Problèmes et méthodes en linguistique française

Introduction aux problématiques du théorique et du quantitatif en sciences humaines

Corpus et sciences des textes

Franck NEVEU

Professeur à Sorbonne Université, Faculté des Lettres

UFR de Langue française

- **Introduction**

- *Vocation du cours*
- *Les mots clés : quantitatif, théorique, théorie, théorisation, épistémologie, historiographie des sciences, observables, observatoires, généralité, type, occurrence, vérité scientifique, objectivité, exhaustivité, rationalité*

- **Première approche : la question des biais cognitifs**

- L'exemple du faux positif
  - Gérald Bronner, *Déchéance de ratio*nalité, Grasset, 2019

- Ex. une maladie qui touche 1 personne sur 1000, détectable par un test qui a un taux d'erreurs positives (faux positifs) de 5%, on dira à tort que le test est fiable à 95%, alors qu'il l'est à environ 2%. Vérité contre-intuitive.
- 5% de faux positifs : sur 100 non-malades il y a 5 personnes reconnues positives au test.
- Imaginons qu'il y ait 99 900 non-malades sur une population de 100 000 habitants, donc 4995 faux positifs, et pour cette même population il n'y a que 100 vrais malades. Calcul : 100/(100+4995)=100/5095 = 0,196, soit environ 2%

- **Première approche : la question des biais cognitifs**

- Biais de négligence des taux de base ou oubli de la fréquence de base
- Amos Tversky (1937-1996), Daniel Kahneman (1934-)
- Ex. Un jour on croise une femme qui vit en région en région parisienne, qui porte un tailleur, et qui lit Le Figaro. Quelle est probabilité la plus importante : 1. qu'elle soit caissière, 2. qu'elle soit DRH d'une grande entreprise ?
- Il y a 20 000 caissières en région parisienne, seule 1 caissière sur 100 porte un tailleur et lit Le Figaro, le nombre d'occurrences de rencontres avec une caissière portant un tailleur et lisant Le Figaro est de 20 000/100 =200
- Il y a 10 femmes DRH d'une grande entreprise en région parisienne, 9 d'entre elles portent un tailleur et lisent Le Figaro. Il y a 209 occurrences possibles 200/209 = 95,7% de chances pour que la femme rencontrée soit une caissière et 4,3% pour qu'elle soit DRH d'une grande entreprise

- **Première approche : la question des biais cognitifs**

- Biais cognitif : distorsion dans le traitement cognitif d'une information, déviation systématique de la pensée logique et rationnelle par rapport à la réalité
- Effet Dunning-Kruger (David Dunning, Justin Kruger)
- Biais de représentativité

- **Seconde approche : le quantitatif comme obstacle épistémologique ?**

- Jean-Claude Milner (2014) , *La Puissance du détail*, Grasset : « *Qui traite les discours comme matière à travailler, entend et voit mieux les réalités. Qui entend et voit mieux les réalités est moins facile à domestiquer* ».
- Catherine Fuchs C. (2014), « Le tournant quantitatif en TAL et en linguistique : enjeux cognitifs », *L'Information grammaticale*, n° 142 : 8-13
- […] ce n'est pas le traitement des données par la machine (sans modélisation préalable) qui, sur le fond, fera surgir des connaissances inédites concernant le système de la langue : tout au plus permettra-t-il de mettre en évidence certaines régularités statistiques ayant trait à un certain type de pratique langagière dans des circonstances particulières. Là précisément peut résider le piège, pour le linguiste tourné quasi-exclusivement vers les ressources. **Le travail sur corpus n'est évidemment pas incompatible avec la réflexion théorique, il peut même être bienvenu pour étayer un raisonnement de linguistique. Mais à la condition de respecter les contraintes d'une démarche scientifique et de savoir construire une véritable problématique théorique.** Or, dans la phase actuelle, le linguiste semble bien souvent transformé en un travailleur de force qui n'en aurait plus ni le temps ni les moyens. A ce compte, la technicisation risque fort de renvoyer aux oubliettes le trésor de descriptions et de théories (non « outillées ») accumulées depuis des siècles (Lazard, 2013), en donnant l'illusion qu'un traitement de surface accompagné de quelques décomptes serait susceptible de révéler *proprio motu* les propriétés de la langue. Or les données langagières (même enrichies d'annotations diverses) ne se confondent pas avec le système de règles – aussi variable et labile soit-il – constitutif de la langue. (Fuchs, 2014 : 12)

• Seconde approche : le quantitatif comme obstacle épistémologique ?

- Paul Watzlawick (1921-2007), Ecole de Palo Alto
- *Comment réussir à échouer* [(en) *Ultrasolutions*], 1986, Norton ; trad. Anne-Lise Hacker, Seuil 1988 ; rééd. poche Points, 2014
- « On sait que les ordinateurs ne font pas seulement des opérations mathématiques, mais aussi des opérations logiques. Autrement dit, ils sont capables de tirer des conclusions. Dans les deux cas, les réponses de l'ordinateur ne sont fausses que s'il y a eu erreur humaine dans la programmation de la machine. En jargon informatique américain, on appelle ce type de complication GIGO (Garbage In, Garbage Out), ce qui veut dire qu'à partir de données erronées, on obtient des résultats erronés. Mais maintenant, GIGO a une autre signification, plus insidieuse, à savoir : « Gospel in, Gospel out » (à partir de l'Evangile on obtient l'Evangile). Et c'est là que ça devient intéressant pour notre propos : ce que l'on croit ou espère nécessairement vrai ou juste revient toujours comme vérité éternelle après être passé à travers le système digestif de l'ordinateur. »

- **Seconde approche : le quantitatif comme obstacle épistémologique ?**

- Robert Nicolaï (2000), *La traversée de l'empirique. Essai d'épistémologie sur la construction des représentations de l'évolution des langues, Paris, Ophrys (Bibliothèque de* Faits de langues*)*
- Robert Nicolaï (2007), *La vision des faits. De l'*a posteriori *à l'*a priori *dans la saisie des langues*, Paris, L'Harmattan
- On développera l'idée selon laquelle « *la science ne peut pas vivre comme pure écoute et accueil fidèle d'une donnée empirique auprès de laquelle elle ne ferait que conduire une exploration indéfinie : précisément parce qu'aller la chercher relève de plusieurs dispositions, de plusieurs lieux, de plusieurs temps* » (Jean-Michel Salanskis, *Crépuscule du théorique ?, Paris, Les Belles Lettres, « Encre marine », 2016*).

Sur la philosophie des mathématiques
À partir de Cavaillès

« La vie, la carrière et le destin de Jean Cavaillès peuvent être présentés en quelques mots. Né en 1903, fils d'officier, de religion protestante, scientifique de formation initiale, élève de l'Ecole Normale Supérieure, professeur au lycée d'Amiens, docteur ès lettres en 1938, maître de conférences de logique et philosophie générale à la Faculté des lettres de Strasbourg; mobilisé en 1939 comme officier de corps franc, puis comme officier du Chiffre, prisonnier des Allemands en juin 1940, évadé, revenu en octobre à l'université de Strasbourg repliée à Clermont-Ferrand, désigné en 1941 par la Faculté des lettres de la Sorbonne comme professeur suppléant de logique, cofondateur du mouvement de résistance *Libération Sud*, fondateur du réseau *Cohors*, arrêté par la police française en août 1942, interné à Montpellier puis à Saint-Paul d'Eyjeaux, évadé en décembre 1942, arrêté par le contre-espionnage allemand en août 1943, révoqué par le gouvernement de Vichy, fusillé par les Allemands et enterré dans la citadelle d'Arras en février 1944, Compagnon de la Libération et Chevalier de la Légion d'Honneur à titre posthume ». Georges Canguilhem, 1967

- **Œuvres de Jean Cavaillès**
- *Correspondance Cantor-Dedekind* (1937)
- *Méthode axiomatique et formalisme. Essai sur le problème du fondement des mathématiques* (1938)
- *Remarques sur la formation de la théorie abstraite des ensembles* (1938)
- *Transfini et continu* (posthume, 1947)
- *Sur la logique et la théorie de la science* (posthume, 1947)

- Ces œuvres, accompagnées d'articles scientifiques, sont réunies dans le volume *Œuvres complètes. Philosophie des sciences,* Paris, Hermann, 1994, 686 pages

- « Vous voyez l'étonnant trajet de la preuve : vous voulez établir que p est vrai, pour cela vous avez vos raisons (c'est votre hypothèse). Dans ce but, vous fabriquez la fiction « non-p est vrai », dont vous espérez qu'elle est fausse ! Et pour nourrir votre espoir, vous tirez des conséquences de cette fiction, vous mouvant ainsi avec une logique implacable dans ce que vous pensez être faux, jusqu'à ce que vous rencontriez une conséquence qui contredit explicitement un énoncé antérieurement démontré comme vrai.
- Cette navigation contrôlée, réglée, entre le vrai et le faux est à mon sens tout à fait caractéristique des mathématiques naissantes, de la coupure qu'elles introduisent avec toute vérité révélée ou dont la force serait uniquement poétique. Or on trouve ce « ton » chez Parménide. Et on le trouve parce que, pour prouver que l'être est, que telle est la vérité première, il établit que le non-être n'est pas. Il raisonne donc par l'absurde. Ma conclusion est claire : la philosophie rationnelle et les mathématiques naissent en même temps, et il ne pouvait pas en être autrement. » A. Badiou, *Éloge des mathématiques*, Flammarion, 2015 (p. 33-34)

- « Le fait que certaines propositions peuvent n'être ni vraies ni fausses dans une théorie donnée nous amène à dire quelques mots de la fameuse querelle du tiers exclu. La règle du tiers exclu, qui remonte à l'Antiquité et a notamment été utilisée par Euclide, consiste en ceci : pour prouver qu'une proposition A est vraie, on ajoute provisoirement aux axiomes de la théorie la négation de A (en disant « supposons que A soit fausse »); et si l'on arrive à une contradiction dans cette nouvelle théorie, on considère qu'on a prouvé que A est vraie dans la théorie initiale!

- C’est la « démonstration par l'absurde ». Ce mode de raisonnement est illégitime, disent les intuitionnistes, car vous avez supposé que A était soit vraie soit fausse, ce qui en général n’est pas le cas. Pas du tout, répliquent les formalistes : ce raisonnement par l’absurde revient finalement à prouver la vérité de la proposition suivante : non A =>A (la négation de A entraîne A) et cette proposition, d’après les règles de la logique, est rigoureusement équivalente à la proposition A elle-même. En d’autres termes, toute démonstration par l’absurde peut toujours être mise sous la forme d’une démonstration qui évite ce recours à l’absurde. En fait, je ne connais pas de mathématicien qui aujourd’hui rejette cette règle du « tiers exclu ». Henri Cartan, « Cavaillès et le fondement des mathématiques », in *Jean Cavaillès, philosophe, résistant,* colloque d’Amiens, 1984, CNDP

- « La mathématique, ce n'est pas du tout la science de la différence entre un feuillage d'automne et un ciel d'été ; elle dit seulement que de toute façon, tout ça, ce sont des multiplicités, des formes qui ont quelque chose en commun, le fait d'être, tout simplement. Et ce sont les formes abstraites de ce « commun » que la mathématique essaie de penser. » (Badiou, *Éloge des mathématiques*, 2015, p. 43)

- Wilhelm Dilthey (1833-1911), et à son *Introduction aux sciences de l'esprit* (1883 : *Einleitung in die Geisteswissenschaften*) traduction française 1992, éditions du Cerf : Œuvres 1 : *Critique de la raison historique. Introduction aux sciences de l'esprit* et autres textes).
- Max Weber (1864-1920), (*Essais sur la théorie de la science*, traduction de Julien Freund, Plon 1965)
- Hourya Benis Sinaceur (*Cavaillès*, Belles Lettres, 2013)

- « Cavaillès est un philosophe de l'épreuve, de la pensée unie à l'action, de la « pensée en acte », comme il l'a écrit, de la pensée comme expérience, et pas seulement pensée de l'expérience. Pour un mathématicien « militant », les mathématiques ne sont pas ou pas seulement objet d'énoncé, de discours et de déduction, mais matière à travailler, à transformer. Les gestes mathématiques sont des actes de pensée sur des objets de pensée […]. Mais, inversement, un objet de pensée, un nombre, une opération (additionner 3 et 4, extraire la racine carrée de 9), une propriété géométrique (être équilatéral par exemple), est le produit d'une séquence de gestes mathématiques. L'objet mathématique est intérieur à la pensée mathématique, la matière coextensive à la forme ». Hourya Benis Sinaceur (*Cavaillès*, Belles Lettres, 2013, p. 75)

- « […] l'un des problèmes essentiels de la doctrine de la science est que […] le progrès ne soit pas augmentation de volume par juxtaposition, l'antérieur subsistant avec le nouveau, mais révision perpétuelle des contenus par approfondissement et rature. Ce qui est après est plus que ce qui était avant, non parce qu'il le contient ou même qu'il le prolonge mais parce qu'il en sort nécessairement et porte dans son contenu la marque chaque fois singulière de sa supériorité. Il y a en lui plus de conscience – et ce n'est pas la même conscience. Le terme de conscience ne comporte pas d'univocité d'application – pas plus que la chose, d'unité isolable. Il n'y a pas une conscience génératrice de ses produits, ou simplement immanente à eux, mais elle est chaque fois dans l'immédiat de l'idée, perdue en elle et se perdant avec elle et ne se liant avec d'autres consciences (ce qu'on serait tenté d'appeler d'autres moments de la conscience) que par les liens internes des idées auxquelles celles-ci appartiennent. Le progrès est matériel ou entre essences singulières, son moteur l'exigence de dépassement de chacune d'elles. Ce n'est pas une philosophie de la conscience mais une philosophie du concept qui peut donner une doctrine de la science. La nécessité génératrice n'est pas celle d'une activité, mais d'une dialectique. » (OC, p. 560)

- Pierre Cassou-Noguès, *Un Laboratoire philosophique. Cavaillès et l'épistémologie en France*, Paris, Vrin, 2017
- Jean-Claude Milner, *L'Universel en éclats*, Verdier, 2014
- Robert Nicolaï, *La traversée de l'empirique. Essai d'épistémologie sur la construction des représentations de l'évolution des langues*, Paris, Ophrys (Bibliothèque de *Faits de langues*), 2000
- Robert Nicolaï, *La vision des faits. De l'*a posteriori *à l'*a priori *dans la saisie des langues*, Paris, L'Harmattan, 2007

Sur la problématique épistémologique des corpus

- **Sur la notion de corpus**

- - textes : objets empiriques de la linguistique
- - texte : unité minimale d'une linguistique évoluée
- - corpus : ensemble dans lequel cette unité minimale prend sens
- - corpus : regroupement structuré de textes intégraux
- - tout corpus suppose une préconception des applications en vue desquelles il a été structuré
- - la linguistique de corpus est objective mais non objectiviste
- - corps vs banque textuelle/base de données
- - corpus vs hypertexte/web
- - corpus vs sac de mots

- **Sur la notion de corpus**

- - structures du corpus : documentaire vs rhétorico-herméneutique
- - codage des variables globales : discours, champ générique, genre, sous-genre
- - l'archive
- - le corpus de référence
- - le corpus d'étude
- - les sous-corpus de travail

- **Le corpus et la distinction langue/parole**

- - « séparation » langue/parole
- - langue vs parole : une distinction non de degrés mais de statut épistémologique
- - les textes sont écrits dans un genre en tenant compte des contraintes d'une langue, et non dans une langue en tenant compte des contraintes d'un genre
- - importance des genres en traduction
- - linguistique de la langue et linguistique de la parole sont unies par l'espace des normes
- - espace normatif des règles et désordre apparent des usages
- - universel de la langue et singularités des emplois

- **Le corpus et la distinction langue/parole**

- - règles de la langue vs performances de la parole
- - pour décrire les normes il faut exploiter les corpus
- - les genres et les discours sont l'espace de variation des structures de langue (morphèmes, mots, phrases, etc.)
- - régularités transgénériques et transdiscursives
- - morphosyntaxe : affaire de règles; sémantique : affaire de normes
- - articulation syntaxe/sémantique
- - lexicologies et syntaxes indépendantes du palier du texte
- - incidence du global sur le local

- **Le corpus et la distinction langue/parole**

- - corrélations entre paliers de complexité
- - mot clé et lexicalisation
- - variables de discours et de genres et variables morphosyntaxiques
- - stéréotypie textuelle et normes de la doxa
- - sections de textes et catégories morphosyntaxiques
- - corrélations entre plan du contenu et plan de l'expression
- - corrélations entre contenus lexicaux et ponctèmes
- - diffusion d'une forme phonique locale au niveau textuel
- - opposition formel/sémantique
- - forme intérieure/forme extérieure

- **Corpus et théorie linguistique**

- - corpus et génération
- - corpus et linguistique computationnelle
- - corpus et humanités numériques
- - l'informatique dans la linguistique de copus
- - nouveau rapport à l'empirique

- **Quantité et qualité en linguistique de corpus**

- - ce qui est mesurable et fréquent n'est pas forcément intéressant
- - les traits de forme n'ont pas de poids statistique
- - les unités rares ou absentes
- - le qualitatif peut échapper à tout dénombrement
- - nouveaux observables
- - un nouvel observable est générateur d'hypothèses nouvelles
- - dépasser la contradiction quantité/qualité