MÉTHODOLOGIE TRANSDISCIPLINAIRE DE GESTION DE CORPUS POUR LES DISCIPLINES DE L'INTERACTION : RECHERCHE DE PRINCIPES DIRECTEURS

Hassan ATIFI, Christophe LEJEUNE, Goritsa NINOVA, Manuel ZACKLAD
Tech-CICO, Université de Troyes, Institut Charles Delaunay

SOMMAIRE

Introduction

Le laboratoire Tech-CICO est engagé depuis plusieurs années dans un travail inter- et transdisciplinaire qui associe des chercheurs en Sciences Humaines et en informatique, en particulier, dans le cas de recherches intervention auprès d'organisations professionnelles. Dans ce contexte, la nécessité de partager des corpus numérisés d'interactions communicatives et leurs commentaires entre différents analystes est de plus en plus manifeste. Pour surmonter les difficultés liées aux choix de corpus, aux modalités de segmentation et aux catégories d'analyse, nous avons décidé de lancer un projet de recherche interne visant à développer une méthodologie transdisciplinaire de gestion de corpus pour les disciplines de l'interaction (sélection des situations de référence, acquisition, retranscription, documentarisation, segmentation, interprétation, valorisation, etc.).

Notre objectif est de partir des pratiques concrètes de constitution de corpus dans plusieurs disciplines afin de parvenir à définir les conditions du partage matériel des sources relevant d'un intérêt commun et à expliciter les catégories d'analyse élaborées dans les différentes disciplines pour permettre un dialogue avec ses pairs. Pour mener cette recherche exploratoire nous avons procédé à la réalisation d'entretiens approfondis avec deux représentants des trois disciplines abordées : la sociologie qualitative, la psychologie ergonomique et la linguistique interactionnelle. Nous complétons ces entretiens par une brève revue de la littérature traitant de cette question.

Comme la linguistique de corpus s'est illustrée comme la discipline de référence pour la gestion des corpus, nous introduirons cet article par l'examen de la place du corpus dans cette discipline avant de développer l'apport des trois autres disciplines.

1. La linguistique de corpus

Les études sur corpus en linguistique se caractérisent généralement par une approche quantitative, sur de grandes masses de données, avec des méthodes (semi)-automatiques qui visent à assurer la reproductibilité, la validation et la généralisation des résultats. Les spécialistes s'accordent à dire qu'il existe un lien entre la linguistique de corpus et l'outil informatique qui fait partie intégrante de la démarche empirique prônée par les défenseurs de la linguistique de corpus (Péry-Woodley 1995), (Habert *et al.* 1997).

La linguistique de corpus insiste sur le caractère restrictif du corpus : « Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques *et extra-linguistiques* explicites pour servir d'échantillon *d'emplois déterminés d'une langue* » (Habert 2000).

En fonction des besoins de recherche, plusieurs distinctions parcourent la discipline : corpus de textes *versus* d'échantillons, corpus de référence (qui vise à représenter toutes les variétés pertinentes d'une langue) *versus* corpus de spécialité (qui est restreint à une situation de communication, un domaine, une langue de spécialité).

La constitution d'un corpus de faits langagiers soulève nécessairement le problème de la représentativité. On s'accorde aujourd'hui encore sur la difficulté, en matière de langages, à donner une « définition positive de la représentativité » (Habert *et al.* 1997). Les exigences à respecter dans tous les cas sont formulées ainsi : « la taille des données doit être suffisante (par

souci de représentativité), elles doivent être diversifiées et leur origine doit être clairement mémorisée » (Habert *et al.* 1998).
Pour les linguistes de corpus, il est nécessaire de créer des ressources linguistiques communes et réutilisables. En conséquence, une attention particulière est portée à la documentation de chaque ressource. En amont, les chercheurs s'accordent sur certaines décisions relatives aux variations de domaines et de registres, sur les types de textes, sur le balisage de la structure du texte et sur l'analyse envisageable.
Une fois le corpus constitué, il fera l'objet d'une série de traitements le plus souvent automatiques de normalisation, de nettoyage, de segmentation, de regroupement, d'étiquetage et d'annotation (Habert *et al.* 1997, 1998). Le plus souvent sont utilisés les corpus étiquetés (chaque mot est assorti d'une étiquette) et les corpus arborés (munis d'arbres syntaxiques).
Les objectifs principaux poursuivis par la linguistique de corpus sont la constitution de ressources, leur mise à la disposition et leur confrontation par la communauté des linguistes (Péry-Woodley 2005).

2. La sociologie qualitative

En sociologie, la notion de corpus renvoie à la méthode dite qualitative. Cette contribution n'évoque par conséquent pas la sociologie quantitative (Fox 1999). Typiquement, un corpus se compose d'une série de transcriptions d'entretiens, donc d'interactions occasionnées en vue de répondre à une question de recherche (Bourdieu 1994). Les analyses sociologiques portent également sur d'autres sources empiriques que les entretiens :

- des notes de terrain ainsi que l'expérience du sociologue ayant procédé à une observation participante.
- des documents écrits, comme, entre autres, les règlements de travail (Foucault 1975), les manuels de management (Boltanski 1999), les inscriptions publiques (Heinich 1995), les lettres (Boltanski 1984), les coupures de presse (Chateauraynaud 1999, Duret 2001), les articles scientifiques (Latour 1995), les affiches (Latour 1992), les manuels de savoir-vivre (Elias 2003), les modes d'emploi (Akrich 1991), les listes de courses (Conein 1994) ou les programmes politiques.
- des photographies (Bourdieu 1979, Trepos 1997, Latour 1985).
- des vidéogrammes (Lee 1993).

La constitution du corpus de référence à partir du terrain soulève la question de sa représentativité. La réponse sociologique à cette question procède par l'épuisement de la diversité du matériau recueilli. Comme en linguistique de corpus, représentativité rime avec exhaustivité. Mais, contrairement à la fixation statistique d'un échantillon en linguistique de corpus, la logique de saturation des différences n'a aucun rapport géométrique avec la population globale (le corpus potentiel). En tant que recensement de toutes les singularités, elle est encore distincte du travail du linguiste interactionnel sur des exemplaires typiques choisis pour leur centralité.
La position du chercheur lors de ce recueil va de la neutralité distanciée de l'entretien (similaire à la posture témoignée en psychologie clinique) à l'immersion par observation participante dans un milieu écologique (comparable à celle des conversationnalistes).
L'exploitation du corpus de référence procède par relevé et analyse du déroulement et du contenu des interactions, peu (ou pas) d'attention est accordée à leur forme d'expression (contrairement aux préoccupations des linguistes). La langue (qu'elle soit orale ou écrite) n'est abordée qu'en tant que médiatrice – plus ou moins fidèle – des idées, des opinions ou des attitudes. La sociologie s'intéresse le plus souvent à « ce que raconte » le corpus, qu'elle tente d'articuler à ses questions de recherche et aux cadres théoriques mobilisés. Bien qu'il s'agisse d'une démarche qui se qualifie de qualitative, les comptages sont loin d'être absents des analyses de ce type, que la méthode passe par l'analyse de contenu, l'analyse structurale des récits ou mobilise des logiciels[1] pour exploiter les documents. La récurrence d'une idée lui attribue une première importance par rapport aux hapax. Ce présupposé n'évacue bien entendu pas les éléments marginaux qui sont également mobilisés dans les analyses.

[1] Ces logiciels (relativement nombreux) regroupent entre autres les CAQDAS (NVivo, Nud*Ist, Kwalitan, The Ethnograph), Alceste, Tropes, Prospéro, Candide, Leximappe.
Pour une recension, http://www.smess.egss.ulg.ac.be/lejeune/logiciels/.

3. La psychologie ergonomique

Bien qu'elle n'emploie pas le terme corpus, la psychologie ergonomique collecte des inscriptions des activités spécifiques de résolution de problème qu'elle étudie. Ces enregistrements regroupent le film de la situation observée (quand il est possible de le réaliser) et des paroles des sujets. Dans le cas des situations de résolution collective de problème, les échanges verbaux constituent simultanément un mode d'action et une trace des activités cognitives des sujets. C'est la raison pour laquelle ces situations sont particulièrement recherchées. Lorsqu'elles ne sont pas disponibles, les sujets sont invités à commenter à voix haute ce qu'ils sont en train de faire – verbalisation simultanée – ou ce qu'ils ont fait – verbalisation *a posteriori*.

La question de la représentativité de cet ensemble de matériaux trouve une réponse différente des autres perspectives. Il n'est question ni d'exhaustivité, ni d'échantillonnage, ni de saturation. C'est le statut d'expert du sujet qui rend pertinent son témoignage. À la différence de la typicité de la linguistique interactionnelle, c'est l'efficacité du comportement qui importe.

À la différence de bien d'autres courants en psychologie, la psychologie ergonomique adopte une posture clinique qui n'exclut pas une démarche d'intervention. En effet, les enquêtes du psychologue ergonome visent à enregistrer les pratiques des sujets afin de mettre en évidence ses méthodes, d'identifier des invariants et souvent de rendre le dispositif plus opérant, plus efficace. Praxéologue, il cherche à connaître pour (mieux) agir.

Pour lui, les dits des sujets sont des intermédiaires qui donnent accès aux opérations cognitives comme le raisonnement ou l'inférence (inobservables par ailleurs). La forme des dits n'est donc pas déterminante en elle-même. Une attention accrue est par contre portée à l'articulation de ce qui est dit et fait par le sujet. La résolution d'un problème sert de dispositif de collecte adéquat par rapport à cette posture.

Lors de l'analyse des matériaux accumulés, le psychologue cherche à identifier les processus inférentiels, les raisonnements. Il emprunte les techniques de l'analyse de contenu. En complément, il mobilise également des statistiques descriptives. L'identité disciplinaire n'est sans doute pas étrangère à ce choix, la psychologie, s'étant historiquement engagée très tôt dans le recours à ces techniques. Le psychologue cherche ensuite à dégager les éventuels invariants et à modéliser les procédures (par définition, spécifiques) mises en œuvre par les sujets.

4. La linguistique interactionnelle

Depuis quelques décennies et sous l'influence des courants interactionnistes anglo-saxons « s'est affirmée de plus en plus fortement en linguistique interactionnelle l'exigence de travailler sur des corpus de données attestées comme alternative à des démarches fondées sur l'introspection de jugements des locuteurs ou sur l'élicitation de jugements des locuteurs » (Mondada 2005). Cette préférence des données naturelles sur des données fabriquées par introspection, par simulation ou par expérimentation s'inscrit dans un mouvement général dans plusieurs disciplines de la nouvelle communication (Winkin 1981).

Une deuxième exigence invite à travailler sur des *enregistrements* – audio ou vidéo – d'interactions sociales, c'est-à-dire sur des données permettant de documenter l'émergence et le déploiement de ces pratiques *dans le temps*. On ne travaille donc *ni* sur les descriptions de ces pratiques (dans les entretiens, dans des notes prises par le chercheur) *ni* sur des produits de celles-ci (par exemple dans des textes issus d'une manière ou d'une autre de l'activité) (Mondada 2005).

Le problème de la représentativité n'est pas nécessairement articulé à une ambition de généralisation. La réponse du chercheur à cette question consiste seulement à dégager ce qui est propre au corpus étudié, ce qui en fait le style ou ce qui s'y manifeste comme phénomènes récurrents (Condamines 2005). Le corpus devra être pertinent (Vincent 2003) ou de « qualité » (Plantin 2005) ; pour ce dernier, le corpus doit manifester les trois dimensions suivantes :

- technique : il faut réaliser des corpus de bonne qualité technique sonore et visuelle pour faciliter leur conservation, leur transcription et leur exploitation manuelle ou informatique.
- juridique : l'enjeu juridique touche trois dimensions : le respect de la vie privée des personnes enregistrées (accord préalable des enquêtés, anonymisation des données), le droit d'auteur (entre collecteurs, transcripteurs et chercheurs) et le recueil et la diffusion de données (préparation et mise en place de l'enregistrement).
- sociolinguistique : un « bon corpus » est souvent décrit dans la littérature comme « naturel », « authentique » et « représentatif » (dans le sens évoqué ci-dessus).

Au niveau de l'analyse, le chercheur s'attelle particulièrement à décrire l'organisation, le fonctionnement et les enjeux des interactions. Il rend compte de manière parallèle des phénomènes verbaux, vocaux et gestuels. Une attention est portée à la dimension comportementale ou pragmatique des échanges dont l'unité minimale est l'acte de langage.
S'appuyant sur ces considérations méthodologiques fortes, quelques équipes internationales (comme le projet CLAPI mené par le laboratoire ICAR à Lyon) se sont engagées dans l'archivage de corpus parlés en interaction. Cet archivage a plusieurs visées :

- patrimoniale : sauvegarder des façons de parler et constituer une documentation historique des usages de la langue en interaction.
- scientifique : permettre les études empiriques de phénomènes linguistiques.
- appliquée : tirer des préconisations ou applications des études.

Discussion

Au terme de cette confrontation disciplinaire, nous pouvons avancer quelques constatations[1]. La posture du chercheur par rapport à son objet varie également d'une discipline à l'autre. Là où la linguistique de corpus vise une objectivation des phénomènes dont elle rend compte, le sociologue tente d'appréhender la subjectivité des différents acteurs. Le recours aux entretiens le positionne dans une neutralité distanciée par rapport à ses informateurs. Il arrive que certains sociologues procèdent à une observation participante. Ils tentent alors une immersion dans le phénomène social qu'ils observent. En linguistique interactionnelle, il existe une double tendance. Certaines recherches impliquent la conversion du chercheur qui acquiert, par immersion ou imprégnation, la compétence des locuteurs étudiés (ce qui est proche de l'observation participante en sociologie ou en anthropologie). D'autres recherches procèdent à l'effacement du dispositif de recueil des données. Dans ce deuxième courant, le chercheur est alors un pur observateur, distant et, idéalement, invisible. Cette position est similaire à la posture clinique témoignée par le psychologue ergonomique. Ce dernier ne s'imprègne pas du terrain (vu qu'il n'apprend pas à effectuer les tâches) mais, comme il n'exclut pas l'intervention, il n'emprunte pas non plus la neutralité distanciée des sociologues.
En bref, la sociologie qualitative procède typiquement par entretiens. La psychologie ergonomique s'intéresse aux situations de résolution de problème. La linguistique interactionnelle privilégie une démarche naturelle d'observation, portant sur des situations qui ne sont pas créées, arrangées, préparées par le chercher pour les fins de son enquête. Le comportement interactionnel est enregistré dans sa totalité verbale et non verbale.
Ces disciplines traitent différemment des dires, des actes et de leur contexte. À la différence de la perspective linguistique, le psychologue et le sociologue s'intéressent plus au contenu des dires qu'à leur forme. Toutefois, là où le sociologue s'intéresse surtout à ce que l'informateur lui dit, le psychologue ergonomique étudie l'articulation du geste et de la parole. Le conversationnaliste prolonge plus loin encore cette attention.
La façon dont les interactions sont transcrites est congruente avec ces différentes démarches : le sociologue transcrit la teneur du propos (ce qui est dit), sans trop se préoccuper (la plupart du temps) des gestes, postures, intonations, pauses et hésitations. Le psychologue ergonomique sera lui particulièrement attentif à transcrire en parallèle le geste et le discours. À travers l'analyse conversationnelle, la linguistique interactionnelle s'est enfin particulièrement illustrée pour ses transcriptions fines et détaillées incorporant les chevauchements, les pauses (chronométrées), les intonations et le non-verbal.
Plus préoccupée de la signification du discours que de sa forme, la psychologie et la sociologie partagent plusieurs de leurs outils d'exploitation et d'analyse du corpus (avec, en bonne position, l'analyse de contenu). À nouveau, cependant, des différences se manifestent dans l'usage de ces techniques. Le sociologue se contente de recenser les différents arguments, idées, logiques ou phases d'un entretien. L'épuisement de la diversité lui suffit. Le psychologue sera, pour sa part, plus enclin à recourir à des techniques d'objectivation comme les statistiques. En linguistique interactionnelle, l'analyse ne se limite pas au contenu, elle prend pour sa part également en considération le niveau formel, à travers la structure et l'organisation des interactions. L'unité d'analyse sera, en fonction du niveau de granularité, la conversation, la séquence, l'échange, l'intervention ou l'acte de langage.

[1] Une présentation plus détaillée a été proposée lors de la communication orale.

En ce qui concerne la question de la représentativité, chaque discipline propose une réponse singulière. La sociologie fait rimer représentativité avec exhaustivité. Si cette acception la rapproche de la linguistique de corpus, la logique de saturation qu'elle emprunte la distingue cependant de la fixation statistique d'un échantillon. L'épuisement des diversités n'a en effet aucun rapport géométrique avec la population globale. La psychologie ergonomique et la linguistique interactionnelle donnent à cette même question de la représentativité une réponse qui ne s'articule pas toujours à une visée d'exhaustivité. Les raisons de chacune de ces disciplines sont cependant différentes. Vu que la psychologie ergonomique s'attelle à enregistrer les procédures efficientes, c'est le statut d'expert du sujet qui l'emporte sur le nombre de sujets rencontrés. La linguistique interactionnelle travaille, pour sa part, sur des exemplaires typiques extraits des phénomènes récurrents et réguliers observés dans les interactions.
En conclusion, cette confrontation nous a permis de prendre conscience des écarts disciplinaires. Aucune discipline ne peut à elle seule épuiser la complexité des phénomènes interactionnels. On s'interrogera sur les conditions de (1) partage de corpus communs entre experts de ces différentes disciplines, (2) leur analyse collective et (3) d'éventuelles transformations des pratiques de chacun pour rendre cette collaboration transdisciplinaire opérante.

BIBLIOGRAPHIE

AKRICH, M. et BOULLIER, D. 1991. Le mode d'emploi : genèse, forme et usage, in D. Chevalier (éd.), *Savoir faire et pouvoir transmettre. Transmission et apprentissage des savoir-faire et des techniques*, Maison des sciences de l'homme.
BILGER, M. (éd). 2000. *Corpus : Méthodologie et applications linguistiques*, Paris, Honoré Champion.
BOLTANSKI, L. et CHIAPELLO, E. 1999. *Le nouvel esprit du capitalisme*, Paris, Gallimard.
BOLTANSKI, L., DARRÉ, Y. et SCHILTZ, M.-A. 1984. La dénonciation, *Actes de la Recherche en sciences sociales*, 51, pp. 3-40.
BOURDIEU, P. 1979. *La distinction. Critique sociale du jugement*, Paris, Minuit.
BOURDIEU, P. 1994. *Raisons pratiques. Sur la théorie de l'action*, Paris, Seuil.
CONDAMINES, A. (dir.). 2005. *Sémantique et corpus*, Londres, Hermès.
CHATEAURAYNAUD, F. et TORNY, D. 1999. *Les sombres précurseurs. Une sociologie pragmatique de l'alerte et du risque*, Paris, École des Hautes Études en Sciences Sociales.
CONEIN, B. et JACOPIN, E. 1994. Action située et cognition. Le savoir en place, *Sociologie du Travail*, 4, pp. 475-500.
DURET, P. et TRABAL, P. 2001. *Le sport et ses affaires. Une sociologie de la justice de l'épreuve sportive*, Paris, Métailié.
ELIAS, N. 2003. *La Civilisation des mœurs*, Paris, Agora.
ERICSSON, K.A., SIMON, H.A. 1993. *Protocol analysis : Verbal reports as data*. MIT, Cambridge.
FOUCAULT, M. 1975. *Surveiller et punir. Naissance de la prison*, Paris, Gallimard.
FOX, W. 1999. *Statistiques sociales*, Paris, De Boeck.
HABERT, B. 2000. Des corpus représentatifs : de quoi, pour quoi, comment ?, in M. Bilger (éd.), *Linguistique sur corpus. Études et réflexions*, Perpignan, Presses Universitaires de Perpignan, pp. 11-58.
HABERT, B., NAZARENKO, A. et SALEM, A. 1997. *Les linguistiques de corpus*, Paris, Armand Colin/Masson.
HABERT, B., FABRE, C. et ISAAC, F. 1998. *De l'écrit au numérique : constituer, normaliser et exploiter les corpus électroniques*, Paris, InterEditions.
HEINICH, N. 1995. Les colonnes de buren au palais-royal. Ethnologie d'une affaire, *Ethnologie française*, XXV, 4, pp. 525-541.
HOC, J.-M. et DARSE, F. (éds.) 2004. *Psychologie ergonomique : tendances actuelles*, Paris, PUF.
KERBRAT-ORECCHIONI, C. 1994. *Les interactions verbales*, Paris, Armand Colin.
LATOUR, B. 1985. Les "vues" de l'esprit. Une introduction à l'anthropologie des sciences et des techniques, *Culture Technique*, 14, pp. 5-29.
LATOUR, B. 1992. *Aramis ou l'amour des techniques*, Paris, La Découverte.
LATOUR, B. 1995. *La Science en action*, Paris, Gallimard [La Découverte, 1989].

LEE, J. et WATSON, R. 1993. Regards et habitudes des passants. Les arrangements de visibilité de la locomotion, *Les annales de la recherche urbaine*, 57-58, pp. 101-109.
MONDADA, L. 2005. L'analyse de corpus en linguistique interactionnelle : de l'étude de cas singuliers à l'étude de collections, in A. Condamines (dir.), *Sémantique et corpus*, Londres, Hermès.
PERY-WOODLEY, M.-P. 1995. Quels corpus pour quels traitements automatiques ?, *TAL*, 36(1-2), pp. 213-232.
PERY-WOODLEY, M.-P. 2005. Discours, corpus, traitements automatiques, in A. Condamines (dir.), *Sémantique et corpus*, Londres, Hermès.
PLANTIN, C. 2005. Pour une archive des langues parlées en interaction. Statuts juridiques, formats et standards, représentativité, in J. L. Lebrave, *La société de l'information et ses enjeux,* actes du colloque de bilan « La société de l'information » 2001-2005, ENS-LSH, Lyon.
PLETY, R. 1993. *Ethologie des communications humaines. Aide-mémoire méthodologique*, Lyon, Arci-PUL.
RASTIER, F. et BOUQUET, S. (éds.). 2000. *Une introduction aux sciences de la culture*, Paris, PUF.
RASTIER, F. 2005. Les enjeux épistémologiques de la linguistique de corpus, in G. Williams (éd.), *La linguistique de corpus*, Rennes, Presses universitaires de Rennes.
RICHARD, J.-F. 2005. *Les activités mentales. Comprendre, raisonner, trouver des solutions*, Paris, Armand Colin.
TRAVERSO, V. 1999. *L'analyse des conversations*, Paris, Nathan (coll. 128).
TREPOS, J.-Y. 1997. Approche méthodologique de l'utilisation de la photographie dans l'enquête sociologique, Notes du séminaire du 17 novembre 1997.
WINKIN, Y. 1981. *La Nouvelle Communication*, Paris, Éditions du Seuil.