# LINGUISTIQUE TEXTUELLE, LINGUISTIQUE DE CORPUS

Guy ACHARD-BAYLE
Université de Lorraine, CREM-Praxitexte

Ondřej PEŠEK
Université de Bohême du Sud, Institut d'études romanes

**DOI :** 10.32725/eer.2022.001

Notre sommaire fait suite à trois journées d'étude organisées à Metz puis à Prague au printemps et à l'automne 2018, enfin à České Budějovice en décembre 2020. La thématique ou la problématique de ces journées portait simplement pour titre : « Linguistique textuelle, linguistique de corpus ».

Justifions pour commencer cet adverbe « simplement » : si ce titre peut paraître simple, voir un peu court, la complexité que nous voulons exposer et traiter réside dans le rôle ou le contenu sémantique complexe – ambigu si l'on veut – assumé par la virgule entre les deux types de linguistiques ici réunies.

En fait, nous sommes partis d'un constat, naïf ou ordinaire au sens où il concerne la communication à l'intérieur de la communauté scientifique, qui voit dans nombre d'appels à colloques, d'annonces de parutions, les deux linguistiques confondues ou assimilées sous l'étiquette : « Text/Corpus Linguistics ».

Or l'assimilation a de quoi, sinon étonner, et c'est en effet le cas si l'on prend pour appui la perspective de la linguistique de texte ou textuelle, du moins interroger, et c'est le cas des spécialistes de l'épistémologie et de la méthodologie de la linguistique de corpus, car, ne serait-ce que par le canal, on ne voit pas en quoi *corpus* = *texte*, ce qui d'une part laisse de côté l'oral ou les corpus oraux, d'autre part, et par conséquent, laisse entendre ou sous-entendre que par ailleurs *texte* = *écrit*[1]. Or telle n'est pas notre conception, nous y reviendrons dans la présentation des contributions ; mais disons dès lors que, si nous entendons *texte* autrement, c'est bien pour inscrire dans le cadre de nos analyses, et au sommaire de ce numéro de revue, le discours…[2]

Partant de tels constats, parfois « triviaux » dans le quotidien de notre discipline, notre intention est, plus que de mettre au jour, d'interroger cette assimilation. Pour ce faire, notre propos est de mettre en contact des spécialistes et de confronter des recherches de laboratoires qui se dédient au traitement linguistique *et* informatique des textes – des *textes* entendus alors comme bases de données textuelles, ou transcrites.

---

[1] Voir les appels aux CMLF (Congrès Mondiaux de Linguistique Française) qui ont une catégorie de communications, et ensuite de publications, qui porte pour titre, comme en 2020 et 2022 : Linguistique de l'écrit, linguistique du texte, sémiotique, stylistique.

[2] Et si, par exemple, ZIKÁNOVÁ *et al*., tout comme RYSOVÁ *et al*., envisagent bien le texte comme productions écrites, c'est pour y distinguer et analyser des marqueurs de cohérence discursive ; de manière générale toutes les contributions se conforment à cette distinction.

Nous soulignons maintenant la coordination *et*. Par là nous voulons signifier que le traitement des données textuelles se fait dans ces laboratoires suivant des phases ou à des niveaux distincts. Pour autant toutes les méthodologies, les épistémologies ne se confondent pas : il s'agit donc bien d'échanger principes, ou cadres théoriques, et pratiques, ou modes de traitement.

Partant de quoi nous pourrions maintenant reformuler notre titre en : « Linguistique de texte *et* linguistique de corpus », en choisissant de substituer la coordination à la virgule précédente. Autrement dit, le titre que nous proposons et que nous proposons d'illustrer n'équivaut pas à : « Linguistique de texte *et/ou* linguistique de corpus », ou moins encore « Linguistique de texte *vs* linguistique de corpus » ; ainsi, il sera question de se pencher et de s'interroger, non plus sur l'assimilation, mais sur la complémentarité entre linguistique de texte *et* linguistique de corpus. Pour cela, nous avons réuni un choix de contributions qui met en avant des modèles de traitements « naturels », c.-à-d. sémantiques, et formels, ou « assistés », qui montrent, sans les confondre, répétons-le, la complémentarité de ces traitements, des points de vue épistémologique et méthodologique.

*

Comme on peut le voir, le sommaire soumis est franco-tchèque, ou tchéco-français. Cela s'explique de deux manières. Tout d'abord, les coordinateurs du numéro appartiennent à deux universités, de Bohême du Sud et de Lorraine, dont les recherches – et les formations – en langue française et sciences du langage, convergent, particulièrement dans le domaine de la linguistique textuelle. Or, et c'est la seconde explication, cette coopération n'est pas le fruit du hasard : les recherches en sciences du langage des deux universités s'ancrent dans une tradition de linguistique textuelle qui se réclame de l'École de Prague (bientôt centenaire). Ainsi, les universités de Metz et de Nancy 2 (aujourd'hui réunies dans l'Université de Lorraine) ont joué un rôle majeur dans la diffusion des travaux pragois en France – et au-delà à l'ouest de l'Europe – dans les années 1970.

Ce rappel « géo-historique » permet de dire une autre fois que, pour nous, linguistique textuelle et linguistique de corpus ne se confondent pas… Mais cela dit, il nous faut également ajouter que, dans le panorama des épistémologies actuelles, la linguistique de texte, telle qu'elle a pu être définie par les Pragois (notamment, depuis la seconde moitié du XX$^{e}$ siècle : voir PEŠEK 2010), et reprise par ailleurs (voir ACHARD-BAYLE 2010, 2013), s'est ouverte à la linguistique du ou des discours. Or, si les deux linguistiques, textuelle et discursive, peuvent sans doute être mises en rapport de continuité, il ne saurait être question de les assimiler, tout comme nous ne le faisons entre texte et corpus. Notre sommaire vise donc cet autre objectif, d'illustrer et interroger, à la lumière de la tendance actuelle de traiter de larges corpus de données, cette continuité (voir ACHARD-BAYLE & PEŠEK 2020, 2022), qui nous semble correspondre à une des caractéristiques majeures de l'évolution des linguistiques post-structurales en Europe après, disons, la Seconde Guerre mondiale (voir ABLALI *et al.* 2018).

Enfin dès lors que nous prenons acte de l'importance prise par les outils numériques dans le traitement des données, notre objectif est de montrer qu'en

termes d'analyse et d'interprétation, le recours aux corpus textuels et discursifs permet d'enrichir ce traitement de manière notable. Qu'il s'agisse de grands ensembles textuels ou au contraire de corpus restreints et constitués *ad hoc*, les outils numériques représentent à l'heure actuelle un élément méthodologique indispensable pour la recherche (voir PEŠEK, 2020). Les ordinateurs font des calculs complexes et précis ; ils relèvent les occurrences des phénomènes analysés d'une manière rapide et exhaustive ; et grâce aux fonctionnalités avancées des logiciels actuels, ils permettent de visualiser les données textuelles dans une qualité inégalée (graphiques, images, animations 3D).

Dans les domaines morphologique, lexical ou syntaxique, un grand nombre d'applications ont été développées en vue de l'étiquetage computationnel des phénomènes observables. Or, s'agissant de la structuration textuelle, notamment aux niveaux méso- et macro-, l'usage d'outils numériques est bien moins fréquent, faute d'outils adéquats qui conviennent aux approches théoriques contemporaines. Ainsi, l'un des défis majeurs de la linguistique textuelle actuelle consiste dans le développement des systèmes d'annotation numériques (manuels ou automatiques), qui fournissent des données d'entrée, indispensables pour les calculs et les visualisations ultérieurs. La linguistique textuelle, qui saisit au niveau théorique la structure, la production et l'interprétation des textes, se joint à l'informatique pour contribuer, d'une manière indirecte mais non négligeable, au développement de l'intelligence artificielle.

**Présentation des contributions**

Les sept contributions, qui concourent à la double problématique générale (la singularité de la linguistique textuelle ; ou la complémentarité des traitements textuel-discursif d'une part, informatique de l'autre, en termes épistémologiques et méthodologiques) peuvent être présentées et rassemblées comme suit :

BACH, MAAZAOUI & GAUTIER s'interrogent d'emblée sur la pertinence d'une approche exclusivement quantitative et automatisée, préférant, disent-ils, « garder la main » sur la constitution de corpus notamment s'il s'agit, comme dans leur cas, d'une analyse socio-discursive, c'est-à-dire de productions langagières situées et resituées dans leur environnement culturel, voire professionnel lorsqu'il faut traiter d'une langue de spécialité.

Dans leur contribution respective, ACHARD-BAYLE et LANDRAGIN s'intéressent à la construction du sens et au maintien de la cohérence via la constitution des *chaînes de référence* : autrement dit, ils se penchent sur la constitution de suites d'expressions coréférentielles exo- puis endophoriques (nominales ou pronominales s'il s'agit de leurs substituts anaphoriques), qui permettent à la fois d'accompagner un même référent au fil du texte et, dans certains cas (les *contextes évolutifs*), de voir son identité se reconstruire ou se modifier. Les deux contributions se distinguent néanmoins par la manière dont est choisi ou réuni le corpus, notamment lorsqu'il vise l'analyse linguistique du texte littéraire.

RYSOVÁ, RYSOVÁ & HAJIČOVÁ traitent d'outils de cohérence discursive : les connecteurs. Elles le font en donnant une dimension didactique à leurs analyses, qui portent sur un corpus d'écrits d'étudiants de tchèque langue étrangère, qui d'un

point de vue textuel correspondent aux grands *genres* discursifs : argumentation, narration, information, description. Ce faisant, l'analyse des résultats permet de donner tout son sens à la notion de *texte* : une production verbale qui tend à être présentée – et donc peut s'analyser et être représentée – comme un tout cohérent.

PEŠKOVÁ exploite également un corpus d'apprenants. Concrètement, elle analyse des textes de traduction rédigés par des étudiants tchèques en traductologie espagnole. Se focalisant sur les phénomènes de coréférence endophorique, elle teste les fonctionnalités de deux outils numériques (Analec et Sketch Engine) dans le domaine de la visualisation et de l'annotation des données de corpus. Elle souligne l'importance des compétences textuelles dans le processus de traduction et, par conséquent, dans la formation de futurs traducteurs.

ZIKÁNOVÁ, MÍROVSKÝ & POLÁKOVÁ s'intéressent à la formalisation des relations de discours, et le font via l'outil informatique qui est propre à leur équipe de l'Institut de linguistique formelle et appliqué à l'Université Charles de Prague (Prague Dependency Treebank). Leur contribution porte sur les phénomènes relatifs à la structuration globale du texte. L'étiquetage de ces phénomènes a été effectué en tant que produit secondaire de l'annotation des relations discursives implicites, les auteurs ne manquent pas de souligner l'importance cruciale des genres textuels.

Pour finir, PEŠEK rassemble des cadres d'analyse d'orientation plutôt textuelle ou plutôt discursive. De même que ZIKÁNOVÁ, MÍROVSKÝ & POLÁKOVÁ, il s'intéresse à la structure globale du texte. En appliquant les perspectives compositionnelle et actionnelle, il tente de délimiter des segments mésotextuels, qu'une annotation computationnelle visant à traiter numériquement la structure textuelle globale devrait prendre en compte.

## BIBLIOGRAPHIE

ABLALI Driss, ACHARD-BAYLE Guy, REBOUL-TOURÉ Sandrine, TEMMAR Malika (2018), (Re-)Penser le texte et le discours dans le paysage actuel des sciences du langage, in : ABLALI Driss, ACHARD-BAYLE Guy, REBOUL-TOURÉ Sandrine, TEMMAR Malika (éds), *Texte et discours en confrontation dans l'espace européen*, Berne, P. Lang, p. 9-32. Présentation sur HAL : https://hal.univ-lorraine.fr/hal-01872096.

ACHARD-BAYLE Guy (2010), Du Cercle linguistique de Prague à une linguistique de texte « à la française », *L'Analisi linguistica e letteraria*, 2/2010, p. 431-436. En ligne : https://www.analisilinguisticaeletteraria.eu/wp-content/uploads/2015/02/201002BayleC.pdf

ACHARD-BAYLE Guy (2013), Perspective fonctionnelle de la phrase : histoire-géographie d'une idée linguistique – Prague & l'Europe, in : ACHARD-BAYLE Guy, CHABROLLE-CERRETINI Anne-Marie (éds), Perspective fonctionnelle de la phrase : du Cercle linguistique de Prague à la linguistique textuelle, *Verbum*, XXXV 1-2, p. 3-10.

ACHARD-BAYLE Guy, PEŠEK Ondřej (2020), Linguistique et texte. Contribution franco-tchèque à l'histoire et au rayonnement de l'École de Prague, in : *Expérience et avenir du structuralisme* (préparé par Tomáš Hoskovec), *Travaux*

*du Cercle linguistique de Prague*, nouvelle série, volume 8, Kanina : OPS ; Praha : PLK, p. 227-247.

ACHARD-BAYLE Guy, PEŠEK Ondřej (2022), Le paragraphe et l'organisation thématico-graphique du texte dans les nouveaux écrits numériques, in : MAGRI Véronique (éd.), *Le français moderne*, 90-1-2022, *Nouvelles textualités ?*, p. 75-106.

PEŠEK Ondřej (2010) La linguistique textuelle tchèque au seuil du XXI$^{e}$ siècle : la genèse d'une discipline et la tradition pragoise, *Verbum* XXXII-2, p. 263-282.

PEŠEK Ondřej (2019), Korpusy a významové struktury: komputační sonda do textů moderního českého jazykovědného myšlení [Corpus et structures sémantiques : analyse computationnelle de textes de la pensée linguistique tchèque moderne], in : PAPOUŠEK Vladimír a kol., *Pokušení neviditelného – Myšlení moderny*. Praha: Akropolis, p. 144-181.