

Introduction

Bénédicte PINCEMIN

C.N.R.S. « Interactions, Corpus, Apprentissages,
Représentations » (Lyon)¹

Convenons, à la suite de Rastier (2001) notamment, que les textes sont l'objet de la linguistique. Un texte est entendu ici comme « une suite linguistique empirique attestée, produite dans une pratique sociale déterminée, et fixée sur un support quelconque » (Rastier, 2001 : 21), ce qui intègre pleinement diverses formes d'expression (orales comme écrites). Le corpus de textes est alors le terrain privilégié de l'observation de la langue. Lors de la collecte des données, lors de leur enregistrement comme lors de leur dépouillement, l'enjeu majeur consiste à préserver les dimensions et les structures significatives des échanges langagiers.

Dans cette livraison de la revue *Corpus*, nous voulons rendre compte de la composante herméneutique, fondamentale, de tout texte et de tout « bon » corpus. Non seulement l'interprétation est inévitable et compulsive, mais aussi l'interprétabilité est un critère de qualité déterminant pour un corpus (Pincemin, 1999). Pour que les résultats d'une analyse, d'un traitement sur un corpus prennent valeur, et puissent devenir éléments de réponse à un questionnement scientifique, il faut une interprétation méthodique des produits du traitement, qui, elle-même, s'appuie nécessairement sur une interprétation / compréhension tant de la procédure d'analyse que de la composition du corpus.

Mais très vite, dans le contexte d'une réflexion sur les pratiques de traitement et d'interrogation sur corpus numériques, la question de l'interprétation appelle au plan pratique celle du codage (comment enregistrer, traduire, transcrire) et au plan théorique celle des contextes (quelles sont

1. Laboratoire mixte CNRS / Université de Lyon, UMR 5191.

les unités textuelles et comment entrent-elles en relation). C'est ainsi, dans l'interaction de ces trois composantes – interprétation, contextes, codage – que peut se formuler le thème de notre numéro : tout corpus prend valeur par une interprétation, qui lui construit un sens. Or le sens se déploie en s'appuyant sur des contextes structurants, et le codage est l'expression technique, déterminante, des structures textuelles, inter-textuelles, et contextuelles.

Examinons la teneur que nous donnons à ces trois termes, et les questionnements associés, qui déjà esquissent leurs interactions.

Concrètement, le *codage* renvoie aux choix d'édition lors de la transcription du corpus dans le format adopté pour l'analyse. Dans le cas d'un format XML par exemple, le codage concerne aussi bien le balisage de structures textuelles (notamment par découpage et emboîtements, avec la délimitation de contextes syntagmatiques) que l'enrichissement par étiquetage (l'assignation de catégories créant complémentaiement des contextes paradigmatiques). À la multiplicité des interprétations possibles répond le besoin de vues alternatives sur le corpus : par exemple, les informations enregistrées par le codage pourront être différentes et se noter différemment, et donc conduire à différentes éditions électroniques du corpus, selon que l'objectif est l'archivage, la diffusion, ou le traitement par tel ou tel logiciel. En matière d'analyse assistée par ordinateur, si la robustesse des outils d'analyse est certainement pertinente, il serait dommageable qu'elle dicte un nivellement par le bas de la structuration des corpus. Car les logiciels d'analyse et d'exploration textuelle, notamment à visée sémantique, ont tout à gagner à savoir tirer parti d'un codage riche – ou du moins non appauvri –, donnant véritablement accès aux informations de contextualisation de tous ordres. Reste à trouver un équilibre vertueux, pour éviter des codages excessifs, trop lourds, ingérables, et étouffant l'interprétation au lieu de la rendre accessible et de la susciter.

L'*interprétation* est présente à toutes les étapes du travail sur corpus. Interprétation *a priori* au moment de la constitution du corpus, et avec la conception des opérations

Introduction

d'analyse à pratiquer ; interprétation *a posteriori* pour l'exploitation des résultats produits. Mais la pratique interprétative procède par retours et ajustements, elle n'échappe pas au cercle herméneutique : ainsi, la lecture des résultats motive très naturellement une reprise du codage et une réorientation des traitements. La *Text Encoding Initiative*² prévoit à juste raison un commentaire du codage, livré avec le corpus (rubrique *tagUsage*), comme du cadre et des objectifs du codage (rubrique *projectDesc*) : une telle explicitation des conventions de sens et du mode d'usage des balises dans le contexte du corpus est éminemment importante pour toute exploitation et réexploitation du corpus, en d'autres temps ou d'autres lieux, y compris par ses éditeurs, mais aussi au moment même du codage ! L'annotation des corpus semble questionner encore plus directement l'alliance entre codage et interprétation : peut-on établir une typologie des annotations, et ce à tous les paliers de contexte ? À l'image d'un cheminement interprétatif, l'annotation peut-elle, voire doit-elle, être dynamique (c'est-à-dire ajoutée, rectifiée, oubliée...) ? Doit-elle être partagée et sédimentée – mais avec quels contextes pour limiter gêne mutuelle des séries d'annotations et surcharge artificielle, inhumaine, de la lecture ?

Quelquefois précisée par la distinction entre contexte et co-texte, la réflexion sur le *contexte* dans son lien au codage et à l'interprétation des corpus se centre ici sur les structures syntagmatiques (qui découpent, emboîtent) et paradigmatiques (qui mettent en lien), dans un texte et entre des textes. C'est donc le contexte linguistique qui nous intéresse, par opposition au contexte situationnel, à l'environnement extérieur. Cette option n'est pas si restrictive qu'il y paraît : Rastier (2001 : 14-18) montre que les réalités externes en prise avec le texte (l'auteur, le monde, le lecteur – les *pôles extrinsèques*) se retrouvent par leur empreinte dans le matériau linguistique et textuel (*pôles intrinsèques*), notamment via le genre du texte. Peut-être aussi la question du contexte rejoint-elle directement

2. Présentation de la *Text Encoding Initiative* : <http://www.tei-c.org/>. La description du format proposé est détaillée dans les *Recommandations* (Sperberg-McQueen & Burnard 2002).

celle, fondamentale, de la bonne constitution du corpus : les critères de clôture ou de réflexivité par exemple (Mayaffre, 2002) ne visent-ils pas la recherche d'une contextualisation globale, sémantiquement stable, nécessaire et suffisante, déterminante ? Complémentairement, les techniques d'analyse de corpus dessinent également des contextualisations glissantes, mouvantes : qu'est-ce qu'un passage, et faut-il le coder ? L'observation des affinités et des attirances lexicales par des calculs de cooccurrences suppose la délimitation de contextes : l'environnement d'un mot, sa sphère d'influence, son rayonnement, se laissent-ils délimiter ? uniformément ? de façon unique ? Pour autant, comment garder sa consistance pratique et significative à la notion de contexte ?

C'est aux croisements des trois composantes qui nous intéressent que nous touchons alors le cœur de notre questionnement, et que nous rencontrons bon nombre de points délicats, encore ouverts, jalonnant les pratiques de constitution et d'analyse de corpus.

Si nous rapprochons *codage* et *interprétation*, nous pouvons les envisager comme alliés, concourant au même objectif, ou opposés, voire antagonistes. Par exemple, nous observons le contraste entre la quête d'une certaine objectivité du codage et la subjectivité assumée de l'interprétation : mais l'explicitation de consignes de codage, la documentation des choix de codage, apparaissent comme des solutions pour articuler objectivité et subjectivité, puisqu'on reconnaît alors la relativité du codage, et qu'on objective une interprétation. Autre point de contraste, le codage peut être valorisé : de par sa portée sur l'analyse et l'interprétation il est alors compris comme un travail scientifique déterminant ; mais il peut aussi être dédaigné, réduit, en raison de sa réalité pratique austère, à un travail de tâcheron, machinal. Le codage peut être perçu comme une atteinte au texte, dont il faudrait respecter l'intégrité : mais a-t-on vraiment un état originel du texte, un état de référence ? La philologie, dont l'objectif est justement d'établir le texte, n'en reconnaît pas moins faire œuvre d'édition. Si donc l'on a toujours affaire à une édition, le respect du texte amène à dénoncer une trop forte projection des attentes interprétatives

Introduction

du codeur / lecteur sur le texte, le sur-codage. Enfin bien sûr, nous rencontrons le cercle herméneutique : si, en théorie, le codage ne saurait précéder l'interprétation, comment en rendre compte en pratique ? Par exemple, en cas de d'hésitation ou de désaccord au moment du codage, la multi-annotation apporte-t-elle une solution intermédiaire et provisoire, en reportant l'arrêt d'une interprétation, voire en permettant d'évaluer statistiquement, par le calcul, chaque possibilité interprétative ? Et comment le codage, statique (éventuellement succession d'états discrets), s'articule-t-il avec la dynamique de l'interprétation, et son évolution quelquefois diffuse et continue ?

La rencontre d'*interprétation* et *contextes* n'est pas moins riche. Elle est d'ailleurs au fondement de certaines théories linguistiques, comme la sémantique interprétative (Rastier, 1987), qui procède par une approche différentielle : le sens naît des rapprochements et des contrastes, qui s'établissent dans les contextes d'unités linguistiques et textuelles de tous ordres. Pour notre problématique de travail sur corpus, dès la constitution de celui-ci se pose la question des effets interprétatifs (parfois sous-estimés ou ignorés) liés à la réunion des textes et à la délimitation de collections, à la structuration du corpus et à la génération possible d'éditions (ou *vues*) partielles ou diversement présentées et organisées. La question se retrouve aussi au plan des méthodes d'analyse et de parcours des corpus, notamment pour les dispositifs de visualisation : quel(s) empan(s) de texte présenter au chercheur, qui lui permette une interprétation juste et efficace ? Pour un corpus oral, avec des enregistrements audio-visuels associés, il faut chercher quels sont les découpages pertinents pour l'alignement avec la transcription et l'accès local à des moments de l'enregistrement ; et une contextualisation est aussi créée par la dynamique de consultation des séquences d'enregistrement (passage en boucle, points d'arrêt possibles, etc.), qui ajoute d'autres effets interprétatifs. Par ailleurs, la multiplicité de paliers irréductibles les uns aux autres (le texte n'est pas la simple somme de ses paragraphes, ni l'intertexte une juxtaposition de textes) conduit à chercher des dispositifs exprimant conjointement des contextualisations sur plusieurs

paliers : par exemple, présenter un extrait de texte et donner une visualisation conjointe du positionnement et de l'ampleur de l'extrait dans le texte et dans l'intertexte du corpus.

Enfin, si la question du *codage* croise celle des *contextes*, c'est déjà au sein même des choix fondamentaux de la linguistique de corpus. Car la linguistique de corpus veut s'opposer à une linguistique sur exemples ponctuels ; elle vise à matérialiser, par la « masse » du corpus, une contextualité tangible, observable, quantifiable, riche. La masse n'y suffit pas, bien sûr, l'accumulation de « données linguistiques » ne se fait pas sans critères, et si elle n'est pas structurée, la matière est informe et comme insaisissable, incernable, insondable. C'est ici qu'intervient la question de la définition des contextes : quels paliers de contexte sont pertinents ? Par exemple, Péry-Woodley (1995) argumente l'importance du palier du texte : le texte est une unité de contexte indispensable, il ne doit pas être démembré, notamment par des techniques d'échantillonnage, si bien qu'un corpus ne devrait jamais être conçu comme *du* texte, mais toujours comme *des* textes. Le palier du genre apparaît également déterminant : ainsi, si l'on veut modéliser les connaissances « moyennes » d'un adulte et que l'on représente cette capacité cognitive par un corpus réunissant des articles de presse, des œuvres de littérature et des ouvrages didactiques, il faudrait à tout le moins veiller à ne pas amalgamer ces textes en une sorte de fonds cognitif uniformément accessible, mais bien distinguer des contextualités distinctes de mobilisation de tel ou tel intertexte. En effet, l'usage de la langue est différent selon les pratiques, et, sauf détour interprétatif particulier, ne sont *a priori* accessibles et pertinents que les contextes relevant du même genre textuel. Par ailleurs, reste ensuite la question de la manière de prendre en compte les contextes dans le codage. Les contextes sont *textuels* (en quelque sorte par définition, car présenter quelque chose comme un texte c'est le poser comme unité pour l'interprétation), mais aussi *infra-textuels* (partie, passage, paragraphe, période, syntagme, etc.) et *supra-textuels* (regroupement de textes par auteur, par période, etc.). Certains contextes se présentent comme des segmentations nettes et se prêtent à une traduction dans les formats de codage actuels ; en revanche, d'autres peuvent être plus mouvants, plus flous, et

Introduction

conduisent à explorer la voie de codages partiels, points d'appui pour des procédures interprétatives qui seules actualisent ces contextes.

Notre numéro comprend six contributions, qui nous feront cheminer dans les différents aspects de notre thème.

Il s'ouvre sur une réflexion linguistique consacrée à la notion de *passage*. Cette forme de contexte est toujours très mobilisée (dans les pratiques pédagogiques, documentaires, herméneutiques, etc.), mais elle apparaît encore mal cernée théoriquement. L'enjeu est donc de se donner une définition linguistique du passage, qui explique son fonctionnement sémiotique, et donne des éléments pour sa modélisation. Dans cet article, Rastier souligne en particulier un changement de perspective radical pour les pratiques de linguistique du corpus : le texte n'est pas analysable en *unités* (prédéfinies, délimitées, stables, combinables), mais ce sont des *grandeurs textuelles* (à construire dans une interprétation, potentiellement variables, diffuses, non uniformes) qui permettent de rendre compte de son fonctionnement linguistique. Rastier analyse et décrit les multiples contextualisations créées par un passage, tant internes (interactions entre signifiant et signifié) qu'externes (effets à l'intérieur du même texte, mais aussi parcours interprétatifs intertextuels suscités par exemple par une citation).

Nous entrons ensuite dans des expériences de constitution de corpus. Ce sont des occasions de construire et de traduire, dans des choix de codage et d'instrumentation de la consultation, la portée remarquable de certains effets de contexte, et la non-univocité de l'interprétation – y compris pour l'établissement même du texte et pour l'application de grilles d'analyse bien définies. L'explicitation des conventions de codage et la documentation de la pratique d'annotation ressortent bien alors comme un élément central pour un rapport scientifique au corpus, où l'inévitable subjectivité est reconnue pour mieux être maîtrisée.

Morgenstern et Parisse constituent des corpus oraux d'enfants apprenant à parler (enfants de 1 à 3 ans) : aux difficultés interprétatives connues de la transcription (écrire la parole) s'ajoute la particularité de ces babillages enfantins, qui

suscitent spontanément une interprétation les projetant dans un équivalent en « langage adulte ». Comment alors rendre compte de l'observable linguistique, et préserver la liberté interprétative des différents chercheurs utilisant le corpus ? Comment favoriser les rapprochements utiles à l'analyse sans réduire, voire dénaturer la parole enfantine, par une normalisation excessive ou une sur-interprétation restrictive ? Dans ce domaine du travail sur l'oral, les corpus sont transcrits comme un faisceau de lignes de codage parallèles, rendant compte de différents types d'analyse (phonétique, orthographique, annotations sémantiques, description de la situation, etc.), elles-mêmes alignées avec l'enregistrement audiovisuel : cela tisse des contextualisations multiples, et quelquefois questionnables. Ainsi, le statut de ligne principale, accordé généralement à la transcription orthographique, et d'abord motivé par des considérations techniques d'alignement, oriente et domine en pratique trop souvent les lectures des corpus.

Rinck et Tutin nous rendent compte d'une expérience d'annotation d'un phénomène linguistique donné. Dans le cadre d'un projet sur les marqueurs linguistiques de la subjectivité et de la polyphonie, il s'agit ici d'attribuer aux passages entre guillemets une valeur sémantique parmi cinq (préalablement identifiées par les linguistes), dans un corpus d'écrit de recherche en sciences humaines. Leur étude analyse les résultats d'un codage multi-annotateurs : quelle est la part des divergences, et comment sont-elles interprétables. Si certaines sont la trace d'imprécisions des consignes (que l'on peut donc affiner), d'autres semblent irréductibles, soit par polysémie (les guillemets relevant de plusieurs emplois simultanément), soit par ambiguïté (le choix entre plusieurs valeurs n'étant pas décidable). Cette plurivocité rémanente est pleinement intégrée à la démarche d'annotation : elle n'est pas pointée comme une difficulté de la modélisation, un problème pour le codage, ou une faiblesse de l'étiquetage ; elle n'est pas non plus, à l'inverse, exploitée comme une facilité permettant de couper court à certaines finesses d'analyse. Enfin, la pratique de l'annotation permet d'observer l'incidence des contextes de multiples niveaux sur l'interprétation, y compris au-delà du paragraphe : ainsi, la valeur d'une occurrence peut se trouver

Introduction

déterminée par une première occurrence bien avant dans le même texte, voire renvoyer à un intertexte, connu de l'expert du domaine mais implicite, et pas nécessairement repris dans le corpus.

Les deux contributions suivantes mettent davantage l'accent sur les mécanismes interprétatifs, et la recherche de contextualités pertinentes.

Lecolle est, comme Rinck et Tutin, confrontée à l'étude des valeurs d'un motif linguistique par l'intermédiaire de son annotation dans un corpus. Sa problématique est l'évolution du sens du nom de lieu *Outreau*, étudiée à travers des articles de presse relatant l'affaire (2001-2006) finalement considérée comme un fiasco judiciaire. Du point de vue de la méthode, il s'agit donc d'expérimenter et préciser des critères linguistiques de repérage des sens de *Outreau*, en mobilisant parallèlement toute une palette d'approches (distributionnelle, causale, etc.). Cette étude est l'occasion d'observer la nécessité du recours à un contexte étendu, complémentirement au contexte étroit, notamment lorsque ce contexte étroit est directement lié à l'angle d'analyse. Par exemple, les constructions du nom propre avec certaines prépositions, telles que *à Outreau* et *d'Outreau*, mobilisent pour leur interprétation un contexte au-delà du syntagme, qui peut monter jusqu'au palier du texte, daté. Et là encore, dans la démarche de Lecolle, l'interprétation est envisagée dans sa dynamique et dans sa possible plurivocité, elle n'est pas réduite à une opération de désambiguïsation.

Desquinabo cherche à rendre compte des processus interprétatifs, alliant indices locaux et attentes globales, liés à la (re)connaissance du genre textuel. Il modélise ainsi l'identification des actes de paroles dans un corpus de transcriptions d'émissions télévisées. Dans son expérimentation, il met en évidence (par des analyses statistiques) des indices de tous ordres pour caractériser quatre genres textuels : éléments péritextuels non linguistiques (comme le décor ou le type d'invités) et linguistiques (caractéristiques lexico-syntaxiques de la présentation en début d'émission), et éléments lexico-syntaxiques du texte lui-même. De même, il avait montré par ailleurs que la distribution des actes de parole (annotés manuellement en corpus) est différente selon les genres.

Mobilisant alors tous ces indices, il peut simuler des parcours interprétatifs locaux, montrant comment ceux-ci tirent parti d'indices locaux comme d'attentes globales du téléspectateur, et observant la dynamique de l'interprétation au fil du texte.

Notre numéro se termine, avec la contribution de Loiseau, sur des considérations innovantes et plus techniques sur la nature des corpus. Disposant à l'heure actuelle de corpus annotés, l'étape suivante est l'accès à des corpus multi-annotés, qui fusionnent des annotations complémentaires (de différentes natures, par exemple morphologiques et sémantiques) ou concurrentes (de même nature mais correspondant à deux analyses différentes, par exemple trois segmentations pratiquées par trois analyseurs morpho-syntaxiques). On crée ainsi une contextualité non plus seulement en *largeur*, sur l'axe syntagmatique, mais aussi en *épaisseur*, avec la possibilité de mettre en relation les différentes annotations entre elles. Et en définitive, tout dans le corpus est annotation, ou du moins a le même statut d'interprétation : il n'y a pas un texte objectif et premier sur lequel viendraient se greffer des informations bien définies ; le corpus doit plutôt être pensé comme un faisceau d'annotations traduisant chacune une interprétation. Loiseau développe alors un logiciel, *CorpusReader*, qui permette à la fois la constitution de corpus multiannotés et leur exploitation. Il se fixe quelques principes directeurs forts, comme le refus d'imposer un format aux étiquetages qui conditionnerait leur intégration, ou encore le souhait de donner au chercheur qui utilise le corpus l'accès le plus libre, le moins prédéfini, aux étiquetages collectés, et bien sûr la meilleure contextualisation réciproque des annotations. Loiseau explique alors comment ces principes et l'état de l'art des techniques et des outils déterminent scientifiquement ses choix d'implémentation (XML, TEI, API SAX, etc.)

Puissent ces pages réaffirmer à la fois l'importance et l'intérêt d'une reconnaissance de la nature interprétative des corpus, et corrélativement du rôle central de contextes de tous niveaux, locaux et globaux, pour l'accès aux textes et la description de la langue.

Introduction

Références

- Mayaffre Damon (2002). « Les corpus réflexifs : entre architextualité et hypertextualité », *Corpus 1* : 51-69.
- Mellet Sylvie (éd.) (2002). *Corpus et recherches linguistiques, Corpus 1*.
- Péry-Woodley Marie-Paule (1995). « Quels corpus pour quels traitements automatiques ? », *Traitement Automatique des Langues 36*, 1-2 : 213-232.
- Pincemin Bénédicte (1999). « Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative », *Atelier Corpus et TAL : pour une réflexion méthodologique*, organisé par Anne Condamines, Marie-Paule Péry-Woodley et Cécile Fabre, Conférence TALN'99, Cargèse, 12-17 juillet 1999, pp. 26-36.
- Rastier François (1987). *Sémantique interprétative*. Paris : PUF.
- Rastier François (2001). *Arts et sciences du texte*. Paris : PUF.
- Sperberg-McQueen C.M., Lou Burnard (eds) (2002). *Guidelines for Text Encoding and Interchange*, Published for the TEI Consortium by the Humanities Computing Unit, University of Oxford, 2002. ISBN 0-952-33013-X.

Comité de lecture spécifique au numéro

- Benoît Habert, Adonis & ICAR, ENS-LSH, Université de Lyon.
- Serge Heiden, ICAR, ENS-LSH, Université de Lyon.
- Damon Mayaffre, BCL, CNRS, Université de Nice - Sophia Antipolis.
- Sylvie Mellet, BCL, CNRS, Université de Nice - Sophia Antipolis.
- Bénédicte Pincemin, ICAR, CNRS, Université de Lyon.
- Céline Poudat, SES, ENST, Paris.
- Yannick Prié, LIRIS, Université de Lyon.
- Laurent Rouveyrol, BCL, Université de Nice - Sophia Antipolis.