Présentation

Peter Blumenthal, Denis Vigier

Peter Blumenthal
Université de Cologne

Denis Vigier
Université de Lyon / UMR ICAR (CNRS, Université Lyon 2, ENS de Lyon)

Présentation

1. POURQUOI ENCORE ET TOUJOURS LES PRÉPOSITIONS ?

Le titre du présent numéro s'inspire des travaux du programme de recherche franco-allemand PRESTO (*Évolution du système prépositionnel du français*, 2013-2017)[1], auquel a collaboré la plupart des contributeurs[2]. Les concepts utilisés dans ce titre, trop génériques pour orienter l'attente du lecteur, nécessitent quelques commentaires préalables pour faire apparaître les objectifs sous-jacents au projet, sa stratégie et ses présupposés méthodologiques. De quels types de *quantité* s'agit-il, comment comprendre la notion de *qualité*, bien plus vague, et par quelles étapes passer de l'une des deux catégories à l'autre ? Avant de cerner le fond de la thématique et les méthodes à appliquer se pose une question plus basique : pourquoi se pencher une nouvelle fois sur les prépositions et leur histoire, alors que le thème a récemment été traité à plusieurs reprises par diverses revues, dont *Langages* et *Langue française*[3] ? Quels phénomènes ou points de vue théoriques inédits justifient une nouvelle initiative dans ce domaine de la recherche diachronique, dont on ne saurait affirmer qu'il soit resté terrain inconnu ou peu exploré ?

À nos yeux, la réponse peut se résumer en un constat préalable : les linguistes ont depuis peu accès à un dispositif radicalement nouveau qui décuple le champ de leurs investigations, en ce qu'il réunit trois ingrédients décisifs : de vastes corpus historiques numérisés de la langue française, des outils d'annotation linguistique pour des états anciens de la langue, des plateformes d'exploration et de calculs statistiques. Ce triptyque rend possible l'application de méthodes d'analyses quantitatives robustes et éprouvées à de très grandes quantités de données langagières annotées, créant du même coup – si l'on en croit S. Girault & B. Victorri (2009 : 150) – les conditions d'émergence d'un véritable « observatoire de la langue »[4]. Une telle situation, si elle constitue un tournant proprement contemporain, n'en demeure pas moins l'héritière d'une longue série d'évolutions, de changements voire de bouleversements qui ont affecté les sciences du langage depuis plus d'un demi-siècle (Cori, David & Léon 2008) et qui ont globalement contribué à construire l'état actuel de ce que J. Léon (2010, 2015) nomme l' « automatisation-mathématisation » de la linguistique. Attardons-nous un instant sur les trois volets de ce triptyque. D'abord, si l'accroissement continu du nombre et de la taille des archives, bases et corpus numérisés mis à la disposition des linguistes[5] est un processus à l'œuvre depuis plusieurs décennies et qui a sensiblement modifié la pratique des Sciences du langage[6], en revanche l'accès à des bases de données textuelles (en particulier libres d'accès et téléchargeables) pour les états anciens de la langue est un phénomène récent. Il

[1] Financé par l'Agence Nationale de la Recherche et la Deutsche Forschungsgemeinschaft, ce programme a pour but l'étude diachronique de l'emploi, des valeurs sémantiques et discursives des prépositions françaises *à, en, par, contre, dès, devant, entre, pour, sans, sur, sous, vers, dans*, de l'ancienne langue jusqu'au français contemporain.
[2] À l'exception de Rossari / Ricci, qui appliquent la même méthode à l'italien.
[3] Cf. *Langages* 173, 2009 (*Approches récentes de la préposition*) et *Langue française* 157, 2008 (*Énigmatiques prépositions*).
[4] L'outillage de l'expérimentation s'avérant en grande partie indépendant des propositions testées par l'expérimentateur. Les auteurs discutent ici J-C. Milner (1989 : 127-128)
[5] A titre d'exemple, on comparera, pour l'anglais américain, le million de mots du *Brown Corpus* (1964) avec les 530 millions de mots du *Corpus Of Contemporary American english* (COCA) de 2008. Pour la France, l'ÉquipEx *Ortolang* (*Outils et Ressources pour un Traitement Optimisé de la Langue*) est à sa manière le témoin de la place croissante (qualitativement et quantitativement) prise par les corpus en Sciences du langage.
[6] En accroissant significativement la place occupée par les linguistiques de corpus, qu'il s'agisse de leur pratique, mais aussi des questions d'ordre théorique et épistémologique qu'elles posent aux Sciences du langage, par exemple au regard de la place de l'usage.

existe depuis quelques années des bases mettant à disposition des textes intégraux d'époque médiévale (Base du Français Médiéval-BFM[7], Nouveau Corpus d'Amsterdam-NCA[8], …) ou plus tardive[9], mais il reste encore à constituer un corpus longitudinal libre de droits couvrant toutes les périodes du français. Comme on espère le montrer (voir Diwersy, Falaise, Lay & Souvay dans ce numéro), le programme PRESTO a permis – grâce à la coopération de plusieurs de ces bases[10] – d'accomplir un pas significatif dans cette direction puisqu'il a permis la constitution d'un corpus longitudinal contrôlé[11] de textes étiquetés morphosyntaxiquement et lemmatisés, couvrant la période 1509-1944, et dont une partie sera mise à la disposition de la communauté courant 2017 sous licence libre. Mais un tel corpus (c'est le second volet), pour vaste qu'il soit, constituerait un dispositif incomplet pour observer les états anciens de la langue si l'on ne disposait pas d'outils d'annotation suffisamment performants à même d'en accomplir un étiquetage morphosyntaxique, une lemmatisation et une annotation syntaxique de qualité. Même si, là encore, le chemin à parcourir demeure important, PRESTO a concouru, grâce à ses collaborations externes mais aussi à ses ressources en ingénierie propre, à faire progresser les choses. Enfin, on voit se développer en France et dans le monde un nombre croissant de plateformes d'exploration et de calcul[12] qui mettent à disposition des linguistes dans un environnement de plus en plus convivial, des fonctions documentaires et statistiques ainsi que des outils de visualisation qui leur permettent de manipuler des données textuelles enrichies afin de poursuivre les objectifs scientifiques qu'il se sont fixés. Dans les lignes qui suivent, nous allons revenir – de manière ciblée et forcément parcellaire – sur certaines facettes des trois thèmes évoqués ci-dessus (corpus, outils et méthodes statistiques au service de l'étude diachronique du français) qui se situent au carrefour de nos propres travaux. Contentons-nous d'insister sur le fait que la linguistique historique se trouve selon nous au seuil d'une nouvelle ère, caractérisée par une énorme expansion de nos connaissances de faits, mais aussi par la nécessité d'une réflexion d'ordre épistémologique sur les nouveaux défis. Parmi eux, l'un des principaux étant peut-être d'éviter la tentation, très sensible dans la recherche linguistique contemporaine, d'un nouveau positivisme qui se contenterait d'accumuler des faits fraîchement établis pour les couler dans de magnifiques diagrammes pleins d'astuces statistiques. Pour éviter tout malentendu, soulignons enfin que nous tenons le recours aux statistiques et, d'une manière générale, à l'analyse quantitative, pour une étape essentielle de la recherche diachronique, mais dont le but final ne saurait être qu'une interprétation qualitative et linguistiquement pertinente des faits, prenant racine dans un humus théorique. L'analyse quantitative n'est qu'une étape ancillaire de toute étude linguistique diachronique ; nul sens ne peut émerger de « comptages » aussi sophistiqués fussent-ils, mirage d'autant plus séducteur que, comme l'observent S. Giraut & B. Victorri (*ibid.*), les outils de visualisation élaborés dont on dispose renforcent parfois le « pouvoir un peu magique » qu'on peut prêter à ces techniques descriptives : « Le sens semble « émerger » de la figure alors que seules nos capacités illimitées d'interprétation sont en cause (le sens est dans nos têtes) » (151). L'exploration statistique des données linguistiques doit être appréhendée comme une démarche heuristique pouvant contribuer à l'épreuve d'hypothèses sur les discours et les textes forgées indépendamment des outils, ou à la mise au jour de nouvelles pistes possibles pour le raisonnement linguistique, le retour au texte constituant toujours le plus sûr des garde-fous aptes à protéger l'analyste des constructions abusivement généralisantes, déconnectées des textes et du corpus. Comme le rappelait M. Tournier, « il n'y a de statistiques que des textes et des corpus, et non de la langue » (Brunet, 2011 : 93). Or, si nous n'avons pas la berlue, il advient qu'en ce moment de l'histoire de notre

[7] http://bfm.ens-lyon.fr/sommaire.php3
[8] http://www.uni-stuttgart.de/lingrom/stein/corpus#nca
[9] Voir note suivante.
[10] Le corpus de référence du projet franco-allemand PRESTO (http://presto.ens-lyon.fr) pour la période XVI^{e} s. – XX^{e} s. a été constitué grâce aux textes issus des bases textuelles suivantes : FRANTEXT (http://www.frantext.fr, V. Montémont, G. Souvay), BVH (*Bibliothèques Virtuelles Humanistes*, http://www.bvh.univ-tours.fr - L. Bertrand, M.-L. Demonet), ARTFL (*American and French Research on the Treasury of the French Language*, http://artfl-project.uchicago.edu – R. Morrissey, M. Olsen) et CEPM (*Corpus électronique de la première modernité*, http://www.cpem.paris-sorbonne.fr). Les ressources et les outils élaborés dans PRESTO ont bénéficié des apports des logiciels LGerM (*lemmatisation de la variation graphique des états anciens du français et lexiques morphologiques*, G. Souvay http://www.atilf.fr/LGeRM) et Analog (M.-H. Lay) ainsi que du lexique Morphalou (http://www.cnrtl.fr/lexiques/morphalou)
[11] Ce contrôle s'étant opéré sur la continuité temporelle (pas de « trous » dans les décennies représentées), sur la variété des (sous-)genres discursifs représentés ainsi que celle des auteurs, sur la taille de chaque tranche décennale du corpus (recours à l'échantillonnage), enfin sur la qualité des éditions des exemplaires utilisés (premières éditions des oeuvres privilégiées ou, à défaut, celles respectant le plus l'orthographe d'époque.)
[12] Les plateformes utilisées dans le programme PRESTO sont BTLC-PrimeStat (développée par S. Diwersy (à l'Université de Cologne puis à l'Université de Montpellier 3) et TXM (http://textometrie.ens-lyon.fr) développée par S. Heiden et son équipe à l'ENS de Lyon. Pour une liste des autres outils réunissant fonctionnalités documentaires et statistiques, on peut se reporter en particulier à B. Pincemin *et al.* (2010) ainsi qu'à la page « *Exploration de corpus. Outils et pratiques* » du consortium « corpus écrits » (http://explorationdecorpus.corpusecrits.huma-num.fr).

discipline, un décalage s'opère entre les fantastiques progrès du traitement informatique et une difficulté pour la communauté des linguistes non spécialistes du TAL et/ou des méthodes quantitatives automatisées de se les approprier au bénéfice de leur pratique. Comme l'écrit C. Favre, « ces méthodes et leurs résultats sont encore mal connus de la communauté des linguistes, alors qu'on pourrait souhaiter qu'ils permettent d'étendre la gamme des outils à leur disposition pour explorer le sens des mots à partir de leurs usages dans les corpus » (2015 : 395).

De cet état des lieux, décrit ci-dessus à gros traits, résultent quelques éléments de réponse à la question sur le sens de la présente publication :

a) l'illustration, pour la majorité des articles regroupés dans ce numéro, d'une collaboration fructueuse (et tenant compte des intérêts des deux parties) entre linguistes et chercheurs spécialisés en informatique ;
b) une contribution à l'exemplification des tâches qui incombent à une nouvelle linguistique historique, laquelle essaie de prendre la mesure des démarches méthodologiques et du cadre théorique appropriés à la nouvelle donne en matière d'informatique.

L'élément a) semble suffisamment documenté par les appariements des auteurs responsables des articles ci-dessous, alors que b) ne saurait se passer d'un commentaire plus étoffé. En effet, les tâches en question couvrent toute une gamme de recherches allant, en l'occurrence, d'une approche d'inspiration empirique conduisant à un affinement précieux de nos connaissances sur des points précis, jusqu'à une discussion autour de la nécessaire mise à jour méthodologique ; celle-ci consistera, à nos yeux, dans le rattrapage des modèles linguistiques sur l'avance prise par l'ordinateur. Précisons pour finir que les recherches empiriques évoquées ci-dessus ne se bornent pas à décrire d'une manière un peu plus précise ce que l'on croyait savoir approximativement, mais fournissent parfois des preuves permettant de trancher, désormais en connaissance de cause, entre des solutions alternatives débattues autrefois ou naguère.

2. MÉTHODES : *CORPUS-BASED* ET/OU *CORPUS-DRIVEN*

Nous voudrions approfondir ici le problème méthodologique mentionné *supra* à l'aide des deux notions contraires indiquées dans le titre de la présente section. Dans les procédures *corpus-based*, le linguiste focalise son attention sur la solution d'une question spécifique et utilise la banque de données comme instrument. Dans la situation inverse, celle de corpus-driven, où d'instrument, le corpus assume un rôle plus actif, jusqu'à constituer un milieu tout particulier où se forment quasi spontanément les hypothèses linguistiques. Or, nous sommes de ceux qui considèrent l'hypothèse d'une opposition radicale entre les deux approches, *based* et *driven*, comme une conception certes légitime, mais ne correspondant pas forcément au travail réel du linguiste. Nous proposerons, comme alternative à ce modèle, l'idée d'une opposition de type scalaire, qui implique la possibilité de situations intermédiaires entre les deux pôles, où linguiste et corpus se trouvent en étroite interaction. Ainsi, nous imaginons un linguiste qui, dans son dialogue avec la banque de données, ne mise pas tout sur une stratégie conçue préalablement, mais lance aussi des requêtes au fil de son intuition et sait tirer profit de résultats inattendus. La qualité première de cet analyste : savoir s'étonner – capacité majeure qui a engendré, dit-on, la philosophie. Cela précisé, pour la majorité des contributions présentées ci-dessous, c'est l'approche *corpus-based* qui prédomine.

3. BASE QUANTITATIVE ET MÉTHODES STATISTIQUES

La comparabilité, caractéristique qualitative des corpus, légitime l'application de méthodes quantitatives, visant à calculer les différences caractéristiques entre corpus appartenant à diverses époques. *Corpus-based* ou *corpus-driven*, ces calculs peuvent, dans un premier temps, avoir un caractère rudimentaire, ne concerner que la fréquence, absolue ou relative, de certains phénomènes lexicologiques ou grammaticaux et avoir pour objectif principal d'orienter l'attention du chercheur vers les zones les plus mouvantes de la langue.

Dans un deuxième temps, il convient de faire la part des choses d'un point de vue statistique et de distinguer les différences de nature aléatoire par rapport aux différences significatives, l'existence de ces dernières pouvant susciter toute une gamme d'analyses et / ou d'interprétations. Sur ce plan, on dispose aujourd'hui d'un nombre substantiel de méthodes d'analyses et de calculs statistiques éprouvés hérités des travaux conduits en particulier

dans le cadre du « contextualisme britannique »[13] et de ses nombreuses ramifications contemporaines, ou de l'école française de lexicométrie, de textométrie et de statistique textuelle[14]. Dans le cadre de PRESTO, nous avons puisé chez les premiers (contextualisme britannique) notamment la conviction que le cotexte distributionnel d'une unité constitue la voie royale d'accès à son sens (on rappellera la formule de J. Firth (1957 : 11) « you shall know a word by the company it keeps », guère éloignée de la phrase de L. Wittgenstein (1953 : § 43) : « La signification d'un mot est son usage dans le langage »[15]). D'où notre attention toute particulière portée à l'étude statistique des collocatifs des prépositions, les changements majeurs intervenus dans les cotextes préférés de ces dernières au cours du temps constituant le signe d'autres changements plus souterrains survenus dans leur « invariant » sémantique. Aux seconds (statistique textuelle), nous devons en particulier notre intérêt pour une démarche statistique d'essence contrastive (calcul des spécificités de P. Lafon (1980, 1984) sur corpus partitionnés, recours à l'Analyse Factorielle des Correspondances (AFC), etc.) apte à repérer « les contrastes dans un tout qui fait système » (Pincemin, 2011). Nous adhérons en outre pleinement à la règle du retour systématique au texte comme seule clef de l'interprétation des résultats quantitatifs. Ajoutons pour finir que, outre ces deux sources, nous nous avons puisé dans d'autres travaux contemporains conduits en linguistique quantitative appliquée aux corpus diachroniques, dont notamment ceux de S. F. Fries & M. Hilpert relatifs à la périodisation automatique (2008, 2012).

Soucieux de ne pas en rester à un niveau trop élevé de généralité, nous proposons d'illustrer avec plus de précision certaines des questions que nous avons eu à traiter dans PRESTO, en recourant à un exemple précis tiré de nos activités de calcul : le calcul des cooccurrences. Quel que soit l'indice probabiliste mobilisé (pour ceux que nous avons utilisés : log de vraisemblance / log-likelihood (Dunning 1993) dans BTLC-PrimeStat ou spécificités de P. Lafon (1980, 1984) dans TXM), ce calcul permet pour chaque cooccurrent[16] d'un pivot situé dans un cotexte de cooccurrence paramétré par avance, de le classer selon la plus ou moins grande « attraction » ou « répulsion » qu'il manifeste vis-à-vis de ce pivot. On conçoit l'intérêt que revêt une telle information pour le linguiste qui veut explorer le cotexte cooccurrentiel d'une unité : elle lui permet de trier et de hiérarchiser les cooccurrents d'un pivot sur le critère statistique de ses préférences combinatoires. La question que nous proposons d'approfondir ici est la suivante : de quoi une préférence combinatoire entre un pivot donné et tel ou tel de ses cooccurrents peut-elle être le signe ? On touche ici au problème de l'interprétation des résultats.

4. VERS LE QUALITATIF

Ainsi que souligne D. Geeraerts (2010 : 177) : « Distributional corpus analysis is primarily a method, not a model. It opens an impressive amount of empirical data, but how exactly those data may be interpreted is not always given by the technique itself. » Les causes d'une attraction statistiquement significative calculée entre un collocatif[17] et un pivot donné sont le plus souvent multiples, et leur effet cumulé peut parfois expliquer le caractère très élevé de l'indice de préférence fourni par la machine. Les mesures statistiques « captent » en effet, par la mesure de la sur-utilisation statistique d'un mot au voisinage d'un autre dans un corpus, un ensemble de phénomènes qui peuvent relever de plusieurs domaines et niveaux d'analyse. Comme l'écrit S. Loiseau (2011 : 73) : « Les divergences systématiques identifiées par les tests statistiques mêlent toujours des choses qui doivent être distinguées dans l'interprétation des résultats. (…) l'interprétation des faits doit distinguer ce qui relève de phénomènes phraséologiques, de normes textuelles, de faits historiques, de différences de variété, etc. Autrement dit, l'utilisation de ce type de données relève d'une herméneutique. » Si l'on en croit l'auteur (cf. « les *faits historiques* »), l'interprétation des résultats statistiques peut donc même conduire l'expérimentateur à sortir des

[13] Pour une présentation générale du contextualisme britannique, voir J.-R. Firth (1957-1968), M. A. K. Halliday (1961), J. Léon (2008, 2015), J. McH Sinclair (1991), M. Stubbs (1993)

[14] Issue des recherches de P. Guiraud (1954, 1960) et de C. Muller (1973, 1977, 1992), la statistique textuelle s'est notamment développée dans le cadre des recherches menées au laboratoire « lexicométrie et textes politiques » de l'ENS Fontenay, pour ensuite essaimer en France et à l'étranger grâce notamment aux travaux des héritiers plus ou moins proches ou lointains de M. Tournier : P. Lafon, L. Lebart, B. Pincemin, A. Salem, … Pour un historique partiel de la textométrie/statistique textuelle, voir V. Baudouin (2016) et J. Léon & S. Loiseau (2016). Comme ouvrage de référence sur la statistique textuelle, voir L. Lebart & A. Salem (1994).

[15] On sait combien J.R. Firth a été influencé par Wittgenstein qu'il cite plusieurs fois.

[16] Ces cooccurrents peuvent être des « tokens » bien entendu, mais aussi des « types » (Pincemin, 2004). *Idem* pour le pivot qui peut être en outre un motif (succession d'unités) linéairement continu ou discontinu.

[17] Dans la terminologie que nous utilisons, le *cooccurrent* d'un pivot est une unité textuelle quelconque (représentée sous forme de token, lemme, …) figurant dans le cotexte gauche ou droit de ce pivot, sans considération d'ordre quantitatif. Le *collocatif* d'un pivot est un cooccurrent dont un calcul statistique a montré que sa distribution cotextuelle revêtait un caractère significatif (qu'il s'agisse d'une sur- ou d'une sous-représentation : voir D. Vigier dans ce volume).

limites du champ de la linguistique « *stricto sensu* » pour gagner les terres plus inconnues des modes, des conceptions dominantes d'une époque et d'une catégorie sociale déterminée dont on peut penser qu'elles ont influé le cours de la langue *via* les usages[18], etc. Nous en resterons cependant ici au champ de la linguistique *stricto sensu* pour proposer une ébauche de quelques « trajets » les plus souvent parcourus dans PRESTO et menant du quantitatif vers le qualitatif.

La mesure statistique d'une préférence combinatoire d'un mot pour un autre peut être le symptôme d'*un figement* plus ou moins avancé, ce phénomène multifactoriel étant intrinsèquement scalaire et nécessitant des retours minutieux au texte et l'usage de tests chaque fois que possible en vue de déterminer à quel point de figement se situe la séquence considérée. On se donne ainsi les moyens d'observer et de problématiser l'émergence, la réorganisation et le déclin de structures figées. Les contributions dans ce numéro de P. Blumenthal sur les indicateurs de perspectives d'une part, de M. Charolles, S. Diwersy et D. Vigier sur les marqueurs de topique d'autre part, donnent une illustration de ce type de phénomènes dans deux sous-corpus spécialisés de PRESTO : les corpus (diachroniques) des *Encyclopédies* et du *Figaro*.

Toute collocation (statistique) n'est cependant pas le signe d'un figement. « La haute spécificité d'une combinaison n'équivaut pas forcément à la présence de figement » (Blumenthal, 2011 : 293). Autrement dit, il y a des combinaisons « libres » significativement récurrentes. Ces combinatoires préférentielles non figées peuvent ouvrir des voies d'interprétation fécondes et parfois inattendues sur la manière dont les locuteurs, à une époque donnée, « conceptualisent » les référents des termes nominaux placés dans la dépendance d'une préposition. A cet égard, l'article de D. Leeman & A. Falaise dans ce numéro, qui traite des noms de régions et de départements dans le régime de *en* et de *dans*, constitue une première étape (non probabiliste) dans cette voie en faisant l'hypothèse que l'évolution qu'ils détectent dans leur corpus (à partir de l'étude de fréquences relatives) serait l'écho, dans les perceptions contemporaines des territoires, d'une conceptualisation nouvelle de leur identité.

La sur-spécificité d'un coocccurrent peut aussi être le symptôme d'une contrainte sélectionnelle pesant sur la syntagmatique des classes de mots dans l'état de langue considéré**.** On touche ici aussi bien aux collocations qu'aux 'colligations' telles que l'entendent J.-R. Firth et J. Sinclair[19]. Si l'on prend l'exemple de *en*, on observe que lorsqu'on accomplit un calcul de cooccurrences (indice des spécificités) sur le cotexte aval (1 mot) de cette préposition pour la période 1900-1944 dans le corpus PRESTO, l'un des *tokens* les plus spécifiques qui apparaît dans la liste est le participe présent *faisant*. Inversement, les mots dotés des indices de spécificité les plus significativement négatifs sont les formes de l'article défini *le, la, les, l'*. Une des informations que l'on peut extraire de ce classement, à savoir que *en* se combine préférentiellement avec des formes du participe présent (notamment) et qu'elle possède une aversion pour les déterminants définis, paraît triviale pour la connaissance du français contemporain. Cela devient nettement plus intéressant quand on adopte une perspective diachronique, et qu'on cherche par ex. à déterminer à partir de quelle époque *en* a cessé d'avoir pour collocatif préféré un déterminant. L'article de D. Vigier dans ce numéro montre quels avantages on peut retirer, pour l'étude de l'évolution de *en* et de *dans*, des informations que livre le calcul des spécificités sur les préférences manifestées dans le corpus par ces deux prépositions vis à vis des déterminant actualisant leur régime nominal.

Une collocation peut aussi s'expliquer pour partie par les influences qu'exercent sur le niveau syntagmatique les discours, les champs génériques, les genres et sous-genres discursifs dont relèvent les textes du corpus. Là encore, les études citées *supra* de P. Blumenthal d'une part, de M. Charolles, S. Diwersy et D. Vigier d'autre part, mettent en jeu ce régime d'explication (influence probable des genres discursifs que déploient respectivement le discours encyclopédique et le discours journalistique de la presse d'information nationale, et qu'il faudrait approfondir par des procédures d'analyse plus ciblées).

Qu'on nous permette de terminer ce rapide et incomplet coup de sonde dans le « feuilleté » des interprétations possibles que le linguiste doit patiemment élaborer à partir des résultats « bruts » que lui livrent les calculs probabilistes, par la notion de « plus-value sémantique » que les contributions réunies dans ce numéro

[18] Tel a été le thème du colloque « Changements linguistiques et phénomène sociétaux » organisé à l'École Normale Supérieure de Lyon en mars 2016 et dont les actes (D. Vigier & P. Blumenthal éds) sont à paraître chez P. Lang.
[19] Voir D. Geeraerts (2010 : 170).

ne mobilisent pas en tant que telle, mais qui nous est progressivement apparue, dans PRESTO, d'un intérêt majeur pour faire franchir aux études linguistiques en diachronie une nouvelle étape.

Par ce terme de « plus-value » sémantique, largement synonyme de la notion de « valeur ajoutée » chez U. Eco (2003 : 384), nous désignons la capacité d'une combinaison de mots à engendrer, au-delà de leurs sens référentiels, un avantage communicatif supplémentaire qui peut porter, entre autres, sur l'expressivité, les associations ou des facteurs d'ordre rhétorique. C'est cette capacité mise au service des besoins communicatifs nouveaux des locuteurs à un moment donné de l'histoire qui, selon nous, permet en certains cas d'expliquer l'origine des phénomènes statistiques discutés ici. Pour illustrer notre propos, rien de plus utile que la reprise d'un exemple déjà cité : *en Gers,* plutôt que *dans le Gers,* peut aujourd'hui permettre de connoter les qualités d'un département qui a su garder son authenticité rurale, puisque la préposition *en* s'avère capable de conférer au département un charme que l'on croyait réservé aux anciennes provinces. Dans ces conditions, il ne paraît pas étonnant que certaines voix, comme les syndicats d'initiative, soient tentées de tirer profit de l'impact publicitaire de la combinaison, voire d'en abuser. Une illustration plus « globale [20]» de cette notion de plus-value pourrait aussi se tirer des analyses conduites, dans deux articles de ce numéro, à partir du constat de l'extension récente de certaines expressions prépositionnelles discursives (*par ailleurs*, *sur le plan de, au niveau de*, *dans le domaine de*, *dans le secteur de*, etc.), qui attirent l'attention sur le thème (topique) du discours ou préparent le lecteur à son changement. Le fait que la fréquence de ces marqueurs dans la presse écrite augmente fortement entre 1900 et 2000 n'a pas manqué de nous intriguer. L'interprétation de cette évolution a fait surgir une discussion sur le statut des marqueurs en question : rendent-ils seulement explicites des contenus véhiculés implicitement par les textes de la tranche chronologique antérieure? Ou bien peut-on élucider, à travers eux, un bénéfice communicatif créé par une nouvelle manière d'organiser les textes ? Selon la réponse apportée, on abordera différemment la phase d'interprétation qualitative, susceptible de s'orienter vers la détermination d'une éventuelle plus-value. Si l'hypothèse d'une telle refonte des structures textuelles s'avérait fondée et si, plus généralement, l'étude diachronique de la combinatoire, figée ou non, pouvait s'établir comme un sous-domaine de l'histoire de la langue, la linguistique serait bien inspirée d'ouvrir un dialogue avec d'autres sciences humaines (sociologie des médias, médiologie, philosophie, histoire des mentalités, sciences cognitives, etc.) pour éclairer les conditions sociétales qui motivent certaines évolutions stylistiques et permettent de cerner les attributs de la plus-value. Car cette dernière ne saurait se définir qu'en fonction des critères, besoin et rêves d'une société donnée, un peu comme la vérité selon Pascal, confinée par les Pyrénées. Une telle réflexion interdisciplinaire devrait conduire à dépasser, d'une part, le clivage entre histoire interne et externe de la langue, et d'autre part, le niveau des explications anecdotiques des changements linguistiques. Voilà une excellente motivation pour les passionnés de l'histoire de la langue : l'espoir de parvenir enfin à ce qui mérite le nom d'« explication » en matière de diachronie semble en principe plus fondé que jamais, grâce aux progrès de l'informatique, – si seulement les linguistes réussissaient à se départir d'un isolement de moins en moins splendide.

5. EN RÉSUMÉ : QUE PEUT APPORTER PRESTO ?

En quoi ce recueil d'articles peut-il apporter du nouveau à notre compréhension de l'histoire de la langue, et plus précisément de la grammaire historique du système prépositionnel ? Le lecteur constatera un enrichissement de nos connaissances à trois niveaux – parfois au sein d'un seul article ; il s'agit en effet

- d'expliquer la configuration des corpus et des outils et calculs qui forment les préalables matériels et méthodologiques de nos recherches, mais virtuellement aussi de bien d'autres projets, car cet ensemble sera librement accessible après la fin de PRESTO, à moins que les droits d'auteur ou d'éditeur ne s'y opposent ;
- de mobiliser ces ressources pour rendre compte ─ de façon bien plus précise que par le passé ─ des époques, des dimensions et des fonctions des changements à l'étude ;
- de rediscuter, à la lumière de nos résultats quantitatifs, les hypothèses formulées antérieurement sur ces problèmes ;

[20] « Globale » parce que portant virtuellement sur un texte dans son ensemble. Nous songeons ici aux contributions de P. Blumenthal et de M. Charolles, S. Diwersy & D. Vigier.

- d'ouvrir de nouvelles perspectives à l'histoire du français, étant donné que nos outils informatiques mettent en relief, pour certaines époques, des évolutions dont le caractère systématique a parfois échappé à la recherche antérieure.

6. PRÉSENTATION DES CONTRIBUTIONS

L'article qui figure en ouverture des contributions permettra au lecteur de se faire une idée précise des choix techniques effectués pour la constitution du corpus PRESTO, son annotation et son exploitation par des méthodes lexico-statistiques. S. Diwersy, A. Falaise, M.-H. Lay et G. Souvay y présentent d'abord plusieurs des difficultés majeures que rencontre toute entreprise d'analyse automatique de textes datant du XVI^{e} et du XVII^{e} siècle, qu'elles soient liées aux changements intervenus dans l'ordre des mots (syntaxe), au lexique, ou encore à l'importante variation graphique des mots et à leur découpage typographique. Les auteurs présentent ensuite les critères ayant présidé au choix des textes du corpus ainsi que les métadonnées qui y ont été associées, avant d'aborder la création de ressources (base lexicale, modèle de langage) et les étapes du processus d'annotation qui les utilisent : tokénisation, étiquetage morphosyntaxique et lemmatisation. Certaines des méthodes lexico-statistiques utilisées dans le présent numéro pour l'approche quantitative des faits linguistiques en diachronie font enfin l'objet d'une brève présentation.

Les processus de grammaticalisation et de lexicalisation dont certaines expressions employées en français contemporain comme marqueurs de topique de discours sont le siège (comme *à propos de*, *sur le plan de*, etc.) font l'objet de l'article suivant. Le corpus mobilisé pour cette étude compte environ 70 millions de mots et réunit un ensemble de numéros du quotidien *Le Figaro* publiés à la fin du XIX^{e} siècle d'une part, au début du XXI^{e} siècle d'autre part. Après avoir regroupé ces expressions topicales en sous-classes suivant des critères morphologiques, syntaxiques et sémantiques, M. Charolles, S. Diwersy et D. Vigier présentent et discutent les résultats d'un codage syntactico-sémantique accompli sur un vaste échantillon d'énoncés prélevés aléatoirement sur le corpus de départ. Après extrapolation des fréquences obtenues pour chacune des sous-catégories de marqueurs de topiques distinguées à l'issue de ce codage, les auteurs étudient l'évolution en diachronie de leur fonctionnement sémantique et syntaxique entre le XIX^{e} et le XXI^{e} siècle, en s'aidant notamment de la méthode statistique des spécificités.

D. Leeman et A. Falaise combinent leurs compétences respectives en morphosyntaxe et en linguistique informatique pour étudier, sous un jour nouveau, un vieux problème de grammaire française, correspondant à une difficulté connue et redoutée dans l'enseignement du FLE. Il s'agit du bon usage des prépositions devant certains types de noms géographiques (départements, régions). Le choix porte essentiellement sur *dans*, *en* et *à* et peut dépendre de facteurs formels, comme le genre grammatical du nom (*la Gironde*, mais *le Gers*, en parlant du département), du nombre (*le Jura*, mais *les Ardennes*), de la formation du nom (*Seine*, mais *Hauts-de-Seine*), du statut de l'article (*La Réunion*, (*la*) *Martinique*, *Mayotte*) – et surtout des croisements de ces facteurs. L'application d'un ensemble de règles grammaticales sur ces données de base aboutit à la production de syntagmes qui ne portent guère à contestation au niveau de la norme et du français courant. Ainsi, tout le monde accepte la différence de la préposition dans des cas de figure comme *dans le Bas-Rhin* et *à La Réunion*. Or, sur ce fond de constructions non marquées, documentées statistiquement, peuvent se greffer des phénomènes plus rares qui possèdent une dimension stylistique et / ou cognitive. Car, d'une part, l'emploi d'une préposition inattendue et stylistiquement marquée (*en Gers* pour *dans le Gers*) influe sur la perspective dans laquelle on parle de ce département ; d'autre part, tel emploi particulier exigé par la norme, mais contraire à une règle plus générale, montre que la norme elle-même peut entériner une certaine perspective. C'est le cas de *à Mayotte* au lieu de **en Mayotte*. Comme on le voit, le passage du quantitatif au qualitatif peut aussi déboucher sur le problème des règles et des exceptions.

B. Fagard et K. Krawczak reprennent une vieille question sur laquelle se sont déjà penchés les synonymistes du XIX^{e} s. : la variation de la préposition (*à* ou *de*) entre les verbes *commencer* et *continuer* et l'infinitif relève-t-elle de la phonétique ou de la sémantique ? Dans le premier cas, elle serait conditionnée par la recherche d'une certaine euphonie et dépendrait du début vocalique ou consonantique de l'infinitif. Dans le second cas, le choix entre les prépositions serait déterminé par l'aspect lexical (télique ou non ?) du verbe à l'infinitif et soulignerait

donc une harmonie entre les propriétés aspectuelles de la préposition et de sa suite verbale. On se doute que cette dernière interprétation a eu les faveurs de synonymistes comme B. Lafaye. À tort, comme le prouvent les deux auteurs sur la base de divers corpus historiques, exploités et analysés avec rigueur. Dans une large perspective diachronique, le conditionnement aspectuel s'avère grosso modo secondaire. Sur un plan méthodologique, on notera que l'approche *corpus-based* conduit, en l'occurrence, à une démarche qui synthétise les phases quantitative et qualitative, séparées dans d'autres contributions. Car ce sont les résultats du calcul de la fréquence de chacune des deux situations qui constituent une réponse qualitative (confirmation de l'hypothèse phonétique).

C'est la naissance de la préposition *dans* au XVI^{e} siècle qui fait l'objet de la contribution de D. Vigier. En s'appuyant sur le corpus historique PRESTO, l'auteur présente d'abord un panorama de l'évolution des fréquences des prépositions concurrentes *en* et *dans* entre 1550 et 1940 environ. Il se focalise ensuite sur la période des premiers emplois de *dans* entre 1550 et 1700, dans le but d'éprouver en corpus l'hypothèse forgée par A. Darmesteter (1885) et qui prévaut encore aujourd'hui, selon laquelle la spectaculaire fortune fréquentielle de *dans* à partir de 1550 serait due à « la décadence des formes contractes » issues des combinaisons de la préposition *en* avec les formes *le, les* de l'article défini. En recourant à des calculs probabilistes implémentés sur la plateforme TXM et à des analyses issues des statistiques textuelles, l'auteur illustre un des apports possibles de l'approche quantitative sur corpus, en montrant comment elle permet de trancher entre plusieurs hypothèses concurrentes élaborées parfois il y a longtemps pour rendre compte de phénomènes bien identifiés en linguistique historique, mais dont l'explication reste incertaine.

Le point de départ quantitatif de P. Blumenthal, qui compare deux encyclopédies (XVIII^{e} et XXI^{e} siècles), est l'observation du remarquable dynamisme de quelques locutions prépositionnelles (entre autres *du point de vue de*, *au point de vue de*, *dans l'optique*, *sous l'angle* (*de*), *dans le domaine de*, *sur le plan de*, *dans le cadre de*, *au niveau de*), rares ou inexistants à l'âge des Lumières, très fréquents à notre époque. Sémantiquement, elles expriment souvent soit une certaine perspective du problème scientifique à présenter, soit elles se chargent d'une classification du réel. Les deux valeurs permettent aux nouvelles tournures de contribuer fortement à la structuration du texte et à lui conférer un type de cohérence inconnu, dans la vulgarisation scientifique, avant le XX^{e} s. Ce résultat qualitatif de l'analyse implique une interrogation sur les causes du changement – ou du moins les circonstances, proprement linguistiques ou extralinguistiques, qui ont pu favoriser l'évolution. Vu la nature des textes en question, l'auteur plaide pour un parallèle avec le cheminement de l'épistémologie moderne.

C. Rossari et C. Ricci s'intéressent quant à elles à certaines expressions prépositionnelles formées avec les mots *conséquent/conseguente* et *conséquence/conseguenza* et qui ont – ou ont eu – un emploi de connecteur en français et en italien. Les auteurs examinent en premier lieu l'évolution quantitative de ces locutions entre le XVI^{e} s. et le XX^{e} s. en s'appuyant sur un corpus diachronique constitué à partir de plusieurs bases de données textuelles de langue française et italienne. Les principales tendances évolutives mises au jour, un focus est établi sur les expressions *en conséquence*, *in conseguenza* et *di conseguenza* qui ont pour spécificité d'avoir développé, entre le XVI^{e} s. et le XX^{e} s., des emplois de connecteurs succédant à des emplois non connecteurs. La description des types d'environnements linguistiques de ces deux emplois fait l'objet d'une analyse qualitative dans le but de caractériser les conditions contextuelles favorisant l'émergence de l'emploi de connecteur. L'étude quantitative et qualitative des environnements dans lesquels l'ensemble de ces expressions apparaît permet ensuite de déterminer si leur pic d'emploi en tant que connecteurs a coïncidé ou non avec un élargissement de leurs variétés de construction.

Nous tenons enfin à exprimer notre gratitude à Janine Schwieres (Cologne), qui a préparé le manuscrit avec compétence et efficacité.

Références

BLUMENTHAL P. (2011), « Le figement : du XVIIe siècle au français contemporain », *in* J.-C. Anscombre & S. Mejri, *Le figement linguistique : la parole entravée*, Paris : Honoré Champion, 283-302.

BLUMENTHAL P., VIGIER D., éds. (à par. 2017), *Changements linguistiques et perspective sociétale*, Actes du colloque CLPS (*Changements linguistiques et perspective sociétale)* ENS Lyon, mars 2016. Berne : P. Lang.

BRUNET E. (2011), « Le viol de l'urne », *in Ce qui compte. Écrits choisis,* tome II. *Méthodes statistiques*, 79-94.

CORI M., DAVID S., LEON J. (2008), « Présentation : éléments de réflexion sur la place des corpus en linguistique », *Langages* n° 171, 3, 5-11.

DUNNING T. (1993), « Accurate methods for the statistics of surprise and coincidence », *Computational Linguistics*, 19-1, 61-74.

ECO U. (2003), *Dire presque la même chose*, Paris : Grasset.

FAVRE C. (2015), « Sémantique distributionnelle automatique : la proximité distributionnelle comme mode d'accès au sens », *Ela*, n° 180, octobre-décembre, 395-405.

FIRTH J. R. (1957), « A synopsis of linguistic theory 1930-1955 », *Studies in linguistic analysis*, Oxford: Blackwell, 1-32.

GEERAERTS D. (2010), *Theories of Lexical Semantics*, Oxford: Oxford University Press.

GIRAULT S. & VICTORRI B. (2009), « Linguistiques de corpus et mathématiques du continu », *Histoire Epistémologie Langage*, 31(1), 147-170.

GRIES S.T. & HILPERT M. (2008), « The identification of stages in diachronic data: variability-based neighbour clustering », *Corpora* 3(1), 59-81.

GRIES S.T., HILPERT M. (2012), « Variability-based neighbor clustering: a bottom-up approach to periodization in historical linguistics », *in* T. Nevalainen & E. Traugott (eds.), *The Oxford handbook of the history of English*, Oxford: Oxford University Press, 134-144.

GUIRAUD P. (1960), *Problèmes et méthodes de la statistique linguistique*, Paris : PUF.

GUIRAUD P. (1954), *Les Caractères statistiques du vocabulaire*, Paris : PUF.

HALLIDAY M. A. K. (1961), « Categories of the theory of grammar », *Word*, 17.3, 241-92.

LAFON P. (1980), « Sur la variabilité de la fréquence des formes dans un corpus », *Mots* n°1, 127-165

LAFON P. (1984), *Dépouillements et statistiques en lexicométrie*, Genève-Paris : Slatkine-Champion.

LEBART L. & Salem A. (1994), *Statistique Textuelle*, Paris : Dunod.

LEON J. (2008), « Aux sources de la « Corpus Linguistics » : Firth et la London School », *Langages* , n° 171, 12-33

LEON J. (2010), « Automatisation-mathématisation de la linguistique en France dans les années 1960. Un cas de réception externe », *Congrès Mondial de Linguistique Française* (CMLF), 825-838.

LEON J. (2015), *Histoire de l'automatisation des sciences du langage*, ENS Editions.

LEON J., LOISEAU S. (éds., 2016), « History of Quantitative Linguistics in France », *Studies in Quantitative Linguistics* 24.

LOISEAU S. (2011), « Les faits statistiques comme objectivation ou comme interprétation : statistiques et modèles basés sur l'usage », Travaux de linguistique, n° 62, 59-78.

MILNER J-C. (1989), *Introduction à une science du langage*, Paris : Le Seuil.

MULLER C. (1973-1992), *Initiation aux méthodes de la statistique linguistique*, Paris : Champion.

MULLER C. (1977), *Principes et méthodes de statistique lexicale*, Paris : Hachette.

MULLER C. (1992), *Principes et méthodes de statistique lexicale*, Paris : Champion.

PINCEMIN B. (2004), « Lexicométrie sur corpus étiquetés », *in Actes des JADT 2004,* 865-873.

PINCEMIN B. (2011), « Sémantique interprétative et textométrie – Version abrégée », *Corpus*, 10, 259-269.

PINCEMIN B., HEIDEN S., LAY M.-H., LEBLANC J.-M., VIPREY J.-M. (2010), « Fonctionnalités textométriques : proposition de typologie selon un point de vue utilisateur », *in Actes des JADT 2010*, 341-353.

SINCLAIR J. McH. (1991) *Corpus*, *concordance*, *collocation*. Oxford: Oxford University Press.

STUBBS M. (1993), « British tradition in text analysis: from Firth to Sinclair », *in* Baker M. et al. (eds.), *Text and Technology. In honour of John Sinclair*, 1-33.

WITTGENSTEIN L. (1953-1961), *Tractatus logico-philosophicus* suivi de *Investigations philosophiques*, Paris : Gallimard (TEL).