Théorie, terrain et techniques : quels recoupements et quelles médiations ?

Adrien Barbaresi

▶ **To cite this version:**

Adrien Barbaresi. Théorie, terrain et techniques : quels recoupements et quelles médiations ? : Un exemple en linguistique informatique. Rencontre méthodologique ENthèSe, " Pour une nouvelle appréhension des textes et données textuelles : pratiques, outils, méthodes ", May 2011, Lyon, France. halshs-00965371

HAL Id: halshs-00965371

https://shs.hal.science/halshs-00965371

Submitted on 25 Mar 2014

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.

Théorie, terrain et techniques
Quels recoupements et quelles médiations ?
Un exemple en linguistique informatique

Adrien Barbaresi

Laboratoire ICAR
ENS Lyon

27 mai 2011

1 Introduction

L'existence d'une zone d'expérimentation est commune à de nombreuses disciplines. Qu'il s'agisse de retours d'expérience ou d'une démarche exploratoire, le domaine empirique peut jouer un grand rôle dans la recherche scientifique. Si les débats d'idées concernant son importance ou sa validité ne sont pas nouveaux, sa thématisation sous le vocable de « terrain » au sens large et en-dehors du champ de la biologie ou de l'anthropologie semble beaucoup plus récente.

Au sens courant comme dans le contexte scientifique [1], on peut dire que la connaissance et la valorisation du terrain en tant que tel s'est construite en partie contre un savoir abstrait, théorique (alors même qu'étymologiquement la théorie a partie liée avec le champ du visible). Ce concept décrit un lien étroit entre démarche scientifique et réalité concrète.

Le sujet choisi se veut une invitation à penser la théorie et le terrain (perçus comme un ensemble et sans exclure leurs pluriels) non en disjonction ni même en relation de subordination mais bien en combinaison. Il s'agit d'interroger leur statut et partant de s'inscrire dans une réflexion plus large sur les rapports entre la science et le réel. Plus concrètement, on s'intéressera à la détermination de l'objet des pratiques scientifiques.

C'est dans ce cadre qu'il peut être utile d'aborder le rôle de la technique, qui représente l'intermédiaire tout désigné face au terrain d'expérience. Or, si elle semble être de prime abord une médiation idéale, par exemple à travers le dispositif expérimental, elle a tendance à prendre une certaine autonomie, au point de remettre en question des modèles

1. Notons par ailleurs que le terme utilisé en anglais ou en allemand, contenant aussi la signification de « champ », évoque un terrain plus fertile : si on parle en français de « préparer » ou de « sonder » le terrain, on perçoit moins l'existence de récoltes ou de résultats.

établis. L'omniprésence de techniques peut changer l'acception du terrain lui-même, que ce soit dans sa perception (par les instruments) ou dans son étendue (restreinte à leur domaine d'application). La légitimation scientifique est l'un des ressorts de ce processus.

La linguistique, et plus précisément la linguistique informatique et ce que l'on appelle depuis peu la linguistique de corpus, illustre à plusieurs égards les oppositions structurantes mais aussi les possibles recoupements entre ces approches.

2 Contexte

2.1 « Bureau » et « terrain »

La dichotomie entre « bureau » et « terrain » recouvre sous une apparente simplicité de nombreux enjeux. De la définition du champ disciplinaire aux implications politiques de la recherche, elle traverse un large spectre qui comporte notamment les critères de scientificité.

La linguistique et la linguistique informatique sont des disciplines qui depuis les années 1950 ont plutôt été ancrées du côté « bureau », dans des descriptions d'abord théoriques des phénomènes, ce qui a eu une influence certaine sur leur évolution récente.

2.1.1 L'ancrage en linguistique

Le champ disciplinaire de la linguistique académique semble s'être d'abord défini par la théorisation de phénomènes avant de s'ouvrir au recueil de connaissances empiriques. Les années soixante ont ainsi vu les débuts de l'institutionnalisation d'une « linguistique de terrain » qui s'est développée à part, en partie contre la « linguistique de bureau ». Ces deux traditions ont surtout évolué en parallèle.

> « L'objet de la linguistique, du structuralisme saussurien au génerativisme, a donc été construit en accordant la priorité à l'invariant, et à la dissociation entre phénomènes "linguistiques" d'une part (le code) et phénomènes communicationnels et socioculturels d'autre part (les usages des codes). [2] »

En effet, selon Philippe Blanchet, la divergence porte sur des modèles épistémologiques. Une tradition intègre les critères de scientificité issus des sciences exactes, parle d'une « théorie scientifique » de la langue, utilise une « méthode hypothético-déductive appliquée à l'analyse atomisante d'une "mécanique interne" » et vise une « recherche logico-mathématique d'objectivité et de prédictibilité ». A l'inverse, l'autre tradition intègre des critères liés aux spécificités des pratiques humaines décrites, utilise une « méthode empirico-inductive d'observation en contexte » et propose une interprétation nuancée [3].

Cette distinction a également été faite dans le domaine anglo-saxon : par exemple à l'image du linguiste « de fauteuil » (*armchair linguist*) [4]. Elle est par exemple particulièrement pertinente concernant la question de l'acceptabilité des énoncés, qui, notam-

2. [Blanchet, 2003, p. 287]
3. [Blanchet, 2003, p. 288]
4. [Fillmore, 1992]

ment dans le cadre de la grammaire générative, a vu jusqu'à récemment s'imposer un certain arbitraire théorique.

Ainsi, la tendance chez certains linguistes à conférer à leur simple faculté d'introspection un caractère universel peut être vue comme le parangon du primat de la théorie sur les données d'expérience [5]. Citant Jean-Claude Milner, Pierre Corbin critique le tri arbitraire d'énoncés jugées acceptables ou non fondées sur des théories de la langue d'un linguiste expert, qui ne veut pas prendre en compte ou ne peut pas expliquer des phénomènes pourtant attestés.

Dans un article sur l'enquête en linguistique, Gabriel Bergounioux propose deux explications à cette partition et au prestige certain du « bureau ». La première est scientifique, elle a trait à l'abstraction des phénomènes :

> « plus les procédures sont abstraites, modélisables par graphes et algorithmes, moins la déperdition provoquée par le philologue ou le linguiste est flagrante (ce pourrait être une des raisons du prestige de la syntaxe dans les études linguistiques). [6] »

La seconde est plus sociologique (voire idéologique), elle est liée aux structures universitaires et à l'administration du savoir en France :

> « Le marché de la linguistique en France, centralisé par le quasi-monopole de Paris sur la circulation des biens symboliques et dominé par les études littéraires dans l'organisation universitaire, favorise objectivement la recherche abstraite, universalisante ou formalisante, au détriment des pratiques de recension descriptives, autrement dit, ceux qui travaillent en bibliothèque plutôt que sur le terrain. La présence dans les bureaucraties d'administration de la science, par exemple, est plus favorable à une carrière dans l'enseignement supérieur que l'enquête. [7] »

On peut donc penser qu'il y a des enjeux politiques derrière un type de recherche donné, de même que dans un cadre plus large des conceptions différentes de l'homme et des phénomènes sociaux.

2.2 Une évolution générale : les technosciences

Hottois retrouve d'ailleurs dans la linguistique formelle héritée des théories de Chomsky une manifestation de ce changement : « La grammaire générative est l'expression la plus achevée du remplacement de l'essence langagière et théorique de l'homme par son autre opératoire. [8] » Peut-être est-ce là une vision trop univoque, étant donné que la grammaire générative est issue par filiation de la linguistique des années 1930, très attachée à l'observation du terrain.

Le réel tel que nous le connaissons et le rencontrons aujourd'hui se manifeste souvent par le biais de la technique, par exemple sous la forme d'archivage distant, d'organisation

5. [Corbin, 1980]
6. [Bergounioux, 1992, p. 6-7]
7. [Bergounioux, 1992, p. 18]
8. [Hottois, 1984, p. 61]

des sources en bases de données ou de chaîne de traitement automatisée. Le terrain se trouve équipé voire quadrillé par un ensemble de techniques, il reste modifiable et malléable bien après sa saisie ou son enregistrement et devient une ressource.

Le contexte de ce que l'on appelle de plus en plus fréquemment les technosciences joue un rôle dans cette médiation par des techniques dont on peut dire au final qu'elle font plus que simplement pérenniser ou classer les informations.

On peut retenir de la description que fait Bernadette Bensaude-Vincent [9] des technosciences trois aspects qui ont ici leur pertinence.

Premièrement, l'entrée en scène des politiques scientifiques et des agences de moyens, c'est-à-dire par exemple la mutualisation des moyens politiques et financiers de la recherche à grande échelle et les financements accordés de manière concurrentielle à des projets planifiés. Ce changement s'accompagne de la montée en puissance d'une logique de rentabilité à court terme des résultats qui joue un rôle dès la sélection des projets.

Deuxièmement, en lien avec cette évolution, la recherche se voit toujours plus orientée vers les applications ce qui confère à la technique le rôle d'outil indispensable dans la production du savoir et la place au premier plan : il ne s'agit plus nécessairement de comprendre en profondeur des phénomènes mais de rendre utilisable voire profitable la nature ou le réel.

Troisièmement, le savoir et les champs scientifiques se voient recomposés notamment à travers la notion de convergence des théories et des disciplines. Les exemples de cette dernière tendance sont présents sous plusieurs formes. Citons le concept de « théorie de l'unification » en physique comme champ disciplinaire, le développement des sciences de la complexité comme une fusion de différentes disciplines et différents modèles théoriques, ou du point de vue des politiques publiques (le financement de la recherche et son administration) l'existence à Lyon de l'IXXI, pour étudier la complexité sous ses différentes formes, ainsi que celle de nombreux « clusters de recherche » dans lesquels la convergence des moyens et des scientifiques est centrale.

On peut également voir dans cette évolution le développement d'une nouvelle attitude face au donné et aux données, et de nouvelles pratiques. Gilbert Hottois, un des fondateurs du concept de technoscience, y voit le règne de l'opératoire. Dès le début, il a considéré qu'il en allait d'un changement de nature du réel et notamment de la réalité scientifique : « est réel ce qui est (re)productible, manipulable, transformable et non plus le visible, l'intelligible ou le compréhensible. [10] » Dans la création et l'utilisation des ressources du chercheur l'accent porterait donc sur l'interopérativité comme critère de réalité.

En revanche, quand on pense en ces termes les changements intervenus depuis les années 1990 dans l'abord aux textes, l'existence de textes numérisés manipulables et non plus simplement textes sources fait sens au-delà du cadre d'une vision mathématique du langage.

9. [Bensaude-Vincent, 2009], cf notamment p. 11 et 80
10. [Hottois, 1984, p. 62]

2.3 Le cas de la linguistique informatique

Le caractère opératoire des rapports entre science et réel étudié se manifeste en linguistique informatique de deux façons différentes, où l'on retrouve une dichotomie entre bureau et terrain en pleine évolution. Les progrès récents dans la numérisation des textes ainsi que dans les outils d'exploration changent progressivement les contours de la discipline, d'autant plus que certains chercheurs venus de la recherche en informatique disposent d'une réelle connaissance des outils en développement.

La divergence décrite plus haut semble se traduire dans ce cas par un schéma d'opposition, tant en termes d'objectifs que de méthodes. On peut ainsi voir d'un côté des objectifs théoriques, de l'autre des visées pratiques, d'une part des méthodes symboliques et de l'autre des méthodes quantitatives ou numériques.

Marcel Cori parle d'un « TAL [11] théorique », incarnant la « vision d'un traitement automatique étroitement associé à la linguistique formelle [12] » et d'un « TAL robuste », dont les critères seraient de traiter le « tout venant », de trouver une seule solution sans risquer de se bloquer et enfin de permettre des procédures d'évaluation quantitatives [13].

Il remarque le poids supérieur des visées pratiques depuis le début des années 90, les méthodes statistiques et probabilistes débouchant sur des outils [14]. En effet, l'étiquetage morpho-syntaxique par exemple, processus important pour toute étude qui se veut linguistique, se fait très souvent en suivant ce type de méthodes.

Les années 1990 ont vu la mise en place de nombreux corpus librement disponibles comme de nombreux outils d'analyse, changeant le statut des preuves disponibles et leur traitement. Dans la lignée du « web comme corpus », on a assisté à un élargissement prodigieux du terrain et du même fait à un renouvellement du problème des données dites attestées.

Les études fondées sur de grands corpus ont donné lieu à de nouveaux paradigmes et à une « linguistique de corpus » que certains chercheurs veulent considérer comme un ensemble autonome, ce qui ne va pas de soi. Si la démarche, les pratiques et les techniques utilisées semblent justifier l'emploi de ce terme, on peut se demander s'il s'agit vraiment d'une discipline expérimentale, d'une part par ses méthodes et de l'autre par ses prétentions. Mathieu Valette estime ainsi que « la linguistique de corpus ne sera, selon toute vraisemblance, jamais établie en discipline académique. [15] ». Il est d'avis que la linguistique doit prendre position face aux nouveaux enjeux mis au jour par l'évolution des techniques.

Marcel Cori pose le problème en termes de caractérisation épistémologique du traitement automatique du langage et de la recherche en linguistique. Il voit dans la linguistique de corpus une caractérisation double en ce qu'elle se considère à la fois comme une recherche sur la langue et comme une linguistique appliquée [16]. En résumé, la pratique fait retour sur la théorie et prétend en modifier l'acception.

11. traitement automatique du langage
12. [Cori, 2008, p. 96]
13. [Cori, 2008, p. 98]
14. [Cori, 2008, p. 95]
15. [Valette, 2008, p. 9]
16. [Cori, 2008, p. 105]

C'est là un phénomène à prendre en compte si l'on considère, comme certains chercheurs, qu'il n'y a plus de séparation hermétique entre les approches fondées sur un corpus (*corpus-based*) et celles partant d'un modèle théorique (*theory-driven*), mais bien une approche hybride fondée sur la connaissance pour évaluer, tester, générer des hypothèses [17].

3 Méthode

3.1 Croiser les approches

Dans cette lignée, mon projet était au départ de tenter de « croiser les approches », de poursuivre conjointement des recherches sur le plan théorique comme sur le plan du développement logiciel. Il y a à cela des raisons pratiques, on peut en effet imaginer tester l'effet des règles choisies et les corriger en conséquence [18]. Or, de fait, je fais moi-même l'expérience du parallélisme des deux démarches et de l'écueil potentiel que représente la double caractérisation décrite plus haut.

À ce stade de mes recherches je pense qu'il est difficile d'établir une continuité entre le savoir linguistique tel qu'il existe par exemple en topologie et le repérage généralisé, dans mon cas, d'un ensemble de phénomènes linguistiques liés à la complexité. Je me trouve confronté à un dilemme connu : « choisir entre des modèles plus adéquats à la représentation des langues mais qui ne donnent pas lieu à des réalisations pratiques acceptables, ou sacrifier l'expressivité linguistique. [19] »

Je pense qu'il faut donc assumer que les « ruses » trouvées sur le terrain ne soient pas sauf exception pertinentes du point de vue théorique. La détermination des phénomènes à observer ne se fait donc pas systématiquement par leur théorisation, mais par l'examen d'un échantillon de textes étalonnés « à la main ». Je procède donc à un repérage empirique des critères, parfois guidé par la connaissance que j'ai d'un système linguistique (notamment pour ce qui est des langues étrangères que j'ai apprises).

Dans ce regard porté sur le terrain c'est bien l'expertise linguistique qui est mise en avant et non l'apprentissage automatique, qui serait l'autre solution couramment pratiquée. Il s'agit en effet de ne pas perdre de vue l'architecture des processus à l'œuvre, comme cela pourrait être le cas avec l'utilisation de techniques d'intelligence artificielle nécessitant un apprentissage ciblé qui conduisent souvent à ce que l'on appelle des « boîtes noires ».

Dans l'ensemble, je compte recourir à une approximation de traitements plus complexes en gardant la main sur le repérage des phénomènes, même si un peu de « mise en instrument [20] » pourrait être nécessaire à la réalisation d'un produit fini.

C'est la raison pour laquelle je préfère parler de recoupements plutôt que de croisement. Sans relever de la pure coïncidence, ces recoupements heureux au fil des recherches peuvent également être suscités par l'examen du texte et de ses composantes (notamment les annotations). Il faut alors scruter des signes dans les textes à la recherche de points

17. [Wallis & Nelson, 2001]
18. [Cori, 2008, p. 99]
19. [Cori, 2008, p. 100]
20. voir les théories de Gilbert Simondon à ce sujet.

d'appui à partir desquels systématiser l'approche d'un phénomène. En bref, trouver des techniques.

3.2 Dispositif et traces d'expérience

Cette démarche est indissociable d'un rapport construit aux données, et ce à double-titre : d'une part par les outils, instruments et techniques employés, et d'autre part par leur produit, l'existence de traces d'expérience, par exemple l'enrichissement du texte par différents logiciels fonctionnant selon différentes approches, dont on peut alors tirer le produit.

Parce qu'elle s'inscrit dans le processus de la « scientifisation » tel que Bruno Latour le décrit, sa conception des traces multipliées, lues, combinées est pertinente pour décrire un mouvement de recherche :

> « mobiliser, fixer immuablement les formes, aplatir, varier l'échelle, recombiner et superposer les traces, incorporer l'inscription dans un texte, fusionner avec les mathématiques [21] »

Cette liste d'étapes permet d'aborder ce qui fait le propre du travail du chercheur, qui tente de dégager de nouvelles connaissances par l'apposition de traces et le jeu avec différentes échelles. Cette tâche n'est d'ailleurs le plus souvent pas l'affaire d'une seule personne ni même d'une même période. Des expériences à l'abstraction, Jean-Michel Berthelot parle de processus scientifique de sédimentation des notes et des données :

> « processus historique par lequel une multitude de comptes rendus, datés et signés, d'expériences et de formules partielles se transforme progressivement en résultats synthétiques anonymes et stabilisés [22] »

Les formats d'annotation et d'export (comme par exemple XML) permettent de rendre disponibles les objets d'étude, et non pas seulement un compte-rendu. Or l'enrichissement d'un texte, l'ajout d'informations, est une activité productrice d'un grand nombre de traces qu'il faut rendre accessibles (y compris à un parcours par des machines). Le texte fait alors figure de réalité expérimentale et d'économie de la preuve [23].

Il ne s'agit pas d'un simple archivage ordonné des textes et des données mais bien de la mise en place d'un dispositif à même d'assurer une productivité scientifique. D'une certaine manière, le terrain inclut son dispositif. C'est notamment ce en quoi il cesse d'être directement « le terrain » pour laisser la place à une restitution organisée qui forme le substrat à partir duquel des hypothèses peuvent être vérifiées et des raisonnements attestés.

L'existence d'un tel ensemble est commune à plusieurs disciplines. On peut par exemple citer cette réflexion de Yann Calbérac dans son étude du terrain en géographie :

---

21. [Latour, 1985]
22. [Berthelot, 2003, p. 29]
23. « il faut donner aux phénomènes une forme qui soit telle que l'on puisse, en la retravaillant, gagner sur eux plus d'informations qu'on y a mis. » [Latour, 1985]

> « La collecte et la mise en circulation des données dans des dispositifs adéquats aboutissent à une mise en ordre du monde qui relève d'une cosmétique : ces données sont triées, classées, archivées. Ces opérations textuelles permettent l'accumulation et donc l'archivage mais ne rendent pas pour autant le terrain intelligible. Cette intelligibilité du terrain est le résultat d'une deuxième opération, celle de mise en ordre des données accumulées, de leur traitement, de leur analyse et de leur restitution. [24] »

Tous les modes d'organisation ne se valent pas, il existe différentes configurations qui s'adaptent à des besoins spécifiques. On peut penser en ce sens à faire une différence entre un corpus « accumulateur » et un corpus d'exploration, entre un objet empirique « complet » mais insaisissable et objet empirique observable, selon une hypothèse et au prix d'une réduction de sa richesse [25].

Mieux, le terrain vu par le truchement des instruments offre des dimensions d'analyse supplémentaires, le couple dispositif-terrain peut s'avérer productif à plus d'un terme, comme le remarque Mathieu Valette qui prend le parti d'une « science des textes instrumentée » :

> « Les grandes masses de données textuelles ou documentaires nécessitent, pour être analysées et décrites, des dispositifs expérimentaux et des instruments *ad hoc*. Cette instrumentation permet de construire de nouveaux observables qui seraient demeurés invisibles autrement. [26] »

Enfin, à la cosmétique peut s'ajouter en linguistique informatique le dispositif expérimental lui-même, les programmes et scripts d'analyse, qu'il est de plus en plus fréquent de rendre non seulement publics mais aussi améliorables au sein d'une communauté scientifique ou technique (voir le rôle grandissant de l'« open-source »).

4 Changements et modifications du terrain

Pour conclure je voudrais revenir sur la nature du terrain, qui peut être par nature mouvant et difficile à fixer, sinon abusivement, par un dispositif scientifique. La nature du langage est ainsi d'être difficile à saisir comme entité. Les larges corpus disponibles font croire à la possibilité d'une analyse en bloc et à l'établissement d'une forme de « moyenne » grâce au nombre et éventuellement à la représentativité des données recueillies. Néanmoins, du côté théorique, la notion même de langue fait débat, un grand corpus peut être vu comme une intégrale d'idiolectes qui, si elle offre des conclusions commodes, n'est qu'un rapport au sens le plus mathématique de la réalité étudiée et non une approche fine et soucieuse de menues différences.

---

24. [Calberac, 2010, p. 104]
Voir aussi : « C'est ce dispositif dans son entier – dans lequel les données sont traitées, interrogeables, comparables... – qui se substitue au terrain inintelligible. » [Calberac, 2010, p. 99]
25. [Loiseau, 2007]
26. [Valette, 2008, p. 11]

4.1 Conséquences sociales et politiques

On retrouve donc ici les problèmes posés par la puissance des outils d'expérimentation. Sous un autre angle, d'autres transformations liées à l'activité scientifique ou ses conséquences sont à prendre en compte. La modification du réel par la mesure ou la prise d'information est bien connue en physique, mais elle peut également exister, peut-être sous une forme moins directement perceptible, en sciences sociales.

> « lors de productions scientifiques de tous ordres, on induit des changements chez les lecteurs ou auditeurs, qu'il s'agisse de spécialistes du domaine ou du grand public. On ne peut nier l'existence de conséquences potentielles, aussi minimes soient-elles. Or, comme le discute E. Morin [E. Morin : Science avec conscience, Paris, Fayard, 1982], pratiquer la science avec conscience, c'est prendre notamment conscience des dangers avérés ou potentiels de certaines recherches scientifiques, et des discours produits. [27] »

Or, Isabelle Léglise conteste justement une attitude commune selon elle à la plupart des linguistes : « une sorte de choix implicite d'éthique scientifique de non-intervention. [28] ». Comme l'explique Sylvain Auroux, « l'existence d'outils, l'existence d'une standardisation de la communication changent effectivement nos aptitudes. [29] ».

Il y a des risques éthiques à l'analyse de la langue, qui sont en partie liés aux enjeux politiques décrits au début de cet exposé. La méthode de recherche et les présupposés quant à une langue ou un idiolecte impliquent souvent plus qu'un simple point de vue linguistique. Les résultats d'une étude peuvent par exemple être utilisés de manière réductrice à des fins politiques.

4.2 Risques juridiques

De plus, il y a des risques juridiques, suscités par la masse de données manipulées et leur statut, tant du point de vue d'éventuels droits d'auteurs que d'éventuelles atteintes à la vie privée qui naissent des recoupements dans un vaste ensemble. Il y a donc une réflexion à tenir sur le fait de rendre disponible en bloc ce qui n'était auparavant disponible que de manière fragmentaire.

L'apport des textes du domaine public est essentiel dans une optique de transmissibilité des corpus et des résultats, d'autant que la question des droits d'auteurs ne se pose pas partout avec la même acuité. Elle est par exemple encadrée beaucoup plus strictement en Allemagne qu'en France.

Aussi une étude portant sur des articles de journaux, par exemple une comparaison entre plusieurs quotidiens, est soumise à caution. S'il est possible d'obtenir les articles simplement en téléchargeant et nettoyant des pages web, s'il est possible de les analyser librement et de citer des extraits de quelques lignes dans des publications scientifiques, il n'est pas permis de rendre disponible le texte enrichi et étiqueté.

27. [Léglise, 2000, p. 3]
28. [Léglise, 2000, p. 4]
29. [Auroux, 1998, p. 95]

À ma connaissance c'est avant tout l'identification des textes ou des personnes qui semble poser problème. D'où la nécessité de recourir à des techniques dites de « masque[30] », par exemple en remplaçant les mots étiquetés par des autres choisis au hasard dans la catégorie correspondante, ou en mélangeant les phrases du corpus au hasard. En ce qui me concerne, ces techniques ôtent toute notion de cohérence et de cohésion au texte obtenu, limitant fortement l'intérêt d'un tel corpus.

Les articles de journaux ne pourraient donc être utilisés que pour un échantillonnage et une analyse « à vue ». Dans une optique de transmissibilité du dispositif scientifique, la composition des corpus dépend fortement des sources librement reproductibles, par exemple de la nature des œuvres rassemblées par le Projet Gutenberg. Les articles scientifiques constituent à ce titre une piste intéressante.

Références

[Auroux, 1998] Auroux S. (1998). Les enjeux de la linguistique de terrain. *Langages*, **129**, 89–96.

[Bensaude-Vincent, 2009] Bensaude-Vincent B. (2009). *Les vertiges de la technoscience*. Paris : La Découverte.

[Bergounioux, 1992] Bergounioux G. (1992). Les enquêtes de terrain en France. *Langue française*, **93**, 3–22.

[Berthelot, 2003] J.-M. Berthelot, Ed. (2003). *Figures du texte scientifique*. PUF.

[Blanchet, 2003] Blanchet P. (2003). Contacts, continuum, hétérogénéité, polynomie, organisation chaotique, pratiques sociales, interventions... quels modèles ? Pour une (socio) linguistique de la complexité. *Langues, contacts, complexité. Perspectives théoriques en sociolinguistique*, p. 279–308.

[Calberac, 2010] Calberac Y. (2010). *Terrains de géographes, géographes de terrain. Communauté et imaginaire disciplinaires au miroir des pratiques de terrain des géographes français du XXe siècle*. PhD thesis, Université Lumière Lyon 2.

[Corbin, 1980] Corbin P. (1980). De la production de données en linguistique introspective. In A.-M. Dessaux-Berthonneau, Ed., *Théories linguistiques et Traditions grammaticales*, p. 121–179. Presses Universitaires de Lille.

[Cori, 2008] Cori M. (2008). Des méthodes de traitement automatique aux linguistiques fondées sur les corpus. *Langages*, **171**(3), 95–110.

[Fillmore, 1992] Fillmore C. J. (1992). "Corpus linguistics" or "Computer-aided armchair linguistics". In J. Svartvik, Ed., *Directions in Corpus Linguistics*, p. 35–60. Berlin, New York : Mouton de Gruyter.

[Hottois, 1984] Hottois G. (1984). *Le signe et la technique : la philosophie a l'epreuve de la technique*. Paris : Aubier.

[Latour, 1985] Latour B. (1985). Les ≪ vues ≫ de l'esprit. *Culture technique*, **14**, 4–29.

30. [Rehm *et al.*, 2007]

[Loiseau, 2007] Loiseau S. (2007). CorpusReader : un dispositif de codage pour articuler une pluralité d'interprétations. *Corpus*, **6**, 153–186.

[Léglise, 2000] Léglise I. (2000). Quand les linguistes interviennent : écueils et enjeux. *Revue Française de Linguistique Appliquée*, **4**.

[Rehm *et al.*, 2007] Rehm G., Witt A., Zinsmeister H. & Dellert J. (2007). Corpus masking : Legally bypassing licensing restrictions for the free distribution of text collections. *Digital Humanities*, p. 166–170.

[Valette, 2008] Valette M. (2008). Pour une science des textes instrumentée. *Syntaxe et sémantique*, **9**, 9–14.

[Wallis & Nelson, 2001] Wallis S. & Nelson G. (2001). Knowledge discovery in grammatically analysed corpora. *Data Mining and Knowledge Discovery*, **5**(4), 305–335.