

Pour des lectures textométriques en sémantique interprétative. Le projet Sittelle, principes et perspectives

Bénédicte Pincemin

▶ **To cite this version:**

Bénédicte Pincemin. Pour des lectures textométriques en sémantique interprétative. Le projet Sittelle, principes et perspectives. Acta Semiotica et Lingvistica, 2022, Perspectives présentes et futures de la Sémantique interprétative, 27 (2), pp.3-30. 10.22478/ufpb.2446-7006.46v27n2.62999 . halshs-03762150

HAL Id: halshs-03762150
https://shs.hal.science/halshs-03762150

Submitted on 26 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

[Version auteur acceptée du 10 novembre 2021]

in Éric Trudel & Maria de Fátima Barbosa de Mesquita Batista
Dossier « Perspectives présentes et futures de la Sémantique interprétative »
***Acta Semiotica et Lingvistica*, 2022**, vol. 27 n. 2, 46, p. 3-30
https://periodicos.ufpb.br/ojs2/index.php/actas/issue/view/2806

POUR DES LECTURES TEXTOMETRIQUES EN SEMANTIQUE INTERPRETATIVE LE PROJET SITTELLE, PRINCIPES ET PERSPECTIVES

TEXTOMETRIC READING FOR INTERPRETATIVE SEMANTICS
PRINCIPLES AND RESEARCH DIRECTIONS OF SITTELLE PROJECT

Bénédicte PINCEMIN
Université de Lyon, CNRS, IHRIM UMR5317
(benedicte.pincemin *à* ens-lyon *point* fr)

Abstract (français)

Cet article rend compte des intuitions initiales, des principes méthodologiques et des premiers résultats du projet Sittelle (*Sémantique interprétative et textuelle*). Cette initiative vise à expérimenter comment un corpus numérique, doté d'outils textométriques, peut donner accès à des parcours pertinents d'écrits linguistiques de François Rastier. Plusieurs caractéristiques de l'œuvre scientifique de Rastier encouragent une telle approche : sa volumétrie, son unité de pensée, ainsi qu'une écriture organisée de façon souple, fonctionnant souvent par évocation ou reprises.

Des possibilités d'interrogations textométriques sont illustrées sur un corpus expérimental de 28 articles. Les *concordances* offrent une implémentation de la technique herméneutique des passages parallèles. Si besoin, un calcul de *cooccurrence* produit une synthèse statistique des contextes. La *répartition* globale des occurrences d'un mot peut être visualisée graphiquement. Les *inventaires lexicaux* permettent d'observer des dominantes, constructions, et variations. Les *analyses factorielles* sont en mesure de produire des cartographies du corpus. Les *spécificités* apportent une caractérisation différentielle d'un texte. La *méthode Reinert* et de nouveaux modes d'interrogation seraient à étudier pour leur capacité à capter des isotopies.

Quatre principes directeurs orientent les choix techniques de Sittelle : *progressivité* d'une expérimentation par étapes ;.sens de l'*intertextualité* construite par le choix des textes ; o*uverture* dans l'esprit de la science ouverte et par des formats standards ; *philologie numérique* pour des représentations textuelles appropriées à l'activité herméneutique.

Mots-clés : sémantique interprétative, corpus numérique, textes scientifiques, François Rastier, textométrie, analyse statistique des données textuelles, analyse quantitative, analyse qualitative, philologie numérique, herméneutique numérique, science ouverte, logiciel TXM, logiciel IRaMuTeQ.

Abstract (English)

This article presents the original intuitions, the methodological principles, and the first results of the Sittelle project (*Sémantique interprétative et textuelle* / *Interpretative and textual semantics*). This action investigates how a digital corpus of texts from François Rastier, powered by textometric software, can provide new relevant means for methodical readings of these texts. Several features of Rastier's scientific works support such an approach: the magnitude of data, their unity of thought, and a flexibly organized writing style, often involving evocation or reuse of key passages.

Textometric analysis features are illustrated through an experimental corpus of 28 articles. *KWIC concordances* implement the hermeneutical technique of similar or parallel passages. If

Ce document est publié sous licence Creative Commons Attribution 4.0
https://creativecommons.org/licenses/by/4.0

needed, a statistical summary of context passages can be drawn through *collocation* computing. *Word lists* are useful to identify the most frequent words and to examine how words combine or vary. *Factorial analysis* provides a map view of main variations structuring the corpus, based on text content. *Specificity* keyword computing can help characterizing any text by its overused or underused linguistic features. *Reinert's method* and search engine advanced features could be tested as regards their ability to model isotopies.

Four principles guide the technical implementation of Sittelle: *progressivity* of multiple experimental steps; *intertextuality* because texts must be picked so as to contextualize one another and form relevant collections; *open-science* dissemination and standards compliance for data reuse and software interoperability; *digital philology* as text encoding must retain information that matters to interpretative tasks.

Keywords : interpretative semantics, digital corpora, scientific writing, François Rastier, textometry, statistical analysis of textual data, quantitative analysis, qualitative analysis, digital philology, digital hermeneutics, open-science, TXM software, IRaMuTeQ software.

1 Enjeu scientifique

1.1 Intuition initiale et motivations

Il n'est pas simple pour un lecteur novice, voire pour un linguiste familier de la sémantique interprétative, de s'orienter dans le large ensemble des écrits de François Rastier, ou de penser suffisamment comprendre dans toutes ses résonances tel terme, concept, ou passage. Ainsi est née l'idée du projet Sittelle, acronyme pour « Sémantique interprétative et textuelle » : expérimenter, de façon aussi ouverte que possible, comment un corpus numérique, doté d'outils textométriques, peut donner accès à des parcours méthodiques et appropriés de textes de Rastier, comme au demeurant il y engage, à propos d'autres auteurs, dans son livre sur la sémantique de corpus (RASTIER, 2011). Notre hypothèse est qu'un tel environnement pourrait répondre à des besoins de recherche scientifique (approfondissement de concepts, recherche d'un passage précis ou d'une citation appropriée, étude systématique de passages) et constituer une ressource pédagogique (éclairage d'un passage par un autre, recherche de formulations canoniques type définitions, repérage de variantes plus développées ou simplifiées, ou dans un contexte plus clair ou avec des exemples plus parlants, etc.), en amont[1] ou en aval[2] des synthèses didactiques.

[1] Un corpus numérique outillé pourrait aider au recueil systématique de matériaux, à leur organisation, sélection, qualification (par leur contexte local — dans le passage — et global — dans l'œuvre de Rastier —, leur fréquence, etc.). Nous pensons par exemple au projet d'envergure de *Dictionnaire sémiotique en ligne* (https://semiotique.org).

Ainsi, à l'occasion du 35e anniversaire de l'ouvrage *Sémantique interprétative* (RASTIER, 1987) et à partir de notre propre parcours de recherche, il nous a paru plus stimulant non pas de centrer notre propos sur des éléments de lecture personnels de la sémantique interprétative, mais plutôt de partager nouveaux moyens de lecture et d'analyse ouvrant à chacun de multiples possibilités pour tracer son propre cheminement dans l'œuvre de François Rastier, en fonction de ses attentes et ses besoins. Il s'agit non seulement de travailler à la réalisation d'un corpus outillé, mais aussi, conjointement, de réfléchir aux types d'interrogation et de calculs analytiques pertinents sur un tel corpus. Et sur ces deux volets (création et exploitation du corpus), une entreprise collaborative et organisée, mettant au point des éditions de texte homogènes et mutualisant des repères méthodologiques et interprétatifs, semble la voie la plus enrichissante.

Peut-être est-il utile de préciser d'entrée de jeu ce qui distingue la textométrie[3] au sein des approches outillées des textes numériques, pour comprendre en quoi elle apparaît particulièrement appropriée. La textométrie offre des visualisations synthétiques, mais elles ne visent pas à remplacer la lecture du texte, bien plutôt à la relancer, selon une dialectique entre quantitatif et qualitatif, résultats de calculs et retour au texte, *distant reading* et *close reading* (voir par exemple HEIDEN, 2004). Elle donne les moyens de rechercher systématiquement les occurrences d'un mot ou d'une formulation, et de confirmer ainsi factuellement des hypothèses que l'on avait explicitées : c'est un apport que l'on apprécie déjà avec les éditions numériques, et dont la textométrie peut donner une version avancée (corpus unifié et rendant compte d'une intertextualité *vs* multiples recherches sur des textes ou chapitres isolés ; recherche de motifs complexes en tenant compte d'informations morphosyntaxiques ou structurelles des textes *vs* recherche de chaînes de caractères). La textométrie offre de plus des outils

[2] Une présentation rédigée et organisée (comme (HÉBERT, 2001) ; voir aussi les introductions à la sémantique interprétative publiées dans la rubrique *Repères pour l'étude* du site Texto!) est un point d'entrée privilégié permettant de poser les grands repères et d'introduire les concepts-clés ; le prolongement peut ensuite être une rencontre directe avec les écrits de Rastier eux-mêmes, un retour à leur lecture, tant traditionnelle qu'assistée par des outils de recherche et d'exploration.

[3] L'ouvrage de référence concernant la textométrie (aussi appelée lexicométrie, logométrie, ou analyse des données textuelles) est le LEBART & SALEM (1994). (NÉE, 2017) offre une introduction pédagogique dans le contexte de l'analyse de discours, et (LEBART *et al*., 2019) propose une synthèse actualisée. Complémentairement, le chapitre (MAYAFFRE, 2007b) introduit aux concepts centraux (corpus, unités, traitements quantitatifs et qualitatifs) et aux caractéristiques de la lecture textométrique (méthodique, heuristique).

heuristiques (formes de présentation du texte, tris) et des calculs statistiques, pour mettre en évidence des régularités ou singularités inaperçues (MAYAFFRE, 2007b) et susciter de nouveaux observables (RASTIER, 2011, 2020). Enfin, elle ne vise pas à automatiser un mode d'analyse des textes avec un objectif de précision et de performance, car elle ne prévoit pas une lecture unique avec l'extraction « du » sens, mais elle confie la conduite de l'analyse et l'élaboration de l'interprétation au chercheur. En tout cela donc, elle se démarque du *text mining*, des éditions numériques standard, comme du traitement automatique des langues ou du web sémantique (BÉNEL, 2017), avec un positionnement et des objectifs différents, globalement plus proches de lectures scientifiques ou d'activités pédagogiques. Attentive au texte et ouverte au travail interprétatif, elle participe à une philologie numérique et une herméneutique numérique (MAYAFFRE, 2007a).

1.2 Caractéristiques pertinentes du corpus des écrits linguistiques de François Rastier

Un certain nombre de caractéristiques de l'œuvre scientifique de François Rastier encouragent une telle approche.

Une première caractéristique assez évidente est sa volumétrie : l'œuvre est prolixe, mais on peut en avoir une connaissance d'ensemble. La totalité des écrits scientifiques de François Rastier, qui a commencé à publier à la fin des années 1960 et n'a cessé d'écrire sur la cinquantaine d'années écoulées, représente actuellement quelques 600 articles ou chapitres d'ouvrages, et 17 ouvrages. Ces écrits explorent cependant différentes problématiques, développant parfois une relative autonomie au sein d'une pensée unifiée : des questions réactivées par les dynamiques des recherches contemporaines — études **sémiotiques**, sciences **cognitives**, linguistique de **corpus numériques**, croisant un engouement général pour les **ontologies** — par rapport auxquelles peuvent s'élaborer les propositions de la **sémantique interprétative** ; la relativisation épistémologique des frontières disciplinaires et un programme unifiant les **sciences de la culture**, intéressant tant la recherche que l'**éducation** ; la relecture de **Saussure** suite à la découverte de ses manuscrits et de leurs écarts au célèbre *Cours de linguistique générale* édité par les élèves, les questions de l'écriture et des relations entre les œuvres — **création**, **traduction**, **transmission**… —, une approche

linguistique méthodique de **textes littéraires** ; une attention particulière à l'écriture de témoins et survivants de l'**extermination** (en particulier **Primo Levi**), et en contrepoint l'analyse de la production de penseurs dont on peut mettre en évidence le lien à ces idéologies destructrices (**Heidegger**), la caractérisation de l'expression de la xénophobie sur le Web (projet Princip.net de filtrage de sites **racistes**), et une dénonciation critique de mouvements contemporains traversant le monde scientifique et la société — **déconstruction**, décolonialisme, *gender studies*, ***cancel culture***. Comme son nom l'indique, le projet Sittelle voudrait se centrer sur la sémantique interprétative et la sémantique des textes en corpus (sans exclure la perspective de s'ouvrir à d'autres auteurs), plutôt que sur la pensée globale de François Rastier. Le corpus envisagé pour Sittelle ne vise donc pas l'ensemble de la bibliographie de Rastier, mais la collection des écrits les plus linguistiques. Bien évidemment, la délimitation du corpus procédera de l'ajustement d'un seuil entre le « suffisamment » et le « pas assez » linguistique, ou plutôt, si l'on pense la construction du corpus comme une dynamique, des priorités sur les textes à intégrer. Si l'on procède à une estimation sur la base des ouvrages, un premier cercle prioritaire pourrait concerner *Sémantique interprétative* (1987), *Sémantique pour l'analyse* (1994), *Arts et sciences du texte* (2001), *La mesure et le grain* (2011). Mais on pourra trouver des développements complémentaires dans la perspective littéraire de *Sens et textualité* (1989) et *Mondes à l'envers* (2018), dans la discussion de sémantiques ontologiques dans *Sémantique et recherches cognitives* (1991) et *Faire sens* (2018), dans les croisements avec la pensée saussurienne (*Saussure au futur*, 2015), dans les études sémiotiques préalables d'*Idéologie et théorie des signes* (1971) et des *Essais de sémiotique discursive* (1973/1974) etc. Á première vue le corpus Sittelle pourrait donc viser à rassembler une centaine de textes (de type article ou chapitre). Et pour parcourir méthodiquement une centaine de textes (voire quelques centaines si besoin), l'outil textométrique semble bienvenu.

Une seconde caractéristique pertinente du corpus Rastier est son unité d'écriture et de pensée, qui donne une cohérence et une valeur interne à l'ensemble. Ce qui fait la difficulté à délimiter le corpus, fait aussi son intérêt : résonances, reprises, affinités dans les concepts mobilisés, la façon de penser, celle de s'exprimer, à travers les différentes questions abordées, les différents terrains étudiés. Cela nous met dans des conditions favorables pour des observations en partie basées sur les mots : le sens d'un mot peut

bien sûr varier en fonction de ses contextes (afférence), mais, sauf cas d'homonymie, on peut faire l'hypothèse d'une certaine stabilité ou continuité de sens entre les occurrences (inhérence), et de la pertinence du rapprochement de leurs contextes pour enrichir leur description sémantique. C'est le principe herméneutique des « passages parallèles », applicable au sein d'un texte ou d'une intertextualité construite et interprétée.

Une troisième caractéristique à évoquer est peut-être moins évidente, mais non moins déterminante pour notre projet. Avec l'œuvre de François Rastier, nous avons affaire à un style d'écriture fonctionnant davantage par évocation, rapprochements, reprises, que par organisation progressive et hiérarchique, formelle et explicite. La liberté de pensée de l'auteur se coule dans un propos souple, structuré sans être piloté par un cadre préétabli. Notre auteur lui-même a décrit les activités d'expression et d'interprétation comme des mécanismes d'ajustement, de correction, d'adaptation (RASTIER, 2001, p.49-50 ; RASTIER, 2011, p. 61). Nous avons des attentes (expressives ou interprétatives) et le sens se construit dans une réélaboration, un ajustement progressif des contenus. Ce serait à l'inverse une conception ontologique de la sémantique, à l'opposé de la praxéologie affirmée par la sémantique interprétative, qui permettrait de penser qu'une « chose » (idée, concept) puisse être entièrement exprimée ou comprise en une seule énonciation, et que l'avancée dans un texte serait une accumulation ordonnée de contenu. Le renouvellement de la pensée et l'enrichissement du sens se font à la croisée d'une expression multiple — ni unique et définitive, ni strictement positionnée et arrêtée dans un système statique qu'il suffirait de déployer. Et c'est justement parce que la lecture de notre corpus ne se satisfait pas d'un parcours linéaire, mais s'enrichit des reprises intertextuelles, que les navigations transverses permises par l'outil textométrique nous semblent un apport majeur, pour amplifier et préciser le travail herméneutique que réalise déjà le lecteur avec sa mémoire. Les formulations locales, ou l'entrée dans l'univers de la théorie sémantique de Rastier, reçoivent un éclairage par l'accès à des repères globaux.

1.3 Remarques

Il ressort déjà que l'approche textométrique opère une forme de mise en abyme de la sémantique interprétative : que le global détermine le local est un principe de sémantique interprétative, mais devient aussi ici un mode d'accès outillé à

l'interprétation du corpus des écrits de Rastier. De fait, les affinités de la théorie de la sémantique interprétative de Rastier d'une part, et de la méthodologie textométrique d'autre part, sont multiples et importantes (PINCEMIN, 2010).

Une telle initiative est également rendue possible par la générosité et l'ouverture d'esprit de notre auteur, qui non seulement accepte de nous accompagner et soutenir dans ce travail (accès à des documents numériques sources, partage de droits, conseils sur la composition du corpus, retours sur les observations), mais aussi consent à exposer ses écrits aux mécanismes crus et potentiellement brutaux des traitements formels, ne laissant échapper aucun détail de la matérialité linguistique des textes, ce qui suppose des usages éclairés et bienveillants, que nous devrons favoriser mais ne pourrons garantir. Ce type de publication est profondément altruiste, bénéficiant avant tout aux étudiants et aux chercheurs.

2 Perspective méthodologique

Concrètement, quels nouveaux modes de parcours peuvent être envisagés avec un environnement textométrique du corpus ? Nous esquissons une première palette de types d'interrogation[4], sur la base d'un corpus illustratif dit « Zéro », constitué rapidement à titre exploratoire pour un premier retour expérimental. Tous les traitements sont effectués avec deux logiciels open-source : le logiciel TXM 0.8.1 (sections 2.2 à 2.6) (HEIDEN *et al*., 2010), et le logiciel IRaMuTeQ 0.7 alpha 3 (section 2.7, sur la méthode Reinert) (RATINEAU & DEJEAN, 2009 ; RATINEAU, 2018).

2.1 Corpus Zéro illustratif

Le corpus Zéro est composé de textes mis en ligne sur le site Texto![5], diffusés au format HTML[6], et développant des considérations linguistiques[7]. Il rassemble 28 textes,

[4] Pour une vue plus complète des types d'analyses textométriques, on pourra consulter (PINCEMIN *et al*., 2010).

[5] Revue électronique *Texto! Textes & Cultures*, http://www.revue-texto.net, ISSN 1773-0120. Il s'agit de textes que François Rastier a choisi de partager avec un lectorat élargi, et pour lesquels il y a la possibilité de le faire (des considérations de droits l'empêchent pour ses ouvrages récents par exemple).

[6] La diffusion des textes sous forme de page HTML a été remplacée au fil du temps par des téléchargements de PDF. L'exploitation numérique de ceux-ci était trop lourde pour la présente expérimentation : pour ces textes on envisagera de préférence de partir d'un fichier auteur de type traitement de texte, et le travail de préparation sera investi pour l'édition XML TEI. Notre corpus Zéro se limite donc à des textes relativement anciens (1992-2006).

et représente une taille de 320 000 mots (pour le dimensionner selon les unités usuelles de la linguistique de corpus).

Le détail de la composition du corpus et des opérations de préparation pour l'édition textométrique est donné en annexe 1. Pour ce premier aperçu en début de projet, nous avons eu recours à une représentation des textes très sommaire et simplifiée, au format texte brut (.txt). L'édition XML TEI sera plus longue à réaliser mais élargira et affinera les possibilités de traitements textométriques. Dans un premier temps, l'expérimentation sur corpus texte brut peut déjà éclairer sur des éléments qui seront importants à prendre en compte dans les choix de structuration TEI.

Lors de l'import dans TXM, un étiquetage morphosyntaxique par le logiciel TreeTagger (SCHMID, 1994) est effectué à la volée. Il associe à chaque mot son lemme (sa forme non fléchie telle qu'on la trouve en entrée dans un dictionnaire) et sa catégorie grammaticale, selon le jeu d'étiquettes du modèle français général construit par Achim Stein diffusé sur le site du logiciel. Ces informations linguistiques peuvent être mises à profit dans la formulation des recherches, comme le montreront nos exemples.

2.2 Concordance

La concordance recueille et présente de façon synthétique l'ensemble des contextes d'emploi d'un mot donné (ou d'un motif de recherche quelconque). Dans un outil de textométrie comme TXM, l'affichage des résultats permet de visualiser efficacement les reprises et variantes, grâce à une présentation en tableau qui aligne et superpose les contextes (figure 1) et à de multiples possibilités de tri adaptées à la nature linguistique des données. Avec ce dispositif d'affichage heuristique, l'exploration des contextes va bien au-delà des possibilités de navigation offertes par des moteurs de recherche des outils de bureautique généralistes, sans compter l'intérêt aussi de procéder à des recherches balayant tout un intertexte, et non cantonnées à chaque texte-fichier.

Les concordances nous semblent une implémentation particulièrement efficace de la technique herméneutique des passages parallèles, si éclairante pour étudier le sens

[7] Il est bien évident que ce dernier critère est tout-à-fait discutable dans le détail des choix opérés ; mais rappelons que l'enjeu est un état « zéro », de première exploration rapide, sans engagement scientifique, afin de préciser ensuite les jalons pour un premier corpus.

d'un mot ou d'une formulation à partir d'une étude systématique de ses emplois (PINCEMIN, 2007a). Elles sont de fonctionnement simple et offrent un nouveau point d'entrée dans l'œuvre — par mot ou motif —, particulièrement intéressant aussi au plan didactique.

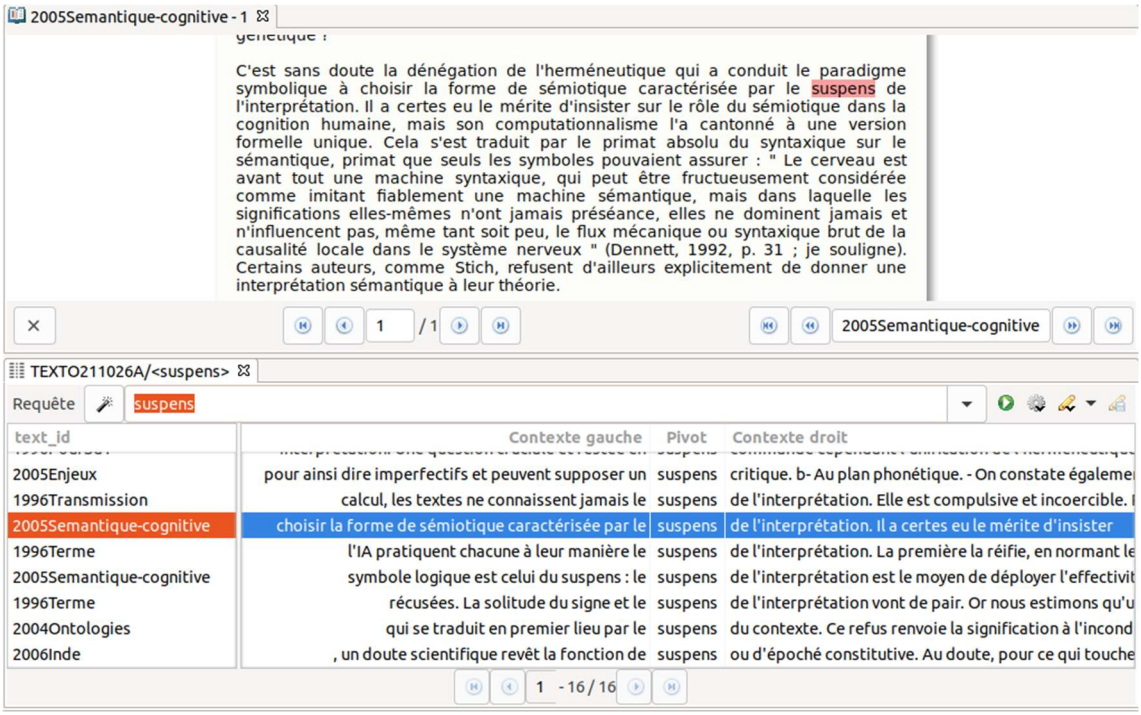


Figure 1 *: Concordance du mot « suspens », triée sur le contexte droit.*

Un premier exemple est donné par une concordance sur le mot « suspens » (figure 1). Toutes les occurrences du mot sont rassemblées dans le tableau des résultats, avec leur contexte immédiat : soit 16 lignes pour 16 emplois du mot (décompte donné en-dessous du tableau). La localisation des passages correspondants est donnée en colonne de gauche : tel que nous avons préparé le corpus, cette localisation indique le code du texte concerné (les indications pourraient avoir une autre forme ou être plus précises, par exemple avec une indication de pagination, si les données sont préparées en conséquence). Ces occurrences sont initialement présentées au fil des textes et dans l'ordre de constitution du corpus (ici l'ordre alphabétique des noms donnés aux textes), mais pour un travail analytique on peut trier ces extraits en fonction des mots qui suivent (ou qui précèdent) notre pivot « suspens » : ainsi nous rapprochons tous les passages évoquant le « suspens de l'interprétation ». Pour bien comprendre ces extraits, nous pouvons avoir besoin d'un contexte élargi, auquel nous avons accès en double-

cliquant sur la ligne considérée : s'affiche alors l'édition complète du texte, positionnée au niveau du passage, et avec le terme cherché mis en évidence.

Le même mécanisme peut être étendu à des recherches de motifs plus complexes. Par exemple, le langage de recherche nous permet de demander tous les passages mentionnant « global » et « local » à moins de 10 mots d'écart.[8] Nous en trouvons 68 dans notre corpus Zéro, et en triant sur le pivot nous pouvons dégager les principales formulations associées[9] : « le global détermine le local » (6 occ. dans 6 textes différents), « la détermination du local par le global » (4 occ.), « la détermination/l'incidence du global sur le local » (6 occ.), « (l'accès/le rapport) du global au local » (7 occ.), soit donc quatre formules qui résument plus d'un tiers des mentions associant global et local. L'examen des contextes fait également bien apparaître que cette détermination du local par le global a valeur de « principe herméneutique » (1996Themes, 2003Semiotique-ontologie, 2005Microsemantique, 2005Mesosemantique).

2.3 Vue d'ensemble de la localisation d'occurrences

La concordance, en indiquant pour chaque contexte où il est situé, permet déjà de se rendre compte par exemple si les occurrences trouvées sont surtout groupées dans un ou deux textes, ou sont réparties tout au fil du corpus. Ce relevé est complet et précis, mais on peut avoir une restitution complémentaire et plus visuelle avec des fonctionnalités dédiées à cette question de la disposition des occurrences.

Prenons le cas de la notion d'isosémie[10]. Une concordance relève 40 occurrences du terme : 39 dans le texte sur la mésosémantique, et une dans celui sur Borges — occurrence isolée avec un contexte de rappel définitoire synthétique, « les isosémies ou accords syntaxiques en genre et nombre ». Pour ce qui concerne les 39 occurrences du même texte, une visualisation nous permet de comprendre qu'elles sont concentrées dans une section du texte (figure 2). Le déroulement du texte est représenté par l'axe des

[8] Requête CQL utilisée : `([word="loca.*"%c] []* [word="globa.*"%c]) | ([word="globa.*"%c] []* [word="loca.*"%c]) within 10`

[9] Entrent aussi dans les décomptes quelques variantes gardant la même structure syntagmatique, par exemple « le global y détermine le local » ou « Déterminations du global (corpus, texte) sur le local (signe) ».

[10] Requête CQL utilisée : `"isosémies ?"%c`

abscisses (les nombres correspondent aux positions des mots dans la succession des textes du corpus), et chaque « marche » traduit une occurrence du terme. Ainsi, l'ascension relativement raide dans le 1er tiers du texte couvre de la section 3.1 (*Les relations de concordance/La construction des fonds sémantiques*) à la section 4.3. (*Exemples et problèmes de description*). La section suivante (5) est consacrée aux formes sémantiques, sans plus aucune occurrence d'« isosémie » (la courbe présente un long palier plat). L'occurrence isolée à la fin du texte est en fait dans une note appelée en section 6.

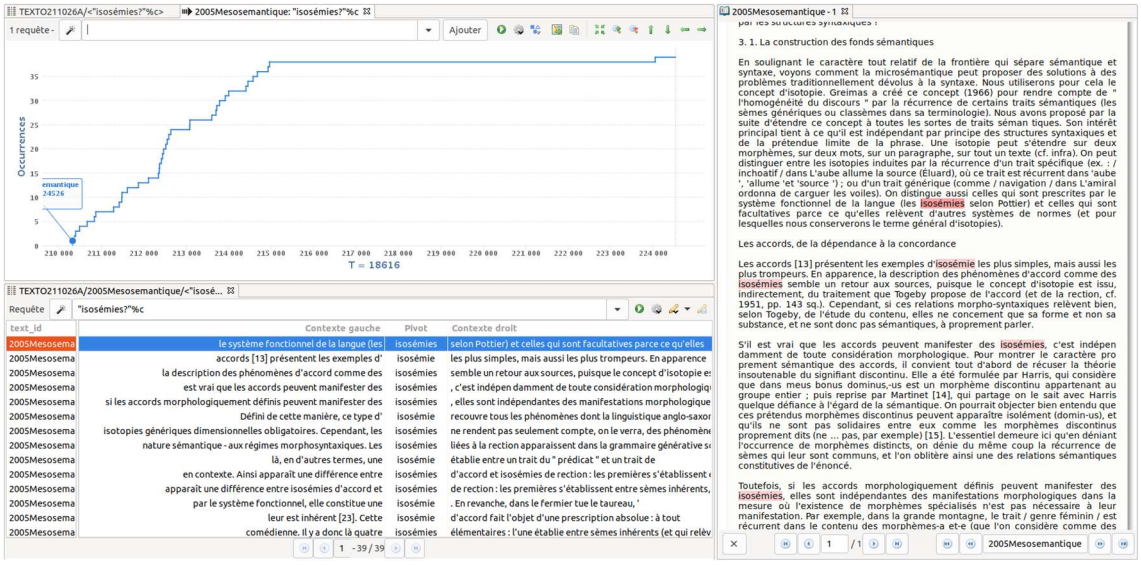


***Figure 2** : Graphique de progression (en haut à gauche) des occurrences d' « isosémie(s) » dans le texte 2005Mesosemantique.*

2.4 Inventaires lexicaux

Les fonctionnalités d'interrogation textométriques permettent également de dresser des inventaires généraux du vocabulaire ou d'observer comment se réalise un motif donné.

Pour illustrer le premier cas de figure (inventaires généraux), nous avons demandé les 20 noms, adjectifs et verbes les plus fréquents de notre corpus Zéro.[11] Cela

[11] Les requêtes CQL utilisées sont : `[frpos="NOM"]`, `[frpos="ADJ"]`, `[frpos="VER.*"]`. Les inventaires sont réalisés en lemmes (propriété `frlemma`), autrement dit les 1420 occurrences de « texte » comptabilisent aussi bien des singuliers et des pluriels, ainsi que des occurrences avec une majuscule comme en début de phrase ; et les 1404 occurrences de « pouvoir » sont reconnues derrière leurs variations flexionnelles (« peut », « peuvent », « pourrait », « pu », « puisse » etc.) sans retenir les

produit des listes de mots hors contexte, mais le logiciel prend soin de prévoir un accès direct (par lien hypertexte) à la visualisation des contextes d'emploi (en concordance). On pourra par exemple vérifier que l'essentiel des 655 occurrences de « genre(s) » renvoient à la notion de genre textuel (avec seulement une dizaine d'occurrences de la locution « ce genre de » et quelques mentions du genre grammatical (masculin/féminin)).

Tableau 1 *: Les 20 noms communs, adjectifs qualificatifs et verbes les plus fréquents du corpus Zéro (avec indication de leur fréquence).*[12]

Noms		***Adjectifs qualificatifs***		***Verbes***	
texte	1420	sémantique	794	être	5469
exemple	814	même	729	avoir	2118
genre	655	autre	658	pouvoir	1404
sens	596	linguistique	495	faire	572
forme	594	textuel	259	permettre	464
langue	586	cognitif	235	rester	427
discours	524	scientifique	233	définir	421
mot	516	interprétatif	230	devoir	350
langage	459	générique	218	relever	288
sémantique	457	général	214	dire	280
théorie	433	divers	202	décrire	262
relation	412	sémiotique	202	trouver	236
corpus	398	propre	195	considérer	217
unité	388	social	191	distinguer	217
signe	362	nouveau	175	falloir	217
sème	349	herméneutique	171	mettre	211
contexte	344	grand	165	prendre	202
linguistique	342	logique	165	constituer	200
concept	341	différent	164	opposer	189
objet	334	humain	155	représenter	182

Ce genre de listes va à la fois confirmer ce dont l'expert se doute déjà (comme la dominance de « texte » et « textuel », et celle de « sémantique » ; et du côté des verbes, la préséance des auxiliaires et des modaux), mais peut aussi rappeler l'importance (*a minima* de simple présence quantitative) d'une notion (avait-on conscience de la place prise par l'« exemple » dans l'écrit scientifique de Rastier), voire peut attirer l'attention sur des points d'entrée qu'on n'attendait pas en si bonne place (ici peut-être les

occurrences nominales de « pouvoir » comme dans « l'hypallage adjectivale a le pouvoir exorbitant de subvertir l'adjectif de nature » (2006Borges) — dans les limites de l'exactitude de l'analyse morphosyntaxique automatique.

[12] Nous avons rectifié les sorties brutes en retirant de la liste des noms les abréviations « cf » et « p », et en réécrivant « sens » le lemme étiqueté de façon inutilement complexe « sen|sens ».

considérations « sociales » qui de fait cumulent les « sciences sociales » (42 occ.), les « pratique(s) sociale(s) » (66), les « norme(s) sociale(s) » (10), la « demande sociale » (8), mais s'étendent aussi à près d'une quarantaine d'autres noms).

On peut en fait dresser des listes de toutes sortes de motifs linguistiques, plus ou moins complexes et plus ou moins ciblés. C'est au chercheur de définir quel « filet » il veut jeter sur le corpus. Ici par exemple, nous pouvons être intéressés d'inventorier des expressions fréquentes pouvant faire terme (avec un patron grammatical de type Nom+Adjectif)[13], ou encore de lister des groupes nominaux formés avec la tête nominale « linguistique » ou « sémantique »[14] (Tableau 2).

Nous pouvons aussi travailler au plan morphologique et explorer les variantes dérivationnelles d'« isotopie »[15] attestées dans le corpus, pour tirer des observations telles que : l'usage dominant concerne le nom (« isotopie(s) ») ; et l'on a par ailleurs de rares adjectifs, « isotope(s) », signifiant en relation d'isotopie (« être isotope », « les contextes isotopes d'un sémème »), et « isotopique(s) », relatif à l'isotopie (« analyse/typologie/faisceaux isotopique(s) »), mais aucun « isotopant(e)(s) » par exemple dans notre corpus Zéro.

[13] Requête CQL utilisée : `[frpos="NOM"][frpos="ADJ"]`. Propriété d'analyse : `frlemma`.
[14] Requête CQL utilisée : `[frlemma="(séman|linguis)tique"%c & frpos="NOM"] ([frpos="ADJ"] | [frpos="PRP.*"] [frpos="DET.*"]? [frpos="NOM"])`
[15] Requête CQL utilisée : `"isotop.*"%c`

***Tableau 2** : Relevés des réalisations de patrons définis par différentes informations linguistiques (morphologiques, lexicales, syntagmatiques).*

Séquences Nom+Adj (fréq. ≥ 2_)	
parcours interprétatif	97
forme sémantique	96
molécule sémique	89
sémantique cognitive	81
champ générique	79
pratique sociale	66
isotopie générique	51
texte scientifique	45
positivisme logique	44
science sociale	42
trait sémantique	38
impression référentielle	37
fond sémantique	34
nom propre	34
relation sémantique	32
traitement automatique	32
discours scientifique	31
sème générique	31
sème inhérent	31
sens littéral	31
unité sémantique	31
roman policier	30
classe lexicale	29
domaine sémantique	29
forme textuelle	29
sème afférent	29
perception sémantique	28
problématique logico-grammaticale	28

sémantique/linguistique + Adj./ Cpl. de nom	
sémantique cognitive	78
sémantique des textes	51
linguistique de corpus	45
sémantique interprétative	20
linguistique historique	17
linguistique générale	15
sémantique linguistique	14
sémantique structurale	14
sémantique lexicale	13
linguistique cognitive	10
sémantique formelle	10
linguistique textuelle	9
sémantique différentielle	9
linguistique du signe	7
linguistique de la parole	6
linguistique des textes	6
sémantique diachronique	6
sémantique textuelle	6
linguistique du texte	5
linguistique saussurienne	5
sémantique procédurale	5

Mots commençant par « isotop »	
isotopie(s)	241
isotope(s)	4
isotopique(s)	3

2.5 Cooccurrences

Le calcul de cooccurrences peut être utile pour opérer une synthèse statistique des contextes d'un mot (ou d'une expression, d'un motif), surtout lorsqu'il est très fréquent et que cela peut aider à avoir une vue d'ensemble. Si par exemple je demande les cooccurrents d'isosémie(s), la liste des résultats du calcul (figure 3) me rappellera un certain nombre de notions associées : isotopie, accord, rection, facultatif/obligatoire, afférent/inhérent, etc.

Illustrons encore cette fonctionnalité avec la recherche des cooccurrents de mots de la famille d'« interprétation » (« interprétation », « interprétatif », « interpréter », etc.) (figure 3). Les résultats rappellent à mon attention la mise en correspondance de

l'interprétation avec la production (déclinée dans les différentes catégories grammaticales : « production », « productif », « produire »), et en allant voir les contextes on pourra confirmer que le propos associe une trentaine de fois « la production ou/et l'interprétation » des textes.

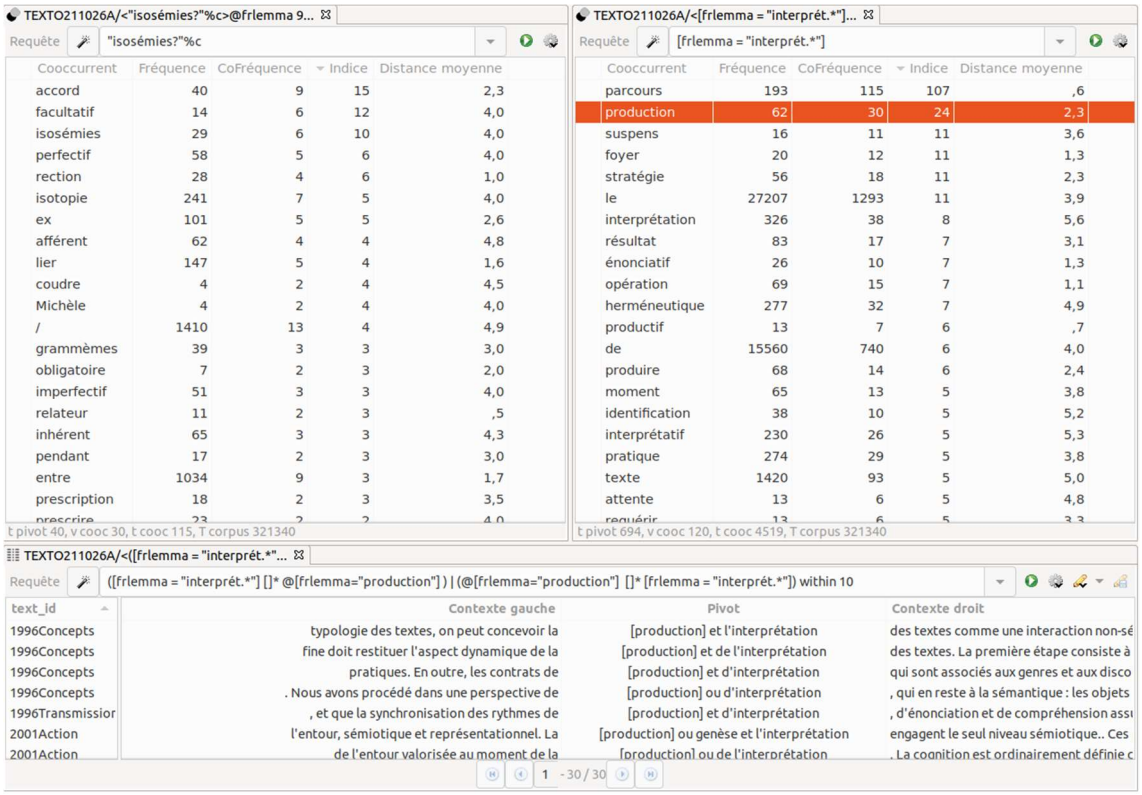


***Figure 3** : Cooccurrents d'« isosémie » (à gauche) ou des mots de la famille d'« interprétation » (à droite).[16] En bas, présentation en concordance des contextes de « production » avec un mot de la famille d'« interprétation », calculée automatiquement par double-clic sur la ligne du cooccurrent (« production »).*

2.6 Cartographie et caractérisation intertextuelle

Du côté des synthèses statistiques, les analyses factorielles sont en mesure de produire des vues d'ensemble d'une collection de textes, sous forme de cartographies, faisant ressortir les principaux contrastes structurant la collection considérée. Selon l'ensemble de textes considéré, on pourrait mettre au jour des déploiements plus chronologiques ou plus thématiques par exemple, ou encore des oppositions de genres

[16] Les paramètres du calcul sont ceux proposés par défaut (notamment le voisinage utilisé : distance de 0 à 9 mots à gauche comme à droite), sauf la propriété des cooccurrents, réglée en lemme (`frlemma`).

textuels. Complémentairement, le calcul des spécificités apporte une caractérisation lexicale de chaque texte, par exemple utile pour guider des choix de lecture: « ce texte-là semble particulièrement aborder ces notions-là ».

Ces calculs d'analyse des données et de statistique, qui apportent une vue d'ensemble sur le corpus entier, sont complémentaires de la lecture d'étude traditionnelle, ainsi que le souligne Étienne Brunet à l'occasion d'un travail sur le temps dans un corpus de littérature française :

> […] l'ordinateur échappe aux pesanteurs du temps qui entravent la mémoire humaine et qui imposent aux souvenirs une perspective fuyante, où les éléments les plus disponibles sont ceux que fournit l'environnement présent. Là où le lecteur parcourt en marchant l'espace littéraire, dans la succession changeante et l'effacement progressif des paysages, l'ordinateur saisit d'un coup le même espace, comme on lit une carte géographique ou stratégique, tous les points étant à plat, offerts à l'œil en même temps. En réalité les textes, si écrasés qu'ils soient par la perspective plongeante d'un observateur posté sur Sirius, acquièrent une lisibilité qu'ils n'ont pas pour l'explorateur engagé dans le maquis de la lecture. Au ras du sol, en enjambant les ruisseaux, on peut difficilement délimiter la ligne de partage des eaux. Mais d'en haut le paysage littéraire se découvre avec l'orientation des chaînes, les pentes, les ruptures et tous les mouvements de terrain produits par l'histoire. (BRUNET, 2016, p. 371)[17]

Cette fonctionnalité va être intéressante ici pour nous permettre d'apprécier la composition de notre corpus et ses équilibres internes. Pour la construction de cette représentation, nous avons choisi de représenter les textes par leur profil d'usage d'une centaine de noms et adjectifs les plus fréquents[18] (figure 4).

[17] L'article dont est tiré cette citation est également un chapitre numérique de (BRUNET, 2016) disponible sur *Texto!* (http://www.revue-texto.net/index.php?id=3756) et sur l'archive HAL (https://halshs.archives-ouvertes.fr/halshs-01275527v1).

[18] La sélection des mots a été opérée avec la requête `[frpos="NOM|ADJ"]`, en `frlemma`, avec une fréquence minimale de 150. Pour notre objectif, on a écarté de cette sélection 16 mots moins thématiques, notamment davantage employés dans des locutions figées : autre, caractère, cas (c'est/ce n'est pas le cas, dans le premier/meilleur/tous les cas), cf, compte (rendre compte, tenir compte), divers, doute (sans doute), effet (en effet), fait (de fait, en fait, tout à fait, de ce fait, du fait), même, p (mention de pagination), part (d'une/d'autre part, (mis) à part, pour sa/notre/leur part), propre (propre à/aux, nom propre), sein (au sein de, en son/leur sein), seul, sorte (en quelque sorte, toutes sortes de, deux/trois/quatre/... sortes). Il reste alors 92 noms ou adjectifs fréquents pour la caractérisation des textes.

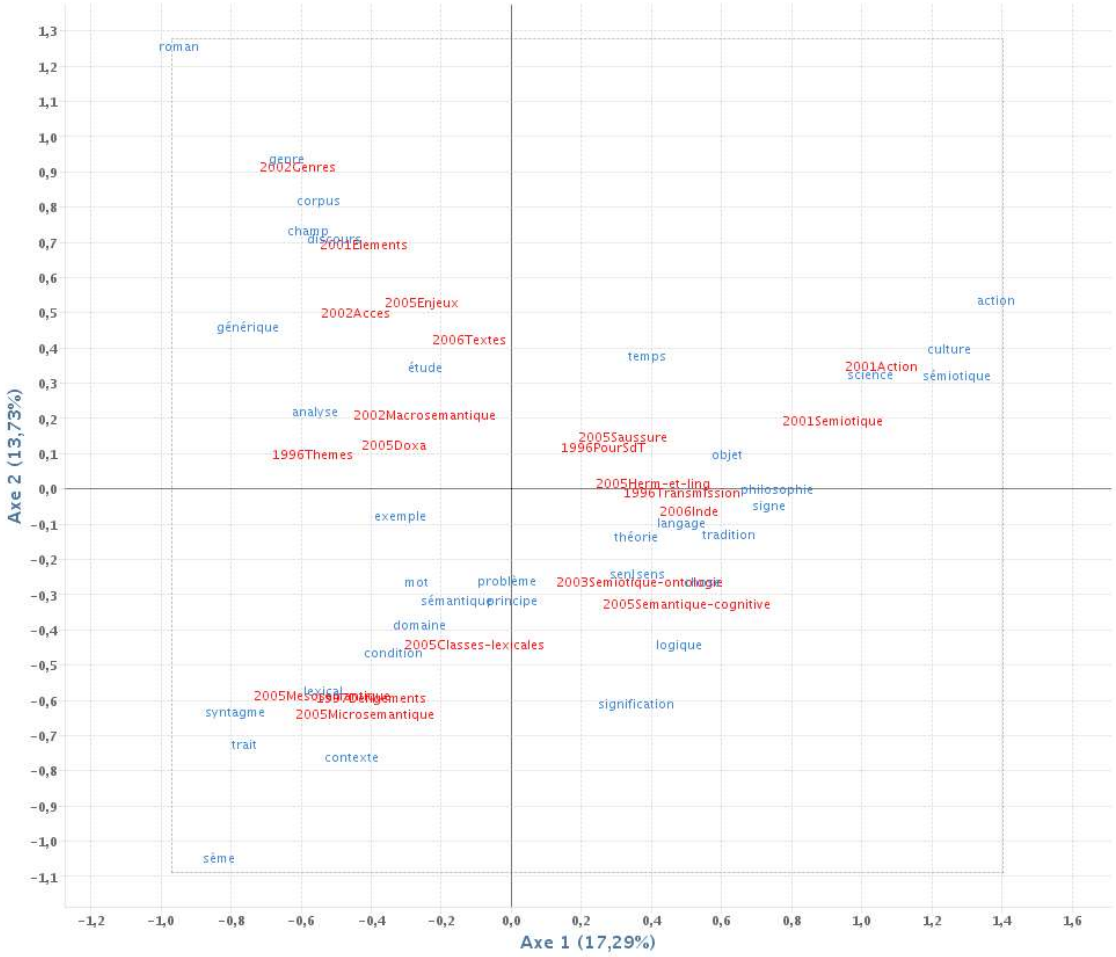


Figure 4 *: Cartographie par analyse factorielle des correspondances des 28 textes du corpus Zéro, caractérisés par 96 noms ou adjectifs les plus fréquents.*[19]

Notre corpus Zéro apparaît se déployer dans trois grandes directions : les descriptions lexicales de la sémantique interprétative (avec les sèmes, les classes, etc.) se distinguent des études plus textuelles et en corpus, et cet ensemble de travaux peut-être plus appliqués et descriptifs contraste avec des développements épistémologiques du côté de la sémiotique et des réflexions sur le signe, notamment en lien avec les sciences cognitives mais aussi les études saussuriennes.

Mettons que je veuille repérer le vocabulaire qui caractérise un texte de façon différentielle, **par contraste avec l'ensemble du corpus** : le calcul des spécificités va

[19] La représentation a par ailleurs été allégée et clarifiée en filtrant les points (mot ou texte) les moins utiles : d'une part mal représentés ($cos^2 < 0,3$) dans cette projection plane, optimisée mais réduite (information de variation conservée = 17,29 + 13,73 = 31,02 %), et également d'autre part peu impliqués dans les grandes variations décrites dans ces deux premières dimensions (ctrb1 < 1 % et ctrb2 < 1 %).

repérer les mots particulièrement surreprésentés dans le texte choisi (ou plus généralement dans une partie quelconque définie dans le corpus : un sous-ensemble de textes, un type de parties, etc.).[20]

Tableau 3 *: Spécificités du texte* 1996Transmission *dans le corpus Zéro, en considérant l'ensemble des mots du corpus.*

Lemme	***Fréquence totale dans le corpus Zéro***	***Fréquence dans le texte 1996Transmission***	***Indice de spécificité***
communication	108	71	70,6
transmission	53	33	32,1
commentaire	38	27	28,8
traduction	63	33	28,8
message	28	23	27,2
code	38	23	22,1
transcodage	18	16	20,2
information	59	25	19,0
modèle	137	33	16,1
translation	10	10	13,9
transmettre	22	12	11,1
interprétation	326	43	10,9
le	27207	1308	10,6
émetteur[21]	20	11	10,3
métalangage	27	12	9,7
tradition	289	35	8,0

Considérons par exemple le texte 1996Transmission (*Communication ou transmission ?*), le tableau 3 ci-dessus donne les mots spécifiques de ce texte avec un score statistique de 8 ou plus. On pourra observer que le calcul statistique relativise la forte proportion des occurrences d'un mot dans un texte en considérant aussi sa fréquence : par exemple, il juge moins remarquable l'exclusivité d'emploi de « translation » (10 occ.) que la concentration de 71 des 108 occurrences de

[20] Le calcul modélise mathématiquement la situation suivante : étant donnée la fréquence totale du mot (s'agit-il d'un mot plutôt fréquent ou plutôt rare) et la taille du texte considéré dans le corpus (quelle proportion du corpus représente-il), quelle serait la probabilité d'observer le mot avec une fréquence aussi élevée si les mots étaient répartis au hasard. Autrement dit, un indice statistique de 5 par exemple, qui traduit qu'on n'aurait qu'une chance sur 100 000 (indice de 5 → 1 suivi de 5 zéros → 1 / 100 000) d'avoir le mot avec une telle fréquence dans le texte, signale le caractère « anormalement » élevé de la fréquence, qui semble par conséquent relever d'un choix particulier de ce texte par rapport à l'ensemble du corpus.

[21] La sortie brute du logiciel proposait « Emetteur ». Pour la lisibilité des résultats, nous l'avons remplacé par « émetteur », défini comme lemme réunissant les 6 occurrences d'« Emetteur » et les 14 d'« émetteur » (à défaut que cela ait été opéré traitement automatique), et nous avons mis à jour les fréquences et le score statistique dans le tableau.

« communication », qui aurait été beaucoup moins possible au hasard. Il permet ainsi d'attirer l'attention sur des cas statistiquement remarquables mais qui auraient pu passer inaperçus à l'œil nu, ou avec une simple règle de trois.

Les spécificités peuvent également être lues du point de vue des mots plutôt que de celui des textes : autrement dit, pour un mot donné, quel est son profil quantitatif (statistique) d'usage par rapport à l'ensemble des textes ? Considérons par exemple la présence plus ou moins forte (et statistiquement remarquable) des trois mots suivants au sein des 28 textes du corpus Zéro : « corpus », « données », « traitements »[22]. Un diagramme de spécificités nous aide à repérer les quelques textes qui semblent développer et allier plusieurs de ces notions : « corpus » et « données » dans *1996Themes*, *2002Genres*, *2005Enjeux* ; « corpus » et « traitements » dans *2002Acces*.

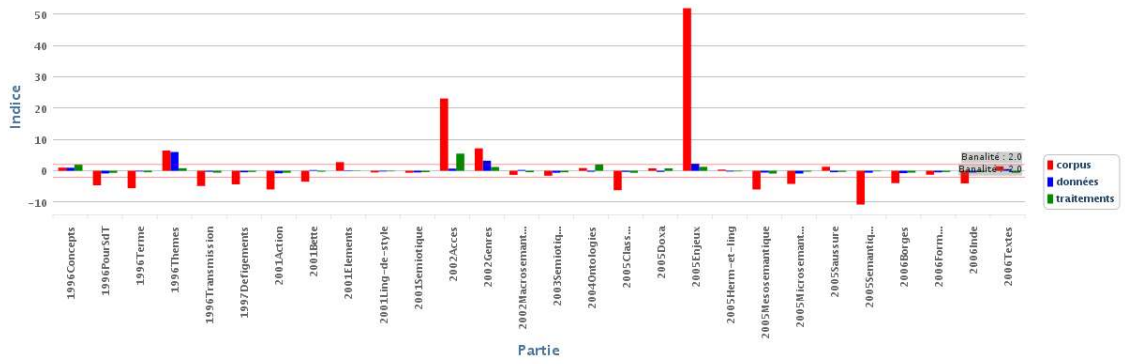


Figure 5 *: Diagramme de spécificités des mots « corpus », « données » et « traitements » sur les 28 textes du corpus Zéro.*

2.7 Méthode Reinert et isotopies

Nous aimerions aussi expérimenter la méthode Reinert (REINERT, 1990 ; RATINAUD, 2018) (algorithme précurseur et proche des *Topic models*) pour évaluer sa capacité à esquisser des isotopies. En effet, le calcul définit des ensembles de mots particulièrement présents dans des fragments textuels qui ont tendance à se ressembler lexicalement par opposition au reste des textes : ces mots pourraient-ils actualiser des sèmes d'isotopies particulièrement importantes et structurantes pour l'ensemble de textes considérés ? Les questions méthodologiques qui se posent concernent notamment

[22] Ce choix de mots nous est inspiré par la caractérisation de la classe 4 construite par la classification Reinert et présentée en figure 6.

le réglage de paramètres comme la taille des segments de texte et le nombre de classes à demander. Nous avons lancé une première expérimentation en adoptant le découpage textuel par défaut (segments d'environ 40 mots) et en demandant une vingtaine de classes. Le résultat obtenu peut être représenté graphiquement par la figure 6. Notre hypothèse est que les mots représentatifs des classes pourraient suggérer des candidats intéressants pour définir des isotopies génériques dominantes des textes.

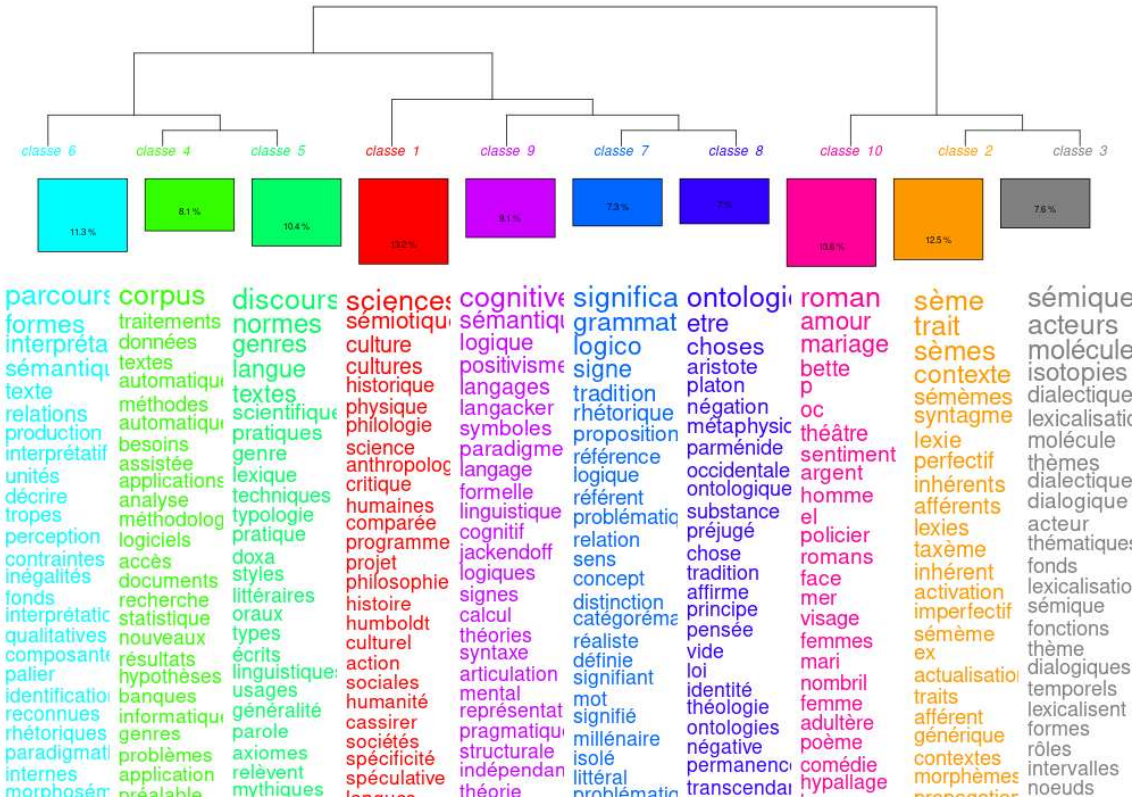


***Figure 6** : Classification Reinert appliquée au corpus Zéro (avec les réglages : segments de 40 mots, 20 classes demandées).*

Une piste complémentaire et inversée serait une recherche des réalisations d'une isotopie. Plutôt que de formuler une recherche sur la base d'un mot ou motif, l'idée serait de travailler sur des rencontres de mots : les mots se contextualisent mutuellement et peuvent être la trace, le support, d'isotopies. Des propositions en ce sens ont été expérimentées sur des paires de mots (MAYAFFRE, 2008 ; BRUNET, 2016, chapitre 17). On pourrait envisager de considérer des ensembles de mots (avec une certaine structuration (PINCEMIN, 1999)) et étudier une façon de s'appuyer sur les possibilités avancées d'un moteur de recherche textométrique pour repérer des manifestations

variables d'un ensemble de lexèmes ou morphèmes qui se contextualisent sémantiquement mutuellement : cela permettrait d'opérationnaliser une forme de recherche moins lexicale, plus sémantique. Une autre piste est d'ores et déjà observable avec la fonction « corpus en couleur » d'IRaMuTeQ (figure 7), qui rapporte chaque segment textuel à une classe, potentiellement représentative de son isotopie dominante.

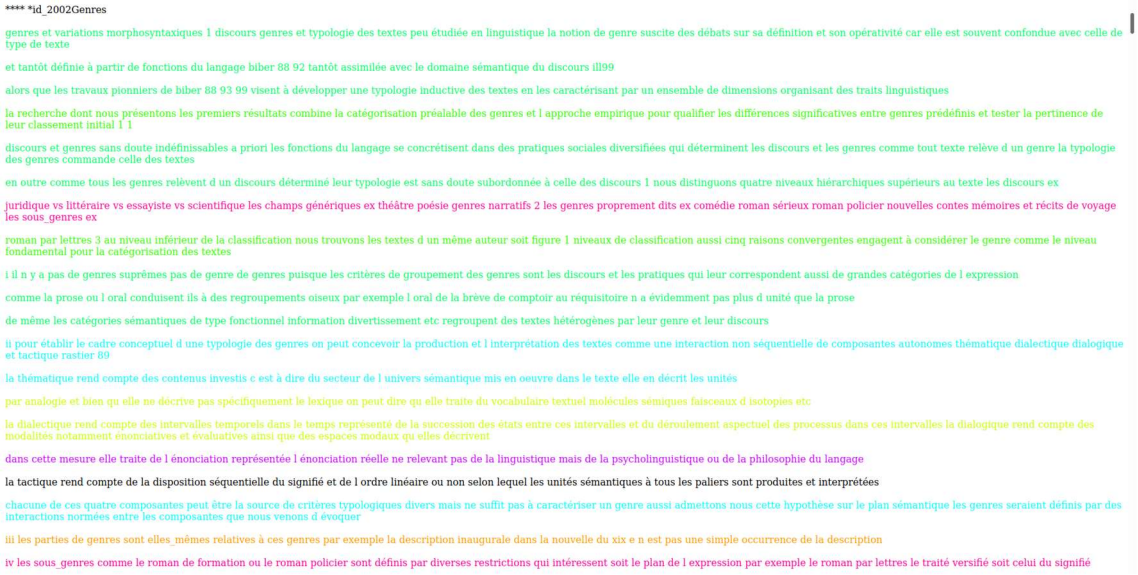


***Figure** 7 : Fonction « Corpus en couleur » d'Iramuteq, résultat pour le début du texte* 2002Genres *: application à chaque segment de texte du code couleur de la classe à laquelle il a été attribué par la classification*[23]

3 Principes directeurs et choix techniques

Ces perspectives passionnantes sont à partager. L'œuvre de Rastier est d'une richesse qui se prête à de multiples points de vue, et l'approche textométrique n'implémente pas un calcul sémantique (qui identifierait « le » sens du corpus) mais elle offre des moyens d'investigation que chaque chercheur mobilise en fonction de ses interrogations et de ses hypothèses. Cela fait sens d'investir collaborativement dans la production d'un corpus de qualité, pour pouvoir ensuite partager de multiples parcours interprétatifs construits selon les expertises diverses d'une large communauté de

[23] La classe grise du dendrogramme en figure 6 correspond au coloriage en jaune-vert en figure 7. Les segments textuels en noir sont ceux qui n'ont pas été classés dans les 10 classes retenues (ici 14 % des segments du corpus).

lecteurs. Il s'agit alors d'expliciter les principes directeurs qui vont préciser nos choix techniques.

3.1 Progressivité

Nous commençons par un petit choix de textes, sur lequel peut déjà s'éprouver l'approche textométrique ; puis cet ensemble pourra être étendu et diversifié progressivement, pouvant donner lieu à des corpus et sous-corpus à géométrie variable. Autrement dit, il ne s'agit pas de produire « le » corpus des écrits de sémantique interprétative de Rastier, mais d'alimenter par étapes, au fil du temps, une collection unifiée à partir de laquelle construire diverses intertextualités pertinentes.

Une même progressivité est à prévoir pour le modèle textuel XML TEI, qui se précisera avec l'expérience. Également, tous les outils textométriques ne pourront être déployés dès le départ : une première étape sera la mise à disposition de corpus compatibles avec des outils que les chercheurs et étudiants pourront installer sur leur ordinateur ; et c'est dans un second temps que nous pourrons envisager un accès outillé en ligne, dispensant d'installation logicielle et ouvrant d'autres possibilités, notamment pour gérer plus finement des questions de droits. La mise en place d'articulations avec des éditions d'ouvrage en ligne sous contrôle d'accès chez les éditeurs fait partie des questions a priori les plus complexes et les plus lourdes et pourra être étudiée une fois franchies les étapes précédentes.

3.2 Texte et intertextualité

Très clairement, nous prenons parti pour un corpus de textes, respectant l'unité contextuelle du texte intégral, par opposition à la possibilité de réunir des extraits (sinon promus au rang de morceaux choisis) voire des « échantillons » : la sémantique interprétative nous invite elle-même à pratiquer la linguistique de corpus comme science des textes instrumentée (VALETTE, 2008).

Par ailleurs, en sémantique interprétative comme en textométrie, le global détermine le local. La construction de la collection de textes numériques sera donc guidée par le choix de textes pouvant se contextualiser mutuellement de façon pertinente.

Une particularité de notre corpus est que nous avons affaire à des textes vivants : retravaillés à l'occasion d'une nouvelle édition, repris dans des publications englobantes. Plutôt qu'une collection de textes singuliers et bien identifiés, nous avons plutôt une sorte de réseau, d'arborescence, avec le développement de filons ou de branches. Une piste serait de construire notre corpus avec des textes parangons, représentatifs du développement d'un sujet, éventuellement eux-mêmes redéfinis à l'occasion du projet Sittelle, et donc correspondant davantage à une période d'écriture-réécriture qu'à une date de publication. Il n'est pas du tout évident qu'il soit possible et intéressant de rassembler des textes correspondant précisément à tout un ensemble de publications assez complet, ou en tout cas cela constituerait un autre type de corpus, avec d'autres possibilités (diachronie plus fine) et d'autres limites (redondances) pour les observations.

3.3 Ouverture

Au plan juridique, l'enjeu serait de rendre accessibles aux interrogations textométriques des ensembles de textes larges et pertinents, en les rendant compatibles avec les exigences légales pour les mettre à disposition publiquement au plus grand nombre. En pratique, il serait beaucoup plus simple de constituer une collection privée d'accès restreint ; mais nous voudrions que Sittelle soit justement l'occasion de chercher et d'expérimenter des solutions pour investir collectivement dans un corpus partagé, disponible à tous. Nous sommes donc directement intéressés par les développements récents de la science ouverte. Certains textes scientifiques deviennent librement utilisables pour des usages non commerciaux. En France, la *Loi pour une République numérique* devrait faciliter la mise à disposition, pour l'enseignement et la recherche, des articles scientifiques.

Pour les textes sous droit, nous pourrions préciser des formes de complémentarité avec l'édition traditionnelle pour ouvrir des accès ciblés en bonne intelligence avec le respect des contraintes économiques des éditeurs, selon une approche constructive et innovante comme celle par exemple d'OpenEdition[24] et de son

[24] « OpenEdition est une infrastructure [française nationale] d'édition numérique au service de la communication scientifique en sciences humaines et sociale. » Voir la présentation complète sur : https://www.openedition.org/6438

programme Freemium[25], parmi les plus avancées. Ainsi, un outil d'interrogation en ligne pourrait varier les accès en fonction des droits attachés à un compte de connexion et aux différents textes ; il pourrait déployer les possibilités d'analyse quantitative, et contrôler plus finement l'affichage de contextes, de passages ou de pages. La version portail de TXM a commencé à développer de telles possibilités pour les besoins de la Base de français médiéval, qui comportait certains textes sous droits.

Au plan technique, l'ouverture consiste à opter pour des formats numériques standards, favorisant la réutilisation des données, l'interopérabilité logicielle, et partant la qualité des échanges scientifiques (transparence, reproductibilité, etc.) (GUILLOT *et al.*, 2018). En effet, ce serait dommage de produire des données élaborées mais d'imposer ensuite un accès médié par un logiciel donné, aussi puissant soit-il ; certains chercheurs pourront vouloir explorer le corpus à travers des outils complémentaires et cela nous paraît très important de le permettre. Le format TEI est précisément prévu pour le partage et l'échange de données textuelles à travers un standard international. Il peut être directement exploité par certains outils (comme TXM par exemple) ; mais pour des outils se basant sur d'autres formats d'entrée et pour faciliter d'autres usages, on peut concevoir des versions du corpus automatiquement dérivées du format TEI.

3.4 Philologie numérique

L'édition numérique des textes invite à problématiser la représentation du texte et à expliciter des choix, car les « données » textuelles ne sont pas données, « les données sont faites de ce que l'on se donne » (RASTIER, 2001, p. 86 ; voir aussi RASTIER, 2021) : concrètement, de quels éléments constitutifs du texte vise-t-on à rendre compte ? Ainsi les corpus engagent-ils scientifiquement à une philologie numérique (RASTIER, 2001 ; GUILLOT *et al.*, 2017). Pour les analyses textométriques, outre le repérage d'unités linguistiques et textuelles nécessaires aux traitements (mots, contextes), les questions de lecture et d'interprétation sont centrales.

[25] « OpenEdition Freemium est un programme pour le développement de l'édition scientifique en libre accès dans le domaine des sciences humaines et sociales. Ce partenariat, que nous proposons exclusivement aux institutions (bibliothèques, campus, centres de recherche), vise à construire un modèle économique innovant et durable. […] les textes sont accessibles en libre accès au format HTML pour tout internaute, et ils sont téléchargeables aux formats PDF et ePub uniquement pour les utilisateurs des institutions partenaires. Aucun DRM ni quota de téléchargement ne sont appliqués. » (https://www.openedition.org/14043)

Les technologiques actuelles sont donc à mettre au service d'une restitution des structures textuelles essentielles à l'activité de lecture et d'interprétation, et notamment les contextualisations de tous ordres (intra- et intertextuelles) (PINCEMIN, 2007b), plutôt que d'appauvrir le texte pour le « faire rentrer » dans les outils.[26]

Ce travail d'édition numérique, qui suppose des traitements experts, assistés mais non automatiques, représente clairement un certain coût. La textométrie a-t-elle besoin d'un tel travail ? N'opère-t-elle pas aussi bien sur le texte « brut » ? De fait, on peut déjà tirer beaucoup d'observations intéressantes à partir d'un corpus non structuré comme notre corpus Zéro. Mais nous rencontrons aussi déjà quelques limites. Par exemple, en figure 1, nous faisons un retour au texte pour comprendre le sens d'une occurrence de « suspens ». Dans cette lecture nous avons une citation pour laquelle François Rastier précise que le ou les soulignements sont de lui. Or l'affichage du texte dans TXM a perdu ces marques de soulignement (ici l'italique). Une édition numérique XML-TEI permettra de préserver cette information et de la restituer pour la consultation du texte. De même, pour les parties que nous avons purement et simplement éliminées, comme la bibliographie : une édition TEI permet de ne pas s'en priver tout en contrôlant leur rôle dans l'analyse textométrique. La vue de progression en figure 2 aurait également pu être plus précise si nous avions pu coder les divisions intratextuelles : nous aurions alors pu tracer sur le graphique les bornes des sections successives, pour lire directement sur la figure ce que nous avons été rechercher dans le texte. Autre élément encore, les corrections de coquilles : est-ce toujours une bonne idée de simplifier en les passant sous silence ? Prenons le cas d'une liste numérotée, dans laquelle on trouve une erreur de numérotation, un même numéro est utilisé deux fois : peut-il être important que le lecteur sache que, dans la source publiée, l'item numéroté 4 dans TXM est en fait numéroté 3 dans une version publiquement diffusée ? L'édition TEI donne les moyens, si on le juge utile, de garder trace de corrections opérées, sans interférer de façon gênante avec les traitements d'analyse.

Un enjeu sera de trouver le bon niveau de modélisation des structures textuelles : cette modélisation vise à la fois à être légère (par pragmatisme, pour que la mise en

[26] Comme nous l'avons fait pour le corpus Zéro expérimental, qui vise un premier aperçu rapide en optant un format pauvre (txt), rapide à mettre en œuvre, à la différence de l'encodage XML TEI, demandant plus de travail mais permettant à moyen terme de capitaliser et partager les enrichissements apportés.

œuvre puisse être déployée à un grand nombre de textes sans nécessiter des moyens d'envergure) et prudente (pour limiter les difficultés de surinterprétation : plus le modèle détaille de cas, plus on rencontre la nécessité de déterminer de façon claire le cas à appliquer dans chaque situation), mais aussi productive (intéressante pour les analyses textométriques).

4 Éléments de conclusion

Comme le projet Sittelle vient compléter d'une façon originale les modes de diffusion de la sémantique interprétative, notre souhait est qu'il aide à étendre et à approfondir l'accès à la celle-ci. Il pourrait permettre à la fois une étude des concepts et propositions de la théorie de Rastier, et peut-être dans une certaine mesure une expérimentation concrète, opératoire, d'éléments de ce modèle (détermination du local par le global, isotopies). L'intérêt du projet dépend également de la mesure dans laquelle la communauté pourra se l'approprier : en appréciant un nouvel observatoire de la sémantique interprétative, mais aussi en y contribuant au plan du corpus ou des propositions d'exploration, de l'infrastructure ou de la documentation, et en développant de nouvelles expertises partagées. En fonction du contexte et de ses interlocuteurs, le projet Sittelle pourrait aussi être une occasion de contribuer à la réflexion sur de nouveaux modèles d'édition numérique, visant à conjuguer des accès pertinents aux écrits scientifiques et un modèle économique sain, durable et équilibré entre les acteurs de l'édition.

Le projet est à la fois ambitieux et modeste : exigeant dans ses principes directeurs, enthousiaste dans ses perspectives scientifiques, mais posant d'entrée de jeu une démarche progressive, par étapes et ajustements, qui assume de communiquer de premières réalisations très partielles et expérimentales — un peu à la façon des méthodes dites « agiles » en développement informatique. D'ailleurs les Humanités numériques ne se prêtent pas à une division du travail entre tâches techniques et activité scientifique (BRADLEY, 2012) : le codage, méthodique et philologique, est une édition scientifique ; l'interrogation du corpus, construite et problématisée, est un parcours interprétatif.

Le corpus outillé élaboré et partagé par le projet Sittelle ne remplacera pas les introductions pédagogiques ni les ouvrages de synthèse sur la sémantique interprétative, et moins encore l'intérêt et le plaisir de la lecture des textes eux-mêmes : car le corpus numérique entre naturellement en dialogue avec la lecture attentive et suivie, l'un et l'autre s'appelant mutuellement.

Remerciements

Cet article doit beaucoup au soutien de François Rastier, sans qui évidemment le projet Sittelle n'aurait pu naître ; il a aussi bénéficié de l'accompagnement patient et expert de plusieurs collègues du laboratoire IHRIM, en particulier Alexei Lavrentiev, Serge Heiden et Matthieu Decorde de l'équipe TXM pour les réflexions sur le codage des textes, et Isabelle Treff et Nadine Pontal du pôle Humanités numériques pour les premiers repères apportés sur les questions juridiques.

5 Références bibliographiques

5.1 Bibliographie générale

BÉNEL, Aurélien. Archives numériques et construction du sens ou « Comment échapper au Web sémantique ? ». Gazette des archives, 245, 2017, p. 173-187. https://doi.org/10.3406/gazar.2017.5524

BRADLEY, John. No Job for Techies: Technical Contributions to Research in the Digital Humanities. In: Collaborative Research in the Digital Humanities, Farnham, Burlington: Ashgate, 2012, p. 11-25.

BRUNET, Étienne. Tous comptes faits. Écrits choisis, tome 3. Questions linguistiques. Paris: Honoré Champion, 2016.

GUILLOT, Céline, HEIDEN, Serge, LAVRENTIEV, Alexei. Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. Diachroniques, 7, 2018, p.168-184.

GUILLOT, Céline, LAVRENTIEV, Alexei, RAINSFORD, Thomas, MARCHELLO-NIZIA, Christiane, HEIDEN, Serge. La « philologie numérique » : tentative de définition d'un nouvel objet éditorial. In: Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013), Section 13 : Philologie textuelle et éditoriale. Nancy: ATILF/SLR, 2017, p.143-154.

HÉBERT, Louis. Introduction à la sémantique des textes. Paris: Honoré Champion.

HEIDEN, Serge. Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex. In: Le poids des mots. Actes des 7es Journées

internationales d'analyse statistique des données textuelles. Presses universitaires de Louvain, 2004, v.1, p. 577-588.

HEIDEN, Serge, MAGUÉ, Jean-Philippe, PINCEMIN, Bénédicte. TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In: Statistical Analysis of Textual Data -Proceedings of 10th International Conference JADT 2010. Rome: Edizioni Universitarie di Lettere Economia Diritto, 2010, p. 1021-1031.

LEBART, Ludovic, PINCEMIN, Bénédicte, POUDAT, Céline. Analyse des données textuelles. Presses de l'université du Québec, 2019.

LEBART, Ludovic, SALEM, André. Statistique textuelle. Paris: Dunod, 1994.

MAYAFFRE, Damon. De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie. Syntaxe & Sémantique, 9, 2008, p. 53-72.

MAYAFFRE, Damon. Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques ? In: Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation. Actes du XXVIIe Colloque d'Albi Langages et Signification, Actes 2007 du colloque 2006. Toulouse: CALS-CPST (Colloque d'Albi Langages et signification-Centre pluridisciplinaire de sémiolinguistique textuelle), 2007a, p. 15-25.

MAYAFFRE, Damon. Analyses logométriques et rhétoriques des discours. In: Introduction à la recherche en SIC. Presses universitaires de Grenoble, 2007b, p. 153-180.

NÉE, Émilie (dir.). Méthodes et outils informatiques pour l'analyse des discours. Presses universitaires de Rennes, 2017.

PINCEMIN, Bénédicte. Semántica interpretativa y textometría. Tópicos del Seminario, 23, 2010, p. 15-55.

PINCEMIN, Bénédicte. Concordances et concordanciers. De l'art du bon KWAC. In: Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation. Actes du XXVIIe Colloque d'Albi Langages et Signification, Actes 2007 du colloque 2006. Toulouse: CALS-CPST (Colloque d'Albi Langages et signification-Centre pluridisciplinaire de sémiolinguistique textuelle), 2007a, p. 33-42.

PINCEMIN Bénédicte. Introduction. Interprétation, contextes, codage. Corpus, 2007b, 6, p. 5-15.

PINCEMIN, Bénédicte. Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ? In: Dépasser les sens iniques dans l'accès automatisé aux textes, Sémiotiques, 17, 1999, p. 71-120.

PINCEMIN, Bénédicte, HEIDEN, Serge, LAY, Marie-Hélène, LEBLANC, Jean-Marc, VIPREY, Jean-Marie. Fonctionnalités textométriques : Proposition de typologie selon un point de vue utilisateur. In: Statistical Analysis of Textual Data -

Proceedings of 10th International Conference JADT 2010. Rome: Edizioni Universitarie di Lettere Economia Diritto, 2010, p. 341-353.

RASTIER, François. Data *vs* corpora. In: L'intelligence artificielle des textes — Des algorithmes à l'interprétation. Paris: Champion, 2021, p. 203-246.

RASTIER, François. Mesure et démesure. Quantité et qualité en linguistique de corpus. Le français moderne, 88 (1), 2020, p. 11-25.

RASTIER, François. La mesure et le grain. Sémantique de corpus. Paris: Honoré Champion, 2011.

RASTIER, François. Arts et sciences du texte. Presses universitaires de France, 2001.

RASTIER, François. Sémantique interprétative. Presses universitaires de France, 1987.

RASTIER, François, CAVAZZA, Marc, ABEILLÉ, Anne. Sémantique pour l'analyse : de la linguistique à l'informatique. Paris : Masson, 1994.

RATINAUD, Pierre. Amélioration de la précision et de la vitesse de l'algorithme de classification de la méthode Reinert dans IRaMuTeQ. In: JADT' 2018, Proceedings of the 14th international conference on statistical analysis of textual data. Rome, Italie: Universitalia, 2018, v.2, p. 616-625.

RATINAUD Pierre, DEJEAN Sébastien. IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. In: Colloque Modélisation Appliquée aux Sciences Humaines et Sociales (MASHS2009). Toulouse, 2009.

REINERT, Max. Alceste, une méthodologie d'analyse des données textuelles et une application: Aurelia, de Gérard De Nerval. Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique, 1990, 26(1), p. 24-54.

SCHMID, Helmut. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, UK, 1994.

VALETTE, Mathieu. Introduction — Pour une science des textes instrumentée. Syntaxe & Sémantique, 9, p.9-14.

5.2 Inventaire des ouvrages de recherche de François Rastier en langue française

Idéologie et théorie des signes. La Haye: Mouton, 1971.

Essais de sémiotique discursive. Paris: Mame, 1973/1974.

Sémantique interprétative. Paris, Presses universitaires de France, 1987 (rééd. augmentées 1996, 2009).

Sens et textualité. Paris: Hachette, 1989 (rééd. augmentée Limoges: Lambert Lucas 2016).

Sémantique et recherches cognitives. Paris: Presses universitaires de France, 1991 (rééd. augmentée 2001).

Sémantique pour l'analyse — De la linguistique à l'informatique. En collaboration avec Marc Cavazza et Anne Abeillé. Paris: Masson, 1994.

Arts et sciences du texte. Paris: Presses universitaires de France, 2001.

Ulysse à Auschwitz — Primo Levi, le survivant. Paris: Éditions du Cerf, 2005.

La mesure et le grain — Sémantique de corpus. Paris: Champion, 2011.

Apprendre pour transmettre — L'éducation contre l'idéologie managériale. Paris: Presses universitaires de France, coll. Souffrance et théorie, 2013.

Saussure au futur. Paris: Les Belles-Lettres/Encre Marine, 2015.

Naufrage d'un prophète — Heidegger aujourd'hui. Paris: Presses universitaires de France, 2017.

Créer — Image, Langage, Virtuel. Paris-Madrid: Casimiro, 2016.

Heidegger, Messie antisémite — Ce que révèlent les Cahiers noirs. Lormont: Le bord de l'eau, 2018.

Faire sens — De la cognition à la culture. Paris: Classiques Garnier, 2018.

Mondes à l'envers — De Chamfort à Samuel Beckett. Paris: Classiques Garnier, 2018.

Exterminations et littérature — Les témoignages inconcevables. Paris: Presses universitaires de France, 2019.

5.3 Autres sources : webographie

[Les URL ont toutes été vérifiées le 10 novembre 2021.]

Base de français médiéval. Céline GUILLOT-BARBANCE (dir.), École normale supérieure de Lyon, Université de Lyon, http://bfm.ens-lyon.fr.

Dictionnaire de sémiotique en ligne. Louis HÉBERT, Université du Québec à Rimouski, http://www.semiotique.org.

Guide d'application de la loi pour une République numérique – Art. 30, site Ouvrir la science. Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation, https://www.ouvrirlascience.fr/guide-application-loi-republique-numerique-article-30-ecrits-scientifiques-version-courte/.

Infrastructure éditoriale *OpenEdition*. OpenEdition Center, Unité de service et de recherche (USR 2004) du CNRS, d'Aix-Marseille Université, de l'EHESS et d'Avignon Université, https://www.openedition.org.

Text Encoding Initiative. https://tei-c.org.

Texto! Textes & Cultures. François RASTIER (dir.), Institut Ferdinand de Saussure, Programme Sémantique des textes, http://www.revue-texto.net.

TreeTagger — a part-of-speech tagger for many languages. Helmut SCHMID, Institute for Computational Linguistics of the University of Stuttgart & Center for Information and Language Processing of the University of Munich, https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

6 Annexes

6.1 Composition et réalisation du corpus Zéro

Le tableau suivant recense les 28 textes retenus pour le corpus Zéro (corpus test exploratoire) :

Code pour le corpus	***Titre du texte***	***Année de publication hors* Texto!**	***Adresse en ligne, après le préfixe* http://www.revue-texto.net/1996-2007/**
1996Concepts	La sémantique des textes : concepts et applications	1996	Inedits/Rastier/Rastier_Concepts.html
1996PourSdT	Pour une sémantique des textes – Questions d'épistémologie	1996	Inedits/Rastier/Rastier_PourSdT.html
1996Terme	Le terme : entre ontologie et linguistique	1995	Inedits/Rastier/Rastier_Terme.html
1996Themes	La sémantique des thèmes ou le voyage sentimental	1995	Inedits/Rastier/Rastier_Themes.html
1996Transmission	Communication ou transmission ?	1995	Inedits/Rastier/Rastier_Transmission.html
1997Defigements	Défigements sémantiques en contexte	1997	Inedits/Rastier/Rastier_Defigements.html
2001Action	L'action et le sens – Pour une sémiotique des cultures	2001	Inedits/Rastier/Rastier_Action.html
2001Bette	La Bette et la Bête – une aporie du réalisme	1992	Inedits/Rastier/Rastier_Bette.html
2001Elements	Éléments de théorie des genres	-	Inedits/Rastier/Rastier_Elements.html
2001Ling-de-style	Vers une linguistique des styles	2001	Inedits/Rastier/Rastier_Ling-de-style.html
2001Semiotique	Sémiotique et sciences de la culture	2001	Inedits/Rastier/Rastier_Semiotique.html
2002Acces	L'accès sémantique aux banques textuelles	2000	Inedits/Rastier/Rastier_Acces.html
2002Genres	Genres et variations morphosyntaxiques	2001	Inedits/Malrieu_Rastier/Malrieu-Rastier_Genres.html
2002Macrosemantique	La macrosémantique	1994 (v1), 2002 (v2 en	Inedits/Rastier/Rastier_Macrosemantique1.html

Code pour le corpus	*Titre du texte*	*Année de publication hors* **Texto!**	*Adresse en ligne, après le préfixe* **http://www.revue-texto.net/1996-2007/**
		anglais)	
2003Semiotique-ontologie	De la signification au sens – Pour une sémiotique sans ontologie	1999 (en italien)	Inedits/Rastier/Rastier_Semiotique-ontologie.html
2004Ontologies	Ontologies	2004	Inedits/Rastier/Rastier_Ontologies.html
2005Classes-lexicales	De la sémantique cognitive à la sémantique diachronique : les valeurs et l'évolution des classes lexicales	2000	Inedits/Rastier/Rastier_Classes-lexicales.html
2005Doxa	Doxa et sémantique en corpus – Pour une sémantique des « idéologies »	2005	Inedits/Rastier/Rastier_Doxa.html
2005Enjeux	Enjeux épistémologiques de la linguistique de corpus	2005	Inedits/Rastier/Rastier_Enjeux.html
2005Herm-et-ling	Herméneutique et linguistique : dépasser la méconnaissance	2003 (en allemand)	Dialogues/Debat_Hermeneutique/Rastier_Herm-et-ling.html
2005Mesosemantique	Mésosémantique et syntaxe	1994 (v1), 2002 (v2 en anglais)	Inedits/Rastier/Rastier_Mesosemantique.html
2005Microsemantique	La microsémantique	1994 (v1), 2002 (v2 en anglais)	Inedits/Rastier/Rastier_Microsemantique.html
2005Saussure	Saussure au futur : écrits retrouvés et nouvelles réceptions. Introduction à une relecture de Saussure	-	Saussure/Sur_Saussure/Rastier_Saussure.html
2005Semantique-cognitive	Sémiotique du cognitivisme et sémantique cognitive – Questions d'histoire et d'épistémologie	1993 + 1996	Inedits/Rastier/Rastier_Semantique-cognitive.html
2006Borges	L'hypallage et Borges	2001	Inedits/Rastier/Rastier_Borges.html
2006Formes-semantiques	Formes sémantiques et textualités	2006	Inedits/Rastier/Rastier_Formes-semantiques.html
2006Inde	Saussure, la pensée indienne et la critique de l'ontologie	2002	Saussure/Sur_Saussure/Rastier_Inde.html
2006Textes	Pour une sémantique des textes théoriques	2005	Inedits/Rastier/Rastier_Textes.html

Les textes ont été collectés sur le site Texto! en octobre 2021. Le contenu de la page HTML contenant le texte a été copié/collé dans un fichier .txt (encodé en UTF-8). Des transformations ont été appliquées en vue de l'intérêt des traitements textométriques.[27]

- Sont retirés : auteur(s) du texte, affiliation, mention de référence originelle, sommaire, bibliographie[28], annexes, indicateurs de pagination, abstract en anglais ; tout-à-fait en fin de page : le contact, la référence de l'article dans *Texto!*.

- Sont conservés : le résumé, les éléments de contenu textuels ou chiffrés issus de tableaux ou figures (mais pas les séquences de ponctuations ou de symboles), l'exergue (même en langue étrangère), les remerciements, les notes.

- Élimination des majuscules de mise en forme : lorsque le titre est en capitales, il est réécrit en typographie courante. Idem pour les intitulés de sections. Nous avons aussi réécrit les mentions des ruptures ou décrochements sémiotiques notées en majuscules par l'auteur (ex. ICI → 'ici'), mais à la réflexion ce n'était sans doute pas utile (on aurait pu garder ce système de notation original car il était régulier dans notre corpus).

- Variations typographiques de certains caractères : les tirets (cadratin, demi-cadratin et double tiret : --) sont tous convertis en tiret unique simple ; les guillemets doubles (et l'éventuel espace associé), présents en trois types différents (droits, chevrons français, américains) sont unifiés vers le guillemet double droit ; le guillemet simple et l'apostrophe, représentés de quatre façons différentes (droit, droit inversé, courbe dans les deux sens), sont tous réécrits en apostrophe droit ; les points de

[27] Ces transformations sont liées au format d'import choisi, de type texte brut (.txt). Pour la suite du projet Sittelle, on envisage plutôt un format de représentation structuré type XML, qui permettra de recoder certaines informations plutôt que de les retirer, selon des principes de philologie numérique plus avancés.
[28] Il nous a semblé que le réseau des auteurs cités restait globalement présent malgré l'ablation de la bibliographie, grâce aux mentions renvoyant à la bibliographie au fil du texte. Pour le texte *2002Genres* pour lequel les renvois bibliographiques utilisent un code, celui-ci a été remplacé par le nom d'auteur complet.

suspension représentés comme trois points successifs sont recodés avec le caractère dédié.

- On insère un espace entre le mot et l'appel de note qu'il porte, le cas échéant.

- Correction de quelques coquilles aperçues au passage (moins d'une dizaine).

6.2 Esquisse de la chaîne de traitement éditorial envisagée pour le corpus XML TEI

L'idée générale du projet Sittelle est de construire un corpus de textes au format **XML** selon les recommandations de la *Text Encoding Initiative* (**TEI**, https://tei-c.org), utilisant tous un même sous-ensemble simple des éléments de la TEI. L'explicitation de ce sous-ensemble, sous la forme d'un **schéma** dans un **document ODD**, vise une modélisation textuelle appropriée pour rendre compte de structures textuelles des écrits scientifiques de François Rastier pertinentes dans un contexte textométrique. La textométrie considère à la fois des informations de paramétrage de certains calculs (ex. paragraphes pour définir des contextes, divisions du texte pour situer des résultats), de type de contenu pour des opérations de sélection (ex. relever du vocabulaire mais pas dans les bibliographies), mais aussi de présentation pour restituer des éléments de mise en page importants dans la lecture du texte (ex. soulignement).

Pour l'élaboration du schéma, nous comptons nous baser initialement sur notre connaissance préalable des textes de François Rastier (et des outils textométriques), et utiliser l'outil **Roma** (https://roma2.tei-c.org). Comme nous voulons une représentation légère, nous partirons d'un schéma minimal et préciserons des éléments nécessaires à ajouter. Nous nous attendons à ce que le schéma ne soit pas d'emblée satisfaisant mais se stabilise progressivement avec l'expérience de sa mise en œuvre, en effet ses premières utilisations sur notre corpus (codage puis interrogations textométriques) seront révélatrices de lacunes ou de choix à rectifier. Ce schéma devra d'ailleurs être documenté avec un **Guide d'encodage** (en principe intégré au document ODD), qui précisera en pratique l'interprétation des éléments TEI dans le contexte du corpus Sittelle, en se basant sur des cas concrets représentatifs. En effet, bien que très contrôlé formellement (liste limitée d'éléments disponibles, combinaisons possibles ou pas), le

codage reste une activité fortement interprétative et experte (c'est un travail d'édition scientifique). Nous pensons d'ailleurs que la mise au point du schéma et de sa documentation sera l'occasion d'une réflexion scientifique sur certains aspects de la textualité, de l'écriture scientifique et du corpus Rastier.

Pour les textes eux-mêmes, dans la mesure du possible, et avec le soutien de l'auteur, il s'agirait de partir de fichiers auteur en format **traitement de texte**. Si besoin, pour les fichiers plus anciens au format .doc, opérer une conversion vers **.docx** (au moyen du logiciel Microsoft Word), format qui en pratique assure une meilleure compatibilité avec l'outil de conversion vers XML TEI, **OxGarage** (https://oxgarage.tei-c.org), maintenu par la communauté TEI. La conversion automatique de docx à TEI P5 (profil « default ») produit un fichier XML TEI basé sur une interprétation standard des marques de mise en page, et pas nécessairement conforme au schéma XML défini pour Sittelle. L'édition du texte TEI avec un logiciel spécialisé pour XML, tel **Oxygen** (https://www.oxygenxml.com – outil commercial mais souvent disponible dans le monde de la recherche grâce à des licences partagées au niveau des universités ou des laboratoires), permettra d'ajuster le codage de façon assistée, pour le rendre conforme au schéma Sittelle et en adéquation avec les indications du Guide d'encodage. Certains pré-traitements généraux pourraient être rassemblés dans des **scenarios** Oxygen facilitant leur application méthodique.

Pour ce qui concerne l'**entête TEI**, on pourra envisager de gérer les informations qu'il contient dans un fichier unique de type **tableur** (cela permet une vue centralisée, plus facile à maintenir). En effet, on peut ensuite utiliser une feuille **XSL** pour alimenter puis mettre à jour automatiquement les entêtes dans le fichier TEI. La Base de français médiéval procède de cette façon (les métadonnées sont gérées dans une base de données).

L'ensemble des textes dans leurs différents états d'avancement dans la préparation du corpus, ainsi que les documents et ressources utilisés pour les traitements et plus généralement par le projet, pourront être gérés de façon partagée à l'échelle internationale via une solution de stockage de données sécurisée en ligne dédiée à la recherche en sciences humaines, le **ShareDocs** (https://documentation.huma-

num.fr/sharedocs-stockage/) de l'infrastructure **Huma-Num** (https://www.huma-num.fr).