La linguistique et le nombre

Véronique Magri

La linguistique et le nombre

Véronique MAGRI

La linguistique est la science statistique type ;
les statisticiens le savent bien ;
les linguistes l'ignorent encore[1]

Depuis cette déclaration prometteuse, en exergue, d'un des pionniers de la statistique linguistique, où en sont les relations entre linguistique et outils statistiques ?

La linguistique qui s'affirme comme discipline scientifique offre, *a priori*, plus d'affinités spontanées avec les méthodes statistiques, issues des sciences dites exactes, que l'analyse littéraire, sur laquelle plane toujours le soupçon de subjectivité. La linguistique a, par ailleurs, cette propension à l'observation des faits qui l'éloigne de la grammaire normative. Loin d'édicter des règles théoriques et de déclarer non conformes les écarts éventuels, la linguistique des textes développe une démarche empirique et expérimentale.

> La linguistique est donc une science empirique, en ce sens qu'elle définit une instance de réfutation et que celle-ci est constituée à partir des données contingentes des langues. C'est une science expérimentale, en ceci qu'elle construit activement les observations qui donneront lieu aux procédures de réfutation[2].

Opposé au rationalisme scientifique, l'empirisme est connoté défavorablement, comme le rappelle François Rastier ici même, comme on le lit dans la citation ci-après :

> L'empirisme est un donjon étroit et abject d'où l'esprit emprisonné ne peut s'échapper que sur les ailes d'une hypothèse[3].

La nature de l'interaction entre le chercheur et son objet d'étude discrimine les sciences dites expérimentales et les sciences d'observation. L'observation est « la constatation pure et simple d'un fait », l'expérience « le contrôle d'une idée par un fait[4] ». L'expérimentation sert à valider ou non les hypothèses formulées pour expliquer un fait observé et en rechercher les causes. Les faits d'expérimentation sont obtenus dans des conditions déterminées par l'expérimentateur. « Dans la méthode

[1] Pierre Guiraud (1959), p. 15.

[2] Jean-Claude Milner (1989), *Introduction à une science du langage*, Le Seuil, cité par Franck Neveu (2016), p. 5.

[3] Claude Bernard [1947], (1987), p. 77.

[4] Claude Bernard, *Ibid*., p. 55.

expérimentale, l'expérience et l'observation marchent toujours de front[5] » cependant et toute science expérimentale commence par être une science d'observation.

La variabilité des paramètres durant les phases de préparation et d'expérimentation permet d'ajuster la définition d'un modèle reproductible en laboratoire. La linguistique de corpus ou des usages[6] a intégré l'avantage du traitement de masses de données, autorisé par l'outil statistique. Le matériau verbal offre une plasticité appréciable en termes de contrôle, de reproduction, de transmission des données[7]. Même si la fréquence d'un fait de langue ne prévaut pas systématiquement sur sa simple attestation pour la linguistique générale, la linguistique de corpus a su tirer parti des outils quantitatifs, en particulier pour la classification des textes, que ce soit selon le genre, défini par des propriétés morphosyntaxiques, selon l'auteur dont l'écriture peut être appréciée en termes de fréquences d'emploi de telles ou telles structures, ou encore selon le moment littéraire, caractérisé par des habitudes scripturales temporellement situées. La massification des données est un atout indéniable dès qu'une variation linguistique doit être observée et analysée. L'émergence de régularités permet alors, au prix d'une simplification heuristique, de construire un modèle. Les singularités ne sont pas occultées pour autant, comme on pourrait le penser, mais elles servent justement de contrepoint au général pour affiner la modélisation.

L'analyse statistique des données textuelles utilise les outils du *Traitement automatique de la langue* (TAL)[8] sans se confondre avec lui[9]. Si les outils sont mutualisés, la méthodologie et la finalité ne sont pas les mêmes. Le TAL vise la performance des algorithmes et des logiciels ainsi que l'automatisation globale des processus. Un corpus d'apprentissage est construit et confronté ensuite à un corpus de travail. Les domaines d'application du TAL sont bien connus : la traduction automatique, la reconnaissance vocale ou encore la fouille de textes. La procédure qui allie apprentissage et reconnaissance est également suivie par le *deep learning* qui connaît actuellement un regain d'attention de la part des linguistes, en particulier pour l'attribution d'auteur.

C'est une démarche expérimentale qu'adopte l'article de L.-M. Ho-Dac et alii (article 3) dont l'enjeu est l'optimisation des outils à même d'identifier les noms capsules ou « coquilles conceptuelles » et qui sont testés en fonction d'une double approche : la première approche est distributionnelle et lexicale et pose l'hypothèse que les noms abstraits sont de bons candidats pour entrer dans la catégorie des noms capsules ; la seconde est sémantique et conjecture que ces noms entrent dans des patrons syntaxiques identifiables, les constructions spécificationnelles et

[5] Claude Bernard, *Ibid*., p. 63.

[6] Franck Neveu (2016), p. 3.

[7] Voir Étienne Brunet (2011).

[8] Voir Ludovic Tanguy, Cécile Fabre (2014) ; Egle Eensoo, Mathieu Valette (2015) ; Catherine Fuchs, Benoît Habert (2004) ; Bénédicte Pincemin, ici même.

[9] Voir Catherine Fuchs (2014) ; Mathieu Valette (2016).

encapsulantes. L'expérimentation a pour but de vérifier ces hypothèses et le protocole est suivi méthodiquement depuis l'établissement des deux corpus contrastifs à même de pointer des différences d'emploi dépendantes du genre dans lequel les noms s'inscrivent. Suivent les étapes du pré-traitement des corpus, de leur exploitation, de la validation éventuelle des hypothèses initiales enfin.

L'article de H. Flamein et I. Eshkol Taravella (article 4) applique le même protocole qui est d'évaluer la performance des instruments de détection automatique des noms de lieux et propose des dispositifs de mesure de cette performance : les outils *Rappel, Précision, F-Mesure*. La méthodologie repose sur la confrontation entre un corpus-norme, annoté manuellement, et un corpus applicatif. Les résultats de l'expérimentation automatique sont comparés aux résultats obtenus par annotation manuelle en vue d'évaluer l'efficacité du repérage automatisé. L'article de O. Kraif et A. Tutin (article 5) pose comme évaluation de la performance des outils de détection automatique des collocations une comparaison entre les résultats automatisés et l'identification manuelle des occurrences. L'article de S. Mejri et L. Zhu (article 6) teste des instruments de reconnaissance et d'extraction des réseaux phraséologiques.

La linguistique quantitative observe un corpus numérique, préalablement outillé et annoté par l'expert, mais vise à la connaissance du fait linguistique, en pariant sur le potentiel heuristique du dénombrement et de la mise en contraste des faits textuels. Le chercheur, à la fois linguiste et usager des méthodes quantitatives, doit mener une réflexion sur les données qu'il construit, sur le traitement qu'il leur fait subir et sur les conclusions qu'il peut tirer de son expérimentation. La linguistique quantitative n'est pas « une théorie mais une méthodologie d'étude du discours qui se veut exhaustive, systématique et automatisée »[10] ; cette méthodologie s'applique à des degrés divers aux trois étapes qui structurent toute expérimentation sur corpus : la sélection des données, leur exploitation et leur interprétation.

La première vise la constitution du recueil de données. Il ne s'agit pas de procéder à un échantillonnage d'items choisis en fonction de leur représentativité décrétée *a priori* mais de travailler sur des faits réunis en fonction de similarités répondant au point de vue du chercheur et à sa perspective de recherche. Ce recueil de données correspond à la définition du *corpus-driven*[11], la source dont le traitement doit faire émerger des régularités, par opposition au *corpus-based*, conçu comme réservoir d'exemples, à vocation essentiellement illustrative, en faveur d'une approche qualitative et fonctionnant comme une archive (F. Rastier, article 1). La démarche du linguiste est systématique en ce sens que les très grands recueils ainsi constitués peuvent être soumis à un questionnement, sinon objectif, du moins stable. L'article de L. Tanguy et J. Rebeyrolle (article 7) fait le choix de travailler, lui, sur des données brutes fournies par un extrait de HAL, sans chercher à échantillonner ou à rééquilibrer l'ensemble selon, en particulier, le domaine de recherche des textes scientifiques.

[10] Patrick Charaudeau, Dominique Maingueneau (2002), p. 342.

[11] Elena Tognini-Bonelli (2001).

Les données de l'encyclopédie collaborative *Wikipédia* offrent les garanties d'un très grand corpus suffisamment varié pour assurer la robustesse de l'expérimentation quand l'étude porte sur un fait de langue (article 3). La variété est assurée par la coexistence d'articles rédigés et de leurs commentaires développés dans le fil de discussion qui se tisse en parallèle ; deux registres différents, un type didactique, un type argumentatif sont ainsi illustrés. Le corpus Eslo2 offre les mêmes garanties sur des données orales cette fois, ce qui soulève les problématiques spécifiques à la transcription d'interventions parlées (article 4). De même, deux sous-corpus sont délimités et contrastés : un corpus d'*Entretiens* et un autre composé d'*Itinéraires*, correspondant à des « micro-trottoirs ». Le dictionnaire informatisé dont l'ambition est de rendre compte de la totalité de la langue est sans doute l'exemple le plus abouti de l'exhaustivité dans la description de la langue (article 6 de S. Mejri et L. Zhu).

Les observations portent sur des unités quantifiables, donc discrètes, au sens mathématique du terme et sont automatisées par les instruments. Le pré-traitement des données lui-même peut être automatisé lorsque l'annotation morphosyntaxique et syntaxique est réalisée par l'analyseur *Talismane* dont l'efficacité repose sur les étapes de l'apprentissage et de la reconnaissance. Ce logiciel sert à l'étiquetage des noms capsules et à celui, morphosyntaxique, du corpus d'étude de l'article 7. L'étiquetage demeure manuel pour le repérage et la classification des noms de lieux mais la double expertise assure plus de robustesse à l'annotation, dans ce cas (article 4).

L'usage de l'instrument permet de changer la nature même des observables, qui se construisent à la croisée du point de vue du chercheur, de son objectif de recherche, du matériau ou des ressources qu'il a à sa disposition[12]. Les objets d'étude sont ainsi corrélés à un degré de pertinence. La mesure quantitative relative de chaque fait linguistique est soumise à l'axiome initial qui veut que la fréquence d'une donnée soit significative. L'article 7 a l'ambition de faire émerger les propriétés lexicales et structurelles des titres des publications scientifiques en considérant qu'une fréquence relative remarquable est un indice de spécificité, de portée possiblement définitoire du fait textuel observé.

Reste à parier sur un saut qualitatif qui pourra attribuer une valeur à cette significativité afin d'élever le fait significatif au fait signifiant et, réciproquement, d'assurer un gain en termes de connaissance de la langue. L'analyse « augmentée » du texte s'assortit d'une pertinence heuristique, pour ne pas sombrer dans la recherche, certes outillée, mais vaine, et ne pas confondre technicité et scientificité. L'enjeu de l'article 6 est bien d'isoler dans le dictionnaire ce qui est automatisable en faveur d'une meilleure connaissance du système linguistique. Le filtrage manuel des résultats est l'intervention minimale de l'expert après le calcul des algorithmes (article 4). Mais les théories linguistiques et grammaticales forment le soubassement de la démarche empirique : la grammaire des constructions, l'analyse en dépendances syntaxiques

[12] Anna Jaubert (2017).

pour ce même article par exemple, si bien que l'intervention humaine, d'ordre qualitatif, est solidaire du quantitatif.

Un partenariat est établi entre le chercheur et les instruments, entre l'homme et la machine. Au concept de « corpus » ou de « conditions de production », les nouvelles textualités proposent l'alternative de la notion d' « environnement »[13] empruntée à la cognition sociale. La linguistique quantitative est contextualisante : elle articule le particulier au global, elle replace le fait à étudier dans son environnement cotextuel ; il est judicieux d'élargir le contexte aux outils d'analyse du texte, de voir, dans ce postdualisme revendiqué par les nouvelles textualités, un dépassement des dichotomies traditionnelles et l'élaboration d'une complémentarité indissoluble entre linguistique et statistique textuelle ou, mieux, d'un continuum entre l'appréciation humaine et l'ingénierie linguistique. La lecture numérique des faits textuels substitue à la linéarité de la lecture cursive et syntagmatique le point de vue surplombant de mise en relation des faits qui mène à une lecture paradigmatique et réticulaire. L'article 5 rappelle que les observations quantitatives reposent sur l'étape préalable du pré-traitement, qui consiste en la catégorisation des observables qui est d'ordre qualitatif. F. Rastier comme B. Pincemin insistent sur la complémentarité du qualitatif et du quantitatif. Le ruban de Moebius[14] pourrait être l'emblème de cette complémentarité pensée et réfléchie, où linguistique textuelle et linguistique quantitative sont la même face d'un objet complexe mais aussi où l'esprit du linguiste et le nombre de son analyse sont solidaires et complémentaires. Loin de s'arrêter aux résultats chiffrés, le linguiste statisticien conserve ce « grand principe expérimental [qu'] est donc le doute, le doute philosophique qui laisse à l'esprit sa liberté et son initiative[15] ».

Le numéro entend développer la réflexion sur la linguistique outillée[16]. Il s'agit de montrer comment les nouveaux outils conduisent à une nouvelle approche de la linguistique : les observables comme leur traitement sont modifiés. Une interaction permanente doit être trouvée entre les outils du chercheur à adapter au corpus d'étude et les observables à modéliser afin de les rendre exploitables ; l'observation des données à l'aide de ces outils, en contexte, conduit à des considérations plus générales sur les structures de la langue.

Deux articles dont l'enjeu est épistémologique ouvrent ce numéro et sont suivis par des articles de visée plus applicative : ceux-ci prennent pour objet d'étude un fait de langue et en proposent un repérage et une extraction à même d'en permettre une meilleure définition et de parvenir à une connaissance augmentée du système linguistique.

L'article de F. Rastier rappelle les fondamentaux de la linguistique outillée dans une appréciation mesurée et approfondie. Des mises au point affinent avec acribie les

[13] Voir Anne-Marie Paveau (2017).

[14] Voir Douglas Hofstadter (2008).

[15] Claude Bernard, *Ibid*., p. 88.

[16] Benoît Habert (2004).

définitions traditionnelles et rebâtissent les binômes méthodologiques traditionnels pour en démontrer la complémentarité (quantité/qualité), pour en redessiner les contours et affiner les oppositions (corpus/archive, data/corpora, prévision/prédiction, général/particulier, types/occurrences, fréquence/pertinence, mesure/grain). Nous est fournie là une manière de *vade mecum* à l'usage du linguiste expert ou amateur en méthodes quantitatives. L'article de B. Pincemin est un plaidoyer en faveur de la textométrie. Il aspire à lever tous les malentendus qui peuvent subsister dans cette querelle moderne entre quantitatifs et qualitatifs, partisans et opposants du dénombrement, en vue d'un concordat optimiste et fructueux.

Des applications sont proposées ensuite qui exposent les contraintes mais aussi la plus-value associée aux approches outillées pour l'analyse d'unités linguistiques. La fouille de données concerne divers faits de langue.

L'article 3 de L.-M. Ho-Dac, A. Miletić, M. Wauquier, C. Fabre s'intéresse aux noms sous-spécifiés en en proposant une détection automatique ; ces noms sont caractérisés sur le plan lexical et syntaxique, tandis que l'objet d'étude de l'article de H. Flamein et I. Eshkol (article 4) est les noms de lieux, distingués sur le plan sémantique. Outre le repérage à visée lexicograpique qui permet l'établissement d'un lexique, et qui vise la performance de l'outil lors des étapes d'annotation et d'extraction des données du corpus oral Eslo2, la recherche vise une meilleure connaissance de ces unités de langue et esquisse un commentaire d'ordre plus sémantique et sociologique : la perception de la ville par ses habitants.

Les assemblages stéréotypés ou collocations font l'objet d'une recherche systématique par O. Kraif et A. Tutin (article 5) tandis que S. Mejri et L. Zhu (article 6) visent une meilleure connaissance de la langue à partir des phraséologismes reconnus et extraits automatiquement dans le dictionnaire, envisagé comme simulation à ambition exhaustive du système linguistique décrit. L'enjeu de l'article 5 est la comparaison d'un repérage manuel des collocations dans un corpus et la détection automatique de ce même objet, ce qui pose des questions complexes d'identification au linguiste. Les phases d'extraction et d'évaluation suivent le protocole des mesures d'association fondées sur la comparaison entre une fréquence de cooccurrence attendue et la fréquence observée. C'est le cœur du « raisonnement expérimental » où il y a toujours « jugement par une comparaison s'appuyant sur deux faits, l'un qui sert de point de départ, l'autre qui sert de conclusion au raisonnement[17] ». L'enjeu est également lexicographique qui consiste à réaliser des ressources lexicales.

Si ces quatre articles, dans une démarche apparentée au TAL, visent le façonnage d'outils performants, l'article 7 de L. Tanguy et J. Rebeyrolle, tout en utilisant les mêmes outils, a pour finalité d'établir une typologie des titres d'articles, au terme d'une enquête qui s'appuie sur des différenciations disciplinaires liées aux domaines scientifiques de rattachement des articles. Les titres des contributions sont observés d'un point de vue lexical et morphosyntaxique, en vue de mettre à jour des schémas

[17] Claude Bernard, *Ibid.,* p. 56.

récurrents de ce type de publications. La perspective est générique et l'analyse fait la part belle à l'interprétation des résultats après la réflexion sur les outils exploratoires. La discipline scientifique illustrée par les articles peut avoir une incidence sur la manière dont ceux-ci sont écrits. On rejoint par là une amorce d'enquête sociologique.

Trois articles finalement se retrouvent au cœur des variations de la phraséologie : l'article 5 qui évalue les collocations en langue, l'article 6 qui, au travers du dictionnaire, découvre la structuration sémantique et phraséologique de la langue, l'article 7 qui postule les associations de mots comme possible invariant générique. F. Rastier, dans l'article inaugural, explicite les atouts et les limites du « système connexionniste » que les recherches quantitatives ont largement exploité, dès que la lexicométrie a glissé vers la textométrie, dès que la prise en compte du cotexte s'est avérée nécessaire à la construction du sens, dès que le texte a été envisagé non plus comme un « sac de mots » mais comme un assemblage, un ensemble réticulaire. La connexion se trouve modulée en cooccurrences, comme contexte minimal, voire en corrélats si on accepte ce saut qualitatif entre proximité matérielle d'unités lexicales et connexions sémantiques, qui mène aux collocations, aux unités polylexicales (article 6)[18], aux motifs[19].

L'explicitation d'un parcours méthodologique réunit tous les articles ici présentés : la modélisation préalable de l'observable, choisi comme objet d'étude, la fouille de très grands corpus, la conception d'outils spécifiques adaptés aux particularités de chacun d'eux, le retour sur la langue en termes de plus-value *in fine.*

Puisse la promesse qui ouvre ce préambule s'être transformée, désormais, en prophétie !

Véronique Magri
Université Côte d'Azur, CNRS, BCL, France
Veronique.MAGRI@univ-cotedazur.fr

Références

Bernard, Claude [1947], (1987), *Principes de médecine expérimentale*, Paris, PUF.

Brunet, Étienne (2011), « Plaidoyer pour la statistique linguistique », Céline Poudat. *Ce qui compte. Écrits choisis* tome II, Champion, pp. 311-329, 978-2-7453-2225-8. hal-01580838.

Charaudeau, Patrick, Maingueneau (2002), Dominique *Dictionnaire d'analyse du discours*, Paris, Seuil.

[18] Salah Mejri (2003).

[19] Dominique Longrée, Xuan Luong, Sylvie Mellet (2008).

Eensoo, Egle, Valette, Mathieu (2015/4), « Associer heuristiques textométriques et méthodes d'évaluation issues du traitement automatique des langues », Paris, Klincksieck, *Éla. Études de linguistique appliquée*, n° 180, pp. 429-436.

Fuchs, Catherine, Habert, Benoît (2004), « Le traitement automatique des langues : des modèles aux ressources », *Introduction*, *Le Français moderne*, LXXII, 1.

Fuchs, Catherine (2014), « Le tournant quantitatif en TAL et en linguistique : enjeux cognitifs. *L'information grammaticale,* Peeters Publishers, pp. 8-13. hal-01382602.

Guiraud, Pierre (1959), *Problèmes et méthodes de la statistique linguistique*, D. Reidel, Publishing Company, Dordrecht, Holland.

Habert, Benoît (2004), « Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs », *Revue française de linguistique appliquée*, vol. IX, n°1, pp. 5-24, https://www.cairn.info/article.php?ID_ARTICLE=RFLA_091_0005#

Hofstadter, Douglas (2008), *Gödel Escher Bach, Les Brins d'une guirlande éternelle*, Paris, Dunod.

Jaubert, Anna (2017), « Linguistique(s) du discours et stylistique : points de vue sur l'observable », *Le Discours et la langue*, 9.2, pp. 35-45.

Longrée, Dominique, Luong, Xuan, Mellet, Sylvie (2008), « Les motifs : un outil pour la caractérisation des textes », http://lexicometrica.univ-paris3.fr/jadt/jadt2008/pdf/longree-luong-mellet.pdf.

Mejri, Salah (2003), « Polysémie et polylexicalité », *in Syntaxe et sémantique*, n° 5, pp. 13-30.

Milner, Jean-Claude (2016), *Introduction à une science du langage*, Le Seuil, 1989.

Neveu, Franck (2016), *Observatoires et observables en linguistique française*, *Le français moderne*, 1, p. 3.

Paveau, Anne-Marie (2017), *L'Analyse du discours numérique. Dictionnaire des formes et des pratiques*, Paris, Hermann.

Tanguy, Ludovic, Fabre, Cécile (2014), « Évolutions de la linguistique outillée : méfaits et bienfaits du TAL », *L'Information grammaticale*, Peeters Publishers, pp.15-23. hal-01057493.

Tognini-Bonelli, Elena (2001), *Corpus linguistics at work*, John Benjamins Publishing Company.

Valette, Mathieu (2016), « Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée ». *International Conference on Statistical Analysis of Textual Data (JADT2016*), Jun 2016, Nice, pp.697-706. hal-01335084.