mais qui doit aller de paire avec une analyse pragmatique.

Le sixième chapitre aborde la question de la résolution des références en contexte notamment les références à des objets ou à des actions, et le cas particulier des références issues de l'historique du dialogue. Pour que le système puisse détecter des énoncés reliés entre eux par des relations d'anaphores et de coréférences, il est utile de mettre en place des processus en plusieurs étapes :

analyse du contexte visuel, analyse des trajectoires gestuelles et analyses linguistiques (sémantique verbale, rôles actanciels, aspects temporels).

Le septième chapitre présente les méthodes utilisées pour la détection automatique des énoncés et, plus précisément, les processus de « haut niveau » qui concernent le sens des énoncés. Généralement, les énoncés portent sur des actes de requêtes, de dialogue ou des questions ; cette diversité rendant leur identification complexe. Des techniques de classification comme la FIPA (*Foundation for Intelligent Physical Agents*) de DAMSL (*Dialog Act Markup in Several Layers*) proposent une classification hiérarchique avec différents niveaux de granularité, contrairement à un annotateur de corpus qui se contentera d'exploiter les types d'actes de premier niveau. Ce chapitre insiste sur l'importance de la qualité de l'identification des actes de langage, essentielle pour que le système effectue les raisonnements nécessaires pour produire une réaction en retour. Cependant, l'interaction peut être faussée car les limites du système sont rapidement perceptibles par l'utilisateur, celui-ci adaptant ainsi ses comportements et ses énoncés. Cette attitude constitue le principal frein d'un DHM car elle résulte d'une faible adhésion de l'utilisateur. Pour cela, on insiste sur l'importance de traiter les tâches de dialogue en priorité plutôt que toute autre technique d'interaction (interaction haptique et/ou visuel, capture des émotions faciales et des gestes de désignation).

Le huitième chapitre aborde cette problématique de traitement en sortie, avec des exemples de stratégies de dialogue pour optimiser les sorties d'échanges vocaux de manière fluide et spontanée. Les aspects techniques de la gestion du dialogue comportent trois phases : le contrôle du dialogue, qui cherche à gérer le processus interactif de manière à déterminer un type de réaction, la modélisation du contexte de dialogue, et l'initiative, qui ajoute un comportement particulier.

Dans le neuvième chapitre, l'auteur aborde les principes généraux pour la conception des modules chargés des sorties des systèmes, notamment les techniques de génération automatique de messages linguistiques et multimodaux. Les méthodes linguistiques sont semblables à celles des traitements des entrées, mais inversées. Lors de la détermination des messages en langage naturel, il s'agit d'appliquer les maximes de Grice, de minimiser les risques d'ambiguïtés en anticipant sur les conditions de production de celles-ci, d'éviter les actes de langage indirect et composite et d'exploiter la structure informationnelle. Il apparaît clairement, tout au long de ces chapitres, que la transcription de la prosodie est une problématique commune dans le traitement des entrées et des sorties, celle-ci étant déterminante dans la reconnaissance et la compréhension automatiques des énoncés. L'analyse des actes de langage est facilitée si l'on repère le contour intonatif du discours. Par exemple, la détection automatique de la fin de l'énoncé permet de privilégier une hypothèse sémantique par rapport à une autre ; le système adoptant ainsi une intonation en lien avec l'intention de communication du système.

Enfin, le dernier chapitre présente les méthodologies des évaluations utilisées en DHM oral et multimodal. L'auteur distingue l'évaluation globale qui s'intéresse à la pertinence des énoncés échangés *via* les traces d'interaction tandis que l'évaluation segmentée se focalise sur les entrées et les sorties de chaque module. La fiabilité des résultats reste néanmoins fragile car les systèmes sont conçus pour une tâche donnée, rendant difficile toute évaluation comparative ou normée.

L'ouvrage constitue une excellente approche pour aborder les différents aspects informatiques, cognitifs et linguistiques des DHM. L'auteur souligne l'importance d'une méthodologie pluridisciplinaire, qui intègre l'expérimentation, les études de corpus et les confrontations de théories pour leur application au DHM. La richesse des travaux scientifiques montre que le DHM n'est pas une impasse et que de nombreuses pistes d'ordre linguistique, dialogique et technique sont à explorer. Les enjeux linguistiques sont, entre autres, une meilleure gestion de la saillance, de la redondance et des références, un meilleur contrôle de la production volontaire des ambiguïtés, et l'analyse en temps réel de la prosodie. Actuellement, il y a une remise en cause des découpages classiques des niveaux d'analyses linguistiques longtemps réalisées en cascade et qui tendent à devenir collaboratives. D'un point de vue architectural, c'est l'hybridation d'approches symboliques et statistiques qui affiche des résultats prometteurs. Nous pouvons tenter de résumer en disant que la principale difficulté des DHM est de pouvoir fonctionner sur un domaine ouvert et connecté au Web : des modèles généralistes, mais orientés, peuvent pallier les difficultés de la reconnaissance de la parole. L'essor du Web sémantique *via* les langages standardisés constitue également une opportunité pour les DHM actuels pour exploiter en temps réel les ressources du Web et améliorer la qualité des données. Une autre possibilité est de formaliser le langage naturel avec les contraintes de la logique modale, temporelle et hybrides, ce qui ouvre la voie à de nouvelles problématiques.

Emily M. BENDER. Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax. *Morgan & Claypool publishers*. 2013. 166 pages. ISBN 978-1-6270-5011-1.

Lu par **Marie CANDITO**

Alpage – Université Paris Diderot, INRIA

Cet ouvrage recense cent points de linguistique considérés comme fondamentaux pour la conception de systèmes de traitement automatique de la langue écrite, en se concentrant surtout sur la morphologie, la syntaxe et l'interface syntaxe et sémantique.

Ce livre a pour objectif de fournir des connaissances linguistiques qui soient « directement pertinentes pour la conception de systèmes de TAL » (*immediately*

relevant to the design of NLP systems), avec une emphase sur les connaissances en morphologie et en syntaxe.

Une première remarque concerne l'emphase sur la morphologie et la syntaxe : l'argument fourni pour cette emphase est que ces niveaux sont particulièrement pertinents pour le TAL. On peut cependant remarquer que les informations sémantiques sont évidemment cruciales et pertinentes pour le TAL, l'argument pourrait plutôt être que les connaissances sémantiques sont, à ce jour, encore difficilement accessibles automatiquement et donc moins utilisées.

Cet ouvrage comporte un chapitre introductif, puis huit chapitres détaillant des informations linguistiques pertinentes pour le TAL et un dernier chapitre sur les ressources. Un point fort du livre est que ses sections, numérotées de 1 à 100, à travers les chapitres, ont des titres très précis et explicites et peuvent presque se lire comme un résumé de l'ouvrage.

Après un chapitre d'introduction générale, le chapitre 2 donne une introduction à la morphologie, en insistant sur des aspects éloignés de l'anglais comme la possible non-contiguïté des phones d'un morphème ou des langues à tons. Une distinction entre morphèmes lexicaux, flexionnels et dérivationnels est présentée. La notion de mot est introduite sous l'angle de la difficulté à réaliser une segmentation en mots. Cependant, cette notion de mot, justement, n'est pas vraiment définie. On dirait que l'auteure nous propose une définition intuitive. Si la notion de clitique est amplement couverte, bizarrement les notions de mot composé et d'unité polylexicale ne sont pas du tout abordées, alors qu'il s'agit pourtant d'un domaine très couvert en TAL. Une typologie grossière des morphologies est fournie et l'autrice introduit le ratio morphème/mot comme outil de typologie morphologique.

Le chapitre 3 traite de la morphophonologie. Des termes comme « paradigme » ou « élicitation » sont écrits en italique, mais n'ont pas de définition précise. Les deux premières sections concernent l'allomorphie, respectivement d'après le contexte phonologique et le contexte morphologique. L'auteure choisit des exemples précis pour amener des notions, comme la marque du pluriel en anglais ou bien le processus d'harmonie vocalique en turc, qui constituent un exemple où la forme d'un morphème est imposée à distance. L'auteure pointe les divergences entre les prononciations et l'orthographe, comme difficultés supplémentaires auxquelles doivent faire face les systèmes de TAL.

Le chapitre 4 traite de morphosyntaxe, au sens restreint des aspects morphologiques ayant un impact en syntaxe. Suit une liste des types d'informations qui peuvent être marquées morphologiquement dans certaines langues. L'autrice fournit succinctement, mais clairement, les grands types de paradigmes flexionnels, avec des exemples empruntés à des langues variées. Les sections suivantes traitent des phénomènes d'accords. Le chapitre se termine avec des considérations d'ordre typologique un peu redondantes.

Le chapitre 5 est une introduction rapide (trois pages) à la syntaxe. On est ici un peu gêné par le fait que l'auteure ne définit pas vraiment les notions utilisées (par exemple la définition de grammaticalité est tautologique) et reste très général : il y a fort à parier que ce chapitre est peu éclairant pour un néophyte en linguistique, tant il faut lire entre les lignes.

Le chapitre 6 concerne les parties du discours. La définition distributionnelle y est présentée. L'autrice présente également une définition fonctionnelle, censée permettre des comparaisons de catégories entre les langues. Elle insiste sur la difficulté à définir un jeu parfaitement universel de catégories, en particulier pour les catégories fermées. Puis est évoqué à mots couverts le fait que la catégorie d'une tête fournit le typage du constituant qu'elle projette, mais de manière vraiment trop simpliste (« un groupe verbal comme *eats tomatoes* est verbal car il contient un verbe ») (même si la définition de la tête est fournie dans le chapitre 5).

Le chapitre 7, intitulé « *Têtes, arguments et ajouts* » commence par poser (sans vraiment la définir) la notion de constituant. L'auteure choisit de voir des constituants implicites y compris dans des arbres de dépendances, ce qui est pour le moins troublant. La possible non-projectivité permise par les représentations en dépendances n'est pas du tout abordée et l'autrice cite plusieurs fois par la suite la notion de « constituant discontinu » sans définir précisément ce qui leur conférerait un statut de constituant et sans expliquer comment les représenter. La section 52 introduit la notion de tête, valable pour « la plupart des constituants ». Tous les exemples sont ici fournis pour l'anglais et, à ce stade, rien ne vient nuancer l'existence des constituants à travers les langues ou les théories linguistiques. Les sections suivantes introduisent la notion de prédicat, d'argument sémantique et la distinction entre argument et ajout, en prenant quasi exclusivement des exemples anglais et en citant des ressources pour l'anglais comme PropBank ou FrameNet. Le chapitre se termine par une discussion sur l'intérêt, malgré la difficulté, de distinguer argument et ajout en TAL.

Le chapitre 8 présente les notions des rôles sémantiques et des fonctions grammaticales et la notion de « *linking* » entre arguments sémantiques et fonctions grammaticales. Pour illustrer son propos, l'autrice prend d'abord des exemples en anglais, cite majoritairement la ressource ERG (*English Resource Grammar*) (Flickinger, 2002 ; 2011) et un peu FrameNet. D'autres ressources comme VerbNet ne sont pas citées. Globalement, le propos semble un peu confus, avec une présentation très rapide des notions et des ressources anglaises, sans que les choix théoriques sous-jacents soient définis précisément. D'autres langues sont citées pour illustrer des cas de « constituants discontinus », dont la définition précise n'est pas donnée, ou plus généralement des langues à ordre (plus) libre, où les cas ou l'accord peuvent marquer les fonctions grammaticales. L'autrice redonne alors, comme dans la partie concernant la morphologie, des évaluations quantitatives de la prévalence de phénomènes à travers les langues. Ce chapitre se termine par une introduction aux changements de diathèse (sans les nommer), qui seront repris aux chapitre 9.

Le chapitre 9 passe en revue une série de phénomènes de divergences entre les positions syntaxiques et les rôles sémantiques : changements de diathèse, montée, contrôle, coordination, dépendances à longue distance, omission d'arguments…

Le dernier chapitre est assez difficile à cerner. Intitulé « *Ressources* », il comprend trois courtes parties sur les analyseurs morphologiques, les analyseurs syntaxiques « profonds » et les bases de données typologiques sur les langues. Les ressources citées, en particulier dans deux premières parties, sont forcément très parcellaires et l'autrice promeut à plusieurs reprises le projet DELPH-IN.

Globalement, l'entreprise était assez difficile : comment, en effet, décider quelle connaissance sera plus pertinente qu'une autre pour le TAL ? L'ouvrage contient finalement les connaissances typiques en morphologie, syntaxe et interface syntaxe et sémantique.

Dans les parties sur la morphologie, il semble que les concepts linguistiques clés sont abordés clairement, avec le souci de couvrir des phénomènes relevant de différentes familles de langues. L'auteure cite souvent des évaluations quantitatives de la prévalence de phénomènes linguistiques à travers les langues, notamment en citant le *World Atlas of Language Structures Online* (Dryer et Haspelmath, 2011). Cependant, la rédaction est « anglo-centrée » : les exemples sont souvent fournis en contraste par rapport à l'anglais. En outre, l'auteure pointe les particularités de l'anglais, par rapport à d'autres langues, lorsqu'elle suppose qu'elles ont eu un impact sur la conception de systèmes de TAL et qu'il est important d'en être conscient pour obtenir des systèmes plus généraux ou plus portables à d'autres langues.

En revanche, pour les parties sur la syntaxe, l'autrice donne beaucoup moins d'exemples non anglais. Elle présente les grammaires de dépendances comme des avatars des grammaires de constituants et donne une place probablement exagérée à la *English Resource Grammar*. Enfin, il manque une partie sur les unités polylexicales, qui constituent pourtant un champ très actif en TAL.