

La pragmatique de corpus comme lieu d'expérimentation des méthodes mixtes : pour une interdisciplinarité focalisée

Jérôme Jacquin
Université de Lausanne
jerome.jacquin@unil.ch

Résumé

La contribution rend compte des développements récents en pragmatique de corpus, champ d'études au sein duquel la pragmatique linguistique tire bénéfice des acquis de la linguistique de corpus et de l'émergence de méthodes mixtes en son sein. Au travers de deux exemples, la contribution plaide pour une interdisciplinarité focalisée où les spécialistes de méthodes informatiques et statistiques permettent indirectement d'enrichir une discipline, la pragmatique, qui trouve ses origines plutôt dans l'analyse détaillée de (collections de) cas.

1 Introduction

En 2013, les sections de linguistique et d'informatique et méthodes mathématiques de la Faculté des lettres de l'Université de Lausanne fusionnent, sous la forme d'une nouvelle entité dénommée « Section des Sciences du Langage et de l'Information ». Cette fusion se réalise par ailleurs à l'occasion d'un redéploiement et d'une redistribution des bureaux de l'ensemble de la Faculté. Ces bouleversements vont entériner une cohabitation à la fois spatiale et institutionnelle allant bien au-delà des échanges scientifiques et personnels ponctuels qui préexistaient à la fusion et qui vont dès lors largement s'intensifier. En 2018, alors que j'occupe désormais un bureau qui jouxte celui de François Bavaud et qui se trouve à une demi-douzaine de mètres de celui d'Aris Xanthos, je dépose un projet auprès du Fonds National Suisse de la recherche (FNS), qui porte clairement la trace de toute une

série de transformations progressives qui se sont opérées en moi et dont une part est précisément imputable à cette cohabitation. Le projet en question, intitulé « Prendre une position épistémique dans l'interaction. Les marqueurs du savoir, du non-savoir et du doute en français » (PO-SEPI), a été financé par le FNS de 2020 à 2024 [100012_188924]. Il entendait proposer une étude systématique des marqueurs épistémiques (par ex. « peut-être », « il me semble », « c'est pas sûr ») et évidentiels (par ex. « je l'ai vu », « Pierre m'a dit »), tels qu'ils émergent dans un corpus de 28 heures d'interactions naturelles vidéo-enregistrées et documentant des débats politiques et des réunions d'entreprise. Près de 4000 marqueurs ont été annotés (Keck, en préparation, Robin 2024), en suivant un schéma d'annotation exigeant de renseigner des variables catégorielles aussi bien morphosyntaxiques que pragmatiques au sens large, c'est-à-dire énonciatives, interactionnelles et multimodales (direction du regard, gestes cooccurrents ; voir *infra*). Les quelques pages qui suivent abordent trois aspects qui permettent de lier ce projet à la présence concrète et aux intérêts de recherche et de réflexion de François Bavaud. Je commencerai par situer la pragmatique de corpus comme champ d'étude émergeant invitant à adopter des méthodes mixtes au sein d'une discipline, la pragmatique, dont les origines s'ancrent plutôt dans l'analyse détaillée de (collections de) cas, mais qui aujourd'hui bénéficie de plus en plus de notions et outils statistiques. Dans la seconde partie, j'aborderai deux exemples plus concrets à l'occasion desquels les méthodes dites quantitatives et le dialogue direct avec François Bavaud et son équipe ont pu résoudre des défis qui se présentaient : le test d'accord inter-annotateurs sur des variables non discrètes et l'utilisation d'analyses factorielles de correspondances multiples pour articuler de manière outillée les distributions fréquentielles d'ensemble au niveau macro et les analyses séquentielles de cas au niveau micro.

2 La pragmatique de corpus comme champ émergeant

2.1 Développements récents de la linguistique de corpus

La linguistique de corpus a longtemps été associée à l'analyse quantitative de données textuelles, essentiellement écrites (voir par ex. Biber et al. 1998; Svartvik 1992 ; voir également Gries 2010, 2017 sur l'aspect

quantitatif). Depuis une vingtaine d'années, on constate toutefois une double ouverture. La première ouverture concerne un au-delà de la seule quantification et de l'analyse centrée sur des observations statistiques. Cela est rendu possible par un double mouvement. Dans un premier temps, cela se réalise par l'émergence d'une approche plus herméneutique et praxéologique des corpus, qui en font des ressources situées, pilotées par des objectifs de recherche et dont la valeur est fonction des questions que l'on pose et des résultats plus ou moins généraux que l'on cherche à obtenir (voir par ex. [Egbert et al., 2020](#); [Mayaffre, 2005](#); [Rastier, 2004](#)). Dans un second temps, on constate un développement de plus en plus soutenu des méthodes mixtes, qui associent des analyses statistiques à des études fines de collections de cas (voir par ex. [Angouri, 2010](#); [Riazi, 2016](#); [Teddlie & Tashakkori, 2009](#)). La seconde ouverture de la linguistique de corpus tient aux supports analysés et relève d'un au-delà de la seule scripturalité. Cette ouverture est favorisée par des développements théoriques, analytiques et techniques autour de la multimodalité du langage. On entend par là le fait que les ressources verbales ne constituent qu'une dimension sémiotique pertinente parmi d'autres et qu'une science du langage doit pouvoir intégrer le rôle du corps, qu'il s'agisse de postures, d'expressions faciales telles que la direction du regard ou encore de conduites gestuelles (voir par ex. [Bate-man et al., 2017](#); [Kress & Van Leeuwen, 2001](#); [Sidnell & Stivers, 2005](#)). Ces développements multimodaux au-delà des corpus uniquement écrits sont aussi imputables à l'émergence, à l'agrégation et à la (encore très relative) mutualisation de corpus vocaux, voire multimodaux (voir par ex. [Adolphs & Carter, 2013](#); [Avanzi et al., 2016](#); [Baldauf-Quilliatre et al., 2016](#)). En définitive, les développements récents de la linguistique de corpus s'articulent aujourd'hui très volontiers avec le souhait d'une linguistique basée fondamentalement sur l'usage, en tant que cet usage exige des méthodes et des approches plurielles appliquées à des corpus sémiotiquement complexes. La complexité des corpus ne constitue finalement que le reflet de celle des diverses pratiques plus ou moins ordinaires du langage que ces corpus renseignent.

2.2 Développements récents de la pragmatique

La pragmatique linguistique constitue *grosso modo* l'étude de la relation bidirectionnelle entre les mots et la situation de communication (voir par ex. [Bublitz & Norrick, 2011](#); [Huang, 2017](#); [Levinson, 1983](#)). La pragmatique peut donc constituer aussi bien l'analyse de la part du sens des mots qui dépend de facteurs contextuels (par ex. les unités de la deixis comme « je », « ici », « maintenant ») que celle des changements contextuels opérés grâce aux mots (par ex. les actes de langage, les termes à contenus procéduraux, comme les connecteurs, ou encore la gestion de la relation interpersonnelle au travers des ressources relevant de la politesse). Si les origines de la pragmatique se situent principalement en philosophie du langage, puis en linguistique structurale postsaussurienne, on peut noter depuis la deuxième moitié du vingtième siècle une explosion de travaux empiriques. Ces développements empiriques se manifestent aussi bien dans un versant expérimental hérité des travaux en psycholinguistique, que dans un versant plus ethnographique relevant de l'observation et de l'enregistrement de terrains authentiques (sur ces deux tendances, voir par ex. [Bublitz et al., 2018](#)). L'étude de la parole en tant que phénomène situé, contextuel, se développe en linguistique puis rencontre, dès les années 1970, l'analyse conversationnelle d'inspiration ethnométhodologique ([Sacks et al. 1974](#); [Sidnell & Stivers 2013](#); voir aussi le développement de la « linguistique interactionnelle », p.ex. [Couper-Kuhlen & Selting 2018](#); [Selting & Couper-Kuhlen 2001](#)). L'analyse conversationnelle propose non seulement des protocoles éprouvés de récolte et de confection de données interactionnelles (notamment des conventions de transcription), mais aussi une certaine manière d'envisager la posture analytique : il s'agit de s'efforcer d'être au plus près de la dynamique séquentielle des échanges tels qu'ils sont méthodiquement et collaborativement élaborés, pas-à-pas, par les participantes et participants à l'interaction. Ces développements conduisent à la confection de données structurées composées d'enregistrements et de transcriptions documentant une large diversité de contextes linguistiques et socioculturels, aussi bien ordinaires que professionnels et institutionnels (par ex. [Arminen, 2005](#); [Berger & Lauzon, 2017](#); [Drew & Heritage, 1992](#)). Ces collections de données documentant des usages

variés sont régulièrement désignées en tant que bases de données, mais aussi corpus, ce qui a posé la question de leur éventuelle ouverture à la linguistique de corpus.

2.3 L'articulation de méthodes dites qualitatives et quantitatives en pragmatique de corpus

De manière analogue à la discussion autour de l'apparent pléonasme de la « socio-linguistique » (par ex. [Encrevé, 1977](#); [Labov, 1972](#)), il pouvait sembler aller de soi que linguistique de corpus et pragmatique étaient destinées à se rencontrer du fait de leur affinité avec la notion d'usage et de contexte, notamment autour de l'idée d'indexicalité (par ex. [Bar-Hillel, 1954](#); [Duranti & Goodwin, 1992](#); [Levinson, 1983](#)). On aurait même pu imaginer que la pragmatique devienne initiative, voire fédératrice pour la linguistique de corpus ; ou en sens inverse que la linguistique de corpus émerge comme une autre manière d'« intégrer la pragmatique », non pas au niveau théorique du dialogue avec la sémantique, à la suite de Culioli puis d'Anscombe et Ducrot ([Anscombe & Ducrot, 1976](#), p. 8) notamment, mais au niveau plus opérationnel des méthodes et des analyses. Ce n'est de loin pas le cas. Pour reprendre les termes de [Rühlemann & Aijmer \(2014, p. 1\)](#), « for a long time pragmatics and corpus linguistics were regarded as 'parallel but often mutually exclusive' ([Romero-Trillo, 2008a, p. 2](#)) ». Dans les années 1980 et 1990, on constate toutefois quelques travaux précurseurs (voir notamment les recensements proposés par [Rühlemann & Aijmer \(2014\)](#); [Jucker \(2018\)](#)), puis une intensification dans les années 2000 (par ex. [Romero-Trillo, 2008b](#)) et surtout 2010 (par ex. [Romero-Trillo, 2013](#); [Rühlemann, 2019](#); [Weisser, 2018](#)), y compris en francophonie (par ex. [Mellet, 2008](#), sur les connecteurs). C'est en 2014 qu'émerge vraiment l'expression « Corpus Pragmatics » [Pragmatique de corpus] en tant que champ disciplinaire davantage constitué et consolidé, du moins dans la tradition anglo-saxonne ([Aijmer & Rühlemann, 2014](#)). À partir de là, non seulement le domaine se développe (voir *supra*), mais la pragmatique prend petit à petit la mesure de cette émergence qui fait bouger les lignes (par ex. [Bublitz et al., 2018](#), part. V). En effet, la pragmatique de corpus invite non seulement à la quantification, mais aussi à l'adoption

de méthodes mixtes (voir aussi *supra*). Cela passe par des analyses statistiques, qui vont de simples fréquences absolues et relatives à des statistiques inférentielles bivariées et multivariées, qu'on exemplifiera *infra* par les analyses factorielles de correspondances multiples¹. Ces analyses de panoramas d'ensemble, qui permettent de dépasser les apories de l'intuition, de la généralisation abusive et de la saillance des exceptions, sont complétées par des analyses dites « qualitatives », qui visent à proposer des analyses détaillées de cas ou de collections de cas, en considérant divers facteurs contextuels s'accommodant mal d'une annotation massive, tels que les facteurs discursifs, interactionnels ou multimodaux. Cela peut notamment se réaliser par la « next turn proof procedure », c'est-à-dire la prise en compte de la réaction de l'allocutaire dans la fixation du sens de ce qui précède (par exemple, identifier une réponse, permet de situer l'apparition d'une question indépendamment de son formatage syntaxique ou prosodique).

3 Deux exemples d'intégration de méthodes statistiques en pragmatique

J'aimerais développer maintenant deux exemples récents où la collaboration avec François Bavaud et son équipe a pu servir le projet de recherche mentionné en introduction, en développant des méthodes et questions innovantes en pragmatique de corpus de manière à trouver des solutions non pas vraiment sur mesure, mais renseignées et alignées sur l'objectif.

3.1 L'accord inter-annotateurs de variables non discrètes

Le projet de recherche POSEPI a impliqué l'annotation de près de 4000 tokens au travers d'un schéma d'annotation nécessitant de renseigner de très nombreuses variables catégorielles (par ex. des catégories sémantiques, morpho-syntaxiques, discursives, interactionnelles ou multimodales). Il était absolument nécessaire que ces annotations soient suffisamment robustes. Pour ce faire, un guide d'annotation d'une cinquantaine de pages a été rédigé et amélioré au fil de différents tests

1 Comme le précise Jucker, « a quantitative perspective requires a very solid foundation in the preparation of the data base and in the analysis and categorisation of the data » (Jucker, 2018, p. 455). Ce point ne pourra pas être détaillé ici.

d'accords inter-annotateurs sur des échantillons tirés aléatoirement (Jacquin et al., 2022). Ces tests ont été réalisés et interprétés via le Kappa de Cohen (Cohen, 1960; Landis & Koch, 1977). Toutefois, le Kappa de Cohen n'est pas prévu pour des variables non discrètes, comme c'est le cas lorsqu'on doit identifier des portions de matériel continu. Par exemple, dans l'énoncé « il est à la maison je pense », le segment « je pense » constitue un marqueur épistémique portant sur le segment « il est à la maison ». Dans un cas de segmentation comme celui-ci, l'accord inter-annotateur serait probablement assuré. Toutefois, si l'énoncé à considérer est « visiblement lui il est à la maison je pense », cela se complique : « visiblement » fait-il partie ou non de la portée de « je pense » ? Et « lui » ? Le Kappa de Cohen ne fait pas la différence et un écart, quel qu'il soit, entre deux annotateurs mènerait à un désaccord complet ; il n'y a pas de proportionnalité dans l'accord. Le projet MODAL, qui portait sur la modalité à l'oral dans différentes langues (par ex. Nissim & Pietrandrea, 2017; Pietrandrea, 2018), s'est posé la même question. La solution trouvée repose sur la fixation de seuils de chevauchements entre les annotateurs (par ex. 10%, puis 50% et enfin 100%), permettant de diviser le continu en segments discrets et d'ainsi appliquer un Kappa de Cohen classique (Ghia et al., 2016). Une telle solution, bien qu'habile, m'a semblé insatisfaisante du fait de la fixation arbitraire des paliers. Une discussion spontanée avec François Bavaud a permis de poser le problème et de remettre en perspective l'objectif. Elle a conduit à différents échanges où j'ai pu préciser mon idée de mobiliser le *timecode* des annotations respectivement associées aux marqueurs et aux portées, pour calculer des pourcentages d'accord en fonction du degré de recouvrement entre les annotations fournies par les deux annotatrices. Ainsi, le résultat du Kappa de Cohen pour chaque token refléchit le degré d'accord sur les portions annotées respectivement par l'annotatrice A et l'annotatrice B, plutôt que de fixer un seuil qui ferait basculer le recouvrement comme un accord ou un désaccord. La formule proposée par François adapte le « coefficient de communauté » de Jaccard (lui-même élaboré dans un premier temps par Grove Karl Gilbert en 1884) et permet ainsi de mesurer le ratio de chevauchement entre deux intervalles temporels en fonction de l'intervalle

total (Bavaud, 2021)². Appliqué à nos données et plus précisément à nos tests d'accord inter-annotateurs, la formule aboutit à un Kappa de Cohen adapté de 0.728 pour les marqueurs et 0.772 pour les portées, ce qui constitue des accords « substantiels » au sens de Landis & Koch (1977). Cette collaboration ponctuelle avec François Bavaud est un exemple où un spécialiste en mathématiques et statistiques aide à traduire en procédure robuste une intuition basée sur un problème pratique émergeant en pragmatique linguistique.

3.2 Méthodes mixtes et identification des (collections de) cas soumis à l'analyse de détails

Le second exemple ne relève pas de l'annotation, mais de l'analyse, et plus précisément de la question relativement délicate de l'articulation entre les distributions statistiques d'ensemble et les collections de cas spécifiques faisant l'objet de l'analyse de détails. En effet, un des risques des méthodes mixtes est que les analyses dites respectivement quantitatives et qualitatives soient disjointes, dans la mesure où le choix des observables soumis aux secondes ne repose pas sur les observations fournies par les premières. Deux méthodes sont ainsi mobilisées dans le cadre du projet POSEPI : la Classification Ascendante Hiérarchique (CAH) et l'Analyse (factorielle) de Correspondances Multiples (ACM). Pour des raisons de place, je vais illustrer l'intérêt de la seconde, mais la plupart des conclusions seront valables pour la première. Les ACM vont consister à sélectionner des variables catégorielles (qualitatives) pertinentes au niveau de la question posée et à étudier et visualiser l'attraction entre elles, sur un plan à deux dimensions au moins (voir par ex. Benzécri & collab., 1973)³. Dans l'exemple de la figure 1, l'idée était de voir dans quelle mesure il y a une interaction entre les variables suivantes : le genre d'événement au sein duquel le marqueur a été annoté (s'il s'agit d'un débat public, d'un débat télévisé, ou d'une réunion professionnelle), le type morpho-syntaxique du marqueur (par ex. un

2 Cette stratégie ne doit pas être confondue avec un Kappa de Cohen dit « pondéré » [*weighted*] (Cohen, 1968), qui consiste à travailler le Kappa de manière à nuancer le poids de certaines différences interpersonnelles dans l'annotation.

3 Je remercie ainsi Guillaume Guex, collaborateur de François Bavaud, de nous avoir présenté cette méthode et de nous avoir aidés à l'appréhender sur nos données.

adverbe, ou un introducteur de complétive), la direction entre marqueur et portée (est-ce que le marqueur intervient avant ou après la portée), la position du marqueur au sein de l'unité de construction du tour (au début, au milieu, à la fin) et enfin la dimension d'épistémicité en jeu (modalité épistémique d'une part, évidentialité de l'autre). Sans entrer

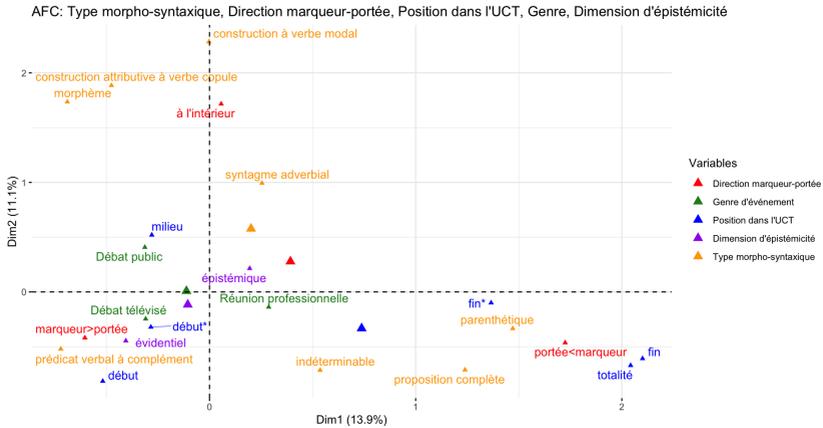


FIGURE 1 – ACM mobilisant cinq variables : le genre d'événement, le type morpho-syntactique, la direction entre marqueur et portée, la position du marqueur au sein de l'unité de construction du tour et la dimension d'épistémicité en jeu. Graphique réalisé avec R (FactoMineR, factoextra et plotly).

dans le détail, cette ACM permet d'identifier des tendances générales, comme le fait que les positions respectives des marqueurs en fonction des portées (en rouge) sont liées aux types de construction morpho-syntactiques mobilisées (en jaune) et à la position du marqueur au sein de l'unité de construction du tour de parole (en bleu). Ces tendances donnent lieu à au moins trois groupes (les trois angles du triangle), qui peuvent dans un second temps être analysés en tant que collections de cas à étudier dans le détail de leurs particularités énonciatives, interactionnelles ou multimodales. Cette manière de procéder répond ainsi à une interrogation fréquente au sujet de la manière dont les méthodes dites quantitatives, statistiques, fréquentielles, peuvent s'articuler aux analyses qualitatives, contextuelles, détaillées de (collections de) cas, et donc à la manière dont l'une peut renseigner l'autre.

4 En guise de conclusion : éloge de l'interdisciplinarité focalisée

La présente contribution souhaitait rendre hommage à l'interdisciplinarité focalisée permise et favorisée par la proximité très concrète de spécialistes de domaines différents. Dans le cas présent, pouvoir bénéficier de la présence, de la disponibilité et des compétences de spécialistes en méthodes statistiques et computationnelles a permis et continue de permettre des développements disciplinaires intéressants au sein de la pragmatique, développements qui rejoignent depuis quelques années un intérêt pour la linguistique de corpus et les méthodes mixtes. À plus long terme, on peut imaginer que le perfectionnement continu des logiciels de transcription (semi-)automatisée de paroles spontanées en interaction autorise le défrichage de masses plus importantes de données et donc favorise une plus grande représentativité des données étudiées. L'application d'analyses statistiques non supervisées plus poussées sur des données plus massives devrait aussi permettre de faire émerger des hypothèses plus originales voire imprévues, hypothèses dont les analyses de détails pourraient dans un second temps tester la validité ou simplement la pertinence. Parmi les sous-domaines qui pourraient bénéficier de l'accroissement des données soumises à l'étude et d'une application plus fréquente de méthodes non supervisées, la multimodalité du langage constitue un exemple qui vient rapidement à l'esprit.

Références

- Adolphs, S. & Carter, R. (2013). *Spoken corpus linguistics: From monomodal to multimodal*. Routledge, London.
- Aijmer, K. & Rühlemann, C. (éd.) (2014). *Corpus pragmatics: A handbook*. Cambridge University Press, Cambridge.
- Angouri, J. (2010). Quantitative, qualitative or both? Combining methods in linguistic research. In Litosseliti, L. (éd.), *Research methods in linguistics*, pages 29–49. Continuum, London.
- Anscombre, J.-C. & Ducrot, O. (1976). L'argumentation dans la langue. *Langages*, 10(42):5–27.

- Arminen, I. (2005). *Institutional interaction: Studies of talk at work*. Aldershot, Ashgate.
- Avanzi, M., Béguelin, M.-J., & Diémoz, F. (éd.) (2016). *Corpus de français parlé et français parlé des corpus / Corpus N°15*. Bases, Corpus, Langage - UMR 6039.
- Baldauf-Quilliatre, H., de Carvajal, I. C., Etienne, C., Jouin-Chardon, E., Teston-Bonnard, S., & Traverso, V. (2016). Clapi, une base de données multimodale pour la parole en interaction : Apports et dilemmes. *Corpus*, (15).
- Bar-Hillel, Y. (1954). Indexical expressions. *Mind*, 63(251):359–379.
- Bateman, J., Wildfeuer, J., & Hiippala, T. (2017). *Multimodality, foundations, research and analysis – A problem-oriented Introduction*. De Gruyter Mouton, Berlin, Boston.
- Bavaud, F. (2021). Similarité entre intervalles. [Manuscrit non publié]. Université de Lausanne.
- Benzécri, J.-P. & collab. (1973). *L'analyse des données. 2 L'analyse des correspondances*. Dunod, Paris.
- Berger, E. & Lauzon, V. (2017). *Pratiques interactionnelles en contexte institutionnel / TRANEL (67)*. Université de Neuchâtel, Neuchâtel.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: investigating language structure and use*. Cambridge University Press, Cambridge.
- Bublitz, W., Jucker, A. H., & Schneider, K. P. (éd.) (2018). *Methods in pragmatics*. De Gruyter Mouton, Berlin Boston.
- Bublitz, W. & Norrick, N. R. (éd.) (2011). *Foundations of pragmatics*, volume 1. Mouton De Gruyter, Berlin.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.
- Couper-Kuhlen, E. & Selting, M. (2018). *Interactional linguistics: Studying language in social interaction*. Cambridge University Press, Cambridge.
- Drew, P. & Heritage, J. (éd.) (1992). *Talk at work: Interaction in institutional settings*. Cambridge University Press, Cambridge.
- Duranti, A. & Goodwin, C. (éd.) (1992). *Rethinking context: Language as an interactive phenomenon*. Cambridge University Press, Cambridge.
- Egbert, J., Larsson, T., & Biber, D. (2020). *Doing linguistics with a corpus*. Cambridge University Press, Cambridge.

- Encrevé, P. (1977). Présentation : Linguistique et socio-linguistique. *Langue française*, 34(1):3–16.
- Ghia, E., Kloppenburg, L., Nissim, M., Pietrandrea, P., & Cervoni, V. (2016). A construction-centered approach to the annotation of modality. In *Proceedings of the 12th ISO Workshop on Interoperable Semantic Annotation*, volume 29. Harry Bunt, Portoroz.
- Gries, S. T. (2010). Methodological skills in corpus linguistics: A polemic and some pointers towards quantitative methods. In Harris, T. & Moreno Jaén, M. (éd.), *Corpus linguistics in language teaching*, pages 121–146. Peter Lang, Frankfurt.
- Gries, S. T. (2017). *Quantitative corpus linguistics with R: A practical introduction*. Routledge, Taylor & Francis Group, New York, 2ème édition.
- Huang, Y. (éd.) (2017). *The Oxford handbook of pragmatics*. Oxford University Press, Oxford; New York, NY.
- Jacquín, J., Keck, A. C., Robin, C., & Roh, S. (2022). Guide d'annotation du projet posepi. Technical Report 1, Université de Lausanne et Fonds National Suisse, Lausanne.
- Jucker, A. H. (2018). Introduction to part 5: Corpus pragmatics. In Jucker, A. H., Schneider, K. P., & Bublitz, W. (éd.), *Methods in pragmatics*, pages 455–466. De Gruyter, Berlin.
- Keck, A. (en préparation). *Le degré de certitude en français. Étude systématique de la modalité épistémique dans un corpus d'interactions sociales*. Thèse de doctorat, Université de Lausanne. Thèse non publiée.
- Kress, G. & Van Leeuwen, T. (2001). *Multimodal discourse*. Arnold, London.
- Labov, W. (1972). *Sociolinguistic patterns*. University of Pennsylvania Press, Philadelphia.
- Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press, Cambridge.
- Mayaffre, D. (2005). Rôle et place des corpus en linguistique : Réflexions introductives. Consulté à l'adresse http://www.revue-texto.net/Corpus/Publications/Mayaffre_Corpus.html.
- Mellet, S. (éd.) (2008). *Concession et dialogisme : Les connecteurs concessifs à l'épreuve des corpus*. Peter Lang, Berne.
- Nissim, M. & Pietrandrea, P. (2017). Modal: A multilingual corpus annotated for modality. In Basili, R., Nissim, M., & Satta, G. (éd.), *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017: 11-12 December 2017, Rome*, pages 234–239, Torino. Accademia University Press.

- Pietrandrea, P. (2018). Epistemic constructions at work. A corpus study on spoken Italian dialogues. *Journal of Pragmatics*, 128:171–191.
- Rastier, F. (2004). Enjeux épistémologiques de la linguistique de corpus. Consulté à l'adresse <http://www.revue-texto.net/index.php?id=543>.
- Riazi, A. M. (2016). Innovative mixed-methods research: Moving beyond design technicalities to epistemological and methodological realizations. *Applied Linguistics*, 37(1):33–49.
- Robin, C. (2024). *Marquer la source de l'information : Approches interactionnelle, énonciative et multimodale de l'évidentialité en français*. Thèse de doctorat, Université de Lausanne.
- Romero-Trillo, J. (2008a). Introduction: Pragmatics and corpus linguistics – a mutualistic entente. In Romero-Trillo, J. (éd.), *Pragmatics and corpus linguistics. A mutualistic entente*, pages 1–10. De Gruyter Mouton, Berlin, Boston.
- Romero-Trillo, J. (éd.) (2008b). *Pragmatics and corpus linguistics. A mutualistic entente*. De Gruyter Mouton, Berlin, Boston.
- Romero-Trillo, J. (éd.) (2013). *Yearbook of corpus linguistics and pragmatics 2013: New domains and methodologies*. Springer, Dordrecht.
- Rühlemann, C. (2019). *Corpus linguistics for pragmatics*. Routledge, New York.
- Rühlemann, C. & Aijmer, K. (2014). Corpus pragmatics: Laying the foundations. In Aijmer, K. & Rühlemann, C. (éd.), *Corpus pragmatics: A handbook*, pages 1–26. Cambridge University Press, Cambridge.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.
- Selting, M. & Couper-Kuhlen, E. (éd.) (2001). *Studies in interactional linguistics*. John Benjamins, Amsterdam.
- Sidnell, J. & Stivers, T. (éd.) (2005). *Multimodal interaction / Semiotica 156*. Walter de Gruyter, Berlin.
- Sidnell, J. & Stivers, T. (éd.) (2013). *The handbook of conversation analysis*. Wiley-Blackwell, Chichester.
- Svartvik, J. (éd.) (1992). *Directions in corpus linguistics*. De Gruyter Mouton, Berlin.
- Teddle, C. & Tashakkori, A. (2009). *Foundations of mixed methods research. Integrating quantitative and qualitative approaches in the social and behavioral sciences*. SAGE, London.
- Weisser, M. (2018). *How to do corpus pragmatics on pragmatically annotated data: Speech acts and beyond*. John Benjamins, Amsterdam.