Linx

Revue des linguistes de l'université Paris X Nanterre

58 | 2008

Aspects de comme

**n° 6 de la revue *Corpus* de l'UMR 6039, 2007,
Interprétation, contextes, codage, sous la direction de
B. Pincemin**

Denise Malrieu

Édition électronique

URL : <http://journals.openedition.org/linx/346>

DOI : 10.4000/linx.346

ISSN : 2118-9692

Éditeur

Presses universitaires de Paris Nanterre

Édition imprimée

Date de publication : 1 juin 2008

Pagination : 217-223

ISSN : 0246-8743

Référence électronique

Denise Malrieu, « n° 6 de la revue *Corpus* de l'UMR 6039, 2007, *Interprétation, contextes, codage*, sous la direction de B. Pincemin », *Linx* [En ligne], 58 | 2008, mis en ligne le 16 février 2011, consulté le 22 septembre 2020. URL : <http://journals.openedition.org/linx/346> ; DOI : <https://doi.org/10.4000/linx.346>

Ce document a été généré automatiquement le 22 septembre 2020.

Département de Sciences du langage, Université Paris Ouest

n° 6 de la revue *Corpus* de l'UMR 6039, 2007, *Interprétation, contextes, codage*, sous la direction de B. Pincemin

Denise Malrieu

RÉFÉRENCE

n° 6 de la revue *Corpus* de l'UMR 6039, 2007, *Interprétation, contextes, codage*, sous la direction de B. Pincemin.

- 1 Ce numéro de la revue *Corpus*, très inspiré par la démarche herméneutique en linguistique de corpus, vise à préciser les modalités de l'activité interprétative lors de la constitution et l'exploitation de corpus annotés.
- 2 Dans son intéressante introduction, B. Pincemin situe l'activité interprétative tout d'abord dans la constitution du corpus (quels sont les textes à réunir et à contraster, toujours fonction d'un point de vue et d'hypothèses), ensuite dans la théorie du texte (quelle définition de la structure et des unités textuelles), enfin dans la sémantique du codage et les choix techniques de format et de représentation du texte : « L'interprétation est présente dans toutes les étapes du travail sur corpus ». B. Pincemin pose la question : quels sont les paliers de contexte pertinents : le texte ? Le genre ? Les contextes infra-textuels (partie, passage, paragraphe, période, syntagme, etc.), les contextes supra-textuels (regroupement de textes par périodes, par auteur etc.) ? Certains contextes peuvent n'avoir aucune marque formelle de frontières comme celui de passage, discuté par Rastier dans ce numéro.
- 3 Passons donc maintenant à l'analyse des différentes contributions.

F. Rastier : *Passages*, p. 25-54

- 4 Affirmant bien fort dès le départ que le texte n'est pas composé de mots et de règles, l'auteur introduit la notion de passage, qui remonte à Hillel l'Ancien quand il formula les règles des passages parallèles : sélectionner des passages comparables pour lever une difficulté d'interprétation. Le concept de passage pose trois problèmes :
 - la contextualité et textualité : les passages sont-ils discrets et contigus ? comment s'articulent valeurs et passage ?
 - La sémiosis : appariement expression/contenu
 - Comment sont discernées les inégalités qualitatives au sein d'un passage et entre passages ?
- 5 Rastier définit la sémiosis comme un processus de codétermination multiple des plans des signifiés et des signifiants à l'intérieur d'un passage où jouent les oppositions forme/fonds, le passage étant caractérisé par un fond d'isotopies thématiques, où l'on observe une forte densité de co-occurents d'un extrait (expression) et de corrélats d'un fragment (contenu). Les formes de codétermination sont elles-mêmes fonction du genre textuel.
- 6 La définition technique du passage est liée à la notion de thème : on recrute un ensemble de co-occurents (signifiants : lexèmes, grammèmes, ponctèmes) qui, dans un contexte défini, par ex. une fenêtre de 20 « mots » (i.e. chaînes de caractères) dépassent un écart réduit seuillé à 3 (en corpus multigenre) voire à 2 (en corpus unigenre). On qualifie ensuite certains co-occurents comme des corrélats sémantiques, dès lors que leur contenu partage au moins un sème en commun avec celui du mot qui supporte l'interrogation initiale. La structure constituée par ces sèmes constitue la molécule sémique du thème.
- 7 La composition des passages est variable et peut aller du mot à des unités plus longues que la phrase. Les passages varient selon les normes des discours et des genres. Le passage « idéal » est caractérisé par une sémiosis dense avec une isotopie mésogénérique propre qui commence et finit avec le passage. Il peut y avoir icônisation et autonomisation de passages comme dans le cas du dicton ou proverbe.
- 8 Il existe divers régimes de passages. Un texte n'est pas plus une suite de phrases qu'une suite de passages : en effet ces derniers sont délimités, qualifiés et hiérarchisés par la pratique interprétative. L'auteur distingue deux régimes de pertinence. La pertinence objective s'appuie sur le fait que tous les passages n'ont pas le même régime de connectivité interne et externe. Ceux qui ont les deux à la fois forment des nœuds herméneutiques, centraux pour l'interprétation. La pertinence « subjective » dépend de la tâche.
- 9 Par rapport aux propositions de F. Rastier, mes remarques seront les suivantes :
 - Le passage n'est défini que par la composante thématique alors que d'autres critères pourraient entrer en ligne de compte. On peut se demander s'il ne serait pas utile d'élargir cette démarche, par exemple à la composante dialogique : ainsi, selon les genres et les passages, l'interprétation des guillemets inclura ou exclura certaines valeurs.
 - dans la perspective d'une linguistique de la parole, qui seule peut être compatible avec une herméneutique matérielle, l'approche méthodologique du global (textuel et intertextuel) n'est pas discutée : à la fois, au niveau du texte, en quoi le genre contraint-il la nature des passages et peut-on faire confiance à la machine pour repérer les passages si l'on ne

distingue pas leur statut dialogique dans le texte ? A quelles conditions des passages de genres différents peuvent-ils être comparés et contrastés ?

M. Lecolle : *Polysignifiante du toponyme, historicité du sens et interprétation en corpus. Le cas de Outreau*, p. 101-126.

- 10 L'auteure propose ici une analyse sur corpus de l'évolution du sens du toponyme *Outreau* et discute les méthodes de repérage des différents sens.
- 11 Après avoir précisé que le Npr de lieu peut normalement prendre différentes valeurs sémantico-référentielles (lieu, habitants, institution qui gouverne le lieu), elle indique qu'il peut aussi désigner des événements liés au lieu (Tchernobyl).
- 12 Pour étudier l'évolution du sens de *Outreau*, l'auteure a partitionné le corpus en cinq périodes, en fonction de connaissances externes (procès), auxquelles correspondent majoritairement les sens suivants : ville et habitants, affaire de réseau pédophile, procès de ce réseau et enfin, fiasco judiciaire. Le contexte pris en compte va du contexte étroit du syntagme à la phrase, au texte (article de presse daté) et à l'intertexte (les cinq sous-corpus).
- 13 L'analyse consiste à étudier les critères de catégorisation des valeurs des occurrences sur un corpus interprété et vérifié par le linguiste, en partant d'une analyse des contextes des occurrences obtenues avec le concordancier *AntConc* : l'analyse syntaxique et distributionnelle, repérage des actants du scénario attaché à *Outreau* événementiel, dialogisme.
- 14 L'auteure discute ensuite l'insuffisance du concordancier pour délimiter les sens du Npr et la variabilité de l'empan du contexte nécessaire à l'interprétation du syntagme ; la sous-détermination des valeurs des prépositions comme *à* et *de* dont la valeur ne peut être souvent définie qu'au palier du texte daté dans le sous-corpus. Il en va de même des expressions comparatives.
- 15 Quantifier, oui, mais quoi ? L'auteure souligne que tous les critères d'analyse ne peuvent être quantifiés. Elle montre toutefois que les acceptions /lieu/ vs /fiasco judiciaire/ définis par analyse distributionnelle ont des répartitions très contrastées dans les périodes 1 et 5, mais qu'il reste une proportion importante (53 %) dans le sous-corpus 5 de cas non décidables par l'analyse distributionnelle ; en fait les résultats du concordancier demandent un travail interprétatif de catégorisation sémantique des contextes, ce que Rastier désigne comme le passage du statut de co-occurent à celui de corrélat. De plus les rangs de fréquence des mots du contexte proche sont aussi significatifs de l'évolution du sens, ce qui correspond à la stabilisation des isotopies contrastées dans les sous-corpus.
- 16 Cet article confirme bien la nécessité de l'activité interprétative du linguiste dans le travail sur corpus, prenant en compte des paliers interprétatifs variables et suggère la possibilité de développer des outils articulant mieux analyse distributionnelle, syntaxique et molécules sémiques liées à une analyse actantielle, à la dimension dialogique du texte et au genre (interaction titre/texte par exemple).

F. Rink & A. Tutin : *Annoter la polyphonie dans les textes : le cas des passages entre guillemets*. p. 79-100.

- 17 Ces auteurs se proposent comme objectif de définir des critères formels de définition des valeurs des passages entre guillemets (PEG) en vue d'une annotation automatique de ces derniers, ceci dans le cadre d'un projet sur les marqueurs linguistiques de la subjectivité et de la polyphonie. La démarche suivie consiste à définir les différentes valeurs, à tester le degré d'accord de 3 linguistes sur l'attribution de ces valeurs en corpus, de discuter les facteurs de désaccord et de proposer un codage et une DTD pour le balisage de corpus.
- 18 Les auteurs distinguent quatre emplois principaux :
 - l'autonymie pure : Ainsi pour un sourd « papa » et « maman » sont pratiquement identiques au niveau de la perception.
 - la polyphonie, qui peut être citation autonome au plan syntaxique ou intégrée au discours citant ;
 - le commentaire modalisant (la langue des signes, langue « parlée » par les communautés sourdes).
 - la dénomination.
- 19 Les indices formels des différentes catégories sont ensuite ébauchés :
 - la citation syntaxiquement autonome est le plus souvent accompagnée de marques typographiques : alinéa, tiret, guillemets, paragraphe. L'inclusion des incises pose toutefois problème.
 - La citation non autonome ne jouit pas de ces marques typographiques : cf le DI, les SN désignant la source (selon N, d'après N).
 - L'autonymie s'accompagne souvent d'introducteurs métalinguistiques (le mot, l'expression, le terme, signifie X,...)
 - Les emplois dénominatifs s'accompagnent de termes classificateurs (le terme, le nom de X, on appelle, nomme X)
 - Les emplois modalisants sont plus difficiles à repérer (ADJ, noms, SN, souvent courts).
- 20 Les emplois modalisants expriment une distance du locuteur par rapport à l'énoncé. Cette distance peut soit exprimer l'insatisfaction du locuteur par rapport à son dire : tel signifiant est utilisé à défaut de plus satisfaisant. (ex 9 : *On peut imaginer la réflexion d'un adolescent confronté à l'incompréhension d'un message émanant de ses idoles, censées avoir partagé les mêmes expériences, les mêmes « galères »*). La valeur ici attribuée se discute.
- 21 Soit exprimer la non adhésion, voire le rejet d'un énoncé prêté à d'autres. Les auteurs ont renoncé à distinguer ces deux valeurs. De plus, ils décident de n'attribuer un double codage de modalisation que lorsque celle-ci s'accompagne de commentaires méta-énonciatifs (*en quelque sorte, pour ainsi dire*) ou lorsqu'on a un terme marqué axiologiquement (cet adolescent en « *intégration sauvage* »). Le cas des proverbes doit être traité à part mais ne fait pas partie de la modalisation. Enfin, la double valeur dénomination + citation est attribuable et assez fréquente dans les écrits scientifiques.
- 22 Les auteurs soulignent enfin que la pratique des guillemets est variable selon les disciplines et les genres de textes (écrits de sciences humaines vs journalistiques) et qu'une analyse des genres reste à approfondir. Ce domaine de recherche a de plus des retombées pédagogiques concernant l'écriture chez les étudiants.

- 23 L'annexe comporte un exemple de texte annoté selon la DTD choisie.
- 24 Je ferai deux remarques à propos de ce travail : si l'objectif est de repérer la polyphonie dans les textes, n'est-il pas aussi intéressant de prévoir le codage des sources autres que le locuteur principal, puisque l'objectif est bien de repérer les positions respectives des locuteurs ? Ce qui impliquerait de distinguer le métalinguistique propre au locuteur (effort sur les signifiants) de la non prise en charge du discours d'autrui.
- 25 On a pu voir que les problèmes les plus épineux sont soulevés par ce qui a trait à la modalisation. Il faut distinguer le processus de désignation de la source (*selon X, P ; X dit que P, comme on dit dans ces cas-là*) et le processus évaluatif (distance épistémique (par exemple le conditionnel de doute), attitude adversative ou affiliative). C'est sur ces deux dimensions que le codage pourrait être enrichi de façon bénéfique.

A. Morgenstern & C. Parisse : *Codage et interprétation du langage spontané d'enfants de 1 à 3 ans*, p. 55-78

- 26 L'article, après un bref historique des études du langage enfantin, analyse les problèmes soulevés par la transcription des activités langagières du très jeune enfant, lors de la constitution de corpus longitudinaux. La transcription implique obligatoirement un part importante d'interprétation, celle-ci n'étant pas forcément décidable. Les auteurs décrivent les choix effectués dans leur système de codage conforme aux conventions CHAT de CHILDES. Ce format permet la transcription phonétique, l'explicitation du non-verbal, l'analyse morpho-syntaxique, et l'alignement avec la vidéo est possible. Un premier constat est que l'accord interjuges est plus fort pour la transcription orthographique que pour la transcription phonétique, ce qui montre l'importance de l'activité interprétative dans le codage. La démarche adoptée est de multiplier les codages (phonologique, pragmatique, sémantique, syntaxique), et les éléments fournis par l'audio et la vidéo (gestes, mimiques, prosodie).
- 27 Les auteurs soulignent abondamment les problèmes lors du codage du découpage en mots et en particulier de l'interprétation en /ə/ ou /a/ dans /aty/ (voiture ou la voiture ?) ; cela a amené à coder de façon différente les « fillers » selon les trois étapes distinguées dans l'acquisition des articles et prépositions ; le principe étant de caler le codage sur ce que le codeur sait du niveau d'acquisition de l'enfant.
- 28 La documentation du contexte de l'interaction est essentielle pour la discussion de l'interprétation et elle est fournie à la fois par un codage de la situation et de l'interaction et par la fourniture des données sonores et vidéo. L'enrichissement des catégories descriptives est laissé au libre choix du codeur, qui peut introduire en lignes supplémentaires des catégories telles que action des interlocuteurs, actes de langage, commentaire du transcripateur, imitation, mimiques, qui est l'adressé, rôle sémantique, gestes etc. Les transcripateurs différents peuvent identifier leur interprétation et leur niveau de certitude.
- 29 Enfin les auteurs discutent la question de l'utilisation des corpus ainsi constitués, par des usagers externes et la nécessité de développer des logiciels d'interrogation plus puissants que ceux qui existent, et remettent en question la notion de ligne principale du codage, tous les éléments pouvant être considérés comme principaux selon le point de vue adopté pour la recherche.

- 30 Dans ce recueil d'articles, celui-ci souligne avec force l'importance des processus interprétatifs impliqués par la constitution des corpus d'étude.

N. Desquinabo : *Intertexte générique et interprétation des actes de parole dans un corpus d'émissions de plateaux télévisés*, p. 127-152.

- 31 L'objectif ici poursuivi est l'attribution des valeurs des actes de paroles en fonction du type d'émission sur un plateau télévisé. Comment définir les propriétés du contexte d'interaction qui vont contraindre l'interprétation de la valeur illocutoire d'un acte local.
- 32 Les activités discursives étant normées à l'intérieur des domaines d'activité, des champs génériques et des genres, l'interprétation des actes de parole se fait à l'intérieur du genre à travers la mise en œuvre de schémas d'actes de parole (à la fois distribution des tours de parole, textures énonciatives, vocabulaires plus ou moins stabilisés et scriptés). L'indexation de l'interaction à un intertexte générique permet d'articuler les indications linguistiques aux instructions globales et d'expliquer l'interprétation des actes de parole mieux que ne le permet l'interprétation séquentielle (Schegloff & Sacks, 1973). Les actes de parole sont catégorisés avec la grille de Chabrol et Bromberg, 1999.
- 33 L'analyse pragmadiscursive, lexicale et morphosyntaxique de 24 interactions de 3 mn, 6 par genre a amené à distinguer 4 schémas génériques d'actes et de tours de parole des animateurs et des invités : controverse, portrait, jugement, polémique, caractérisés par des poids différents des actes de parole. Afin de mettre à jour les indices morphosyntaxiques de ces genres, une analyse multidimensionnelle a été menée sur le péritexte des émissions, le péritexte des interactions et le texte des interactions. On peut cependant se demander si l'analyse morpho-syntaxique est la dimension pertinente de description et si l'analyse n'aurait pas gagné à donner une définition plus linguistique des actes de parole.

S. Loiseau : *CorpusReader : un dispositif de codage pour articuler une pluralité d'interprétations*, p. 153-186.

- 34 S. Loiseau décrit ici un outil qu'il a lui-même développé de multi-annotation de corpus de textes écrits – l'idée étant que des corrélations entre niveaux de description peuvent permettre de caractériser des normes linguistiques des discours ou des genres. Le dispositif de codages multiples qui travaillent sur des unités de tailles différentes doit permettre d'analyser des interactions entre traits de niveaux hétérogènes. Il doit permettre aussi de cumuler des informations partielles liées par exemple à des analyseurs concurrents.
- 35 Concernant le choix du format de représentation, l'auteur argumente les avantages du format XML comme format d'accumulation ainsi que la standardisation liée à la TEI et ses avantages pour documenter l'annotation ; cependant à la différence de la TEI, CorpusReader ne fait pas découler du format arborescent d'XML que les données sont hiérarchiquement ordonnées. Les difficultés rencontrées par l'enchevêtrement des

hiérarchies sont discutées et l'auteur prône non l'annotation débarquée (stand-off) mais l'annotation en nœuds-bornes (balises milestones) qui privilégie l'axe de la précédence sur celui de la dominance, ce qui transforme l'arbre en graphe.

- 36 Contextualiser revient à fusionner les annotations liées au corpus et celles produites par l'analyseur syntaxique.
- 37 L'auteur discute le choix d'une API entre DOM (relation de dominance privilégiée) et SAX (relation de précédence) et se prononce pour les avantages de la seconde (taille de corpus non limitée grâce au traitement en flux, possibilité d'user des nœuds-bornes)
- 38 Je ferai quelques remarques et questions à propos de cet article. La démarche prônée, qui vise à caractériser de façon statistique et inductive les normes des discours et des genres, repère les corrélations entre propriétés annotées a priori des textes ou parties de textes et fréquences de traits morpho-syntaxiques. Outre qu'aucune discussion des unités textuelles liée à une théorie des genres n'est proposée, on ne peut que se poser la question de savoir à quelles conditions l'ACP (qui est le traitement visé et privilégié par CR) qui travaille toujours sur des formes (lexèmes ou grammèmes catégorisés en parties du discours), et non sur la valeur ou le sens, peut par l'annotation contextuelle favoriser l'émergence du sens : il est clair que l'interprétation de l'analyse factorielle doit massivement réinjecter du sens par une lecture sémantique très intuitive des oppositions sur les axes, car le sens local n'est pas déductible statistiquement de la cooccurrence de traits morpho-syntaxiques et de genres. Autrement dit, il existe un gouffre herméneutique entre les méthodes inductives quantitatives et une linguistique de la parole qui vise à élaborer une théorie des corrélations entre valeur locale et traits globaux relevant de la linguistique des genres. Une linguistique de la parole ne travaille pas sur des indices morphosyntaxiques (fréquence de tel lexème, de tel pronom personnel ou de tel morphème du temps verbal), mais sur la valeur locale en contexte de ces derniers. Autrement dit, le parcours interprétatif travaille toujours *au niveau de l'énoncé dans son contexte*, comme le montre l'analyse de Missire & Rastier sur le passage de *Madame Bovary*. C'est peut-être le mot de « corrélat » qui prête à confusion : en linguistique de la parole, il s'agit de corrélation entre valeur locale d'un morphème et traits de genre plus ou moins globaux (tant il est vrai que dans la chaîne de la parole et dans un genre textuel donné, le trait global peut être restreint à l'énoncé : cf. les modalités d'énonciation), alors qu'en linguistique quantitative, il s'agit de corrélation de formes locales et globales. Les analyses quantitatives des « indices morphosyntaxiques » peuvent confirmer des différences massives de catégories morpho-syntaxiques entre textes ou sous-corpus mais sont de peu d'utilité pour l'analyse des contraintes globales sur l'interprétation du local. Le travail de définition précise de l'empan des contraintes selon les paliers de globalité sur le local s'avère nécessaire à l'intérieur de chaque genre et de façon distincte mais coordonnée selon les composantes textuelles.
- 39 Pour conclure je dirai que ce numéro de la revue Corpus, très riche d'expériences diverses de travail sur corpus est à mettre entre toutes les mains et soulève de multiples questions sur la définition des contextes et sur le statut des variables dans l'approche quantitative des contextes. Ayant moi-même largement pratiqué les approches quantitatives et contrastives de corpus multigenres, il m'apparaît de plus en plus nettement que l'on ne peut plus se contenter de cette approche quantitative des sorties d'analyseurs et qu'un travail de définition plus précise de la variation de la textualité selon les genres s'avère nécessaire pour développer une linguistique de la

parole qui rende compte des multiples formes de contraintes du global sur le local, si toutefois l'objectif est d'intégrer le syntaxique et le textuel dans ses contraintes de genre.

AUTEURS

DENISE MALRIEU

MODYCO, Université Paris X