CORPUS

Corpus

18 | 2018
Les petits corpus

Le corpus de pommes d'Isaac Newton était-il un Grand Corpus ?

Was Newton's apple corpus a Large Corpus?

Pierre-Yves Raccah

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/3116
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Référence électronique
Pierre-Yves Raccah, « Le corpus de pommes d'Isaac Newton était-il un Grand Corpus ? », *Corpus* [En ligne], 18 | 2018, mis en ligne le 09 juillet 2018, consulté le 08 novembre 2018. URL : http://journals.openedition.org/corpus/3116

Ce document a été généré automatiquement le 8 novembre 2018.

Le corpus de pommes d'Isaac Newton était-il un Grand Corpus ?

Was Newton's apple corpus a Large Corpus?

Pierre-Yves Raccah

1 Depuis une vingtaine d'années, la linguistique se détourne progressivement de ses objectifs premiers, qui sont la compréhension des lois qui régissent les langues humaines, au profit de l'accumulation de faits de discours et de communication, pour lesquels des descriptions causales invoquant différents domaines non linguistiques sont privilégiées (cognitifs, psychologiques, sociologiques, logiques… et même statistiques) : tout se passe comme si la linguistique ne pouvait être une science et que, par conséquent, il fallait l'encadrer en la plongeant dans des disciplines dont la scientificité serait plus reconnue. Un indice de cette mode est le passage, de plus en plus fréquent, du substantif féminin « linguistique » au syntagme « sciences du langage » (au pluriel), et même « sciences du langage et de la communication », dont les anciens départements universitaires de linguistique sont maintenant qualifiés. S'il est indispensable de tenir compte des interfaces entre la linguistique et ces autres disciplines, cette nécessité n'implique pas l'abandon de l'étude des langues en tant que telles, ni la réduction de la linguistique à l'accumulation d'occurrences de discours.

2 Ce mouvement réducteur, observable dans toutes les branches de la linguistique, est particulièrement nocif en sémantique, et ce, pour deux raisons principales.

3 (i) Comme nous le verrons plus en détail au paragraphe 1.1, les faits de la sémantique ne concernent pas que la forme des occurrences d'unités de langue, mais aussi leur(s) interprétation(s) dans différentes situations. Il en résulte qu'un corpus d'occurrences, qui ne contiendrait donc aucune indication sur les interprétations qu'elles ont occasionnées, ne peut pas fournir d'indication sur la manière dont les unités de langue utilisées contraignent l'interprétation.

4 (ii) Prétendre analyser des discours sans recourir à la sémantique des langues légitime l'idée selon laquelle expliquer l'interprétation des discours ne reposerait que sur l'intuition de l'analyste, éventuellement guidée par des « constantes » de comportement,

sans que les unités de langue utilisées dans lesdits discours n'interviennent de manière stabilisée : l'analyse des discours (et, du coup, l'Analyse du Discours) ne pourrait alors prétendre à rien de plus objectif que les explications de textes pratiquées au collège.

5 Un tel abandon constitue donc un obstacle à la prise en compte de ces interfaces : c'est, en effet, parce que les langues ne sont observables qu'indirectement, par leurs mises en discours, qu'il est permis de penser que l'étude des langues peut être à l'interface de nombreuses sciences de l'homme et de la société.

6 Cette mode, institutionnelle plus qu'intellectuelle, qui s'appuie sans doute sur la croyance populaire selon laquelle le raisonnement scientifique par excellence serait l'induction, exige, de plus en plus, des chercheurs[1] qu'ils accumulent des « faits », à partir desquels, si leur nombre est assez grand, une règle émergera d'elle-même, sans qu'il soit nécessaire de proposer ni de justifier des hypothèses théoriques, lesquelles – toujours selon cette mode – perturberaient cette émergence. Pour suivre cette mode, les chercheurs en linguistique (ou plutôt en sciences du langage et de la communication) sont donc mis en demeure de « travailler sur des corpus », si possible « sur des grands corpus », et, encore mieux, « sur des Très Grands Corpus », sans idées préconçues : un nombre croissant de chercheurs en arrivent à cette position absurde selon laquelle les objectifs expérimentaux ou théoriques, considérés comme des idées préconçues, devraient susciter une grande méfiance…

7 Je montre en détail pourquoi ce mouvement est réellement nuisible en sémantique des langues, et propose une méthodologie plus saine, faisant recours à l'expérimentation, comme c'est le cas dans les autres disciplines empiriques, puisque c'est elle qui permet de tester rigoureusement des hypothèses théoriques. L'expérimentation en sémantique, si elle peut utiliser des corpus, n'a pas, pour autant, besoin qu'ils soient très grands, ni même grands : ce n'est pas la taille de l'outil qui importe, mais la manière dont on s'en sert… J'examinerai les principales difficultés méthodologiques auxquelles est confrontée une démarche expérimentale en sémantique, et, pour illustrer les moyens de les surmonter, proposerai deux tests expérimentaux permettant d'évaluer des hypothèses théoriques concernant la description sémantique des mots d'une langue.

1. Du corpus en sémantique

8 Dans cette partie, je montre pourquoi un corpus d'occurrences ne peut pas, sauf dans un cas (que j'examinerai), servir à étayer une argumentation concernant une description sémantique, mais a généralement pour effet de masquer le recours à l'introspection du linguiste, qui doit fournir lui-même les interprétations desdites occurrences. Après quelques rappels sur l'induction et sur l'abduction, je décris une configuration très spécifique, dans laquelle un corpus d'occurrences peut être utilisé dans une argumentation rigoureuse en faveur d'une description sémantique. Pour conclure cette partie, j'examine ce qu'il conviendrait d'ajouter à un corpus d'occurrences pour que le corpus enrichi qui en résulterait évite les défauts décrits.

1.1. De la nuisibilité, en général, des corpora d'occurrences

9 La notion de *corpus* en linguistique a évolué rapidement depuis 1996 et le rôle des corpora dans les sciences liées aux langues a suivi cette évolution. Alors que, en 1996, un corpus

était seulement considéré comme « a collection of *pieces of language* that are selected according to explicit linguistic criteria in order to be used as a sample of the language » (Sinclair 1996 : 4), dès 2000, un corpus devient « une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extralinguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue » (Habert 2000 : 13). Et même si, comme le signale Geyken (2008 : 77, se référant à Kilgarriff & Grefenstette 2003) : « Comme il est impossible de mesurer ou de vérifier qu'un corpus est représentatif – il faudrait en effet connaître les proportions d'usage des genres dans la langue considérée – on affaiblit la contrainte de la représentativité, et on la remplace par la notion d'équilibrage, *balancedness*, par rapport aux types de textes », le passage de *exemples de langage* à *échantillons d'emplois déterminés*, lui, n'est pas remis en question, pas plus que celui de *sélectionné selon des critères linguistiques* à *sélectionnées et organisées selon des critères linguistiques et extralinguistiques*. Ainsi, alors que les corpora étaient des artéfacts destinés à *montrer*, ils sont devenus, pour certains du moins, des faits représentatifs (ou presque), destinés à *démontrer*.

10 La mode institutionnelle qui a pour effet de tenter d'imposer à toute recherche empirique en linguistique de fonder son argumentation exclusivement sur l'utilisation d'un (grand) corpus est déjà insolite (nous le verrons plus en détail au paragraphe 1.2) lorsqu'elle concerne la morphologie ou la syntaxe ; elle devient complètement incompréhensible lorsqu'elle concerne la sémantique : en effet, si l'on admet que la sémantique étudie ce que les langues donnent comme indications pour que les interlocuteurs construisent du sens, un *fait sémantique* est constitué d'un ensemble de triplets :

<unité de discours, situation, interprétation>

11 dont l'observation permet d'étayer ou de réfuter une description sémantique qui peut prendre la forme

<unité de langue, contraintes sur l'interprétation>

12 Il est donc incontestable que l'accumulation d'occurrences, aussi nombreuses soient-elles, ne peut pas être directement utile à la sémantique : pour chaque occurrence, il faudrait y ajouter, au moins, une description d'une situation et une description de ce qui a été réellement compris (par les sujets observés, et non pas par le sémanticien) dans cette situation. Ces *corpora d'occurrences*[2], même s'ils sont indispensables dans d'autres disciplines (comme, par exemple, les études littéraires), sont donc nuisibles lorsqu'on prétend les utiliser en sémantique des langues. En effet, leur utilisation réintroduit nécessairement l'introspection sémantique de l'observateur (en la maquillant derrière un appareil statistique ou parfois seulement présenté comme tel), puisque, par définition, un corpus d'occurrences ne contient pas d'indications sur la manière dont chaque item a été interprété, ce qui conduit l'observateur à se croire légitimé à se fonder sur l'interprétation que lui-même considère comme appropriée.

13 Il ne sert à rien de tenter de légitimer, comme pour les sondages d'opinion, les choix d'inclusion dans un corpus de tel ou tel échantillon par des contraintes de représentativité de ces échantillons : les critères de représentativité dépendent, entre autre, de l'interprétation, qui, rappelons-le, n'est pas fournie dans les corpora d'occurrence.

14 On voit en quoi cette mode est nuisible à la sémantique, et pas seulement saugrenue : alors que les chercheurs qui utilisent l'introspection savent bien qu'ils ont recours à leurs propres jugements sémantiques et se sentent contraints de justifier ces jugements, ceux

qui utilisent uniquement des corpora ont, eux aussi, recours à leurs propres jugements sémantiques, mais s'affranchissent de l'obligation de les justifier. Le fait qu'on puisse trouver un million d'occurrences d'une certaine forme linguistique n'entraine pas que ce million d'occurrences ont reçu la même interprétation. Et puisque le corpus d'occurrences ne fournit pas d'indications sur les interprétations, c'est le linguiste qui « doit » trancher, et, en général, il le fait à sa convenance, tout en conservant le nombre élevé d'occurrences comme argument en faveur de sa description.

15 Il y a cependant un cas où un corpus d'occurrences relativement abondant peut être utile dans une argumentation rigoureuse. Sa présentation est l'objet de la sous-section suivante.

1.2. De l'utilisabilité, dans quelques cas, des corpora d'occurrences

16 Les cas où un corpus d'occurrences peut servir dans une argumentation rigoureuse en sémantique, malgré l'absence d'indications sur les interprétations des éléments du corpus, sont des cas où la nature des interprétations ne joue pas de rôle dans l'objectif de l'argumentation théorique. Nous sommes donc amenés à examiner d'abord dans quels cas un corpus peut fournir un argument rigoureux, en général, et dans quel cas il ne le peut pas. Les outils intellectuels pour examiner cette question peuvent être fournis par une réflexion sur les rôles de l'*induction* et de l'*abduction* dans la démarche scientifique.

1.2.1. Induction et abduction dans la démarche scientifique

17 L'induction qui, pour des raisons difficilement explicables, passe, à tort, pour *le* raisonnement scientifique par excellence s'appuie sur l'inférence suivante :

> I : « plus on observe de cas dans lesquels un A est aussi un B, plus la croyance que le prochain A que l'on pourra observer sera aussi un B est justifiée »

18 Ainsi, selon cette doctrine, si l'on a vu 853 corbeaux noirs la croyance selon laquelle tous les corbeaux seraient noirs est plus justifiée que si l'on n'en a vu que 85.

19 Nous allons voir, d'abord, que cette doctrine est irrationnelle (ce qui n'empêche pas d'y croire, mais invalide tout raisonnement fondé sur elle…), puis, on verra que, fort heureusement, ce n'est pas du tout ce genre de croyance qui guide la démarche scientifique, mais un raisonnement qui y ressemble très grossièrement, pour un esprit peu attentif ou peu motivé par la rigueur scientifique.

20 Pour montrer l'irrationalité de la doctrine I, j'utiliserai deux arguments distincts (dont un seul serait suffisant) : un argument de bon sens, et un argument logique.

21 a) Argument de bon sens contre la doctrine I

22 Supposons que le lecteur et moi décidions de jouer à *pile ou face*. Nous choisissons une pièce de monnaie, que nous contrôlons dument : pile d'un côté et face de l'autre. Après un certain temps de jeu, la pièce est tombée 853 fois du côté *pile*, et zéro fois du côté *face* (le fait que cette situation soit peu probable n'est, bien entendu, pas pertinent ici). Nous avons donc 853 cas dans lesquels la face visible de la pièce est aussi la face pile. Cette observation justifierait-elle que l'on croie que le prochain lancer donnera aussi *pile* ? La réponse, bien sûr, est non. On voudrait même ajouter « au contraire », mais ce serait une erreur : les lancers sont indépendants et il n'y a aucune raison que les résultats des lancers précédents influencent le résultat suivant.

23 J'ai certes choisi un cas éminemment défavorable à la doctrine I, mais, pour montrer qu'une proposition générale est invalide, il suffit d'exhiber un cas où elle est fausse : c'est ce que j'ai fait, et peu importe si les gens qui étaient attachés à la doctrine I se sentent un peu ridicules. Par ailleurs, il s'agissait d'un argument de bon sens, et chacun sait que le bon sens peut être caricatural...

24 b) Argument logique contre la doctrine I

25 Cet argument, parfois connu sous le nom de « paradoxe de l'induction » utilise l'équivalence logique entre une implication et sa *contraposée* :

> A ⊃ B est logiquement équivalent à ~B ⊃ ~A

26 où le signe ⊃ est celui de l'implication, le signe ~ celui de la négation. Lorsque deux expressions sont logiquement équivalentes, elles ont les mêmes valeurs de vérité, quelle que soit l'interprétation de leurs composants[3].

27 Ainsi, la proposition selon laquelle « tous les corbeaux sont noirs » (pour tout x, si x est un corbeau, alors x est noir) est logiquement équivalente à « toute entité non noire est distincte d'un corbeau » (pour tout x, si x n'est pas noir, alors x n'est pas un corbeau). La doctrine I, appliquée à cette dernière proposition, suggère que

> si l'on a vu 853 entités non noires qui ne sont pas des corbeaux la croyance selon laquelle toutes les entités non noires seraient distinctes d'un corbeau serait plus justifiée que si l'on n'en avait vu que 85

28 Mais comme « toutes les entités non noires sont distinctes d'un corbeau » est, comme on vient de le voir, logiquement équivalent à « tous les corbeaux sont noirs », en suivant I, on doit admettre que l'observation de 853 draps blancs serait un meilleur argument en faveur de la thèse selon laquelle tous les corbeaux sont noirs, que ne le serait l'observation de 85 draps blancs[4].

29 Ce qui fait que notre 853e corbeau noir semble rendre plus plausible la croyance selon laquelle tous les corbeaux seraient noirs n'est pas pris en compte dans la doctrine de l'induction ; c'est d'ailleurs aussi ce qui rend un peu ridicule la croyance que la 853e pièce qui tombe sur *pile* rendrait plus plausible la croyance que, au lancer suivant, la pièce tomberait sur *pile*. Il s'agit d'une hypothèse cachée (implicite, inconsciente), valable pour les corbeaux mais pas pour les lancers de pièces, selon laquelle la propriété que l'on prétend tester aurait une cause générale, et que tous les échantillons observés et à observer ne seraient que des illustrations de cette relation causale. Dans le cas des corbeaux, il pourrait s'agir de l'hypothèse selon laquelle une propriété génétique de l'espèce *corbeau* serait responsable d'un plumage noir. Dans l'exemple des lancers de pièce, si les joueurs n'ont pas pris la précaution de vérifier la pièce au préalable, le 853e lancer donnant *pile* pourrait effectivement (et rationnellement) faire penser que tous les lancers donneront *pile*, à condition de faire l'hypothèse que la pièce est truquée et que ses deux faces sont *pile*.

30 Cette hypothèse cachée, lorsqu'elle n'est pas dissimulée, est une hypothèse *abductive*. L'abduction consiste donc à formuler une hypothèse causale destinée à expliquer un ensemble d'observables. Comme l'induction, une inférence de ce type n'est pas certaine, mais elle peut être testée. C'est ainsi que la démarche scientifique consiste à formuler des hypothèses visant à expliquer les observables, puis à tester ces hypothèses expérimentalement. Muni d'une hypothèse abductive, le chercheur conçoit alors des expériences visant à *réfuter* cette hypothèse : tant qu'il n'y arrive pas, et s'il a fait de nombreuses tentatives pour la réfuter, l'hypothèse peut être considérée comme

acceptable. C'est probablement cette multiplication des tentatives de réfutation qui a permis de faire passer le recours aux grands corpora pour prétendre valider une hypothèse comme une garantie de scientificité : il y a donc un contresens à l'origine de cette doctrine : la multiplication des observations ne doit servir qu'à tenter de réfuter l'hypothèse, et ne peut en aucun cas la valider. Newton n'a pas eu besoin d'un corpus de 853 pommes pour émettre l'hypothèse abductive de la gravitation universelle… Il a juste trouvé quelque chose à expliquer et cherché une explication : c'est l'observation de la chute d'une (et une seule) pomme qui lui a suggéré l'explication étrange qu'il a proposée ; l'accumulation d'observations (y compris, à travers les siècles) ne visait pas à *prouver* la gravitation universelle mais au contraire (si l'on peut dire) à tenter de la réfuter. Et, si cette dernière a résisté vaillamment (jusqu'à la relativité générale, en 1915), ce n'est pas grâce à un Très Grand Corpus de pommes ou d'autres choses, mais parce que personne n'avait réussi à la réfuter malgré les tentatives expérimentales ingénieuses.

31 Plus généralement, dans les disciplines empiriques, les échantillons observables servent soit (i) à donner des idées pour imaginer une hypothèse causale (abduction), et dans ce cas, ils peuvent être extrêmement petits (un corpus d'une pomme, par exemple), soit (ii) à tenter de réfuter une hypothèse, et dans ce cas, ce qui est utile, ce n'est pas la taille, mais l'intelligence : il faut aller chercher les échantillons dont on pense qu'ils pourraient réfuter ladite hypothèse.

32 Ainsi, indépendamment du cas de la sémantique, que nous avons vu au § 1.1, le corpus comme accumulation d'échantillons dans l'objectif de valider une hypothèse descriptive constitue, comme on vient de le voir, une grossière erreur de raisonnement.

1.2.2. Le corpus d'occurrences comme outil de réfutation

33 Un corpus d'occurrences, donc, ne peut valider aucune hypothèse descriptive en sémantique, d'abord parce qu'une collection d'observations, aussi grande soit-elle, ne peut pas valider une hypothèse générale (§ 1.2.1 ci-dessus), et ensuite, parce que, de toutes façons, les échantillons d'un tel corpus ne contiennent aucune indication sur la manière dont les formes linguistiques ont été interprétées (§ 1.1 ci-dessus).

34 Mais rien n'empêche d'utiliser un corpus pour tenter de réfuter une hypothèse descriptive, en réfutant une de ses conséquences. En effet, exhiber un certain nombre d'échantillons attestés, même sans aucune indication sur la manière dont ils ont été interprétés, permet de réfuter l'hypothèse selon laquelle l'assemblage d'unités de langue correspondant à ces échantillons serait ininterprétable[5] : pour réfuter une hypothèse descriptive en sémantique, il suffit donc de tirer une conséquence logique de cette hypothèse, conséquence selon laquelle un certain assemblage d'unités de langue serait ininterprétable, et de réfuter cette conséquence au moyen d'un corpus approprié. En effet, si l'on trouve un certain nombre (pas nécessairement très grand) d'énoncés attestés contenant la structure que l'hypothèse à tester considère comme ininterprétable, on peut en inférer que, puisque ces énoncés sont attestés, ils ont été interprétés et, par conséquent, leur structure linguistique n'est pas ininterprétable.

35 Les défenseurs de l'hypothèse contestée, s'ils veulent tenter de maintenir cette hypothèse, devront prouver que dans chacun des échantillons du corpus, l'interprétation est différente de celle qui était prévue par l'hypothèse. Ils ne pourront pas le faire parce que, d'une part, le corpus ne contient pas d'indications sur les interprétations (corpus d'occurrences), et, d'autre part, la conséquence de leur hypothèse qui était à tester ne

spécifiait pas non plus d'interprétation particulière. Pour une évaluation plus fine de l'hypothèse descriptive, il faut, d'une part, que la conséquence à tester précise *ininterprétable dans tel ou tel sens*, et, du coup, que le corpus qui servira à la tester contienne des indications sur la manière dont les énoncés attestés ont été interprétés.

36 On remarquera que, contrairement à la règle générale que nous avons expliquée au § 1.2.1, une thèse d'ininterprétabilité peut être *confirmée* (jusqu'à nouvel ordre…) par certains types de corpora, mais il faut qu'ils soient vraiment Très Petits… En effet, si, malgré des recherches sérieuses et zélées, le nombre d'échantillons trouvés dans l'ensemble des productions accessibles de la langue étudiée est zéro, il est raisonnable d'en inférer que (jusqu'à preuve du contraire) la structure testée *pourrait bien être* ininterprétable. Un cas de cette nature a été trouvé à propos du lexème hongrois *nő*. Il s'agissait de défendre la thèse selon laquelle ce lexème, traduit généralement par *femme* en français, imposait le point de vue de la jeunesse[6]. Un des arguments, un peu inattendu, a été que, dans le thesaurus de la langue hongroise (153,7 millions d'occurrences de mots), les seules occurrences de *öreg nő* (en principe, *vieille femme*) renvoyaient à un poème surréaliste et à ses quelques commentaires (alors que d'autres expressions désignant des vieilles femmes y étaient abondantes).

37 Cette situation inattendue fait irrésistiblement penser à une situation analogue qui, elle, n'est pas inattendue mais, au contraire, est utilisée pour pervertir le « paysage scientifique ». Supposons qu'une dizaine de chercheurs considèrent, séparément, que les travaux de X sur le sujet Y sont très mauvais, et que chacun de ces chercheurs publie un article dans lequel il montre la mauvaise qualité de ces travaux, les erreurs méthodologiques, les fautes de raisonnement, les insuffisances, etc. des travaux de X sur Y. Il en résultera que le nombre de citations que les travaux de X sur Y auront reçues aura été accru d'au moins dix, augmentant ainsi la notoriété de X sur la thématique de Y, ainsi que le pouvoir de nuisance de ses travaux. L'analogie tient au fait que, là encore, la base statistique à partir de laquelle on tente d'inférer des jugements qualitatifs est incomplète : dans le cas des corpora d'occurrences, il manque les indications concernant les interprétations ; dans le cas des indices de citations, il manque les raisons pour lesquelles les travaux ont été cités[7].

2. Du *corpus enrichi* à l'expérimentation en sémantique

38 Aux paragraphes précédents, nous avons vu quelques-unes des limitations découlant du manque d'indications concernant les interprétations des échantillons exhibés dans les corpora d'occurrences. J'examine ici les moyens de contourner ces limitations.

2.1. Enrichir un corpus

39 Pour qu'un énoncé figurant dans un corpus puisse être exploité, l'idéal serait qu'il soit accompagné de l'interprétation qu'il a reçue dans la situation dans laquelle il a été recueilli. Un tel corpus fournirait un recueil de *faits sémantiques* (<unité de discours, situation, interprétation>), grâce auxquels il serait possible de tester les propositions de descriptions sémantiques (<unité de langue, contraintes sur l'interprétation>). Malheureusement, un tel enrichissement « direct » n'est, en général, pas possible, pour des raisons pratiques, mais aussi et surtout pour des raisons conceptuelles. En effet, même si l'on pouvait interroger les destinataires des énoncés recueillis, même s'ils

acceptaient de répondre, et même s'ils étaient capables de se souvenir de ce qu'ils ont compris des énoncés en question, ils ne pourraient pas nous livrer leur interprétation, mais seulement un discours censé la décrire. Le sens construit est privé et n'est pas accessible à l'observation.

40 En revanche, des indications concernant ces interprétations peuvent être fournies par les réactions des destinataires : ces réactions peuvent être interprétées par l'observateur, de manière à lui procurer des indices de ce que le destinataire avait compris. Ces réactions peuvent être verbales ou non verbales.

41 Les premières posent un problème de circularité : pour que l'observateur puisse justifier d'avoir interprété la réaction verbale du destinataire de telle ou telle façon, ledit observateur doit disposer d'une description sémantique lui permettant de traiter le discours du destinataire. Or, justement, les outils de description sémantique de l'observateur sont en train d'être testés… Dans la plupart des cas, l'observateur devra avoir recours à l'introspection pour expliquer pourquoi il a interprété les réactions du destinataire de telle ou de telle façon. La circularité est moindre qu'un recours direct à l'introspection sur les échantillons du corpus, mais il est probable que l'observateur doive fournir des interprétations pour des expressions utilisées par le destinataire pour lesquelles il n'a pas de description sémantique validée. Même si on peut espérer que, avec beaucoup d'efforts l'observateur arrivera à un compromis efficace entre la richesse expressive nécessaire au destinataire pour décrire ce qu'il a compris de l'échantillon et la limitation des descriptions sémantiques déjà validées, il reste deux difficultés : d'une part, la nécessité de recourir à la sollicitation du destinataire par l'observateur limite beaucoup les corpora constituables et, d'autre part, ce que le destinataire dit de ce qu'il a compris ne reflète pas nécessairement ce qu'il a vraiment compris.

42 Les réactions non verbales des destinataires des énoncés d'un corpus doivent, elles aussi, être interprétées, mais il n'y a pas de circularité : il ne s'agit pas du même système de signes. Par ailleurs, il est possible de recueillir les réactions non verbales du destinataire d'un énoncé sans lui demander de repenser à ce qu'il avait compris, ce qui élimine les deux difficultés supplémentaires concernant ses réactions verbales.

43 Pour que ce type d'enrichissement soit utile, il faudra bien sûr élaborer et justifier les principes méthodologiques qui gouverneront le passage des indices non verbaux aux hypothèses sur ce qui aura été compris. C'est d'ailleurs le même genre de questions que celles qui se posent à qui veut produire un corpus annoté.

44 Moyennant une explicitation et une justification rationnelle des indices non verbaux que l'on choisit et des règles permettant d'en inférer certains aspects de ce que le destinataire a compris, un corpus enrichi peut être utilisé pour faire plus que tenter de réfuter des hypothèses d'ininterprétabilité : il permet de réfuter des thèses de la forme *telle structure de la langue ne peut pas être interprétée de telle manière*, ce qui est, bien sûr, beaucoup plus utile pour tester des hypothèses descriptives en sémantique.

2.2. Quelques principes pour une expérimentation en sémantique

45 On vient d'entrevoir une façon d'enrichir un corpus d'occurrence, de telle sorte qu'il fasse plus que signaler que telle structure de la langue a été utilisée et doit donc pouvoir être interprétée : en y incluant des indices non verbaux de ce que les destinataires des échantillons du corpus ont compris.

46 Mais un tel enrichissement n'est pas à la portée de n'importe quelle méthode de recueil. En particulier, la méthode très largement la plus utilisée, le recueil *aveugle* (sans prise en compte des locuteurs, des destinataires, ni des caractéristiques des situations dans lesquelles les échantillons ont été proférés), qui passait pour une garantie d'objectivité, ne permet pas un tel enrichissement. D'autre part, mettre en œuvre les énormes moyens nécessaires à la constitution de grands corpora enrichis serait un gaspillage intolérable : seule une infime partie d'un corpus enrichi est utile pour tester une hypothèse descriptive : il faut, en effet, que les échantillons (i) concernent les structures à tester, (ii) que les situations d'interprétation correspondent aux interprétations à tester, et (iii) que les indices concernant l'interprétation effective permettent de décider si l'interprétation prévue par l'hypothèse à tester est conforme à l'interprétation effective (suggérée par ces indices). Pour faire comprendre ce gaspillage, une analogie peut se révéler bien utile : un tel programme de corpus enrichi correspondrait, en physique, à la constitution d'un corpus de chutes d'objets (par exemple de pommes), recensées en y indiquant tous les paramètres pouvant *éventuellement* intervenir : jour, heure, lieu, température, altitude, degré d'humidité de l'air, ... et peut-être même l'âge de l'expérimentateur (tous ces paramètres *pourraient* s'avérer pertinents, puisque, précisément, nous sommes encore au stade où nous *testons* la théorie selon laquelle ils n'interviennent pas). Le physicien qui voudrait tester une description théorique de la chute des corps devrait alors s'encombrer de ces millions d'observations qui, pour la plupart, n'ont aucun intérêt pour tester son hypothèse. La stratégie à utiliser consisterait ainsi à créer des bases de données gigantesques construites au hasard, en espérant que, par chance, les cas qui nous intéresseront auront été recensés dans ces bases de données. Une telle stratégie révèle beaucoup sur la mentalité de ceux qui la préconisent, mais est surtout une abominable gabegie. Le physicien évite ce gaspillage en constituant des séries d'observations dans lesquelles seuls les paramètres concernant les descriptions à tester varient. C'est le principe de l'*expérimentation*. Dans le corpus de Newton, il n'est pas nécessaire d'inclure des représentants pour chacune des (plus de) dix mille variétés de pommes. La situation est donc la même en physique qu'en sémantique (et, plus généralement dans toutes les disciplines empiriques – même dans les sciences de l'homme et de la société). Le physicien ne cherchera pas à collectionner un corpus d'événements bruts apparaissant dans la nature, en espérant que, par hasard quelques-uns soient pertinents, mais construira des situations (artificielles, donc) pour lesquelles il a *calculé* que l'observation des événements qui s'y dérouleront apportera des réponses à ses questions ; alors que le sémanticien, lui, s'il suit la tendance actuelle, devra se contenter de corpus, de plus en plus grands, d'occurrences de formes, sans pouvoir contrôler les situations dans lesquelles ces occurrences ont été proférées et sans pouvoir choisir celles qui sont pertinentes pour son étude.

47 Un tel principe d'économie (autant intellectuelle que financière) peut pourtant s'appliquer à la sémantique : les corpora à constituer seront limités aux échantillons susceptibles de réfuter l'hypothèse descriptive à tester, ce qui nécessite qu'ils soient enrichis d'indices permettant de faire des hypothèses sur les interprétations qui leur ont été assignées dans les situations dans lesquelles ils ont été proférés. Ces indices devront donc faire référence à la manière dont les échantillons ont été interprétés et aux situations dans lesquelles ces interprétations ont été assignées. Le corpus ainsi limité est donc considérablement plus petit, mais infiniment plus utile dans l'objectif de tester des

hypothèses descriptives. Mais comment incorporer de tels indices aux corpus d'échantillons ?

2.3. L'intérêt de recourir à des indications non verbales

48 En effet, si l'on peut, à la rigueur, admettre que les situations dans lesquelles les échantillons ont été interprétés sont, au moins partiellement, accessibles à l'observateur, considérer que l'interprétation de ces échantillons dans ces situations lui est, elle aussi, accessible reviendrait à supposer la description sémantique déjà fournie, *avant que l'observateur ne l'ait étayée.* En effet, d'une part, l'interprétation construite par les destinataires d'un énoncé est privée, donc non observable directement et, d'autre part, l'objet de la description est l'ensemble des contraintes que les unités de langue imposent à la construction du sens, et, si nous considérons que nous les connaissons suffisamment pour « deviner » l'interprétation des destinataires, nous supposons le problème déjà résolu. Par ailleurs, si nous demandons aux destinataires de nous fournir leur interprétation, l'énoncé par lequel ils nous fourniront cette interprétation (en admettant qu'ils acceptent et qu'ils soient en mesure de le faire) devra être lui-même interprété par l'observateur pour qu'il puisse caractériser ladite interprétation : cette démarche reposerait donc, en dernière analyse, sur l'introspection de l'observateur, et le recours au corpus imposerait, paradoxalement, une démarche introspective.

49 Il est possible de sortir de ce dilemme (circularité ou introspection nécessaire), en évitant soigneusement de s'appuyer sur des commentaires verbaux des interprètes et en limitant l'observation aux indications non verbales qu'ils fournissent. Ces indications sont, en général, très limitées et ne permettent pas de se faire une idée précise de ce que les destinataires ont compris, mais, parmi les indications non verbales que les destinataires de discours peuvent fournir en ce qui concerne leur interprétation des échantillons qui leur ont été soumis, figure la simple indication d'incompréhension[8].

50 Pour qu'une telle indication soit utile pour tester une hypothèse descriptive en sémantique, un travail préparatoire de l'observateur est nécessaire : il faut qu'il conçoive une conséquence logique de son hypothèse descriptive, conséquence qui stipule que certains assemblages soient incompréhensibles, et ce sont ces assemblages qui seront présentés à des destinataires. Le corpus enrichi, constitué d'échantillons de ces assemblages associés à la réaction des destinataires (*incompréhension* ou *non-incompréhension*), permet alors de décider si l'hypothèse d'incompréhensibilité est réfutée ou (pour le moment) non. On voit que, pour ce type d'expérimentations, un corpus est nécessaire : ce corpus doit être enrichi (des indications de compréhensibilité) ; il peut par ailleurs être petit, ou même Très Petit : dès que l'hypothèse testée est réfutée le processus expérimental peut s'arrêter. Tant que l'hypothèse descriptive n'est pas réfutée, elle est considérée comme provisoirement valide.

3. Deux exemples d'expérimentation en sémantique

51 Pour illustrer la faisabilité et l'efficacité de ce type d'expérimentation, je montrerai comment une hypothèse de description lexicale peut faire l'objet d'une expérience sémantique. On verra, à l'occasion de deux tests sémantiques, comment, en tirant une conséquence logique de l'hypothèse descriptive, on peut concevoir des assemblages qui, si cette hypothèse est valide, devraient être ininterprétables, conséquence qui constitue

l'objet du test sémantique. On verra aussi en quoi cette méthodologie est beaucoup plus économique qu'une étude statistique sur très grand corpus, et incomparablement plus efficace. Je conclurai en montrant que ce procédé simple permet aussi une construction fiable de savoirs concernant la description sémantique des mots de chaque langue.

3.1. Test en *donc*

52 Ce test s'appuie sur une différenciation entre deux rôles sémantiques de *donc*, différenciation que je commence donc par expliciter et justifier. Il est important de garder à l'esprit que les descriptions de ces deux emplois de *donc* ne font pas l'objet du test, mais sont, au contraire, utilisées comme outils pour concevoir le test : elles ont été justifiées au préalable (*cf.* Raccah 2002 : 257-260) et sont considérées comme admises.

3.1.1. Rappels sur la description sémantique de *donc*

53 Le mot français *donc* peut être utilisé dans deux constructions syntaxiques distinctes, dans lesquelles il joue deux rôles différents. Il peut être à la charnière entre deux segments constituant chacun une phrase : il apparaît donc comme un connecteur, comme dans

(1) Jean est riche, donc il invitera Max à dîner

54 Ce premier type de construction correspond à l'usage le plus étudié (peut-être pas le plus fréquent…) de *donc*, celui qui renvoie à la formulation d'une sorte de raisonnement. L'idée spontanée (et un peu naïve) selon laquelle *donc* introduirait une relation de conséquence logique[9] peut être « dénaïvisée » en affaiblissant le type de relation :

Hd_1
donc indique que le segment qui le suit est une formulation de l'orientation argumentative du segment qui le précède[10].

55 Nous avons dit que *donc*, dans ce type d'emploi, *apparaît* comme un connecteur ; mais cette apparence est trompeuse. En effet, si l'on accepte la description générale Hd_1, on doit admettre que le premier segment est un *énoncé* (et non pas une *phrase* car ces dernières n'ont évidemment pas d'orientation argumentative) : les *donc* de ce type ne sont donc[11] pas des connecteurs au sens habituel du terme, puisqu'ils ne relient pas deux phrases[12].

56 Un énoncé contenant « donc »[13] dans ce rôle de quasi-connecteur affirme ainsi, en suivant Hd_1, que son premier membre est utilisé comme argument pour son deuxième membre. Une conséquence de cette description est que, d'une part, un tel énoncé doit paraître bizarre à un interlocuteur lorsque son premier membre ne lui semble pas pouvoir servir d'argument pour le second ; c'est effectivement ce qui peut être observé grâce aux signes non verbaux d'incompréhension : les énoncés de

(2) ? Jean est riche, donc la lune est pleine

57 sont difficiles à interpréter (Hd_1 rend compte de cette observation en suggérant qu'il faut imaginer des situations très artificielles pour les comprendre). Mais il s'ensuit aussi, d'autre part, qu'un tel énoncé paraîtrait redondant et bizarre à un interlocuteur lorsque son premier membre ne lui semble pas *ne pas* pouvoir servir d'argument pour le second : on observe (toujours grâce aux signes non verbaux d'incompréhension) que, lorsque le rôle argumentatif du premier membre est évident, un énoncé qui formule explicitement

ce rôle est difficile à comprendre (Hd_1 rend compte de cette observation en suggérant qu'il est alors perçu comme redondant, et donc lourd).

58 Ainsi, alors que (et justement *parce que*) les énoncés de

(3a) Jean est riche, il a de quoi vivre

59 sont perçus comme clairs et ne posant pas de problèmes d'interprétation, ceux de

(3b) ? Jean est riche, donc il a de quoi vivre

60 paraissent redondants et ne sont acceptables que dans des situations particulières[14].

61 Dans ces cas dans lesquels le premier membre est manifestement un argument pour le second, c'est un autre type de *donc* que l'on peut employer, qui n'est plus un quasi-connecteur, mais un opérateur du syntagme verbal du second membre du segment énoncé : le « *donc* inversé » (*Cf.* Raccah 2002 : 259-260), qui s'intercale entre l'auxiliaire et le verbe du second membre ou, à défaut d'auxiliaire, entre le verbe et ses compléments. Ainsi, pour (3a), au lieu de (3b), on dira :

(3c) Jean est riche, il a donc de quoi vivre

62 qui, lui, ne produit pas d'effets de redondance.

63 En résumé, les deux « donc » que nous venons de voir se distinguent par le fait que, là où le premier (le quasi-connecteur ou « "donc" direct ») indique que l'énoncé qui le contient *asserte explicitement* que son premier membre est un argument pour son second membre, le deuxième (le « "donc" indirect ») indique que l'énoncé qui le contient *suppose implicitement* que son premier membre est un argument pour son second membre. Il résulte de ces propriétés que, lorsque le second membre d'un énoncé ne fait qu'expliciter l'orientation argumentative de son premier membre, c'est le « "donc" indirect » qui est utilisable sans hypothèses particulières sur la situation, tandis que, lorsque le second membre d'un énoncé constitue une orientation argumentative dérivable du premier membre mais non contenue en lui, c'est le « "donc" direct » qui est utilisable sans hypothèses particulières sur la situation.

3.1.2. Formulation générique du test en *donc*

64 Je vais maintenant montrer que ces propriétés des deux emplois de *donc* permettent de concevoir un test pour des hypothèses de description lexicale de la langue française, test utilisant les descriptions déjà admises pour les deux emplois de *donc*, et l'indication non verbale éventuelle de difficulté de compréhension, fournie par les destinataires du test.

65 Lorsqu'on a fait l'hypothèse que la description sémantique d'un mot-de-langue M de la langue française doit contenir l'instruction I_M, pour tester cette hypothèse, il suffit de procéder de la façon suivante :

a. On construit une phrase S(M) contenant le mot M.
b. On construit une formulation $F(I_M)$, en français, d'une application de cette instruction I_M à ce dont parle la phrase S(M).
c. On construit la transformée $F(donc_i, I_M)$ de $F(I_M)$, en y insérant un « "donc" indirect » à la place appropriée.
d. On teste la compréhensibilité des énoncés de
   $\mathbf{\Phi}_1$ = .S(M), « donc » $F(P_M)$. et de
   $\mathbf{\Phi}_2$ = .S(M), $F(donc_i, P_M)$..

66 D'après les descriptions admises pour les deux emplois de *donc*, on sait déjà que, si les énoncés de l'une de ces deux phrases complexes sont compréhensibles sans difficulté, les énoncés de l'autre susciteront des signes de difficulté de compréhension. Il n'y a donc que trois cas possibles.

67 Le test s'interprète alors comme suit :

- Cas n° 1 (favorable) :

Si les énoncés de $\mathbf{\Phi}_1$ suscitent des signes d'incompréhension et ceux de $\mathbf{\Phi}_2$ n'en suscitent pas, alors, la description proposée, I_M, est compatible avec l'observation et peut être conservée.

- Cas n° 2 (défavorable) :

Si les énoncés de $\mathbf{\Phi}_2$ suscitent des signes d'incompréhension et ceux de $\mathbf{\Phi}_1$ n'en suscitent pas, alors, la description proposée, I_M, n'est pas compatible avec l'observation et doit être rejetée.

- Cas n° 3 (défavorable) :

Si les énoncés de $\mathbf{\Phi}_1$ et ceux de $\mathbf{\Phi}_2$ suscitent des signes d'incompréhension, alors, la description proposée, P_M, n'est pas compatible avec l'observation et doit être rejetée[15].

3.1.3. Illustration du test en *donc* : application à une hypothèse de description

68 Pour mieux comprendre le fonctionnement du test, voyons ce qu'il donne à propos de l'hypothèse de description sémantique du mot français *riche*, description que j'ai proposée dans Raccah (2010). Il s'agissait de décrire le mot *riche* dans le cadre de la sémantique des points de vue, de telle sorte que la description rende compte d'un certain nombre de phénomènes que j'avais signalés[16].

69 Pour illustrer le fonctionnement du test en *donc*, j'utiliserai une description moins détaillée et moins formelle que celle qui avait été étudiée en 2010 (la précision de la description n'intervient pas dans le test, mais en rend plus difficile l'illustration). Ce que je propose de tester est l'hypothèse selon laquelle le mot français *riche* induit le point de vue du pouvoir[17], c'est-à-dire, l'hypothèse selon laquelle la description sémantique du mot « riche » doit contenir le point de vue de la *possibilité d'agir*.

70 Pour tester cette hypothèse au moyen du test en *donc*,

71 a) On construit une phrase contenant le mot *riche*. Par exemple :

(4) Jean est riche

72 b) On construit une formulation, en français, d'une application du point de vue qui devrait entrer dans la description sémantique de *riche* (le pouvoir d'action) à ce dont parle la phrase (4). Par exemple :

(5) Il a de quoi vivre[18]

73 c) On construit la transformée de (5), en y insérant un « "donc" indirect » à la place appropriée :

(6) Il a donc de quoi vivre

74 d) On teste la compréhensibilité des énoncés de
$\mathbf{\Phi}_1$ = ‚(4), *donc* (5) [Jean est riche donc il a de quoi vivre], et de
$\mathbf{\Phi}_2$ = ‚(4), (6) [Jean est riche, il a donc de quoi vivre].

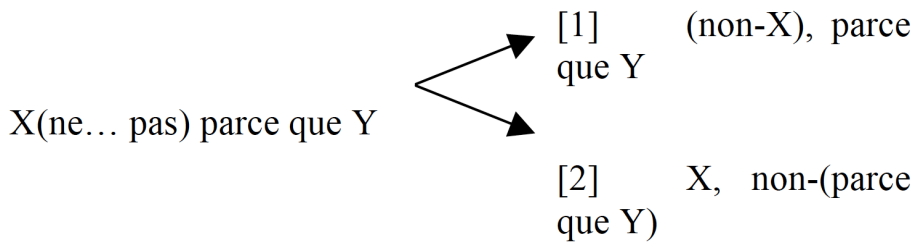
75 Le résultat du test est que $\mathbf{\Phi}_1$, la phrase construite avec un *donc* direct, conduit à des énoncés dont la réception est accompagnée d'une réaction non verbale

d'incompréhension, tandis que Φ_2, la phrase construite avec un *donc* indirect, conduit à des énoncés dont la réception n'est pas accompagnée d'une telle réaction. L'hypothèse selon laquelle le mot français *riche* impose le point de vue du pouvoir peut donc être conservée. On peut, bien entendu, consolider ce résultat en appliquant ce test à de nombreux exemples : cela ne prouvera pas de manière définitive que l'hypothèse est bonne (comme on l'a vu, la loi de la recherche empirique est dure, mais c'est la loi…) mais, si le test réussit à chacune des tentatives, l'hypothèse testée peut alors être considérée comme difficile à réfuter, ce qui permet de justifier qu'on l'admette jusqu'à une éventuelle preuve du contraire.

3.2. Le test en *parce que*

3.2.1. Description de *parce que*

76 En s'appuyant sur les travaux publiés dans [Ducrot *et al.* 1975], Chmelik (2007 : 358-363) propose une description de « parce que » rendant compte de l'ambiguïté des phrases contenant « parce que » et dont la proposition principale contient une négation en « ne… pas ». Chmelik décrit l'ambiguïté introduite par « ne… pas… parce que… » d'une manière que l'on peut résumer comme suit :



77 Ainsi,

(7) Jean n'a pas chanté parce qu'il avait bu

78 peut signifier [1] que Jean n'a pas chanté et que la raison pour laquelle il n'a pas chanté est qu'il avait bu, ou [2] que la raison pour laquelle Jean a chanté n'est pas qu'il avait bu. Dans la première interprétation, la négation de X est assertée et le lien causal entre Y et la négation de X est affirmé ; dans la seconde interprétation, X (et non pas sa négation) est présupposé et la négation du lien causal entre Y et X est assertée.

79 Chmelik fait remarquer que cette ambiguïté n'est pas systématique : certaines phrases contenant *parce que* et dont la principale contient *ne… pas* ne sont pas ambiguës et ne peuvent recevoir que l'interprétation que j'ai nommée [2]. Ainsi, en suivant les remarques de Chmelik et pour prendre un exemple proche de l'un de ceux qu'elle a fournis, les énoncés de la phrase (8) :

(8) Je ne fais pas confiance à Jean parce qu'il est honnête

80 sauf hypothèses très particulières sur la situation, ne permettent que l'interprétation dans laquelle le locuteur présuppose qu'il fait confiance à Jean et affirme que ce n'est pas parce qu'il est honnête (mais pour une autre raison), et semblent interdire l'interprétation dans laquelle le locuteur affirmerait qu'il ne fait pas confiance à Jean et que cette disposition négative est due au fait que Jean est honnête. Ce phénomène curieux mérite que l'on tente d'en trouver la raison : pourquoi donc certaines phrases semblent ne pas être ambiguës alors que les autres le sont ?

81 Sans entrer dans les détails de la description technique que Chmelik propose pour *parce que*, on en retiendra qu'une description adéquate indique qu'un énoncé [B parce que A]$_S$ de la phrase *B parce que A*, dans une situation S indique que, dans cette situation S,

i. le locuteur adhère aux orientations argumentatives suggérées par [A]$_S$,

ii. qu'il adhère à celles qui sont suggérées par [B]$_S$ et

iii. que son énoncé [B]$_S$ est motivé argumentativement par son énoncé [A]$_S$.

82 L'ambiguïté, lorsqu'elle apparaît, provient du champ d'application de la négation : cette dernière peut

(a) soit concerner (ii) seulement ; en ce cas, B est nié, A est maintenu tel quel, mais la clause (iii) est modifiée en remplaçant B par sa négation :

(iiia) pour le locuteur, l'énoncé de non-B est motivé argumentativement par son énoncé de A

(b) soit concerner (iii) seulement ; en ce cas, B est maintenu (l'énoncé ne dit rien sur A) et la clause (iii) devient :

(iiib) pour le locuteur, l'énoncé de B *n'est pas* motivé argumentativement par son énoncé de A

83 Dans le cas des phrases, comme (8), à propos desquels Chmelik faisait remarquer que leurs énoncés, sauf hypothèses particulières sur la situation, ne peuvent recevoir que l'interprétation [2], c'est la clause (iii) qui est concernée par la négation. Pour comprendre pourquoi les phrases comme (8) ne sont pas ambiguës, il suffit donc de comprendre ce qui fait que, dans les énoncés de ces phrases, ou bien B ne peut pas être nié ou bien (iiia) ne peut pas être accepté. Dans le cas de la phrase (8), il faut donc expliquer ou bien pourquoi un locuteur ne peut pas nier qu'il estime Jean ou bien pourquoi il ne peut pas considérer le fait que Jean soit honnête comme un argument pour ne pas l'estimer. Or, le mot français « honnête » est *euphorique*[19], c'est-à-dire que tous les énoncés qui le contiennent introduisent un jugement positif sur ce à quoi renvoie le substantif qu'il qualifie : il est donc facile d'expliquer pourquoi un locuteur ne peut pas considérer le fait que Jean soit honnête comme un argument pour ne pas l'estimer.

84 D'une manière générale, toujours en suivant Chmelik, les phrases contenant *parce que* et dont la principale est niée perdent leur ambiguïté lorsque la subordonnée constitue un argument indiscutable pour la principale non niée, ce qui est toujours le cas lorsque la principale non niée formule un point de vue lexicalisé dans un mot-de-phrase de la subordonnée.

3.2.2. Préparation au test

85 Pour le type de tests que je propose ici, ce qui est à retenir dans la description des mots ou des syntagmes qui servent à composer le test (ici, *parce que*), c'est ce qui peut avoir pour effet de rendre un énoncé incompréhensible ou difficile à comprendre. Il n'est donc pas suffisant de prévoir une absence d'ambiguïté des phrases en *ne... pas parce que* : il faut, en tenant compte de cette propriété, concevoir des phrases dont les énoncés pourraient être difficiles à comprendre. Lorsqu'une phrase est ambiguë, ses énoncés peuvent entrer dans des constructions argumentatives s'appuyant sur l'une ou l'autre des significations qu'elle permet. Ainsi, la phrase

(9) Paul a fini son livre

86 qui peut signifier que Paul a fini de *lire* ou d'*écrire* le livre, peut être utilisée dans un enchaînement comme

(10) Paul a fini son livre : il va pouvoir le rendre à la bibliothèque

87 ou dans des enchaînements comme

(11) Paul a fini son livre mais il ne l'a pas encore envoyé à son éditeur

88 Lorsqu'on modifie la phrase en spécifiant l'une de ses significations possibles, les enchaînements ne correspondant pas à cette signification deviennent difficiles, voire impossibles, à comprendre :

(12) ? Paul a fini d'écrire son livre : il va pouvoir le rendre à la bibliothèque

89 Ce sont des énoncés utilisant des constructions de ce genre qui pourront être soumis aux tests en *ne... pas parce que*.

90 Le test en *parce que* étant un peu plus complexe que le test précédent, je propose, dans un premier temps, de familiariser le lecteur à ses différentes étapes, en décrivant la procédure sur un exemple pour lequel le test n'est pas concluant : il sera, par la suite, plus facile à comprendre ce qu'apporte le test lorsqu'il est concluant.

91 a) Considérons la phrase

(13) Jean est musicien

92 b) On se place dans une situation dans laquelle il est question qu'il aille au concert avec Marie. Une des questions que l'on envisage, sachant (13) est de savoir s'il a les moyens d'inviter Marie (au concert).
Considérons la phrase

(14) Jean a les moyens d'inviter Marie

93 et sa négation

(15) Jean n'a pas les moyens d'inviter Marie

94 La phrase **Φ** = (15) *parce que* (13), c'est-à-dire :

Φ Jean n'a pas les moyens d'inviter Marie parce qu'il est musicien

95 est ambiguë et peut signifier

(i) soit que Jean n'a pas les moyens d'inviter Marie,
et cela, parce qu'il est musicien,
(ii) soit que ce n'est pas parce que Jean est musicien
qu'il a les moyens d'inviter Marie.

96 c) C'est ainsi que **Φ** peut être complétée, par exemple, soit par

(16) Elle va devoir payer son billet

97 soit par

(17) Il connaît bien le directeur de la salle

98 chacun de ces deux compléments sélectionnant l'une des interprétations et éliminant l'autre.

99 On remarquera que les énoncés de (16), dans la situation que nous avons envisagée sont coorientés avec ceux de (15) dans cette même situation

100 d) Construisons **Φ'** = **Φ** : (16), c'est-à-dire :

Φ' Jean n'a pas les moyens d'inviter Marie parce qu'il est musicien :
elle va devoir payer son billet

101 On peut remarquer que **Φ'** n'est pas ambiguë : l'ajout du segment coorienté avec la principale de **Φ** élimine l'interprétation (ii) et impose l'interprétation (i).

102 Le principe du test est de sélectionner les différents segments de manière à ce que, si l'hypothèse à tester est correcte, l'interprétation (i) soit elle aussi impossible, ce qui doit conduire le destinataire du test à proférer un signe non verbal d'incompréhension.

3.2.3. Formulation générique du test en *ne… pas parce que*

103 Lorsqu'on a fait l'hypothèse que la description sémantique d'un mot-de-langue M de la langue française doit contenir l'instruction sémantique I_M, pour tester cette hypothèse, il suffit de procéder de la façon suivante :

a. On construit une phrase S(M) contenant le mot M.

b. On construit une formulation $F(I_M)$, en français, d'une application, à ce dont parle la phrase S (M), de l'instruction pour laquelle on teste si elle devrait entrer dans la description sémantique de M.

c. On construit la négation, $\sim F(I_M)$, de $F(I_M)$.

d. On construit une phrase P dont les énoncés sont co-orientés avec $\sim F(I_M)$ dans les situations dans lesquelles le test sera effectué.

e. On teste la compréhensibilité des énoncés $\mathbf{\Phi'}_S$ de $\mathbf{\Phi'} = \sim F(I_M)$, « parce que » S(M) : P

104 Si les énoncés de **Φ'** sont compréhensibles, l'hypothèse testée est réfutée : la présence du mot à décrire n'impose pas l'instruction sémantique que l'on proposait pour le décrire. Si les énoncés de **Φ'** sont incompréhensibles ou difficilement compréhensibles, l'hypothèse testée est confortée : la présence du mot à décrire impose bien l'instruction que l'on proposait pour le décrire.

3.2.4. Exemple

105 Modifions l'exemple de 3.2.2 en y remplaçant *musicien* par *riche*, pour tester l'hypothèse descriptive selon laquelle le mot *riche* de la langue française donne comme instruction de voir la possession comme source d'un certain pouvoir. En remplaçant ce que nous avons à remplacer, on obtient :

a. S(M) = S(*riche*) =
(18) Jean est riche

b. Une des applications possibles de l'instruction *voir la possession comme source d'un certain pouvoir* à la situation du concert est de voir les possessions de Jean comme lui donnant les moyens d'inviter Marie : on peut donc choisir $F(I_M)$ comme suit :
(19) Jean a les moyens d'inviter Marie

c. Du coup, $\sim F(I_M)$ = (19') = Jean n'a pas les moyens d'inviter Marie

d. Les énoncés de (16), dans notre situation, sont coorientés avec ceux de (19')

e. On a : $\mathbf{\Phi'} = \sim F(I_M)$, « parce que » S(M) : P = (19') « parce que » (18) : (16)
Φ' = Jean n'a pas les moyens d'inviter Marie parce qu'il est riche :
elle va devoir payer son billet

106 Les énoncés de **Φ'** suscitent des réactions non verbales d'incompréhension, qui corroborent l'hypothèse selon laquelle l'interprétation qui n'est pas éliminée par l'ajout de (16) étaient déjà éliminée dans (19') « parce que » (18), ce qui, à son tour, corrobore l'hypothèse que le mot français *riche* intervenant dans (18) indique que la possession doit être vue comme source d'un certain pouvoir (ici, celui d'inviter Marie).

107 Le test en *parce que* corrobore donc aussi l'hypothèse descriptive concernant *riche*, ce qui, comme pour le test en *donc*, ne la prouve pas, mais indique seulement que, malgré des tentatives sérieuses de réfutation, elle n'a pas (encore) été réfutée.

Conclusions (provisoires…)

108 Dans cette étude, j'ai rappelé pourquoi un corpus d'occurrences ne peut servir qu'à réfuter des hypothèses descriptives dont une des conséquences logiques serait que tel ou tel assemblage d'unités de langues ne pourrait pas conduire à des énoncés interprétables, ce qui est déjà beaucoup, mais qui a pour conséquence que la taille du corpus n'intervient pas dans le processus rationnel : pour ces activités scientifiques, les « grands corpus » n'ont pas plus d'intérêt que les petits corpora (mais ils coûtent plus cher…).

109 J'ai ensuite montré qu'un corpus enrichi, contenant – outre les occurrences – des indications sur la manière dont elles ont été interprétées dans la situation de chaque échantillon, ne servait toujours qu'à réfuter des hypothèses descriptives, mais permettait de tester des propositions plus élaborées (comme, par exemple, *tel assemblage d'unités de langue ne peut pas être interprété de telle manière*). Nous avons remarqué, cependant, que des indications précises sur la manière dont les échantillons ont été interprétés étaient difficiles (voire, parfois, impossibles) à établir : pour conserver l'avantage de travailler avec des corpora enrichis, j'ai proposé d'utiliser comme information élémentaire sur l'interprétation, les indications non verbales d'incompréhension (ou de difficulté à comprendre). J'ai montré comment l'utilisation de ces indications permet de construire des tests permettant de tenter de réfuter des descriptions sémantiques, et j'ai illustré cette méthode générale en proposant deux tests sémantiques, eux-mêmes illustrés par des exemples.

110 Le lecteur attentif se demande sans doute pourquoi ma réflexion sur l'enrichissement des corpora me conduit à parler de tests sémantiques, comme s'il s'agissait de la même thématique. La réponse est la suivante : les deux thématiques sont effectivement liées et les corpora enrichis, comme on l'a vu, permettent de mettre à l'épreuve des hypothèses descriptives plus élaborées que les corpora d'occurrences. Et comme il n'est pas possible d'enrichir un corpus en y incluant toutes les interprétations auxquelles tous les échantillons qu'il contient ont donné lieu, il est nécessaire de concevoir des dispositifs permettant de tester les hypothèses descriptives en disposant d'un enrichissement minimal. Ce sont ces dispositifs qui constituent les expérimentations sémantiques.

111 Il y a longtemps que, en physique, l'observation passive n'est plus utilisée pour mettre à l'épreuve des hypothèses théoriques, parce que l'information fournie par une telle observation est insuffisante. Ce n'est pas en observant cinq mille pendules au hasard que le physicien peut tester les équations des mouvements périodiques : il est nécessaire de faire varier les paramètres afin que ces observations fournissent un maximum d'informations utiles, et donc de construire des pendules, tout aussi naturels, qui exhibent ces variations des paramètres. De même, en sémantique, c'est en faisant varier les paramètres des observations, selon un plan déterminé par les besoins du processus de réfutation, que l'on peut mettre à l'épreuve rigoureusement les hypothèses descriptives que l'on est amené à proposer : la taille du corpus ne peut jouer aucun rôle dans ces mises à l'épreuve.

112 Pour enrichir un corpus, il est nécessaire que l'observateur interagisse, même minimalement, avec les destinataires des échantillons, ce qui ne peut se faire que si le corpus est constitué, lui-même, en interaction avec les destinataires des échantillons. Les inconvénients évidents d'une telle exigence méthodologique (par exemple, le fait qu'un corpus préexistant à une recherche scientifique ne peut pas être utile) est compensé, on l'a vu, par l'avantage de pouvoir se contenter de petits corpora sans que cela ne nuise à la rigueur des démonstrations. Une recherche scientifique n'est pas de même nature qu'une fouille de données sur un moteur de recherche : les grands corpora ne sont pas pertinents pour une recherche scientifique en sémantique.

113 Une des propriétés particulièrement intéressante de la méthodologie que je propose d'appliquer à la sémantique est sa récursivité : toute unité de langue dont la description a « réussi » un ou plusieurs tests peut, elle-même, être utilisée, avec sa description, pour construire un nouveau test. Cet avantage est énorme : plus on décrit rigoureusement, plus on peut décrire rigoureusement. Plusieurs autres tests sémantiques ont été conçus et d'autres sont en cours d'élaboration : la boîte à outils va donc se remplir de plus en plus rapidement...

BIBLIOGRAPHIE

Bruxelles S. et Raccah P.-Y. (1987). « Information et argumentation : l'expression de la conséquence », *Actes du colloque COGNITIVA 87*.

Chmelik E. (2007). *L'idéologie dans les mots. Contribution à une description topique du lexique justifiée par des tests sémantiques. Application à la langue hongroise*. Thèse de doctorat, Université de Limoges.

Geyken A. (2008). « Quelques problèmes observés dans l'élaboration de dictionnaires à partir de corpus ». *Langages* 171, (3), 77-94.

Habert B. (2000). « Des corpus représentatifs : de quoi, pour quoi, comment ? », *Cahiers de l'Université de Perpignan* 31, 11-58.

Kilgarriff A. et Grefenstette G. (2003). « Introduction to the Special issue on the Web as Corpus », *Computational Linguistics* 29 (3), 333-348.

Raccah P.-Y. (2002). « Lexique et idéologie : les points de vue qui s'expriment avant qu'on ait parlé », *in* Carel M. (éd.) *Les facettes du dire : Hommage à Oswald Ducrot*. Paris, Kimé, 242-268.

Raccah P.-Y. (2008). « Contraintes linguistiques et compréhension des énoncés : la langue comme outil de manipulation », *in Entretiens d'orthophonie*. Paris, Expansion Formation et Éditions, 61-90.

Raccah P.-Y. (2010). « Racines lexicales de l'argumentation : la cristallisation des points de vue dans les mots », *Verbum* XXXII : 1, 119-141.

Sinclair J. (1996). *Preliminary Recommendations on Corpus Typology*. *Rapport technique, EAGLES* (Expert Advisory Group on Language Engineering Standards). Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale. Pise.

NOTES

1. Les mots n'ont pas de sexe : je dispenserai le lecteur de la corvée BCBG qui consiste à féminiser/masculiniser les mots masculins/féminins renvoyant à d'éventuel(le)s per(mer)sonnes. Lorsque le sexe de l'individu(e) ou du groupe d'individu(e)s au(x)quel(s) je renvoie n'aura pas d'importance, j'utiliserai systématiquement la forme non marquée de la langue française (celle qui est aussi utilisée aussi au masculin, faute de marque spécifique).

2. Corpora ne contenant pas, pour chaque occurrence de chaque corpus, d'indications sur les interprétations effectives que les (éventuellement différents) destinataires de ces occurrences ont construites, *et* les justifications (ou au moins la documentation) qui permettent à l'utilisateur du corpus d'affirmer que telle ou telle occurrence a bien été interprétée de telle ou telle manière par tel ou tel destinataire dans telle ou telle situation.

3. L'*implication* est une opération du calcul des propositions, qui transforme deux propositions en une troisième, qui n'est fausse que si la première est vraie *et* la seconde fausse. Ainsi, A ⊃ B est fausse seulement si A est vraie *et* B fausse ; par ailleurs, sa contraposée, ~B ⊃ ~A, est fausse seulement si ~B est vraie *et* ~A fausse, c'est-à-dire, seulement si B est fausse *et* A est vraie (donc, exactement dans les mêmes conditions que A ⊃ B).

4. Ce « résultat », peu acceptable, n'est cependant pas aussi absurde qu'il y paraît : voyant arriver par les airs quelque chose de blanc, l'observateur pourrait penser qu'il s'agit peut-être d'un corbeau et commencer à mettre en doute sa croyance dans la couleur des corbeaux ; puis constatant qu'il s'agit d'un drap, il reviendrait, soulagé, sur sa mise en doute.

5. Ce raisonnement s'appuie sur le présupposé méthodologique selon lequel, en général, les énoncés attestés ont été compris (au moins d'une certaine manière), dans la situation dans laquelle ils ont été produits : ce présupposé est à la base de toute intention de travailler sur corpus en sémantique. Il peut, bien entendu, être contesté – et défendu – sur le plan méthodologique, mais une telle discussion est inutile dans le cadre de ces réflexions, qui s'adressent à qui envisage déjà d'utiliser des corpora pour appuyer des recherches en sémantique.

6. Voir Chmelik (2007).

7. On ne s'étonnera donc pas que j'évite de citer des travaux qui illustrent les défauts que je souligne…

8. L'incompréhension peut être de très courte durée, les sujets parlants parviennent assez vite à faire des hypothèses supplémentaires qui leur permettent de comprendre. Mais, même lorsque l'incompréhension dure très peu, l'expérience montre que, en général, les destinataires ont le temps de laisser échapper un signe d'incompréhension. Ces signes peuvent être spontanés (regard, froncement de sourcils, gestes…) ou décidés conventionnellement dans le protocole de l'expérimentation.

9. Même si la morale voudrait que les riches invitent les moins riches, on ne voit pas en quoi la richesse de Jean serait une condition suffisante pour qu'il invite Max… et il ne semble pas, non plus, que le locuteur de (1) le prétende.

10. Cette formulation permet de « récupérer » l'interprétation logique dans les énoncés appropriés : il suffit de tenir compte du fait que, dans un texte d'apparence scientifique, les arguments ne sont, en principe, acceptables que s'ils sont déductifs (*cf.* Bruxelles *et al.* 1987) pour une description de « si…alors » selon cette ligne.

11. Le lecteur peut entrevoir, dans mon métadiscours, l'usage d'un autre « donc » : on en reparlera donc…

12. On trouvera, dans (Raccah 2002), une étude plus détaillée du statut sémiotique de « donc ».

13. Pour éviter d'embrouiller le lecteur, j'applique la convention typographique suivante : la mention d'un mot de langue se fait par l'écriture de ce mot en italique, tandis que le renvoi à une occurrence d'un mot dans un énoncé ou un discours s'indique au moyen des guillemets.
14. Le fait qu'un énoncé apparaisse comme redondant ou exige une situation particulière pour être compréhensible ne le rend pas sémantiquement déviant : au contraire, lorsque cela arrive, on est en présence d'un fait sémantique dont il faut tenir compte.
15. Le cas n°3 survient lorsque l'instruction I_M, proposé pour décrire M, n'est pas seulement étrangère à la description sémantique de M, mais, de plus, n'est pas dérivable des points de vue associés à M : ce cas de figure sanctionne un manque d'intuition difficile à imaginer, mais néanmoins concevable...
16. Voir (Raccah 2010 : 132-133) pour les phénomènes, et 134-136 pour une présentation du dispositif de description et son application à la description de *riche*.
17. Ce point de vue concerne, d'une manière générale, la *possibilité d'action* et pas seulement la sphère politique.
18. L'application choisie ici n'utilise qu'une faible partie du potentiel ouvert par le point de vue du pouvoir, mais elle suffit pour illustrer le fonctionnement du test. Le lecteur pourra reprendre *mutatis mutandis* cette même illustration pour tester une application plus « riche », qui donnerait, par exemple, (5') Il a les moyens d'inviter Max à dîner.
19. Voir (Raccah 2008 : 83-84) pour une définition systématique de la *phoricité*.

RÉSUMÉS

Cet article fournit, dans sa première section, une analyse méthodologique de l'utilisation de corpora, au cours de laquelle je détaille, de manière argumentée ce qu'on peut en attendre et pour quels objectifs, précise ce qu'on ne peut pas en attendre et montre pourquoi. La première partie conclut sur l'inutilité des *corpora d'occurrences* en sémantique des langues, même s'ils sont indispensables dans d'autres disciplines (comme, par exemple, les études littéraires) ; j'insiste sur le caractère nuisible de leur utilisation en sémantique, sauf dans un cas que je spécifie, utilisation qui réintroduit l'introspection sémantique en la maquillant derrière un appareil technique inspiré des statistiques. Dans la deuxième section, je montre qu'il existe un moyen de sortir du carcan pseudo-méthodologique que la mode présentant les corpora comme unique accès à l'observation empirique tente d'imposer ; ce moyen, qui part de l'idée d'enrichir les corpora par des indications sur l'interprétabilité des échantillons, comporte deux aspects principaux : (i) réhabiliter l'expérimentation empirique et (ii) rechercher des hypothèses théoriques permettant d'engendrer des hypothèses descriptives à tester au moyen des expériences empiriques. On y insiste sur le fait que, conformément à ce que requiert une démarche scientifique rigoureuse, les tests empiriques ne peuvent pas *vérifier* une hypothèse, mais seulement la *réfuter* : c'est l'échec de la réfutation qui constitue un indice de validité (provisoire) de l'hypothèse. La troisième section présente, en détail et avec des illustrations, deux tests empiriques permettant des expérimentations destinées à tester les descriptions sémantiques que l'on peut être amené à proposer. On y insiste sur le fait que les corpora nécessaires à ces expérimentations peuvent être petits, mais nécessitent une interaction minimale entre l'observateur et les destinataires des échantillons : ils sont enrichis par une indication non verbale, fournie par les destinataires des échantillons, à propos de l'éventuelle difficulté qu'ils

ont eues à les interpréter. Les tests sémantiques utilisant des unités de langue déjà décrites, le chapitre conclut sur l'avantage d'une incrémentation de la descriptibilité des unités de langue s'appuyant sur la récursivité des moyens de tester les descriptions : toute description non réfutée par les tests existants permet de construire de nouveaux tests.

This chapter provides, in its first section, a methodological analysis of the use of corpora, in which I detail what one can expect of them and for what objectives, and specify what one cannot expect of them, and show why. The first part concludes on the uselessness of occurrences corpora in the semantics of human languages, even if they are indispensable in other disciplines (like, for example, literary studies); I insist on the harmful character of their use in semantics, except in one case that I specify, a use that reintroduces semantic introspection, disguising it up behind a technical apparatus inspired by statistics. In the second section, I show that there is a way out of the pseudo-methodological straitjacket that the fashion presenting the corpora as sole access to empirical observation tries to impose; this means, which involves enriching corpora with indications on the interpretability of the samples, has two main aspects: rehabilitate empirical experimentation; and search for theoretical hypotheses which may generate descriptive hypotheses to be tested using empirical experiments. It is emphasized that, as a rigorous scientific approach requires, empirical tests cannot validate a universal hypothesis, but can only possibly falsify it: it is the failure of the refutation that constitutes an indication of the (provisional) validity of the hypothesis. The third section presents, in detail and with illustrations, two empirical tests allowing experiments to test the possible semantic descriptions that one may be led to propose. It is emphasized that the corpora required for these experiments may be small, but require a minimal interaction between the observer and the recipients of the samples: they are enriched by a non-verbal indication, provided by the recipients of the samples, about the possible difficulty they have had in interpreting them. Since semantic tests use language units already described, the chapter concludes on the advantage of incrementing the description of language units using the recursivity of the means for testing descriptions themselves: any one of the descriptions that have not been refuted by the existing tests allows to build new tests.

INDEX

Mots-clés : sémantique empirique, instructions sémantiques, validation des descriptions, méthodologie abductive, enrichissement de corpus, expérimentation sémantique, incrémentation des moyens descriptifs
Keywords : empirical semantics, semantic instructions, descriptions testing, abductive methodology, corpora enrichment, semantic experimentation, descriptive means increment

AUTEUR

PIERRE-YVES RACCAH

CNRS / LLL Orléans