Notes de lecture

Rubrique préparée par Denis Maurel

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

Kristiina JOKINEN, Michael McTEAR, Spoken Dialogue Systems, *Morgan & Claypool publishers*, 2010, 151 pages, ISBN 9781598295993.

Lu par **Sophie ROSSET et Anne VILNAT**

LIMSI – CNRS, Orsay

Le livre de Kristiina Jokinen et Michael McTear « Spoken Dialogue Systems *» présente un panorama très didactique du dialogue oral en domaine restreint. La présentation en est très claire, avec des guides de lecture, des introductions et récapitulatifs à la fin de chaque chapitre.*

Un premier chapitre introductif situe le problème, avec des exemples de systèmes et la présentation des chapitres suivants. En particulier sont présentés rapidement les aspects suivants : architecture simple d'un système de dialogue oral, les domaines d'application, les problématiques liées à la collecte de données et aux corpus de dialogue, l'évaluation des systèmes de dialogue. Quelques outils disponibles pour le développement de systèmes de dialogue sont également décrits.

Le chapitre 2 situe le cœur de la recherche dans ce domaine sur la gestion du dialogue, en distinguant gestion et contrôle. Ce chapitre met bien en perspective l'importance de cet aspect.

Différentes approches sont présentées pour le contrôle du dialogue, dont les graphes et les schémas qui sont les plus utilisés depuis de très nombreuses années. Pour ces deux familles d'approches, de nombreuses références sont listées et parfois commentées. Ce chapitre distingue également les deux interprétations possibles qui se cachent sous les termes de « modélisation du dialogue », à savoir la modélisation du processus interactif (allant de la gestion des tours de parole aux compétences communicatives) et la gestion des informations nécessaires à une interprétation contextuelle (en contexte de dialogue donc) et au contrôle du dialogue. Ce dernier point fait l'objet d'une présentation détaillée dans ce chapitre. Les notions d'historique du dialogue, de modèle de la tâche et de modèle du domaine sont présentées. Les exemples sont toujours simples et s'appuient sur des systèmes, décrits dans la littérature, fonctionnant en domaine limité. Un exemple détaillé, s'appuyant sur VoiceXML est présenté.

Pour finir ce chapitre sur la gestion du dialogue, des approches plus récentes s'appuyant sur des modélisations statistiques sont présentées. En particulier les travaux fondés sur de l'apprentissage renforcé (*reinforcement learning*) et son intérêt dans le cadre de l'utilisation fondée sur des processus markoviens de décision (*Markov Decision Process*) sont clairement décrits. Naturellement, cette description amène les auteurs à présenter les modèles basés sur des processus de décision markoviens partiellement observables (*Partially observable Markov decision process*, POMDP) qui sont actuellement très en vogue.

Le chapitre 3 est consacré à la gestion des erreurs, qui sont un verrou important du domaine : comment détecter qu'une erreur s'est produite et comment y remédier. Après avoir présenté rapidement les deux stratégies possibles, à savoir la demande explicite de confirmation et la demande implicite de confirmation, les auteurs concluent que l'approche la moins coûteuse (en terme de temps et de satisfaction utilisateur) est la demande implicite de confirmation. Celle-ci implique que les systèmes soient en mesure de détecter les erreurs et de les « réparer ». Les approches classiques pour traiter ces deux aspects sont présentées. Ensuite, une approche plus récente, qui consiste à utiliser des stratégies équivalentes à celles des humains en situation de dialogue, est décrite. Elle est abondamment illustrée par des exemples tirés de la littérature.

Le chapitre 4 entre dans le cœur de systèmes qui mettent en pratique les principes de gestion de dialogue présentés au chapitre 2. En premier lieu sont introduits les systèmes fondés sur les états mentaux, ensuite ceux issus du domaine de l'intelligence artificielle (IA) et plus particulièrement de la planification, développés autour d'Allen à Rochester. Les auteurs s'intéressent ensuite aux travaux autour des agents logiciels. Tous ces aspects sont abondamment illustrés par des exemples de systèmes décrits dans la littérature. Ceci permet de mettre en évidence une typologie des différents systèmes existants.

Le chapitre 5 présente les pistes de recherche actuelles, à savoir l'amélioration des compétences du système, sa meilleure perception de l'utilisateur et l'introduction de la multimodalité. Les pistes explorées par les auteurs concernent donc essentiellement celles qui s'appuient sur des objectifs en terme de modélisation : coopération et compétence communicationnelle, adaptation et modélisation des utilisateurs, multimodalité et interaction naturelle. On peut ici être surpris de voir que la naturalité de l'interaction n'est pas considérée par les auteurs comme étant liée à la compétence communicationnelle.

Le dernier chapitre est consacré à l'évaluation, qui constitue un problème difficile dans ce domaine. Les difficultés principales sont listées ainsi que les différentes méthodes d'évaluations de systèmes complets et autonomes. La problématique de la collecte de données est clairement distinguée de celle de l'évaluation (même si, dans les faits, il y a souvent un rapport). Les différentes métriques, tant pour l'évaluation automatique et objective (évaluation de résultats bruts des différents modules) que pour l'évaluation dite subjective (par des

utilisateurs plus ou moins naïfs), sont clairement décrites. Les principaux cadres d'évaluations sont également présentés et discutés, en particulier PARADISE et l'évaluation de la qualité de l'expérience (*Quality of Experience Evaluation*). Deux aspects importants sont également abordés : la semi-automatisation des évaluations (notamment avec la présentation du concept d'utilisateurs simulés) et la standardisation des évaluations (une telle standardisation des procédures et des métriques permettrait de réellement comparer les qualités et les défauts des différentes approches et méthodes).

La conclusion évoque des points à creuser encore pour permettre un développement plus important des systèmes de dialogue. Le premier évoque le fait de dépasser les systèmes en domaine fermé, en évoquant les descendants d'Eliza, le second concerne quant à lui les applications industrielles.

Globalement, le panorama dressé par cet ouvrage est donc intéressant. Les principaux reproches qui peuvent être faits sont de deux ordres : il occulte presque complètement les systèmes de dialogue en domaine ouvert, et il laisse supposer que les travaux sur le dialogue ont commencé dans les vingt dernières années.

Pour le premier point, il est dommage que jamais ne soit évoqué l'ensemble des problèmes que pose le fait d'ouvrir le domaine d'application. Tous les exemples qui illustrent l'ensemble des chapitres ne portent que sur des systèmes de dialogue en domaine limité. Dans ce cadre, il est évident que la gestion du dialogue (le cœur de cet ouvrage) se trouve simplifiée puisqu'elle peut s'appuyer sur un modèle du domaine. Que se passe-t-il lorsqu'on décide d'ouvrir le domaine ou simplement de l'élargir ? Ces points ne sont absolument pas évoqués et cette absence d'évocation pourrait laisser penser qu'il n'y a pas de problèmes à un tel passage. Or il y a eu de nombreux travaux, en Europe, aux États-Unis et au Japon qui ont montré au contraire toute la difficulté de cette opération. Certains ont proposé de s'appuyer sur des systèmes fondés sur un routage préalable des demandes de l'utilisateur qui permet ensuite de s'adresser à un système spécialisé. Cette méthode présente l'avantage de ramener le problème de modélisation dans un domaine connu comme dans les systèmes précédents mais impose à un utilisateur de n'avoir qu'une catégorie de demandes à la fois. D'autres ont tenté de proposer des modèles d'un plus haut niveau, s'appuyant davantage sur un modèle de la tâche (par exemple la recherche d'information) que sur un modèle du domaine. Dans ce cadre se pose toute la question de la complexité de la langue naturelle. Il est dommage que cet aspect, pourtant crucial pour l'avenir, ne soit même pas évoqué.

Pour le second, on ne peut pas forcément faire un historique complet d'une discipline dans ce type d'ouvrage, mais il ne faudrait pas oublier que les ancêtres de ces systèmes ont été développés bien avant le milieu des années 90. Certes, ces travaux étaient développés plutôt dans le cadre du traitement de l'écrit que de l'oral. Il n'empêche qu'on ne peut les ignorer complètement, ce que tend à laisser penser en particulier le chapitre 2 de ce livre.

En résumé, il s'agit d'un livre à recommander pour des étudiants ou des chercheurs qui veulent s'initier à ce domaine, mais en présentant bien les limites : uniquement le dialogue en domaine fermé, et sur les 20 dernières années.

Claire BLANCHE-BENVENISTE, Approches de la langue parlée en français, *Éditions OPHRYS*, 2010, 175 pages, ISBN 2-7080-1278-9.

Lu par **Fabrice MAUREL**

GREYC, université de Caen

Le propos de l'auteur est de nous rappeler ou de nous convaincre que si l'oral a mis longtemps à s'imposer comme un objet d'étude linguistique à part entière, il est aujourd'hui devenu incontournable ; à la fois grâce au développement de nouveaux outils spécifiques et à une évolution des mentalités : réorientations des théories, nouvelles grammaires, taille des corpus disponibles, prise en compte de nouvelles dimensions (cognitives, interactives, pragmatiques, sociolinguistiques) du langage. La conclusion générale est que l'étude du français parlé fait progresser l'étude du français en général. Faut-il encore pouvoir analyser de très grands recueils de productions orales très diversifiées. Le développement des outils du TAL et l'importance prise par le dialogue oral homme-machine permettent d'espérer un investissement accru de la recherche dans ce sens. Claire Blanche-Benveniste est une des principales contributrices à cette prise de conscience. Elle regroupe, dans cette nouvelle version d'un ouvrage publié sous le même titre en 2000, une vingtaine d'années d'expériences, d'analyses et de réflexions sur le sujet, menées avec le Groupe Aixois de Recherche en Syntaxe (GARS), à l'origine de la revue Recherche Sur le Français Parlé.

Dans un premier chapitre, Claire Blanche-Benveniste extrait et développe certaines conclusions de différents auteurs (Cardona, Desbordes et Olson) : sur l'autonomie de l'oral – sans l'écriture, pas de représentation possible des différentes formes de conscience linguistique qui permet une analyse de la langue parlée ; sur les discriminations de l'oral – références fréquentes à l'écrit produites dans le discours oral au lieu d'une intonation équivalente (« entre parenthèses », « point final », « entre guillemets »...) ou pour lever une ambiguïté (par exemple sur le genre ou le nombre d'un mot). L'oral possède également des qualités qui n'existent pas à l'écrit : moins d'ambiguïtés que prévu grâce aux éclaircissements en glosant et à des ressources propres pour séparer ou regrouper des termes (pauses, intonation) ; sur les modes de production de l'oral – le langage parlé est à rapprocher de l'écriture en mode brouillon : la recherche de la formulation par le scripteur ou le locuteur peut conduire à casser la linéarité continue de sa production (accumulation de mots, retour arrière, incidente). Lorsque ces phénomènes n'existent pas c'est que nous sommes dans un mode de production montrant le discours sous la forme d'un produit fini : écriture ou parole professionnelle. Cette remarque amène Claire Blanche-Benveniste à proposer une notation intéressante pour mettre en lumière cette caractéristique de manière simple et visuelle. Sur la transcription de l'oral,

après avoir énuméré les principales difficultés qui peuvent survenir dans cette opération, Claire Blanche-Benveniste expose les conventions proposées par le GARS.

L'auteur précise son propos dans un deuxième chapitre en affirmant qu'un continuum de genres relie la production langagière orale à celle de l'écrit. La notion de « fautes » est discutée selon que celles-ci sont intégrées, spécifiques à une condition sociale ou régionale, conjoncturelles... Cette réflexion amène à une présentation rapide des erreurs communes imputées à la langue parlée. L'hypothèse sous-jacente est que leur étude apporte un nouvel angle de vue pour la compréhension de la construction du discours. Claire Blanche-Benveniste montre de manière convaincante à travers trois phénomènes (inachèvement, autocorrection, interlocution) qu'il ne s'agit pas de véritables erreurs, mais de comportements spécifiques au monde oral, à la fois fondamentaux et naturels, suivant des modèles établis. La remise en cause d'une spontanéité inhérente à la communication orale est un des apports de la réflexion de l'auteur. La démonstration est faite de l'intérêt pour le français d'un corpus informatisé, à la fois écrit et oral, assez conséquent pour faire émerger de son analyse les caractéristiques de la langue qui ne se comprennent pas sans sortir de la simple dichotomie oral *vs* écrit.

Le troisième chapitre aborde les domaines linguistiques qu'il est nécessaire d'inspecter pour aborder l'étude du langage parlé. Ceux-ci sont nombreux et ce point soulève des questions concernant la manière de faire cohabiter des méthodes spécifiques complémentaires ou conflictuelles. Ainsi, l'étude des productions parlées alimente et fait progresser des notions issues des théories de l'information, des actes de langage et d'énonciation... Les possibilités récentes offertes par les traitements automatiques de grands corpus sont démontrées par la présentation d'exemples et de documents extraits d'études menées sur le recueil de français parlé du GARS (environ deux millions de mots). Les conclusions principales sont évoquées puis développées dans les chapitres suivants : une extension des notions syntaxiques classiques de catégories et de fonctions suffit à intégrer dans la grammaire les phénomènes observés ; même si deux grammaires distinctes ne sont pas nécessaires, les notions de morphologie doivent être considérées séparément en raison de grandes différences typologiques, selon la modalité envisagée. La prise en considération de la prosodie, pour ses fonctions de désambigüisation et de regroupement, permet de remettre en question l'unité d'analyse phrastique classique.

Claire Blanche-Benveniste développe dans le quatrième chapitre la remise en cause de l'idée d'une incohérence qui serait inhérente au langage parlée. Au contraire, elle y voit les traces de l'émergence de la syntaxe dans le discours et la mise en place d'innovations par extension du fonctionnement syntaxique. Cette partie est riche de discussions donnant un éclairage nouveau à des phénomènes souvent observés : fonctionnement du verbe *y avoir*, des formes du relatif *qui*, des constructions clivées ou pseudo-clivées, des structures corrélatives, des indices syntaxiques des variations de registres, du recours aux paroles rapportées plus fréquent et varié que dans les écrits.

Le niveau d'organisation défini comme macro-syntaxique est justifié dans le chapitre suivant par l'observation d'unités et de relations qui ne dépendent pas directement des catégories grammaticales (un rôle crucial est donné à l'intonation dans cette fonction de regroupement non pris en charge par la syntaxe). Les propositions du GARS sont décrites à travers les notions de noyau, autonome du point de vue de l'intonation et de la syntaxe, de compléments que l'on peut associer avant ou après le noyau et que l'on peut décrire en fonction de statuts prosodiques particuliers, ou encore d'incises ou de parenthèses dont les caractéristiques semblent plus faciles à décrire au niveau de la macro-syntaxe.

Le sixième chapitre sensibilise aux méthodes et aux résultats obtenus dans ce cadre autour de phénomènes peu étudiés : recherche, transmission et répétition du lexique, réflexions lexico-syntaxiques sur la monotonie du discours parlé et l'utilisation des contrastes, organisation des productions à plusieurs voix. Enfin, soulignons l'enjeu particulier du dernier chapitre qui se pose en défenseur d'une différence entre écrit et oral : la morphologie. Cette hypothèse invalide l'idée d'une forte présence des fautes dans le français parlé et amène à penser que les nombreux choix que l'écrit nous impose pour chaque mot quant à son genre, son nombre ou sa conjugaison, sont simplement de natures différentes dans le monde oral. Cela pour des raisons d'évolution de la langue : certaines règles d'élision, de liaison, d'enchaînement ont changé à l'oral mais sont conservées à l'écrit. La répartition des marques diffère, l'oral ayant par exemple tendance à éviter les pluriels et à n'utiliser qu'une seule marque pour le syntagme entier. L'organisation des marques peut être « typologiquement » différente : alors que, par exemple, l'écrit nous donne à penser que le genre d'un mot se conçoit d'abord au masculin, l'ajout d'un suffixe permettant le passage au féminin, ce même genre à l'oral se construit plutôt par raccourcissement du mot au féminin en le privant de sa consonne finale !

Je conclurai par quelques remarques générales. Le livre présente un réel intérêt pour la communauté du TAL car, en donnant des arguments et des exemples originaux pour défendre l'étude de l'oral, il suscite la réflexion de tous afin de rattraper le retard pris par la recherche sur le français parlé ; ce retard pénalisant les avancées sur le français en général. Le point de vue de Claire Blanche-Benveniste est volontairement et essentiellement tourné vers la prise en compte des caractéristiques de l'oral, car elle regrette une prédominance trop importante de l'étude de l'écrit. Cette approche, nécessaire, n'est peut-être pas suffisante dans le sens où ce qui est observé dans beaucoup d'études actuelles est souvent extrait de recueils certes écrits, mais dans lesquels a été éliminé ce qui est susceptible de « parasiter » l'analyse linguistique ou statistique (typographie, mise en page...) transformant les ressources en productions langagières « nettoyées », « amodales ». Par exemple, la remise en question de la phrase comme unité d'étude linguistique pourrait être faite sans même considérer les productions orales : la littérature autour de la prise en compte dans les outils de TAL de l'architecture visuelle des textes écrits (typographie, mise en page…) montre déjà les limites d'une telle base d'analyse. Même dans les études de transcriptions de corpus oraux, ne rate-t-on pas

des points de réflexions essentiels sur la construction du discours en ne tenant pas compte de la syntaxe des gestes ou des mimiques ? Il s'agit donc de continuer de militer non pas pour une meilleure reconnaissance de l'oral, mais pour une meilleure prise en compte dans l'étude des langues de la modalité et du support de communication.

Gaston GROSS, Ramona PAUNA, Freiderikos VALETOPOULOS, Sémantique de la cause, *Éditions Peeters*, 2009, 365 pages, ISBN 978-2-7584-0105-6.

Lu par **Efi LAMPROU**

Département d'études françaises, université de Chypre

L'ouvrage Sémantique de la cause *est une typologie d'un type particulier de constructions causales du français. Il s'agit des constructions qui opèrent sur des arguments phrastiques et qui s'opposent à celles qui figurent dans une phrase simple. Selon cette analyse, qui se veut originale du point de vue théorique et méthodologique, les relateurs ou connecteurs exprimant la causalité sont considérés comme des prédicats et, de ce fait, sont traités comme tels. Ils ont, par exemple, un schéma d'arguments comme les autres catégories prédicatives (noms, verbes, adjectifs ou prépositions). Ce schéma d'arguments, appelé ici « spectre argumental », constitue la base de cette description. Gaston Gross et ses collaborateurs mettent en évidence la diversité de l'expression de la cause et confirme, pour le traitement de ces constructions, la nécessité d'une typologie établie selon des critères linguistiques. Cette typologie est fondée sur la description exhaustive des propriétés syntactico-sémantiques des constructions exprimant la causalité.*

Dans cet ouvrage, la cause est traitée d'un point de vue linguistique. Son objectif est de décrire les relations causales telles qu'elles sont exprimées dans les langues naturelles. Cette description a été réalisée à partir d'un corpus constitué de deux bases de données : les dix dernières années du journal *Le Monde* et *Frantext*.

Définition de la cause. Le cadre théorique

D'après le cadre théorique proposé dans cet ouvrage, la causalité linguistique se divise en deux catégories selon que la cause est exprimée dans une phrase simple (*L'enfant a renversé la chaise*) ou dans une phrase complexe (*Cette nouvelle a provoqué un malheur*). Dans ce second cas, la cause est exprimée par un prédicat de second ordre, *i.e.* un prédicat dont les arguments sont eux-mêmes des prédicats (*nouvelle*, *malheur*). Seules les causes de ce type (causes externes) sont analysées dans cette étude qui suit le modèle théorique de Z. Harris, selon lequel une phrase est constituée d'un prédicat (le pivot) et d'arguments.

Les causes externes sont définies comme des phrases complexes comportant trois prédicats : le causatif et deux autres prédicats, chacun de ces deux prédicats

constituant le pivot de chaque phrase-argument. Ainsi, la construction causale *Les fortes pluies de cette nuit ont provoqué des inondations* sera présentée schématiquement comme suit : *provoquer (pluies, inondations)*. Le prédicat *provoquer* sélectionne ses arguments de nature phrastique (*pluies* et *inondations*). Ces derniers partagent les propriétés des « prédicats nominaux événementiels » sauf que, dans ce cas de figure, ces prédicats ne sont pas actualisés (ils sont privés d'informations sur le temps et l'aspect). *Provoquer* est un prédicat de second ordre, car il a des arguments qui sont, eux-mêmes, de nature prédicative.

Du point de vue morphologique, la cause n'est pas uniquement exprimée par des verbes. Cette étude met en avant l'idée qu'il existe d'autres catégories morphologiques, telles que le nom (*cause, raison*), la préposition (*car, puisque, parce que*) ou une locution (*en raison de, par dépit*), susceptibles de traduire une relation causale entre, par exemple, deux événements. Toutes ces formes sont regroupées ici sous la dénomination traditionnelle de « relateurs » ou de « connecteurs ». Cependant, contrairement à la tradition grammaticale qui définit les connecteurs comme des éléments servant uniquement à relier la phrase principale à la subordonnée, Gaston Gross et ses collaborateurs considèrent ces derniers comme des prédicats (introduction et chapitre 1).

Les outils mis au point pour décrire les prédicats en général (*i.e.* les prédicats ayant comme arguments des noms élémentaires) s'appliquent également aux relateurs de causation externe. Dans un premier temps, Gaston Gross et ses collaborateurs mettent au point des outils théoriques pour décrire la causalité dans sa totalité. Dans le premier chapitre de l'ouvrage, les auteurs proposent une série de paramètres qui permettent de dégager les grandes classes (appelées ici « hyperclasses ») des prédicats de cause. Trois grandes classes sont définies en fonction de l'opérande (autrement dit, selon que les causes opèrent sur des événements, des actions ou des états) : les prédicats des *causes événementielles* (*causer, provoquer*, etc.), les prédicats des causes du *faire* (*faire, inciter, demander, seconder*, etc.), et les prédicats des causes d'*états* (*rendre, devenir, mettre, améliorer*, etc.). Leur description détaillée fait l'objet des chapitres 2, 5 et 6. De nombreuses sous-classes et leurs propriétés sont étudiées à l'intérieur de chaque chapitre.

Diversité des expressions causales

Toutefois, l'application systématique de ces paramètres montre l'hétérogénéité et la complexité des expressions causales. Cette complexité est due à plusieurs facteurs.

Tout d'abord, il s'avère que la cause n'est pas toujours explicite. À côté de certains relateurs tels que *provoquer, causer, à cause de* qui traduisent la cause de façon explicite, il existe un très grand nombre de cas où le connecteur ne peut être interprété comme causal que si l'on reconnaît un emploi métaphorique. C'est le cas, par exemple, des verbes *attiser, éradiquer, conduire à*, des locutions prépositives *à la source de, à l'origine de*, etc. ou bien des constructions complexes comme *donner*

naissance à, faire naître, être le fils de, etc. Bien qu'exprimant la cause de façon univoque, ces relateurs traduisent aussi des valeurs supplémentaires. De ce fait, ils sont le résultat d'un surcodage. Le chapitre 3 est consacré aux causes métaphoriques du mouvement, de l'origine et du monde végétal.

De même, en dehors d'un codage normal ou adéquat, on trouve également la cause inférée. Cette étude systématique des connecteurs exprimant une cause montre que l'expression causale peut ne pas être explicite, mais être le résultat d'une inférence (chapitre 7). D'après les auteurs, dans ce cas de figure l'interprétation d'une cause peut être liée à notre connaissance du monde. On parle aussi de causes « inférées » lorsque la cause est rendue par des connecteurs qui indiquent habituellement une fréquence (*chaque fois que*), un temps (*quand, lorsque*), une condition (*si*). Ces causes ne sont pas des causes pures et ne véhiculent pas directement une idée causale, mais d'autres circonstances. *Chaque fois que* dans la phrase *Chaque fois qu'on lui parle un peu rudement, il se met à pleurer* traduit l'itérativité (*i.e.* la fréquence), et non pas directement la causalité ; le lien causal est inféré et fondé sur la cooccurrence de deux événements (la principale est considérée comme une conséquence). Les auteurs analysent également les cas où l'inférence est maximale (par exemple la parataxe : *La pente est raide. La descente est difficile*). Aucun élément lexical ne prend en charge la relation de cause. L'analyse montre qu'à l'instar des causes inférées les causes liées à un fait d'énonciation relèvent, elles aussi, d'un sous-codage.

Enfin, les causes linguistiques sont très souvent exprimées par des faits d'énonciation qui échappent à la définition classique de la cause. C'est le cas des causes du *dire* étudiées dans le chapitre 8. L'expression de la cause dans ces cas de figure n'est pas indépendante du locuteur. On parle alors d'une certaine intrusion qui s'observe dans deux types de constructions différentes. Dans le premier cas, cette intrusion est dénotée par une *conjecture* lorsque nous sommes devant des événements inattendus dont l'interprétation est incertaine (*La voiture a été anéantie. Les experts ont fait la conjecture que le chauffeur roulait trop vite*). Pour les comprendre, il faut développer une stratégie explicative. Le second cas d'énonciation concerne le point de vue de la personne qui prend en charge le discours comme nous pouvons le constater dans *Pierre vient subitement de s'enrichir. Je le soupçonne d'avoir fraudé le fisc*. Dans d'autres cas, le locuteur se fonde sur ce qui a été dit sur l'événement et laisse à d'autres le soin d'asserter la cause (*Alexandre a démissionné. On dit qu'il n'a pas pu supporter la pression*). Une analyse très intéressante des connecteurs *parce que*, *puisque* et *car* est proposée dans ce chapitre. Ces connecteurs sont considérés comme porteurs des causes « justificatives », ce qui explique leur sous-codage.

Le dernier paramètre qui rend compte de la diversité de l'expression des causes est lié à l'aspect (chapitre 4). Les auteurs analysent les aspects externes des prédicats duratifs d'actions ou d'événements. Les causatifs d'états, qui apportent des informations sur une phase de l'opérande, sont analysés dans le chapitre 6 en corrélation avec leurs autres propriétés (actualisation, détermination, etc.). À côté

des causes « globales », Gaston Gross et ses collaborateurs distinguent des causes inchoatives (*déclencher*, *instaurer*), progressives (*perpétuer*, *maintenir*) et terminatives (*clore*, *éradiquer*).

Traitement de la diversité des causes. Méthodologie

Pour remédier à cette hétérogénéité, Gaston Gross et ses collaborateurs proposent une analyse détaillée du spectre argumental de chaque connecteur selon des classes sémantiques fines appelées « classes d'objets ». La multiplication des classes est indispensable pour pouvoir décrire la subordination de manière adéquate.

Les prédicats de ces classes sont décrits en extension. La description du spectre argumental de tous les connecteurs est réalisée à partir d'un corpus de grande ampleur (sont recensés 250 relateurs et 6 000 substantifs sur lesquels opèrent les prédicats de la causalité). Les compléments sont donnés par ordre de fréquence. Ce travail méthodique révèle que chaque prédicat de cause sélectionne un nombre d'opérandes qui lui est propre, les opérandes représentant des classes sémantiques spécifiques. Le verbe *provoquer* est, par exemple, compatible avec 800 substantifs et sélectionne comme arguments-objets i) des événements tels que les <accidents-incidents>, les <catastrophes naturelles>, etc., ii) des états événementiels <conflits>, <maladies>, etc., et enfin iii) des actions événementielles <attitudes et comportements>, <opérations intellectuelles>, <dire>, etc. À côté de leur combinatoire (opérateurs appropriés : verbes, adjectifs et adverbes) sont également analysées toutes les transformations possibles (thématisation, passivation, nominalisation, etc.).

Conclusion

En conclusion, la cause est une notion complexe qui ne peut pas être réduite à une définition conceptuelle unique impliquant un événement du monde des phénomènes responsable de l'existence d'un autre événement du monde des phénomènes. Cette étude, qui constitue une illustration parfaite de cette complexité, pourrait avoir des applications non seulement dans le traitement automatique des langues naturelles, mais aussi dans l'enseignement du français comme langue première et seconde. Elle pourrait aussi contribuer aux recherches en analyse du discours.

Benoît HABERT, Construire des bases de données pour le français, Tome 1. Notions, *Éditions OPHRYS*, 2009, 200 pages, ISBN 2-7080-1218-5.

Lu par **Marie-Paule JACQUES**

LIDILEM, Université Stendhal/IUFM, Université Joseph Fourier, Grenoble

---

Avec cet ouvrage et ses riches compléments en téléchargement gratuit, Benoît Habert continue un travail entrepris il y a plusieurs années : doter le linguiste – et plus largement le

chercheur en sciences humaines – d'outils et de ressources informatiques. Il présente ici les notions essentielles pour l'utilisation de bases de données à partir de réels cas de recherche.

Avec cette parution, Benoît Habert poursuit la démarche entreprise dans de précédents ouvrages : familiariser le linguiste – et plus largement le chercheur ou étudiant en sciences humaines – avec les possibilités offertes par les outils et ressources informatiques.

Il se focalise ici sur l'utilisation de bases de données pour recueillir, stocker, organiser, interroger les données support d'une recherche. L'une des forces de l'ouvrage est de présenter les opérations permises par l'utilisation d'un système de gestion de bases de données (SGBD) à partir de trois études de cas, chaque étude éclairant diverses manipulations en fonction de leur pertinence pour l'étude elle-même et ses objectifs. Les opérations sont ainsi ancrées dans les besoins réels d'interrogation et de sélection des données et non exposées *in abstracto*.

L'auteur n'a pas lésiné sur les moyens pour aider le lecteur à se forger une véritable compétence en matière de bases de données : le tome 1, payant, est accompagné non seulement d'un tome 2 en téléchargement gratuit, mais aussi des bases de données exemples elles-mêmes, dans des formats variés pour en permettre l'utilisation par divers gestionnaires de bases de données[1]. On ne peut que louer ce souci de diversité logicielle qui offre au lecteur le choix du SGBD proprement dit. Avec les deux tomes et les exemples, un lecteur néophyte soucieux d'élargir son champ de compétences a à sa disposition les outils nécessaires.

Après une introduction qui situe clairement le propos, il ne s'agit pas de la formation à un logiciel précis, mais d'un exposé détaillé des concepts principiels des bases de données, le tome 1 est divisé en 11 chapitres. Trois études de cas, chacune courant sur plusieurs chapitres, offrent à l'auteur les illustrations nécessaires à la présentation des notions et opérations générales des SGBD.

Les quatre premiers chapitres s'appuient sur des données issues d'une enquête dans un service de prématurés, constituées aussi bien de texte (par exemple, les remarques inscrites par les infirmières sur un cahier de suivi) que de données chiffrées (par exemple l'ancienneté des infirmières dans le service, leur âge...). Le chapitre 1 « prépare le terrain » en présentant ces données avec une focalisation sur des notions utiles ultérieurement : l'association *individu* (bébé, infirmière) et *caractères* (sexe, âge, ancienneté, commentaire rédigé, etc.). Le chapitre 2 reprend ces notions sous la terminologie usuelle *enregistrement/attributs* et introduit à la notion de *table*, cruciale pour les bases de données, en faisant le détour par d'autres outils informatiques susceptibles de créer et manipuler des tables : traitement de texte et tableur. Sont ainsi mises en évidence les limites des opérations que ces outils réalisent et, par contraste, la puissance des SGBD. Les chapitres 3 et 4 débutent

1. Ce compte-rendu se restreindra à la présentation du tome 1 sans s'interdire des allusions au tome 2 (http : //www.toeditions.com/Sources/Habert_Bases-de-dpnnees-2.htm).

l'inventaire des traitements possibles : sélection des données répondant à certaines caractéristiques, liste des combinaisons d'attributs, tris successifs, déductions de nouvelles informations par regroupements, tris, combinaisons d'opérations élémentaires qui permettent d'obtenir de nouvelles données, notamment chiffrées (par exemple, le nombre d'enregistrements pour lesquels l'attribut *X* présente la valeur *n*).

Les chapitres 5 à 7 poursuivent cette présentation d'opérations possibles, en prenant comme support une analyse de la pièce *Phèdre*. Les données sont ici le texte de la pièce, découpée en vers, mais aussi un premier niveau d'observations et de relevés effectués sur ce matériau, comme par exemple la transcription phonétique du texte, qui permet de s'affranchir des graphies et de saisir la dimension sonore, seul accès au phénomène de la rime, ou encore la position de chaque syllabe au sein d'un vers. La spécificité d'une pièce de théâtre de présenter une triple structure, dramatique, syntaxique, métrique, fournit une parfaite illustration de la notion de *relation* au sein d'une base de données. Les chapitres 6 et 7 explicitent donc les divers aspects liés à cette notion, comme les concepts de *clé primaire* et *clé étrangère*, les types de *jointures*, en n'oubliant pas de réutiliser les opérations exposées précédemment pour montrer des exemples d'exploitation des fonctions d'un SGBD et des exemples de résultats produits par de telles opérations.

Les chapitres 8 à 10 changent de perspective en s'appuyant sur un nouveau « cas » : l'étude du morphème *-esque*. Là encore, le premier des trois chapitres présente les données de façon détaillée, mais les deux chapitres suivants ne sont pas, autant que précédemment, consacrés à l'exposé de nouvelles opérations de traitement des données. Sont expliquées des combinaisons des opérations vues auparavant pour la réorganisation des données, notamment l'éclatement de la table initiale de l'étude en plusieurs tables pour améliorer la cohérence. Mais ces deux chapitres « prennent de la hauteur » en s'attaquant au niveau de la modélisation elle-même, c'est-à-dire à la détermination des entités, des attributs, des relations, en amont de leur implémentation dans une base de données. Cette implémentation, la création concrète de la base et de ses objets, ainsi que le peuplement en données de la base sont traités dans le chapitre 10. Le choix d'aborder ces questions plus abstraites après que le lecteur a touché du doigt la « matière » d'une base de données nous paraît particulièrement judicieux. Il est en effet plutôt difficile de construire un objet sans comprendre ni sa destination, ni son contour final.

Aide supplémentaire à cette construction, le chapitre 11, dernier chapitre de l'ouvrage avant la conclusion, prodigue d'utiles recommandations de « bonnes pratiques » pour la prévention des incohérences, la justesse des données, l'exploitation de la base comme maillon d'une chaîne à la fois de traitements et de chercheurs qui peuvent avec cet outil échanger des données et des résultats d'analyse.

Tout au long de l'ouvrage, l'ancrage sur des études authentiques, pour précieux qu'il soit, ne suffit pas à gommer totalement l'aridité intrinsèque du propos. Nous

voulons dire par là que la création et la manipulation de bases de données sont pour un linguiste « lambda » un sujet plutôt rébarbatif et que même le talent pédagogique d'un Benoît Habert ne suffit pas à le rendre attrayant. Il faut déjà à la fois un bagage informatique minimal et une conviction enthousiaste de l'utilité de l'outil informatique dans la recherche en sciences humaines pour digérer le sujet. L'ouvrage ne se prête guère à une lecture « de découverte » mais s'adresse à tout chercheur ou étudiant prêt à y consacrer du temps. Ce tome 1, qui recèle la quintessence des notions, gagne fortement à être utilisé en conjonction avec le tome 2, qui n'en est pas une suite ou une « application » mais un complément entremêlé. D'ailleurs, la table des matières du tome 1 fait état, pour chacun des 11 chapitres, des compléments et prolongements à trouver dans le tome 2, ainsi que de quatre chapitres additionnels. Aucun des deux tomes n'a véritablement de sens sans l'autre. Mais le chercheur ou l'étudiant qui acceptera d'ouvrir devant lui les deux tomes et les bases de données proposées afin d'en décortiquer le fonctionnement verra s'ouvrir aussi une nouvelle dimension de la recherche en linguistique et en sciences humaines, tant il est vrai que l'outil change la perception et la compréhension des données.