Qu'est-ce qu'une machine linguistique ? Epistémologie de la technique et linguistique instrumentée

Pierre-Yves Modicom

Version auteur. Pour accéder aux figures : https://zenodo.org/record/1137788/files/annexes.pdf

Qu'est-ce qu'une machine linguistique ?
Épistémologie de la technique et linguistique instrumentée

Pierre-Yves Modicom

Le "tournant empirique" de la linguistique comme tournant mécanique

Introduction

> « Le corpus : la notion et l'objet risque d'être victime aujourd'hui en France de son succès. Plus une discipline, plus un comité scientifique, plus un chercheur qui n'y fasse référence ; plus un linguiste, surtout, qui ne le manipule, le caresse ou le maltraite. »

C'est par ces mots que Damon Mayaffre, un des pionniers du traitement quantitatif des discours (politiques, en l'occurrence) en linguistique, ouvre un article de réflexion générale sur le statut du corpus dans la linguistique actuelle[1]. Au même moment, François Rastier lui fait écho en déclarant que le corpus est l'ensemble dans lequel « le texte », conçu comme unité de référence de la « linguistique évoluée », peut seul « prendre son sens »[2].

Comme le relève Mayaffre dans son article, le renversement de la dichotomie saussurienne entre langue et parole, dichotomie à laquelle on pourrait d'ailleurs adjoindre les oppositions équivalentes plus pratiquées dans la sphère anglo-saxonne, par exemple entre système et usage ou entre compétence et performance, aboutit à un primat radical des données authentiques et débouche pour finir sur l'affirmation selon laquelle « il n'y a aucune linguistique qui ne soit de corpus[3] », ce qui revient à prendre le contrepied exact de la phrase de Chomsky selon laquelle « la linguistique de corpus ne veut rien dire[4] ».

Le sentiment d'un changement radical est aujourd'hui très répandu dans le champ des sciences du langage, ce qui se traduit par une floraison de littérature méthodologique tentant de cerner ce que beaucoup qualifient de tournant, ou même de changement de paradigme[5]. Dans ces débats, la notion de linguistique « outillée » ou « instrumentée » est souvent mobilisée pour illustrer la thèse d'une forme particulière, et potentiellement supérieure, de scientificité de la linguistique « de corpus » en tant que linguistique fortement liée à l'usage des machines, et donc supposément moins tributaire de la « théorie » ou de la spéculation. Le présent article entend justement analyser et questionner ce rapport de la linguistique de corpus à ses machines.

Place des grands corpus dans les revendications de scientificité des linguistes

Dans ce que nous qualifierions pour notre part davantage d'épistémologie spontanée des savants que de méthodologie, le corpus est fréquemment posé comme susceptible de connaître deux statuts plus ou moins antithétiques :

Tandis que la linguistique « sur » corpus (*corpus-based*) passe par des corpus d'illustration, qui permettent de mettre à l'épreuve une théorie, et par conséquent pourrait se revendiquer de la méthode hypothético-déductive (selon un mouvement fréquemment dit descendant, *top-down*), la linguistique « de » corpus (*corpus-driven*) se revendique de la méthode inductive (selon un mouvement souvent dit ascendant, *bottom-up*). C'est dans ce dernier cadre que s'est développée la plus grande partie de la réflexion sur le caractère

---

1 MAYAFFRE Damon, « Rôle et place des corpus en linguistique : réflexions introductives », *Texto!*, n°10, vol. 4, p.5.

2 RASTIER François, « Enjeux épistémologiques de la linguistique de corpus », *Texto! Inédits*, en ligne : http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html (dernière consultation le 1er juin 2018, 10:22) ; 1re publ. *in* G. WILLIAMS (dir.), *La linguistique de corpus*, Rennes, Presses universitaires de Rennes, 2005, p. 31-46.

3 CHARAUDEAU Patrick, « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique », *Corpus* n°8, 2009, p. 60 (Référence signalée par Sara Benoist, c. p.).

4 « Corpus linguistics doesn't mean anything. It's like saying suppose a physicist decides, suppose physics and chemistry decide that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world and they'll collect huge videotapes of everything that's happening and from that maybe they'll come up with some generalizations or insights. Well, you know, sciences don't do this. » (ANDOR József, « The master and his performance : An interview with Noam Chomsky », *Intercultural Pragmatics* n° 1, vol. 1, 2004, p. 97).

5 Ainsi Glynn parle-t-il de « major paradigm shift in linguistics, from theory-driven to empirical research. » (GLYNN Dylan, « Corpus-driven cognitive semantics: an introduction to the field », *in* Dylan GLYNN & Kerstin FISCHER (dir.), *Quantitative methods in cognitive semantics: Corpus-driven approaches*, Berlin, De Gruyter, 2010, p. 1-40)

supposément incontournable du recours aux corpus et aux statistiques. On en trouve un bon exemple dans les ouvrages de D. Glynn, figure majeure de la sémantique « de corpus » et auteur de nombreux travaux sur les problèmes de méthodes en linguistique « empirique » :

« La recherche empirique quantitative est-elle une voie praticable pour la recherche en sémantique ? Plus spécifiquement, pouvons-nous utiliser des données de corpus pour produire des résultats testables et falsifiables de description sémantique ? »[6]

Dans ses formes les plus radicales, la linguistique de corpus suggère un parallélisme ou une équivalence entre les outils techniques et mathématiques à la disposition du linguiste et les mécanismes cognitifs (et donc, en dernière instance, du moins si l'on adopte la perspective spécifiquement neuronale qui prévaut souvent en science cognitive, cérébraux) qui caractériseraient le langage comme faculté, ou bien une langue singulière comme « code ».

> « Ainsi, nous pouvons dire que la fréquence de co-occurrence, qui est fondamentale en recherche de corpus, est une opérationalisation quantitative des théories de base de la Linguistique Cognitive – l'enracinement (*entrenchment*) et la catégorisation. Ces théories, l'enracinement et la catégorisation, expliquent la grammaire et le sens[7]. »

Même sans aller jusqu'à cette forme nouvelle de parallélisme psychophysique, on peut dégager, dans les discours sur le « tournant instrumenté » de la linguistique, un réseau d'arguments dont les deux plus significatifs nous paraissent être le recours à la notion de falsification et le réinvestissement de la catégorie d'« empirie » : tout d'abord, par la convocation, fût-elle tacite, de la figure de Karl Popper, le simple concept de falsifiabilité a déjà valeur de proclamation de scientificité au bénéfice de la méthode quantitative, fût-ce au prix de présupposés discutables (on ne saurait considérer le falsificationnisme poppérien comme une théorie épistémologique allant de soi, et en outre, le rapport entre reproductibilité et falsifiabilité, dans le cas d'une linguistique « de » corpus, appellerait un retour précis sur ces deux notions, qui interdit de les tenir pour acquises en droit). La notion de « tournant empirique » ou de « recherche empirique » trouve ici une valeur polémique et représente à nouveau, en creux, un argument pour la nouvelle linguistique ou plutôt contre l'ancienne, présentée comme théoriciste.

Chez certains auteurs, on retrouve également l'idée que le recours à l'analyse quantitative, comprise comme analyse mécanique, permettrait de minorer la part de spéculation du linguiste, et notamment du sémanticien, qu'impose l'analyse dite qualitative – ce qui implique effectivement de (se) poser des objectifs sensiblement différents dans la description des faits de langue. Cette idée, la linguistique quantitative la partage avec les approches ethnométhodologiques de l'interaction : la fidélité aux données brutes doit préserver le linguiste du danger de la spéculation interprétative. Au cœur du discours de la nouvelle linguistique outillée, il y a donc selon nous une revendication de fidélité absolue aux données. Or cette revendication semble liée à un postulat, celui de la transparence de l'instrument, ou de sa neutralité. Les discussions méthodologiques en linguistique de corpus portent volontiers sur les opérations statistiques mobilisables pour l'analyse et l'interprétation des « données », mais l'ontologie des données et celle des instruments utilisés ne semble avoir donné lieu à un nombre d'études nettement moins fourni. Parmi les quelques travaux méthodologiques s'aventurant dans ce domaine, on peut relever les travaux de Dalud-Vincent et Kalampalikis & Moscovici sur le logiciel Alceste[8][9], et l'article plus général de Busse & Teubert sur la méthode empirique en sémantique historique des discours[10]..

---

6 "Is quantitative empirical research possible for the study of semantics? More specifically, can we use corpus data to produce testable and falsifiable results for the description of meaning?" (*op.cit.*, p.1).

7 "Thus, we can say that frequency of co-occurrence, which is fundamental to corpus research, is a quantitative operationalisation of the basic theories of Cognitive Linguistics – entrenchment and categorisation. These theories, entrenchment and categorisation, explain grammar and meaning". GLYNN Dylan, « Corpus-driven cognitive semantics: an introduction to the field », art. cit., p. 8.

8 DALUD-VINCENT, Monique, « Alceste comme outil de traitement d'entretiens semi-directifs : essai et critiques pour un usage en sociologie », *Langage et Société*, 135, 2011, p. 9-28. disp. sous https://www.cairn.info/revue-langage-et-societe-2011-1-page-9.html (dernière consultation le 1er juin 2018, 10h25). L'auteur remercie un relecteur anonyme pour cette référence et la suivante.

9 KALAMPALIKIS, Nikos & MOSCOVICI, Serge, « Une approche pragmatique de l'analyse Alceste », *Cahiers internationaux de psychologie sociale* 66, 2005, p. 15-24 ; disp. sous https://www.cairn.info/revue-les-cahiers-internationaux-de-psychologie-sociale-2005-2-page-15.htm (dernière consultation le 1er juin 2018, 10h30).

10 BUSSE, Dietrich et TEUBERT, Wolfgang, « Ist Diskurs ein sprachwissenschaftliches Objekt ? Zur Methodenfrage der historischen Semantik », *in* Dietrich BUSSE, Fritz HERMANNS & Wolfgang TEUBERT (dir.), *Begriffsgeschichte und*

Pour une critique du neutralisme machinique

C'est cette ontologie des objets de la linguistique « de/sur corpus » en tant que linguistique outillée qui sera au centre de mon propos. Plus précisément, les deux opérations constitutives d'une linguistique de corpus étant l'annotation et l'exploration, la question centrale est celle du statut théorique des instruments techniques utilisés pour constituer et explorer un corpus.

Pour une ontologie de l'objet technique « corpus » en linguistique, il me semble souhaitable de se tourner vers les catégories proposées par l'épistémologue des techniques Gilbert Simondon dans son ouvrage *Du mode d'existence des objets techniques*[11]. Au cœur de la réflexion de Simondon, ou trouve l'idée du processus de concrétisation, c'est-à-dire de l'évolution d'un objet technique au fil de l'histoire, menant à une adéquation totale à sa ou ses fonctions, qui peut aller jusqu'à la surdétermination (par laquelle l'outil, devenu hyperspécialisé, a besoin du renfort d'autres outils pour assurer des fonctions que ses « ancêtres » remplissaient seuls). Dans cette perspective, l'outil « ne saurait être considéré comme un pur ustensile[12] ».

La thèse centrale défendue ici, en application des théories de Simondon sur l'ontologie des machines, sera que les corpus sont des *milieux techniques mixtes*, constitués par les textes du corpus et les objets techniques eux-mêmes complexes qui servent à leur exploration. Les « ensembles techniques » constitués par un corpus et son outil d'exploration sont l'objet d'un processus d'individuation technique propre, qui débouche sur une pluralité de configurations possibles, questionnant ainsi le postulat sous-jacent d'une neutralité de l'instrument technique conçu comme simple révélateur.

L'épistémologie de la machine chez Simondon

Le processus de concrétisation

La caractéristique majeure de l'individuation comme processus de concrétisation de l'objet technique est l'accroissement de son unité interne, c'est-à-dire que l'adéquation croissante de l'objet à sa/ses fonction(s) va de pair avec l'adaptation de ses composants les uns aux autres, tandis que les effets secondaires ou les pertes d'énergie sont peu à peu éliminés ; la concrétisation est un processus de réduction de la contingence et de resserrement des liens entre les composants : « Dans un moteur actuel, chaque pièce importante est tellement rattachée aux autres par des échanges réciproques d'énergie qu'elle ne peut pas être autre qu'elle n'est » [13]. L'exemple du moteur illustre en quoi la concrétisation de l'objet technique obéit à une dynamique historique ou généalogique :

> « Dans le moteur ancien, chaque élément intervient à un certain moment dans le cycle, puis est censé ne plus agir sur les autres éléments ; les pièces du moteur sont comme des personnes qui travailleraient chacune à leur tour, mais ne se connaîtraient pas les unes les autres[14]. »

Dans cette dynamique généalogique, la technique est investie de part en part par la science[15], qui se saisit des caractéristiques fonctionnelles des outils et des composants, pour en tirer tout ce qui peut en être tiré et en exclure tout ce qui n'est pas désiré. Ainsi, la mécanique comme branche de la physique (étude des mouvements et des résistances entre corps) informe la mécanique comme technique (de production et de mise en fonctionnement de machines). On retrouve là un topos déjà souligné par Kant dans *Théorie et pratique* : la technique est une dépendance de la science, et une technique moins pleine de science est une technique moins efficace à l'aune de ses propres critères de technicité. À cet égard, on peut d'emblée relever qu'il n'y a pas de sens à opposer une linguistique « outillée » et une linguistique « spéculative » si par « spéculative » on entend *theory-driven* : un outil, c'est de la théorie matérialisée, et en ce sens il n'y a donc rien de plus *theory-driven* qu'une étude conduite en s'en remettant à un outil.

Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historische Semantik. Opladen: Westdeutscher Verlag, 1994, p. 10 - 28.

11 SIMONDON Gilbert, *Du mode d'existence des objets techniques*, Paris, Aubier, 1958 (rééd. 2012).
12 *Ibid.*, p. 16.
13 *Ibid*., p.23.
14 *Ibid*., p.24.
15 *Ibid*., p.43.

L'objet technique et son milieu

Mais la concrétisation n'est pas qu'un processus de transformation interne. Elle a aussi un impact sur le rapport de l'objet technique à son dehors, qui est double : l'extérieur de l'objet technique est à la fois l'espace où il se trouve, avec ce qui emplit cet espace et qui présente des caractéristiques (pression de l'air, luminosité, environnement acoustique…), ce que Simondon appelle le « milieu géographique », et les autres objets techniques avec lesquels l'objet est susceptible d'interagir ou d'être combiné, le « milieu technique[16] ». Pour Simondon, la caractéristique de l'objet technique est justement d'être « au point de rencontre[17] » de ces deux milieux, et d'instaurer, dans, par et pour son fonctionnement, un milieu mixte associant les deux facettes de son dehors. Ce milieu mixte, lorsqu'il est stabilisé et pleinement intégré au fonctionnement de l'objet technique, Simondon l'appelle le « milieu associé ».

« On pourrait dire que l'invention concrétisante réalise un milieu techno-géographique [...] qui est une condition de possibilité du fonctionnement de l'objet technique. L'objet technique est donc la condition de lui-même comme condition de ce milieu mixte. [...] L'évolution des objets techniques ne peut devenir progrès que dans la mesure où ces objets techniques sont libres dans leur évolution et non nécessités dans le sens d'une hypertélie fatale. Pour que cela soit possible, il faut que l'évolution des objets techniques soit constructive, c'est-à-dire qu'elle conduise à la création de ce troisième milieu techno-géographique, dont chaque modification est auto-conditionnée[18]. »

L'individuation

Sur ces bases, Simondon peut distinguer trois stades d'individuation de l'objet technique :

1) l'individu technique à proprement parler, celui qui présente un milieu associé sans lequel il ne peut pas fonctionner. L'objet technique instaure alors un certain régime d'existence de tout ou partie de son environnement, régime qui dépend du fonctionnement de l'objet et dont ce fonctionnement dépend en retour.
2) l'ensemble technique (ou « ensemble de formes techniques ») : il ne présente pas de milieu associé unifié. Corrélativement, il se compose d'objets techniques relativement autonomes les uns des autres et n'interagissant que faiblement. Il est donc supra-individuel et faiblement concrétisé.
3) Le composant technique, enfin, est infra-individuel. Il n'a pas non plus de milieu associé, mais peut être partie intégrante tant d'un ensemble que d'un individu technique.

Les outils, au sens restreint que le terme a chez Simondon (un marteau par exemple), et leurs cousins, les instruments (un microscope), sont des éléments techniques.

> « L'outil prolonge l'organe, et est porté par le geste. Le 18e siècle a été le grand moment du développement des outils et des instruments, si l'on entend par *outil* l'objet technique qui permet de prolonger et d'armer le corps pour accomplir un geste, et par *instrument* l'objet technique qui permet de prolonger et d'adapter le corps pour obtenir une meilleure perception ; l'instrument est outil de perception[19]. »

À ce titre, la linguistique de corpus n'est donc pas « outillée » mais bien « instrumentée », puisque les objets techniques que nous allons maintenant passer au crible sont bien censés permettre de mieux voir, ou percevoir, les données. Simondon est d'ailleurs très explicite quant au fait que pour lui, la pratique de la science repose sur le recours à des instruments[20].

Dans ce qui suit, nous examinerons quatre instruments, en deux temps. Nous commencerons par deux instruments textométriques, c'est-à-dire des logiciels permettant de procéder à des relevés systématiques d'occurrences d'une forme dans un corpus et d'en dégager les schémas de cooccurrence (c'est-à-dire de lister les formes les plus susceptibles d'apparaître dans le voisinage du terme étudié). Nous nous pencherons dans

16 *Ibid*., p.64.
17 SIMONDON Gilbert, *Loc. cit.*
18 *Ibid*., p.69.
19 *Ibid*., p. 159.
20 *Ibid*., p. 160.

un second temps sur deux corpus en ligne, où les « données » textuelles sont directement intégrées dans l'instrument d'exploration.

Implications pour les instruments textométriques de la linguistique de corpus

AntConc et TXM : présentation

AntConc, un concordancier

AntConc est un logiciel pionnier de l'exploration quantitative des textes. Ce logiciel peut donner lieu, au-delà de ses fonctionnalités de départ, à des usages additionnels via un travail supplémentaire en combinaison avec d'autres logiciels. Nous laisserons ces usages de côté : sans aller jusqu'à parler de bricolage, ces usages sont en effet extérieurs à la problématique de l'adaptation de l'objet technique aux fonctions en vue desquelles il a été créé.

AntConc nécessite le chargement d'un corpus composé d'un ou plusieurs fichiers .txt (le corpus). L'exploration s'effectue à partir de la saisie d'une requête (on tape une forme dans le moteur de recherche). On obtient alors :

(i) toutes les occurrences de cette forme (et les indications de fréquence et de distribution) dans le corpus .txt préalablement compilé et chargé par l'analyste[21];
(ii) la liste de l'ensemble des mots présents dans le corpus, classés par nombre d'occurrences[22];
(iii) les collocations les plus fréquentes pour la forme recherchée au départ[23].

De ce fait, le logiciel présente trois caractéristiques qui seront pertinentes dans la comparaison avec TXM :

(i) c'est à l'analyste de charger directement les fichiers .txt ; à charge pour lui, par exemple, de vérifier l'encodage des caractères faute de quoi le logiciel rencontra un problème de traitement face aux signes diacritiques (pour le français : cédille, accents ; pour l'allemand : Umlaut essentiellement ; à noter également la rétivité du logiciel au ß) ;
(ii) AntConc est en fait aveugle à la langue ; le logiciel isole les mots à partir des espaces et signes de ponctuation dans le fichier ;
(iii) le corpus ne contient aucune annotation : il se réduit à un texte brut.

Le travail d'AntConc relève donc de la mise en ordre de données brutes et de la quantification d'occurrences. Tout travail interprétatif se fait en amont ou en aval de l'utilisation de l'instrument.

TXM

TXM peut à bien des égards être présenté comme un descendant d'AntConc : ce logiciel permet de lancer toutes les recherches vues pour AntConc... et davantage ; si l'on compare TXM à AntConc en gardant en tête les caractéristiques précédemment relevées pour celui-ci, la différence la plus importante concerne l'annotation : TXM travaille impérativement sur un corpus annoté. Si le corpus chargé n'est pas annoté, par exemple si c'est le même fichier .txt qu'AntConc peut explorer, TXM commencera par procéder lui-même à l'annotation (morphosyntaxique) lexicale avant qu'il soit possible de lancer la première requête et de commencer l'exploration du corpus. En l'occurrence, TXM va notamment réunir les différentes formes d'un même lexème sous un chapeau commun et étiqueter les formes par partie du discours : Nom, adjectif, verbe – en distinguant alors le plus souvent entre verbes pleins, auxiliaires, modaux… conjugués ou non. Nous écrivons « le plus souvent », car si le composant technique qui procède à l'annotation morphosyntaxique est à première vue toujours le même (c'est le logiciel TreeTagger, utilisé comme module), en réalité cette annotation ne peut bien sûr pas être « aveugle à la langue », comme nous l'écrivions pour AntConc, c'est-à-dire qu'il existe au moins une version de TreeTagger par langue susceptible d'être traitée par TXM, et que le logiciel est inutilisable pour examiner les textes d'une langue pour laquelle il n'existe pas de version de TreeTagger. Or, d'une langue à l'autre, non seulement bien sûr le « dictionnaire » utilisé pour regrouper

21 Voir annexe en ligne, figure 1. L'annexe en ligne est consultable soushttps://zenodo.org/record/1137788 (DOI 10.5281/zenodo.1137787 ; dernière consultation le 1er juin 2018 à 11:01).
22 Voir annexe en ligne, figure 2.
23 Voir annexe en ligne, figure 3.

plusieurs formes d'un même lexème n'est pas le même, mais les étiquettes (*tags*) morphosyntaxiques, notamment pour distinguer les parties du discours, ne sont pas les mêmes et les possibilités d'exploration peuvent varier.

Dès l'abord, il apparaît donc qu'avec TXM, le corpus est « digéré » par le logiciel d'exploration et d'analyse. Les recherches passent impérativement par un « langage de requête » qui impose de signaler à quel(s) niveau(x) d'annotation on se situe : on peut demander les occurrences d'une forme, d'un lemme[24], d'une partie du discours, ou d'une structure complexe associant des informations de différents niveaux, par exemple tous les passages du corpus où un adverbe est immédiatement suivi d'un verbe conjugué, ou tous les adjectifs apparaissant immédiatement avant le lemme *Entscheidung* dans le corpus, etc. Pour toutes ces requêtes, outre l'inventaire des occurrences, on peut obtenir des relevés de collocations, ou des indications de fréquence ou de distribution dans le corpus[25]. L'image ci-dessous correspond ainsi au relevé des collocations pour la structure associant un verbe modal conjugué suivi à moins de cinq mots de là par un adverbe, sur mon corpus DRKORP[26].

{TXM_synt_colloc.tif}

Figure I : Liste de cooccurrents pour une requête « verbe modal conjugué suivi d'un adverbe avec un écart compris entre 0 et 5 mots » sur TXM, corpus DRKORP.

En réalité, même si l'on charge un fichier texte, le logiciel travaille sur un fichier .xml. Dans un usage « naïf » de TXM, on peut travailler avec le logiciel sans jamais accéder à ces fichiers tableurs ni les modifier. Mais il est également possible de rechercher dans l'ordinateur l'endroit où TXM stocke les fichiers XML correspondant à une base de fichiers .txt, pour intervenir directement sur ces fichiers XML.

A titre d'illustration, je reprends ici à Hardie[27] l'annotation dans son standard minimal Modest XML de l'énoncé *The cat sat on the mat.*

```
<w pos="ART"  lemma="the">The</w>\\
<w pos="NOUN" lemma="cat">cat</w>\\
<w pos="VERB" lemma="sit">sat</w>\\
<w pos="PREP" lemma="on" >on</w>\\
<w pos="ART"  lemma="the">the</w>\\
<w pos="NOUN" lemma="mat">mat</w>\\
<w pos="PUNC" lemma="."  >.</w>\\
```

L'annotation à des niveaux supplémentaires (sémantique par ex.) se fait encore essentiellement à la main sur les fichiers XML, même s'il existe des systèmes balbutiants d'annotation sémantique automatisée à la manière de TreeTagger en syntaxe (cf. p.ex. le corpus SALSA à Sarrebruck).

Il existe également des logiciels d'annotation comme ANALEC ou ELAN/CLAN pour limiter la confrontation directe au code XML.

Du point de vue ontologique

Si l'on se pose la question du « milieu associé » de ces instruments, TXM présente un degré d'assimilation du milieu nettement supérieur à AntConc. Dans l'usage de TXM, il n'y a plus de privilège de la ligne de texte "première", qui fait l'objet de requêtes formulées de la même manière que n'importe quelle autre ligne d'annotation. Le système implique une transformation de l'input textuel en une mini-base de données codée spécifiquement pour la recherche sur corpus ; pour les fonctionnalités de base, cette transformation est opérée par l'outil lui-même. Enfin, la syntaxe des recherches implique la soumission aux catégories notionnelles de l'outil.

Une façon radicale de formuler cette situation serait de dire qu'avec TXM, il n'y a plus de « données » de corpus distinctes de l'objet technique, puisque le corpus ne se réduit plus au texte-source : le vrai corpus de TXM, c'est l'ensemble des données XML, qui ne distinguent pas entre le texte source et les annotations,

24 Voir figure 4 de l'annexe en ligne.

25 Voir figure 5 dans l'annexe en ligne.

26 Voir MODICOM Pierre-Yves, *L'énoncé et son double : recherches sur le marquage de l'altérité énonciative en allemand,* Paris, Université Paris-Sorbonne, 2016.

27 HARDIE Andrew, « Modest XML for Corpora: Not a standard, but a suggestion », *ICAME* n°38, 2014, p. 73-103.

lesquelles procèdent de choix qui sont soit ceux des programmateurs des modules utilisés pour une annotation automatique, soit ceux de l'annotateur individuel. Le corpus peut être qualifié de milieu associé de la machine, puisqu'il est modifié par celle-ci dans sa nature même.

Le degré d'individuation de TXM est également assez élevé du point de vue de la structure interne : le logiciel se caractérise par une forte co-adaptation à ses différents composants (modules d'annotation, d'exploration, d'exportation etc., qui sont intégrés dans l'interface d'utilisation alors qu'ils peuvent avoir été conçus séparément). Cela étant, le système n'est clos qu'en première instance, le corpus lui-même étant modifiable manuellement *ad libitum*.

Bases de données linguistiques (« corpus ») en ligne : DDD et DeReKo

À côté des instruments permettant d'analyser un corpus que l'analyste constitue soi-même, l'une des évolutions contemporaines majeures est l'apparition des grandes bases de données textuelles en ligne (« grands corpus ») comme le *Deutsches Referenzkorpus*, le *British National Corpus* ou le *Corpus of Contemporary American English*. Dans ce qui suit, nous nous intéresserons à deux bases bien différentes, le corpus *DeutschDiachronDigital* et le *Deutsches Referenzkorpus.* Nous le verrons, l'étiquette de « corpus » est en un sens réductrice pour qualifier ces bases.

Généralités

DeutschDiachronDigital (DDD) est le nom d'une base de données sur le vieil-haut-allemand en accès libre, hébergée par le portail de corpus de la Humboldt-Universität à Berlin et adossée à l'outil d'exploration et d'annotation Annis, comme tous les corpus de la HU.

Le *Deutsches Referenzkorpus* (DeReKo), pour sa part, est un portail d'exploration, qui donne accès à une base de données textuelles en accès libre une fois franchie l'étape de l'enregistrement gratuit. Le DeReKo est développé et hébergé par l'Institut für Deutsche Sprache (Mannheim). Il est adossé à l'outil d'exploration Cosmas II. De façon assez caractéristique, de même que beaucoup de linguistes appellent « Annis » le portail de corpus de la HU, « Cosmas » est souvent traité comme le nom du corpus plus que comme celui de l'interface d'exploration. A chaque fois cette interface d'exploration préexiste au corpus. Le site est conçu pour elle. Les données textuelles annotées et/ou lemmatisées ne sont exportables qu'après une requête et à partir de l'interface.

DeutschDiachronDigital

Le DDD repose donc largement sur l'outil, ou plutôt l'instrument, Annis (pour « Annotation of Information Structure »). Cet outil d'exploration de corpus implémenté sur un navigateur web a été mis au point par l'ancien *Sonderforschungsbereich* berlinois en linguistique théorique, dont le site héberge également la banque de corpus en ligne où figure la base DDD.

Il faut toutefois relever qu'Annis est téléchargeable seul, et donc utilisable indépendamment du site. Il peut tout à fait être exploité pour d'autres corpus.

Annis permet une annotation syntaxique (sous la forme d'arborescences dépendancielles) et donc des recherches comme « donne-moi tous les GN incluant 3 constituants dont un GPREP » (pour une étude sur la valence nominale, par exemple). Mais toute recherche implique la maîtrise du langage de requête (*Query Language*) propre à Annis. Il n'y a pas d'assistant, juste une liste de requêtes-types à imiter, et un manuel.

L'une des questions posées dans la section précédente était celle d'un éventuel privilège (ou d'une antériorité) de la ligne de texte d'origine (notée *edition*). Ici, la seule spécificité de cette ligne de texte est qu'on peut y réclamer une forme directement sans recourir au langage de requête. Il s'agit en fait d'un artifice de programmation : cette ligne étant la seule qui puisse être explorée sans que l'on formule à quelle ligne d'annotation doit se trouver l'information recherchée, Annis interprète l'absence de consigne sur ce point comme une consigne de chercher dans la ligne « edition ». Toutes les lignes d'annotation peuvent indifféremment faire l'objet d'une requête.

À bien des égards, le DDD fait figure de pièce rapportée pour Annis : les fonctionnalités de recherche syntaxique (dépendancielles) ne sont pas encore disponibles dans toutes les bases de données, et en particulier pas pour le DDD. La recherche « donne-moi tous les GN incluant 3 constituants dont un GPREP » n'y est donc pas encore possible. Cela impliquerait qu'une équipe de linguistes diachroniciens et d'informaticiens mettent au point un système de traitement automatique des relations syntaxiques dans les textes en vieil-haut-allemand.

{DDD_Tatian1.tif}

Figure II : Aperçu d'une fenêtre de travail sur le DDD, avec les annotations d'un énoncé du Tatian.

Du point de vue du mode de coexistence de l'instrument et du corpus, il n'en demeure pas moins que l'on observe une forte incorporation du texte à l'instrument ; le corpus n'a plus d'existence en-dehors du système d'analyse et d'exploration. Il y a donc une forme d'asymétrie : l'instrument préexiste au texte et le texte doit s'adapter à l'instrument.

L'objet technique présente un fort degré de clôture, mais il est permis de se demander si le corpus est véritablement un milieu associé de l'individu technique, ou bien si l'on n'a pas affaire à un seul ensemble technique organisé autour de l'instrument d'exploration et dont le corpus serait un élément.

DeReKo - Cosmas II

Le DeReKo n'est pas à proprement parler « un » corpus, mais une base de corpus. Par défaut, lorsque l'on parle du DeReKo, il s'agit du W-Archiv, le corpus général de l'allemand écrit, qui sert de base aux autres corpus.

On observe ici une certaine coadaptation du corpus et de l'outil : Cosmas I puis II ont été mis au point à l'IDS pour le projet DeReKo ; Cosmas n'est pas en open source, n'est pas téléchargeable ; Cosmas I/II n'est donc utilisable que pour le DeReKo. Symétriquement, le DeReKo n'est interrogeable et exploitable que par Cosmas II.

Comme pour le DDD, le privilège du texte d'origine se manifeste par la possibilité de demander à Cosmas toutes les occurrences d'une forme donnée en se contentant de la taper. Mais il s'agit d'une fausse évidence, puisque même dans ce cas, l'interrogateur se voit proposé de choisir entre toutes les variantes typographiques de la forme en question. Autre artifice de programmation : une simple espace entre deux formes dans une requête est interprétée comme une façon de demander que ces deux formes soient immédiatement consécutives. Ici aussi, l'absence de métalangage est liée au fait que le non-recours à une méta-consigne concernant la succession des formes n'est possible que dans un cas de figure (la consécution immédiate), ce qui a permis de programmer Cosmas pour qu'il assigne à cette absence de consigne une valeur équivalant à une consigne.

Hormis ces requêtes élémentaires, toutes les requêtes incluant plusieurs termes ou plusieurs formes d'un terme impliquent un minimum de codage. Ici aussi, on peut uniquement s'appuyer sur un petit répertoire de requêtes-types à imiter, et un manuel.

Cosmas permet ainsi d'obtenir des listes d'exemples apparaissant au format KWIC[28] exactement comme AntConc ou TXM[29], mais aussi des profils de co-occurrence.

Au sein du DeReKo, il convient d'isoler les « archives C et T », deux corpus reprenant une partie des textes de l'archive W mais avec une annotation syntaxique. Celle-ci est réalisée par deux instruments différents : soit Connexor (C), soit TreeTagger (T), que nous connaissons.

{dereko_treetagg_coocc.tif}

Figure III : table des cooccurrences pour une requête « verbe modal conjugué suivi d'un adverbe, avec un écart compris entre 0 et 5 mots » pour TreeTagger sur l'archive T du DeReKo, 7 juin 2017.

Ici, il y a un assistant dans la formulation des requêtes, qui permet de ne pas avoir à maîtriser le langage de requête comme avec TXM, si bien que le texte de la requête apparente ne sera pas le même pour le TreeTagger allemand sur TXM et pour le TreeTagger du DeReKo-T-Archiv. À noter, dans le même état d'esprit, que Connexor et TreeTagger ne permettent pas les mêmes requêtes : Connexor permet de faire des recherches par parties du discours en distinguant également certaines catégories grammaticales (tous adjectif vs. adjectifs au degré zéro vs. adjectifs au comparatif vs. adjectifs au superlatifs ; toutes formes verbales vs.

28 *KeyWord In Context* ou « mot-clé en contexte » : format de concordancier permettant de faire apparaître le terme recherché en pivot central de la fenêtre d'affichage, précédé et suivi de son contexte immédiat, défini par exemple comme l'énoncé ou comme les X mots graphiques précédant ou suivant le pivot. Le format KWIC fournit notamment une aide précieuse aux phases d'indexation et d'annotation manuelle des occurrences d'un terme dans un corpus.

29 Voir figure 6 dans l'annexe en ligne.

infinitifs vs. participes vs. formes fléchies ; toutes formes verbales fléchies vs. fléchies à l'indicatif vs. fléchies au subjonctif 1 vs. fléchies au subjonctif 2 ; fléchies à tout temps de l'indicatif vs. fléchies au présent vs. fléchies au prétérit, etc.). TreeTagger pour DeReKo, de son côté, ne distingue pas les degrés de l'adjectif mais oppose les épithètes et les autres ; ou ne reconnaît pas les temps ni les modes du verbe, mais isole les verbes à l'impératif. Le choix de l'archive instrumentée (archive C ou archive T) sera donc déterminé par la nature de la requête visée.

Fonctionnalités avancées

Du point de vue de l'utilisation technique, la base de données textuelles et l'outil n'ont plus d'existence distincte et forment un seul outil. Cet outil ne se qualifie pas forcément pour autant comme individu technique : on voit mal ici ce qui jouerait le rôle de milieu associé. Il s'agit plutôt d'un seul ensemble technique faiblement individué comprenant une pluralité de composants.

En première instance, l'ensemble COSMAS / DEREKO maintient un fort privilège du texte-source, y compris dans sa matérialité typographique. Les « données » n'en sont toujours pas vraiment, puisqu'elles procèdent toujours des choix d'annotation conscients ou inconscients liés au recours à tel ou tel instrument, ou à la formulation de la requête, mais la part d'autonomie conservée par la base textuelle dans l'ensemble et le biais théorique en faveur de l'analyse des cooccurrences maintiennent le degré d'assimilation technique des textes à un niveau relativement bas comparé au DDD par exemple.

Conclusion : La linguistique, de l'atelier à la fabrique ?

Quelle leçon tirer de ce bref parcours ? Il nous semble ressortir de ces premiers coups de sonde que la linguistique instrumentée n'est ni plus ni moins « empirique » que celle reposant sur l'analyse dite qualitative d'exemples arbitrairement sélectionnés. Cela ne signifie pas qu'elle ne constitue pas un apport : elle permet effectivement un accroissement des types de recherche possibles, et permet de faire dire autre chose au corpus (et de constituer des corpus d'un nouveau type pour leur faire dire ces autres choses). Elle induit un nouveau rapport aux objets de la recherche, qui restent plus que jamais des construits, et non des « donné(e)s ». Nulle catastrophe à cela : la surdétermination théorique de la démarche du scientifique étant sans doute de toute façon inévitable, qu'il y ait ou non recours aux machines, l'essentiel est qu'elle soit consciente et avouée.

Mais il n'en demeure pas moins que la linguistique des machines induit effectivement une nouvelle position de l'analyste, qui s'intègre à son propre dispositif de recherche et abandonne (même sans s'en rendre compte) l'omni-intervention démiurgique qui caractérisait parfois l'exercice avant le recours aux machines. Pour nommer ce changement et conclure cet itinéraire, laissons parler une dernière fois Simondon :

> « Ce n'est pas essentiellement par la dimension que la fabrique se distingue de l'atelier de l'artisan, mais par le changement du rapport entre l'objet technique et l'être humain : [...] la fabrique utilise de véritables individus techniques tandis que, dans l'atelier, c'est l'homme qui prête son individualité à l'accomplissement des actions techniques. [...] L'ingénieur, *engineer*, l'homme de la machine, devient en fait l'organisateur de l'ensemble comprenant des travailleurs et des machines. Le progrès est saisi comme un mouvement sensible par ses résultats, et non en lui-même par l'ensemble d'opérations qui le constituent[30]. »

30 SIMONDON Gilbert, *Du mode d'existence des objets techniques, op. cit.,* p. 163-164.