Revue Sciences/Lettres

2 | 2014
Épistémologies digitales des sciences humaines et sociales

La linguistique est-elle soluble dans la statistique ?

Thierry Poibeau

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/rsl/402
DOI : 10.4000/rsl.402
ISSN : 2271-6246

Éditeur
Éditions Rue d'Ulm

Référence électronique
Thierry Poibeau, « La linguistique est-elle soluble dans la statistique ? », *Revue Sciences/Lettres* [En ligne], 2 | 2014, mis en ligne le 07 octobre 2013, consulté le 20 avril 2019. URL : http://journals.openedition.org/rsl/402 ; DOI : 10.4000/rsl.402

La linguistique est-elle soluble dans la statistique ?

Thierry Poibeau

La chaise longue et l'instrument

1 La linguistique a été assez largement dominée depuis la fin des années 1950 par le courant chomskyen. Ce courant, même s'il a connu d'importantes variantes (Chomsky, 1957, 1965, 1981, 1995), s'appuie sur plusieurs points fondamentaux : l'analyse porte avant tout sur la syntaxe, c'est-à-dire sur la façon dont les mots se combinent au sein de la phrase, laissant de côté (au moins en partie dans la version standard) la sémantique. La théorie s'intéresse à la compétence, c'est-à-dire à la connaissance abstraite d'une langue particulière et l'intuition est l'élément essentiel pour attester de la validité d'une séquence linguistique donnée. En fait, au-delà de l'analyse de langues particulières, c'est la recherche de principes généraux sur la langue, plus ou moins abstraits, qui est visée.

2 Au même moment, d'autres courants, fondés essentiellement sur la collecte et l'observation minutieuse de données attestées, ont continué à se développer (McEnery et Wilson, 1996 ; McEnery *et al.*, 2005). Ainsi, la linguistique de corpus (c'est-à-dire le courant de la linguistique partant de l'observation de données langagières attestées, rassemblées au sein d'ensembles représentatifs, le plus souvent sur support informatique) a eu une influence majeure dès les années 1950 (Léon, 2005), même si cette approche a parfois été décriée, notamment parce qu'un ensemble de données, aussi grand soit-il, n'épuisera jamais la créativité langagière (autrement dit, il y aura toujours des phrases possibles mais non attestées en corpus[1]). De façon plus prosaïque, on peut constater que les données particulières accumulées sont souvent complexes, disparates et font parfois obstacle à l'élaboration d'une théorie, qui justement souhaite faire abstraction du particulier pour découvrir des lois générales.

3 Cette description est évidemment très schématique : être chomskyen ne veut pas dire que l'on se désintéresse des données, tandis que les linguistes de corpus ne refusent pas l'intuition, qui est utile pour formuler certaines hypothèses de travail ou compléter un

jeu d'exemples incomplet. Il n'empêche, ces deux courants se sont assez largement ignorés : on a parlé de « linguiste en chaise longue » (Fillmore, 1992) par opposition au « linguiste à l'instrument » (Habert, 2005), qui utiliserait un dispositif outillé (notamment au niveau informatique) pour explorer les données[2].

4 Si on va un peu plus dans le détail, on observe que certains chercheurs, au sein du courant chomskyen, ont rapidement (dès les années 1960) mis en avant la nécessité de s'intéresser de plus près à la sémantique et au lexique (Harris, 1995). L'organisation des mots au sein de la phrase dépend fondamentalement du lexique et seule une description minutieuse et exhaustive de celui-ci peut permettre de « coder » l'information linguistique nécessaire, surtout dans une perspective de traitement automatique qui oblige à rendre compte explicitement de toutes les variations de surface. Par ailleurs, l'importance de la fréquence (des mots, des constructions ou des structures) dans les phénomènes d'acquisition et d'évolution des langues peut difficilement être ignorée et oblige à avoir accès à des données attestées, quel que soit le paradigme que l'on choisisse.

5 Cette description rapide permet d'expliquer l'intérêt soutenu des recherches sur le lexique à partir des années 1970, une fois constatées les limites des méthodes formelles face à la diversité des données. On peut citer deux exemples, dans le domaine du traitement automatique des langues :

- En France, Maurice Gross prend conscience dès la fin des années 1960 de la nécessité de développer des lexiques riches et exhaustifs pour analyser des textes de façon automatique (Gross, 1975). L'initiative de Gross restera longtemps pionnière, à une période avant tout marquée par la recherche de principes théoriques généraux sur la langue. De fait, le français a probablement été une des langues les mieux dotées dans les années 1980 en ce qui concerne les lexiques syntaxiques.
- Le développement d'analyseurs automatiques à la fin des années 1970 (dans le monde anglo-saxon notamment, à travers le développement des formalismes fondés sur l'unification) montre très rapidement que l'élément fondamental pour l'analyse est le lexique (Abeillé, 1993). Les années 1980 seront alors marquées par le développement d'analyseurs dits « fortement lexicalisés » (par exemple *LFG*, Kaplan et Bresnan, 1982) puis, logiquement, par le développement des recherches sur le lexique lui-même (Evans et Gazdar, 1990 ; Pustejovsky, 1991). On remarquera toutefois le décalage entre les recherches dans le monde anglo-saxon et dans le monde francophone dans ce domaine, la France ayant précédé le mouvement sans pour autant s'imposer ni garder sa position avancée.

6 En dehors du traitement automatique des langues, les lexicographes[3] ont toujours été intéressés par la notion de corpus, en premier lieu pour la recherche d'exemples, mais aussi pour observer les mots en contexte et ainsi avoir accès à leurs usages essentiels (Teubert, 2002). La linguistique de corpus s'inscrit donc dans cette tradition soucieuse de partir de la diversité des données pour élaborer des modèles de la langue (qu'il s'agisse de morphologie, de syntaxe ou de sémantique).

7 Le fait nouveau depuis une quinzaine d'années est l'arrivée massive de données grâce au développement du web. Deux éléments ont plus particulièrement joué un rôle majeur :

- D'une part, la masse de données disponibles est incomparable avec tout ce qu'on a connu jusque-là, et se développe à un rythme effréné (plusieurs dizaines de milliards de pages web sont aujourd'hui accessibles, dans de très nombreuses langues, même si peu de langues sont finalement bien représentées).

- D'autre part, les moyens de calcul à disposition grâce au développement de l'informatique (puissance des processeurs, taille de la mémoire, disponibilité des réseaux puis des grilles de calcul) offrent là aussi des possibilités insoupçonnées jusqu'à récemment.

8 Ces évolutions majeures ont révolutionné le traitement automatique de l'information. On assiste ces dernières années au développement de méthodes automatiques, souvent probabilistes ou statistiques, qui ont pour première caractéristique d'être efficaces face à cette masse de données. D'une certaine manière, l'efficacité l'a d'abord emporté sur la précision. On ne peut traiter des milliers ou des millions de documents en quelques dizaines de secondes en mettant en œuvre une analyse linguistique profonde, même si la puissance de calcul des ordinateurs augmente régulièrement. Du coup, le plus simple est de compter (des formes, des mots, des séquences[4] et des corrélations entre ces formes, ces mots et ces séquences). Mais ces comptages ne sont pas anodins et donnent accès à une information beaucoup plus profonde qu'on ne pourrait de prime abord imaginer.

9 En effet, la langue, et donc la linguistique, est avant tout affaire d'usage et l'usage n'est pas une abstraction théorique : il correspond fondamentalement à l'emploi des mots (des formes, des structures) en contexte. La linguistique de corpus répète à l'envi l'expression empruntée à Wittgenstein : « on connaît un mot à son contexte » (*« you shall know a word by the company it keeps »* Firth, 1957, p. 11 – *« Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache »*, Wittgenstein, 1953). Il ne s'agit donc plus alors de décrire le contenu sémantique du mot ou du syntagme, mais de calculer celui-ci à partir de l'entourage du mot ou du syntagme. La notion de contexte est donc fondamentale, même si les calculs obligent à le réduire drastiquement, sous risque d'explosion combinatoire (si l'analyse repose sur un contexte trop large, les calculs ne sont plus possibles à grande échelle du fait de la complexité induite). En pratique, il s'agit le plus souvent des mots les plus proches ou d'annotations sur ces mots, suivant le modèle utilisé.

10 Les sections qui suivent présentent quelques travaux qui se situent, directement ou indirectement, dans cette perspective, à commencer par la lexicographie qui s'attache à l'étude du sens des mots.

Quand l'intuition n'est plus suffisante…

11 La lexicographie a depuis toujours eu recours aux corpus : comment définir un mot si ce n'est en examinant son usage, c'est-à-dire en essayant de rassembler autant d'emplois que possible, à partir de contextes et de genres textuels différents ? L'intuition permet d'avoir accès à un usage particulier mais celui-ci est souvent biaisé et rarement représentatif (le lexicographe ne décrit pas sa connaissance propre de la langue mais vise dans le cas général la langue courante partagée par une communauté). Au-delà de la lexicographie, la linguistique adopte de plus en plus souvent une démarche partant des données, maintenant que celles-ci sont largement disponibles, donnant accès à la fois à la quantité et à la diversité (Teubert, 2002).

12 Rassembler un ensemble d'exemples pertinent et représentatif a toujours été une des tâches les plus importantes et les plus difficiles du lexicographe. Les années 1960 ont vu le développement des premiers corpus informatisés mais il fallu attendre les années 1990 pour disposer de grands corpus annotés de l'anglais (cf. le BNC – British National Corpus, Burnard et Aston, 1998) et les années 2000 pour le français (Abeillé *et al.*, 2003), qui ne dispose d'ailleurs toujours pas de corpus de référence en tant que tel, c'est-à-dire de

corpus diversifiés, faisant une place à différents genres textuels, ainsi qu'à l'oral. Le web permet aujourd'hui de disposer de corpus de plusieurs milliards de mots, lemmatisés et annotés morphosyntaxiquement pour une dizaine de langues bien représentées sur la toile (outre l'anglais, les principales langues occidentales, le chinois, le japonais, etc.) (Baroni *et al.*, 2009).

13 La disponibilité de tels corpus a, comme on peut aisément s'en douter, complètement changé la donne : le lexicographe est maintenant soumis à une avalanche d'exemples à partir duquel il faut extraire les données les plus représentatives, sans omettre les usages les moins fréquents mais en évitant aussi les idiosyncrasies. Le linguiste, lui, peut profiter de ces nouvelles ressources pour examiner des exemples ou des constructions rares, évaluer de nouvelles hypothèses ou proposer de nouvelles explications face à des questions réputées difficiles. Prenons quelques exemples bien connus.

14 L'étude des quasi-synonymes a beaucoup bénéficié de la disponibilité de corpus sur support électronique. Il est dorénavant aisé de déterminer les contextes les plus fréquents d'un mot donné. Prenons l'exemple de *strong* et *powerful* en anglais. Si on interroge le grand public, voire un public plus spécialisé (l'expérience a été faite avec un ensemble d'étudiants de niveau master en linguistique à l'université de Cambridge), on voit que les locuteurs de l'anglais sont incapables de dire en quoi *strong* est différent de *powerful*, même s'ils savent choisir l'un ou l'autre des adjectifs en fonction du nom considéré (ainsi, *« powerful tool »* et *« powerful engine »* sont attestés, tout comme *« strong support »* et *« strong opposition »* ; à l'inverse, *« strong tool »* et *« powerful support »* sont très peu attestés, ce qui est aussi très significatif). Church et Hanks (1989) ont montré qu'une formule statistique simple (le *t-test*) permet de repérer les collocations fréquentes de ces deux adjectifs (c'est-à-dire les principaux couples adjectif-nom apparaissant en corpus), de les pondérer (en tenant compte de la rareté des formes isolées par rapport à la fréquence du couple considéré) et d'obtenir ainsi deux listes de noms, l'une plutôt associée à *strong*, l'autre à *powerful*. Les auteurs proposent ensuite en guise d'explication une opposition interne *vs* externe : *powerful* s'appliquerait à une force tournée vers l'extérieur (on pourrait parler de finalité : dans *« powerful tool »*, la puissance de l'outil est à mettre en relation avec la finalité de cet outil, de même pour *« powerful engine »*, le moteur produit de l'énergie pour autre chose que pour lui-même) tandis que *strong* qualifierait une qualité inhérente de l'objet vu en lui-même (dans *« strong opposition »*, l'opposition est forte en elle-même, elle ne vise pas à appuyer autre chose). Le corpus permet ici d'avoir immédiatement accès à un immense ensemble de données qui permet de mieux cerner la sémantique du mot étudié.

15 Le corpus électronique permet aussi d'étudier les emplois non classiques, déviants ou nouveaux de mots ou d'expressions linguistiques diverses. Atkins et Levin (1995) montrent que des verbes anglais réputés intransitifs (comme *quake* ou *quiver*, équivalents de *trembler* ou *frémir*) sont en fait susceptibles d'être employés transitivement, avec des nuances de sens diverses difficilement prédictibles sans exemples concrets (*« she saw a glint, it quaked her bowels, the steel of a cut-throat razor [...] »*, *« the brave young hedgedog did not quiver a quill »*). Ces exemples, même s'ils peuvent être qualifiés de littéraires, n'en révèlent pas moins des emplois remarquables de ces verbes qu'un panel d'étudiants qualifie « d'indubitablement intransitifs » (« un emploi transitif n'est pas possible / serait fautif »). Les mêmes étudiants, face aux exemples, se montrent surpris mais n'ont aucun problème de compréhension. Dans le même genre d'idées, Lamiroy et Charolles (2008) examinent le cas des verbes de parole en français, leurs usages transitifs et intransitifs et

surtout des cas de « transitivation » de verbes à l'origine intransitifs. Le corpus (en l'occurrence la base Frantext) permet d'avoir accès aux exemples suivants, pour les verbes *glapir* ou *maugréer*, généralement considérés comme intransitifs :

> [...] il se mit à **glapir** d'une voix affreusement enjouée des fragments de son répertoire de caf'conc', dont le refrain intéressant et quelque peu énigmatique est resté dans ma mémoire [...] (R. Gary, *La Promesse de l'aube*, 1960)
> Elle **maugréa** les pires injures à son intention, le traita de jean-foutre, de ruchemerde, de cornecul et de godebillard, de propoune et de bandouillette. (J. Lanzmann, *La Horde d'or*, 1994)

16 Les données jouent évidemment un rôle essentiel dans ce cas, elles permettent de voir la langue telle qu'elle est, ou dans son évolution quand on dispose de corpus couvrant une large période de temps.

17 Il est de plus possible de concevoir des dispositifs automatiques repérant certaines constructions remarquables, ce qui permet au lexicographe de se focaliser sur ces emplois non classiques. On a ainsi pu montrer (Messiant *et al.*, 2010) que le verbe *essaimer* dans le journal *Le Monde* est régulièrement transitif (même si cet emploi demeure évidemment minoritaire par rapport à l'ensemble des occurrences du verbe) : « Cuba a essaimé les effets de son syncrétisme culturel au gré des modes et des engouements », « Il n'en a pas moins essaimé son séjour chinois de proclamations optimistes pour l'avenir [du] pays », etc. Les analyseurs récents permettent d'obtenir une analyse syntaxique qui n'est pas toujours juste mais qui est suffisante pour repérer ce type de construction « déviante ».

18 Il existe enfin des questions de linguistique qui ne se laissent pas facilement cerner sans recours à un dispositif outillé. Comment expliquer l'alternance dative en anglais (« *He gave a book to John* » / « *He gave John a book* ») (Bresnan *et al.*, 2007) ? Quels sont les critères qui permettent d'expliquer la position de l'adjectif (antéposée ou postposée) en français, quand cette position est supposée libre (« une souplesse de gestion réelle », « une réelle souplesse de gestion ») (Thuillier *et al.*, 2010) ? Il est possible de montrer que chacun de ces phénomènes s'explique par une multiplicité de critères, qui doivent être pondérés et hiérarchisés dans la mesure où ils n'ont pas tous la même importance. Par exemple, Bresnan et ses collègues (2007) relèvent que de nombreuses hypothèses ont été formulées pour l'alternance dative en anglais. Une des plus populaires dit que les deux structures correspondraient à deux idées différentes, dans un cas un changement de possession (*event causing a change of state – possession)*, dans l'autre cas un changement de lieu (*event causing a change of place – movement to a goal*). Les auteurs montrent sans ambiguïté que ces hypothèses, qui ont fait l'objet de publications reconnues et abondamment citées (par exemple Gropen *et al.*, 1989), ne tiennent pas face aux faits. Ils examinent ensuite un ensemble d'autres hypothèses (caractère animé ou non des compléments ; caractère nouveau ou non des compléments – autrement dit, possibilité de pronominalisation ; longueur relative des compléments, etc.). Les auteurs montrent que certains de ces critères sont liés (cf. les pronoms sont généralement plus courts qu'un groupe nominal complet) et que tous sont nécessaires pour expliquer l'ensemble des cas envisagés. Leur modèle (fondé sur une analyse à base de régression logistique) permet en outre de pondérer les paramètres selon leur importance pour expliquer l'usage observé des deux constructions concurrentes. On voit ici que le recours aux données ne se limite pas à un simple comptage d'occurrences mais qu'il permet d'éliminer certaines hypothèses et d'établir des modèles mêlant plusieurs éléments que les humains ont du mal à hiérarchiser, tant les données sont complexes.

La statistique au secours de la sémantique

19 La sémantique est extrêmement difficile à aborder du point de vue du traitement automatique, pour au moins deux raisons : la sémantique dépend fortement de la syntaxe et demande donc que les paliers « inférieurs » de l'analyse soient résolus (découpage en mots et en phrases, analyse morphosyntaxique puis analyse syntaxique proprement dite) ; la sémantique met enfin en interaction tous les éléments de la phrase, les mots mais aussi leurs connexions et leurs contextes, ce qui rend l'analyse infiniment complexe.

20 De fait, les premières tentatives de traitement se sont rapidement heurtées au « mur » de la sémantique. Les raisons de cet échec sont à la fois simples et complexes. Tout d'abord, nul ne sait représenter la sémantique d'un mot (et encore moins d'une phrase). Un dictionnaire va donner des définitions mais on voit bien que les définitions des dictionnaires usuels, destinées à des humains, n'ont guère d'intérêt pour une machine dans la mesure où elles devraient elles-mêmes êtres interprétées pour être utilisables. Dès lors, quelle représentation choisir ? Beaucoup de propositions ont été faites : définir un jeu de primitives sémantiques plus ou moins réduit, rassembler les mots en familles de sens, les organiser en arbre, en réseau ou en treillis, etc. Le même problème de représentation se pose au niveau du syntagme ou de l'énoncé. De fait, un humain sait fournir des synonymes ou formuler des paraphrases mais nul ne sait « représenter » la sémantique d'un mot ou d'un énoncé (Kayser, 1987 ; Poibeau, 2011).

21 De plus, on a vu dans la section précédente la difficulté à distinguer les usages d'un mot. Combien de sens faut-il considérer pour un verbe comme *couper* ? Suivant les dictionnaires, le mot connaît de quatre ou cinq nuances de sens mais si on se réfère à l'équivalent anglais *to cut*, on voit que Wordnet[5] distingue une quarantaine de (nuances de) sens différents (Gayral et Saint-Dizier, 1999) ! Des expériences d'annotation manuelles révèlent de plus que ces sens (même quand la granularité de la ressource est faible) ne sont pas exclusifs : plusieurs sens distingués par le dictionnaire peuvent, à des degrés divers, s'appliquer pour un emploi donné (Erk et McCarthy, 2009 ; Erk *et al.*, 2009), alors qu'un autre ne semble correspondre à aucune entrée du dictionnaire. Ces questions ne peuvent pas être résolues par l'ordinateur : la granularité de description dépend de choix complexes (connaissance de la langue, finalité de l'application, choix de codage) qui sont rarement explicites. De fait, il y a un certain paradoxe à essayer de découvrir le nombre de sens d'un mot si on admet que le sens varie en fonction du contexte.

22 Si l'ordinateur ne peut résoudre le problème du sens, il peut néanmoins donner à voir le sens suivant différentes perspectives. Nous évoquerons rapidement deux domaines complémentaires, le premier étant celui de la désambiguïsation sémantique, le second celui de la traduction automatique.

23 La désambiguïsation est la pierre angulaire du traitement automatique. À partir du moment où l'on dispose de dictionnaires très complets d'une langue donnée, il faut se résoudre à constater l'ambiguïté des formes rencontrées (ceci amène d'ailleurs à se demander si l'exhaustivité est toujours la meilleure solution : faut-il des dictionnaires très complets, ce qui a pour conséquence de multiplier les sources d'ambiguïtés ou des dictionnaires moins complets mais plus faciles à gérer ? On voit ici les choix cornéliens que pose le traitement automatique de la langue). Par exemple, *danse* peut être un nom ou un verbe, *voler* peut concerner un oiseau ou un malfaiteur, etc. Que l'ambiguïté concerne la catégorie grammaticale *(danse)* ou le sens *(voler)*, l'ordinateur est aveugle : tant qu'on

n'a pas fourni à la machine des algorithmes de désambiguïsation (c'est-à-dire des techniques permettant de faire un choix, en fonction de l'entourage proche ou lointain, parmi les différentes possibilités attachées au mot dans le dictionnaire), celle-ci n'est évidemment pas capable de faire un choix (sinon à choisir « en aveugle » la solution la plus probable, technique parfois utilisée en dernier recours, quand rien d'autre n'est disponible pour un résultat moins aléatoire).

24 Trouver la catégorie grammaticale d'un mot (savoir si *danse* est un nom ou un verbe en contexte) est une tâche relativement bien maîtrisée (Schmid, 1994). Ainsi, les systèmes étiquettent correctement environ 97 % des mots d'un corpus écrit journalistique (par exemple le journal *Le Monde*) principalement en se fondant sur les *n* étiquettes précédent le mot à analyser[6]. En ce qui concerne la désambiguïsation sémantique, les résultats sont bien moindres. Le taux de réussite dépend de la granularité des sens distingués : ce taux dépasse rarement 65 % avec une granularité fine de type Wordnet (Snyder et Palmer, 2004) et peut aller jusqu'à 82 % pour les meilleurs systèmes avec une granularité beaucoup plus grossière (Navigli *et al.*, 2007). Ces chiffres sont surtout parlants si on les met en regard de performances obtenues par un système peu sophistiqué (ce qu'on appelle la *baseline* et qui permet de mesurer l'avantage d'un algorithme donné par rapport à une technique « bas de gamme ») : si on dispose d'un dictionnaire où les sens sont classés suivant leur fréquence, ce qui n'est pas rare (cf. Wordnet) et qui permet donc de connaître le sens le plus probable d'un mot ambigu, choisir systématiquement le sens le plus fréquent permet d'atteindre des performances de l'ordre de 60 % pour une granularité fine et 78 % avec une granularité grossière (ces chiffres sont donnés par les publications déjà citées). On constate donc que les algorithmes de désambiguïsation sémantiques obtiennent des résultats seulement légèrement meilleurs que la *baseline* (comparer 60 % avec 65 %, 78 % avec 82 %). Autrement dit, la stratégie bas de gamme fondée sur l'étiquette la plus probable est difficile à battre et les algorithmes actuels sont loin d'être parfaits.

25 Pourquoi ne sait-on pas faire mieux en dépit de plusieurs dizaines d'années de recherche ? Plusieurs raisons viennent à l'esprit : les sens ne sont pas nettement séparés (voir les références déjà citées, Erk et McCarthy, 2009 ; Erk *et al.*, 2009). Il est aussi important d'avoir accès au taux d'accord entre annotateurs, qui se situe autour de 70 % quand l'annotation est faite à partir de Wordnet[7] (baseline : 60 %, performance du meilleur système : 65%). Ces chiffres confirment l'observation de Erk et de ses collègues : comme plusieurs nuances de sens peuvent souvent être observées en contexte, le taux d'accord entre humains lors de tâches d'annotation n'est pas fameux. Enfin (et surtout), les contextes d'usage ne peuvent être tous décrits à l'avance du fait de la créativité des locuteurs, ce qui limite les performances des systèmes automatiques.

26 Cette dernière remarque pourrait sembler aller dans le sens de Chomsky (cf. le début de cet article) : à quoi bon s'intéresser aux données et aux corpus si ceux-ci sont toujours incomplets ? C'est justement cette question qui nous intéresse : en mettant au point des algorithmes explicites, les linguistes informaticiens (ou tout au moins une partie d'entre eux) visent, au-delà de l'aspect applicatif de leurs recherches, à étudier les mécanismes rendant possible la compréhension des langues. Il ne s'agit pas de reproduire les mécanismes de compréhension directement par ordinateur mais de mettre au point des dispositifs montrant les possibilités et les limites d'un apprentissage fondé sur la seule confrontation avec des données réelles.

27 Le traitement automatique des langues essaie depuis quelques années d'aller plus loin dans cette direction en mettant au point des systèmes inductifs visant à se passer du dictionnaire initial donnant un inventaire de sens *a priori* (Lin et Pantel, 2002 ; Navigli et Crissafulli, 2010 ; van de Cruys *et al.*, 2011). Il s'agit en quelque sorte de dépasser la tâche de désambiguïsation sémantique pour essayer de déterminer automatiquement un ensemble de sens à partir de l'emploi des mots en corpus (il ne s'agit donc pas de déterminer un inventaire de sens dans l'absolu mais de déterminer automatiquement, pour un corpus donné, quels sens émergent en fonction des contextes d'emploi). Ces recherches sont encore préliminaires mais, au-delà des simples aspects applicatifs, elles semblent importantes car elles permettent d'approcher une position plus naturelle. Même dans une conception du langage faisant une large part à l'inné, il est évident que l'homme acquiert une partie de ses connaissances sur la langue à partir de ce qu'il entend (notamment la connaissance des mots et de leur combinatoire, dont on a vu que c'est un élément à la fois essentiel et particulier à chaque langue, que l'on peut donc difficilement poser comme inné).

Le cas de la traduction automatique

28 L'application phare du traitement automatique est évidemment la traduction. Dès les années 1950 avec l'apparition des premiers ordinateurs, la traduction est apparue comme une application essentielle, surtout dans le contexte de guerre froide qui rendait primordiale la capacité de traduire de grosses quantités de documents du russe vers l'anglais (et inversement pour le monde soviétique) (Hutchins, 1986, 2000, 2001 ; Léon, 2000).

29 Il n'est pas possible de retracer ici l'histoire de la traduction automatique, mais il faut rappeler l'importance du rapport ALPAC (*Automatic Language Processing Advisory Comittee*) de 1966, jugeant les applications de traduction automatiques inaccessibles à court et moyen terme, du fait de la complexité de la langue, qui avait été largement sous-estimée jusque-là. Le rapport pointe la nécessité de produire des analyses sémantiques fines pour pouvoir traduire, et souligne que celles-ci ne pourront fonctionner si elles ne mettent pas en œuvre des connaissances fines sur le monde.

30 Les exemples de mauvaises traductions pullulent, et chaque magazine examinant la question rapporte son lot d'exemples de traduction amusante ou navrante. Hutchins (1986) rappelle certains exemples célèbres, ainsi *Out of sight, out of mind*, et *The spirit is willing but the flesh is weak* qui auraient été traduits par certains outils automatiques (en passant par exemple de l'anglais au russe, puis à nouveau du russe à l'anglais) par *« invisible insanity »* pour le premier et par *« The whiskey is all right but the meat has gone bad »* pour le second (Hutchins rappelle les autres traductions colportées par la légende, comme *« Invisible and insane »* et *« The vodka is good but the meat is rotten »* ou encore *« invisible lunatics »* et *« the ghost is willing but the meat is feeble »*).

31 Ces exemples mettent bien en évidence la primauté du contexte : impossible de traduire mot à mot, choisir la traduction exacte demande de préciser la signification des mots voire des expressions complexes par un examen minutieux de leur voisinage. Cette analyse est naturelle, évidente, voire inconsciente pour un humain, mais elle est infiniment complexe pour une machine du fait de la combinatoire de l'ensemble des possibilités de signification, et du fait que cette analyse exige d'activer un ensemble de

connaissances sur le monde que la machine n'a pas. Encore une fois, ces processus très complexes sont quasi inconscients pour l'humain : on ne prend généralement conscience du problème que quand on pointe une réelle ambiguïté, ce sur quoi joue souvent la publicité[8]. Les grammaires de construction (Goldberg, 1995, 2008) prennent partiellement en compte cette complexité en attachant des significations à des éléments complexes (syntagme, etc.) : ainsi, la signification d'un énoncé n'est plus seulement la somme de la signification de chacun des composants de l'énoncé (Col *et al.*, 2010).

32 La recherche s'est alors focalisée sur la « compréhension de texte » et plusieurs programmes de recherche se sont attachés à décrire et formaliser des « connaissances sur le monde » ou « connaissances de sens commun » nécessaires aux inférences à effectuer pour comprendre le sens d'un texte (un des problèmes majeurs pointés par le rapport ALPAC). Doug Lenat avec sa société CYCorp a bénéficié dès les années 1980 d'importants moyens pour développer manuellement une base de connaissances sur le monde (Guha et Lenat, 1990 ; Lenat, 1995). Plusieurs groupes de chercheurs ont depuis cherché à dériver ce type de connaissances automatiquement à partir de textes en langage naturel disponibles en ligne (Rumshisky *et al.*, 2006). L'idée est simple : il y a tellement de documents disponibles qu'un nombre infini de connaissances pourrait en être extrait semi-automatiquement afin de résoudre le problème initial du manque des connaissances sur le monde *(knowledge acquisition bottleneck)*. Le résultat de ces recherches est des plus mitigés, probablement parce qu'il est difficile de déterminer les connaissances nécessaires au processus de compréhension et qu'il n'est pas évident que celles-ci soient explicites au sein des textes disponibles en ligne (il est même évident qu'une grande partie de ces connaissances n'est pas disponible).

33 De fait, le rapport ALPAC en 1966 a eu pour conséquence un quasi arrêt du financement et donc des recherches en traduction automatique aux États-Unis puis dans le monde (Hutchins, 1986). La traduction automatique qui était un programme fleuve des années 1950 jusqu'au milieu des années 1960 est alors passée au second plan, la tâche étant généralement jugée trop difficile en l'état des connaissances d'alors (même si des entreprises et des laboratoires ont continué à travailler sur la question depuis les années 1950). Les programmes donnant les meilleures performances dans les années 1990 étaient souvent les programmes les plus anciens (par exemple Systran, une entreprise fondée dans les années 1960), c'est-à-dire les programmes ayant accumulé le plus de données afin notamment de choisir la bonne traduction en fonction du contexte.

34 L'existence de données langagières massives et multilingues sur la toile à partir des années 1990 va relancer les recherches, guidées par des besoins applicatifs clairs. Notons qu'on ne cherche pas obligatoirement une traduction exacte quand on est sur internet : une traduction vague donnant une idée du contenu du document original peut parfois être une aide précieuse (par exemple pour décider de demander une traduction plus exacte à un traducteur professionnel).

35 Ainsi, au tout début des années 1990, une équipe de chercheurs d'IBM propose une approche à l'opposé de toutes les présuppositions sur le domaine (Brown *et al.*, 1990). Cette équipe, menée par F. Jelinek au centre Watson d'IBM, ne s'intéresse ni aux problèmes d'ambiguïté ni aux connaissances sur le monde, mais utilise simplement le corpus bilingue Hansard (transcriptions des débats du parlement canadien, disponibles en version anglaise et française « alignées » de façon quasi parfaite, c'est-à-dire qu'il est possible de mettre en regard le texte anglais et français au niveau de la phrase, voire du syntagme ou du mot). L'idée des concepteurs du système (qui venaient du monde de

l'analyse de la parole) est simple : trouver automatiquement des équivalences de langue à langue et voir jusqu'où peut aller une analyse purement automatique fondée sur une approche statistique s'appuyant sur la notion de *n*-grammes (des séquences de *n* mots).

36 L'équipe d'IBM met alors au point un modèle de traduction n'ayant aucune connaissance explicite sur la langue ou sur le monde mais essayant juste de repérer les équivalences de langue à langue à partir du corpus Hansard. Le modèle d'IBM[9], bien qu'en apparence simpliste au vu de la tâche visée, obtint alors des performances tout à fait remarquables en comparaison de systèmes plus élaborés, qui avaient bénéficié d'années de développement par des équipes de linguistes. La raison en est relativement simple (mais contre intuitive) : le langage a une dimension statistique fondamentale. Il est ainsi possible de trouver non seulement des traductions de mots simples en ayant recours à de grands corpus alignés, mais on peut aussi identifier des expressions idiomatiques pourvu qu'elles soient locales et relativement fréquentes (ceci grâce à un « alignement asymétrique », qui autorise une correspondance de 1 mot dans la langue source vers *n* mot dans la langue cible et *vice versa* ; l'algorithme peut par exemple identifier le fait qu'en face de *potatoe* se trouve très fréquemment l'expression *pomme de terre*). Les *n*-grammes constituent une modélisation extrêmement simple de la notion de contexte et permettent de désambiguïser les mots naturellement quand le contexte local le permet. Dans les autres cas (mot simple apparaissant dans un contexte inconnu, c'est-à-dire dans un *n*-gramme non identifié jusqu'alors), la traduction la plus probable du mot est appliquée, ce qui permet déjà d'obtenir des performances non négligeables comme on l'a vu dans la section précédente consacrée à la désambiguïsation sémantique[10].

37 L'approche est frustre mais elle fonctionne aussi bien qu'une autre car les méthodes automatiques permettent de couvrir un grand nombre de structures très fréquentes là où le linguiste décrira sans doute plus précisément certains cas qui risquent d'être finalement peu productifs. Jelinek fera scandale en déclarant sur le ton de la blague (sans que ce soit foncièrement faux à la base) qu'à chaque fois qu'il « vire » un linguiste, les performances de son système s'améliorent (*« Every time I fire a linguist, the performance of our speech recognition system goes up. »*, Workshop on Evaluation of NLP Systems, Wayne, Pennsylvania, US, déc. 1988). La citation vaut pour le contexte de la reconnaissance de la parole mais elle a été fréquemment reprise pour la traduction automatique.

38 L'approche sera surtout reprise, améliorée et complexifiée par Ochs à la fin des années 1990, pour se trouver aujourd'hui à la base de Google Traduction, un des outils de traduction automatique les plus populaires et les plus performants sur le web (voir par exemple Ochs et Ney, 2000). L'approche initiale proposée par IBM est complexifiée pour pouvoir prendre en compte des alignements *m*-*n* (*m* mots pouvant correspondre à *n* mots, alors que les premiers algorithmes proposés par IBM devaient obligatoirement avoir un mot simple comme point de départ). Il ne s'agit pas ici de rentrer dans les détails techniques des algorithmes actuels qui se sont évidemment largement complexifiés, mais de retenir l'idée qu'une approche à base de corpus est généralisable et efficace même pour des tâches aussi exigeantes que la traduction automatique.

Les limites des statistiques

39 Il ne faudrait pas déduire de ce qui précède que tout dans la langue est statistique ou peut se résoudre par le simple fait de disposer d'un jeu de données représentatif. Les algorithmes à base de traitement statistique sont efficaces pour deux raisons : ils

permettent de bien couvrir les cas les plus essentiels[11] et de nombreux problèmes linguistiques peuvent finalement se traiter localement. Cependant, les succès récents de ce type de méthodes ne doivent cacher leurs limites.

40 Les mots peuvent le plus souvent être catégorisés morphosyntaxiquement sur la base d'un contexte relativement local. Cela est beaucoup moins vrai pour la désambiguïsation sémantique pour lequel on obtient des résultats beaucoup moins fiables, comme on l'a vu *supra* (les méthodes statistiques ou probabilistes obtiennent des résultats un peu supérieurs à la *baseline* seulement). L'analyse sémantique demande de prendre en compte un contexte plus large, qui échappe en partie aux systèmes à base d'apprentissage.

41 Enfin, il faut aussi relativiser les succès récents de systèmes comme Google Traduction : même si le système est impressionnant au vu des techniques employées, il est toujours aisé de le mettre en défaut et Google Traduction n'échappe pas aux mauvaises traductions dès que l'on s'éloigne des données sur lesquelles le système a été « entraîné » (c'est-à-dire le type et le domaine du corpus d'apprentissage : dans le cas de Google Traduction, il s'agit majoritairement de textes non techniques provenant d'internet). Une phrase complexe sera quasiment toujours mal traduite (ainsi le système a du mal à traiter les dépendances longue distance), sans parler de textes issus de domaines spécialisés (Google Traduction est peu fiable quand il s'agit d'aider à traduire un article scientifique ; l'expérience montre qu'il vaut souvent mieux repartir de zéro plutôt que d'essayer d'utiliser les éléments de traduction fournis par Google). Il faudrait alors « ré-entraîner » le système (c'est-à-dire lui faire acquérir un nouveau jeu de règles à partir de données proches de celles visées) mais cela est souvent impossible du fait précisément du manque de données parallèles (c'est-à-dire en situation de traduction) dans des domaines spécialisés.

42 En effet, la qualité de la traduction est directement liée aux ressources parallèles disponibles. Pour les couples de langues où relativement peu de données alignées ou alignables existent, les performances restent très faibles (traduction du finnois au grec par exemple, ou même du finnois au français). La recherche s'est récemment focalisée sur les corpus dits comparables (Zweigenbaum *et al.*, 2011) : il s'agit de corpus qui ne sont pas des traductions mais des textes dans deux (ou plusieurs) langues sur un sujet proche, qui peuvent laisser espérer trouver des équivalents linguistiques d'une langue à l'autre (Morin et Daille, 2004). Ceci est surtout vrai pour les termes techniques, mais ne peut guère aller au-delà.

43 Enfin, la traduction automatique reste aujourd'hui encore faite à partir de séquences supposées être en situation de traduction et mises ensemble bout à bout dans l'espoir de produire ainsi une phrase bien formée dans la langue cible. Aucun système ne sait reformuler un propos, ce qui est pourtant souvent nécessaire afin d'échapper à un mot à mot peu compréhensible ou peu naturel (cf. l'exemple souvent cité par Martin Kay : « veuillez ne rien oublier dans le train » traduit par *« be sure to take all your belongings with you when you leave the train »* : le verbe *oublier* n'a pas de correspondant direct dans la version anglaise).

44 Il faut alors revenir à notre point de départ et reconnaître que les statistiques ne peuvent pas tout, et que le linguiste peut finalement apporter des connaissances utiles pour faire progresser les systèmes. Jelinek, au-delà de son « bon mot » sur les linguistes, reconnaissait lui-même que l'approche à base de traitement statistique n'était qu'une première étape faisant efficacement « un travail de gros » mais ne pouvait produire à elle

seule une traduction de qualité professionnelle *(« we must put language back into language modeling »)*. On peut toutefois faire le constat que la recherche en traduction automatique est encore aujourd'hui une recherche de type statistique et/ou algorithmique, et que la linguistique y a toujours la portion congrue. De fait, les systèmes de traduction sont complexes : leur amélioration repose sur de nombreux facteurs hétérogènes et les algorithmes peuvent encore améliorer les résultats obtenus automatiquement, ce qui s'avère souvent plus prometteur que l'intégration de connaissances linguistiques fines.

Une linguistique déstabilisée

45 Plus fondamentalement, il nous semble que le succès des méthodes empiriques a déstabilisé le domaine du traitement des langues et, par contrecoup, la linguistique elle-même. En effet, comment ne pas être déstabilisé quand des systèmes fondés sur des techniques relativement simples et prenant le contre-pied des principes de base d'un domaine bien établi obtiennent des performances supérieures aux systèmes élaborés depuis des décennies par de bonnes équipes du domaine ? Les modèles mis au point chez IBM et repris par des systèmes comme Google Traduction n'incluent aucune information sémantique en tant que telle et n'incluent pas de système de désambiguïsation sémantique, alors qu'il s'agissait d'un point supposé essentiel jusque-là.

46 Dans cette section, je citerai abondamment Yorick Wilks (notamment Wilks, 2008, 2011), qui a travaillé depuis la fin des années 1960 sur différents programmes d'analyse sémantique automatique (y compris sur la traduction automatique, l'extraction d'information et le dialogue homme-machine). Ses réflexions sont souvent partisanes voire provocatrices mais aussi riches d'enseignements.

47 Le premier point que souligne Wilks est l'incrédulité qu'ont d'abord suscitée les approches empiriques, si opposées à la vulgate de l'époque. Wilks a même eu des doutes face au système présenté par Jelinek et ses collègues d'IBM : « Je n'étais pas sûr que l'"équation fondamentale de la traduction automatique" d'IBM pouvait suffire à expliquer leurs résultats et j'ai suggéré qu'ils développaient autre chose pour produire leur taux de réussite remarquable, à savoir environ 50 % de phrases bien traduites. Maintenant que leur méthode s'est répandue dans toute la communauté de la linguistique informatique, je retire mes critiques : IBM avait bien sûr raison, et avait tout à nous apprendre. »[12]

48 Un des enseignements des recherches empiriques est évidemment la nécessité de consulter les données, et si possible de consulter des données massives[13]. Ceci pose évidemment question face à l'approche traditionnelle en linguistique qui se fonde le plus souvent sur la base d'un nombre limité d'exemples, parfois fabriqués par le linguiste lui-même. Wilks parle en particulier de sémantique formelle et va assez loin dans sa critique en prétendant que celle-ci s'intéresse le plus souvent à des faux problèmes, comme le cas fort étudié des *donkey sentences* (des phrases « dans lesquelles les interactions de portée entre un quantificateur universel et un indéfini posent un problème de compositionalité », pour reprendre la définition donnée par la Sémanticlopédie[14], cf. « *Every farmer who owns a donkey beats it* »). Voilà ce qu'en dit Wilks : « Les *donkey sentences* n'existent pas, à part bien sûr dans les corpus de linguistique. Elles n'existent pas, pas plus qu'on ne dit « John veut épouser une Norvégienne » pour signifier une Norvégienne particulière – on ne parle tout simplement pas comme cela ! Donc ces phrases ne sont pas ambiguës, contrairement à ce que dit la théorie. »[15] (Wilks, 2011). On ne peut pas donner tout à fait tort à Wilks, qui conclut ainsi : « Le problème n'est pas lié à ceux qui

s'intéressent à un nombre limité de données, mais il est lié à toute une tradition qui fabrique et discute d'exemples qui ne correspondent en aucune façon à ce qui pourrait se dire dans la vie réelle.[16] » (Wilks, 2011).

49 Cette critique vigoureuse de la sémantique formelle laisse ouverte la question de savoir quel type de sémantique, et plus généralement de linguistique serait à la fois utile et réellement en prise avec le réel. K. Spärck Jones, autre éminent chercheur du domaine, parlait ainsi des risques mais aussi des opportunités à saisir pour la linguistique :

> « Donc, nous devons être vigilants. Pas simplement parce que nous pourrions nous retrouver à mettre la charrue avant les bœufs. Nous pouvons devenir obsédés par les roues et finir par réinventer sans fin des roues carrées, ou trop perfectionnées, ou inadéquates, ou juste des monocycles. Ce qui compte, c'est la façon dont la charrue, les bœufs et le chargement font ensemble un trajet cohérent [...]. C'est la façon dont on intègre langage et calcul d'une part, symbolique et statistique de l'autre. Et pour réussir en cela, nous ne devons pas oublier que la linguistique générale peut avoir des choses à nous offrir, même si ce n'est pas tout ce que les linguistes eux-mêmes peuvent imaginer.[17] » (Spärck Jones, 2007)

50 K. Spärck Jones parle de ce que la linguistique peut offrir au traitement des langues mais il nous semble que le propos peut s'appliquer de manière plus générale, sans avoir de visée pratique particulière.

Conclusion : la victoire du langage ordinaire ?

51 Les propos de Wilks que nous avons rapportés dans la section précédente doivent être relativisés. La sémantique formelle ne se résume pas aux *« donkey sentences »* et il est évident que nombre de recherches ayant une base formelle ont permis des avancées, y compris pour une approche informatisée. À l'inverse, ce n'est pas parce qu'un chercheur utilise un ordinateur ou des données massives que ses résultats auront un intérêt ou une validité supérieure à celle de travaux plus théoriques. De fait, la linguistique de corpus n'a pas toujours permis des avancées décisives. La tendance descriptive et non explicative soulignée par Chomsky n'est pas sans fondement, de même que la tendance à s'intéresser à la surface des énoncés plutôt qu'à leur structure sous-jacente, qui est pourtant fondamentale. Le langage est structure, même s'il nous parvient avec une certaine linéarité : cf. la critique de Manning (2003), citée et traduite par Habert (2002) : « la linguistique de corpus [...] – ou "modèles de la grammaire basés sur les usages" – [...] emploie intégralement la rhétorique vertueuse sur le fait de constituer une science empirique, falsifiable, objective mais elle échoue parce qu'elle se limite elle-même largement à des faits langagiers de surface au lieu d'utiliser des modèles formels sophistiqués de la grammaire faisant un appel fort à des structures cachées (comme des arbres syntagmatiques ou d'autres niveaux abstraits de représentation.[18] »

52 Le point qui nous semble le plus intéressant dans les recherches récentes en linguistique est le souci de revenir à la langue réelle, au « langage ordinaire », pour reprendre le terme employé en philosophie du langage. Il s'agit avant tout de s'interroger sur les données manipulées, s'assurer qu'elles sont plausibles, s'appuyer sur des corpus réels si possible, et prendre en compte le point de vue de locuteurs natifs. La linguistique de corpus mais aussi la psycholinguistique occupent une place particulière dans ce cadre. La dimension cognitive n'est pas ignorée non plus (il faut cependant garder en tête qu'il y a toujours eu des linguistes partant des données et non des théories, mais leurs écrits

n'étaient alors pas aussi populaires que les travaux dits théoriques ou formels[19], nous semble-t-il).

53 S'intéresser à la langue telle qu'elle est ne veut pas dire que l'on est obligé de s'intéresser à des données massives en corpus. Par exemple, nous avons nous-mêmes étudié récemment le cas de l'instrumental sujet (c'est-à-dire la possibilité d'avoir un instrument en position sujet), qui pose des questions redoutables au linguiste, même si ces questions peuvent en soi sembler anecdotiques. Pourquoi a-t-on l'alternance suivante *Jean ouvre la porte avec la clé* *la clé ouvre la porte* mais pas *Jean mange sa soupe avec une cuillère* **la cuillère mange la soupe* ? Pourquoi la phrase *la grue a déplacé la terre* semble-elle plus acceptable que *la pelle a déplacé la terre* ? Il nous semble qu'il s'agit de questions légitimes, purement linguistiques, qu'il faut poser. Il faut les poser car ces phrases ne sont pas problématiques pour un locuteur français : elles semblent naturelles et des exemples peuvent facilement être trouvés dans des corpus réels, sur la toile notamment. Ces exemples sont intéressants car ils posent la question du rapport entre connaissances linguistiques et connaissances sur le monde, problème récurrent en analyse linguistique, surtout quand on s'intéresse aux données de façon contrastive : le grec et le néerlandais n'acceptent pas d'instrument en position sujet, contrairement à l'anglais ou au français. Ceci amène à s'interroger sur l'organisation des catégories dans la langue, ou plutôt dans les langues.

54 L'exemple de l'instrumental sujet nous semble poser d'autres questions remarquables, au cœur de la recherche en linguistique actuelle. En particulier, les jugements d'acceptabilité sont nuancés : les phrases examinées sont souvent possibles mais jugées plus ou moins acceptables les unes par rapport aux autres. De fait, les linguistes s'interrogent de plus en plus sur ce que recouvre cette notion d'« acceptabilité », au cœur du travail linguistique. Jusqu'à récemment, dans beaucoup d'études, les exemples étaient jugés de façon binaire (acceptable / inacceptable), et le lecteur était parfois dubitatif face aux jugements d'acceptabilité proposés. Pour juger de l'acceptabilité d'exemples, il est nécessaire d'interroger différents informateurs, de vérifier l'existence de phrases similaires dans un corps type (éventuellement le web), voire de procéder à des expériences psycholinguistiques.

55 Ce souci des données n'est pas une révolution. Il s'agit plutôt d'un retour à des pratiques qui n'ont jamais été complètement abandonnées mais que l'arrivée de données massives a remis au goût du jour. Il ne s'agit pas d'une remise en question de la théorie, qui reste nécessaire, mais le traitement automatique des langues et les besoins applicatifs liés ont bien montré l'intérêt d'approches « brutes » pour des domaines réputés difficiles comme la sémantique ou la traduction automatique.

56 Cependant, comme nous espérons l'avoir montré, les problèmes sont encore loin d'être résolus. Gageons que la sémantique résistera encore longtemps à tout traitement automatique et à toute formalisation théorique.

BIBLIOGRAPHIE

Abeillé, A., *Les Nouvelles syntaxes*, Paris, Armand Colin, 1993.

Abeillé, A., Clément, L. et Toussenel, F., « Building a Treebank for French », in A. Abeillé (éd.), *Treebanks,* Dordrecht, Kluwer, 2003.

Atkins, B. T.S. et Levin, B., « Building on a Corpus : A Linguistic and Lexicographical Look at Some Near-synonyms », *International Journal of Lexicography*, n° 8, 1995, p. 85-114.

Baroni, M., Bernardini, S., Ferraresi, A. et Zanchetta, E., « The WaCky Wide Web : A Collection of Very Large Linguistically Processed Web-Crawled Corpora », *Language Resources and Evaluation* , n ° 43, vol. 3, 2009, p. 209-226.

Bresnan J., Cueni, A., Nikitina, T. et Baayen, H., « Predicting the Dative Alternation », in G. Boume, I. Kraemer, et J. Zwarts (éd.), *Cognitive Foundations of Interpretation*, Amsterdam, Royal Netherlands Academy of Science, 2007, p. 69-94.

Brown, P. F., Cocke, J., Della P., S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R. L. et Roossin, P., « A Statistical Approach to Machine Translation », *Computational Linguistics*, n° 16, vol. 2, 1990, p. 79-85.

Burnard, L. et Aston, G., *The BNC handbook : exploring the British National Corpus*, Edimbourg, Edinburgh University Press, 1998, p. xiii.

Charniak, E., « A Maximum-Entropy-Inspired Parser », *Proceedings of the 1st Annual Meeting of the North American Association for Computational Linguistics (NAACL'2000)*, Seattle, 2000.

Chomsky, N., *Syntactic Structures*, La Hague, Mouton, 1957.

—, *Aspects of the Theory of Syntax*, Cambridge, The MIT Press, 1965.

—, *Lectures on Government and Binding : The Pisa Lectures*, Holland, Foris Publications, 1981.

—, *The Minimalist Program*, Cambridge, The MIT Press, 1995.

—, « The master and his performance : An interview with Noam Chomsky », entretien avec J. Andor, *Intercultural Pragmatics*, n° 1, vol. 1, 2004, p. 93-111.

Chklovski, T. et Mihalcea, R., « Open Mind Word Expert : Creating Large Data Collections with Web Users' Help », *DLib Magazine*, n° 6, vol. 2, 2002.

Church, K. et Hanks, P., « Word Association Norms, Mutual Information and Lexicography, », *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics (ACL'89),* Vancouver, Canada, 1989.

Col, G., Aptekman, J., Girault, S. et Victorri, B., « Compositionnalité gestaltiste et construction du sens par instructions dynamiques », *CogniTextes*, n° 5, 2010, (http://cognitextes.revues.org/372).

Copestake, A. et Lascarides, A., « Integrating Symbolic and Statistical Representations: The Lexicon Pragmatics Interface », *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97)*, Madrid, 1997, p. 136-143.

Erk, K., McCarthy, D. et Gaylord, N., « Investigations on Word Senses and Word Usages », *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL'2009)*, Singapour, 2009.

Erk, K. et McCarthy, D., « Graded word sense assignment », *Proceedings of the Empirical Methods in Natural Language Processing Conference47th Annual Meeting of the Association for Computational Linguistics (EMNLP'2009)*, Edimbourg, 2009.

Evans, R. et Gazdar, G., *The DATR Papers*, University of Sussex, Cognitive Science Research Paper CSRP 139, Brighton, 1990.

Fillmore, Ch. J., « "Corpus linguistics" vs "Computer-aided armchair linguistics" », *Directions in Corpus Linguistics*, La Hague, Mouton de Gruyter, 1992, p. 35-60, (Proceedings from a 1992 Nobel Symposium on Corpus Linguistics, Stockholm.).

Firth, J. R., *Papers in Linguistics 1934-1951*, Oxford, Oxford University Press, 1957.

Gayral, F. et Saint-Dizier, P., « Peut-on couper à la polysémie verbale ? », *Actes de la conférence Traitement automatique des langues naturelles* (TALN'99), Cargèse, 1999.

Goldberg, A., *Constructions : A Construction Grammar Approach to Argument Structure*, Chicago, University of Chicago Press, 1995.

—, *Constructions at Work : The Nature of Generalization in Language,* Oxford, Oxford University Press, 2006.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R. et Wilson, R., « The learnability and acquisition of the dative alternation », *Language*, n° 65, 1989, p. 203-257.

Gross, M., *Méthodes en syntaxe. Le régime des constructions complétives*, Paris, Hermann, 1975.

Guha, R. et Lenat, D., *Building Large Knowledge Based Systems*, Reading (Mass.), Addison Wesley, 1990.

Habert, B., « Outiller les linguistes / outiller la linguistique : par où, par qui commencer ? », *Actes de la table ronde de la conférence Traitement automatique des langues naturelles* (TALN'2002), Nancy, 2002.

—, « Portrait de linguiste(s) à l'instrument », *Texto !*, n° 104, 2005, (www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html).

Harris, R. A., *The Linguistics Wars*, Oxford, Oxford University Press, 1995.

Hutchins, J., *Machine translation : past, present, future*, Chichester, Ellis Horwood (Ellis Horwood Series in Computers and their Applications), 1986.

—, *Early Years in Machine Translation : Memoirs and Biographies of Pioneers*, Amsterdam, John Benjamins, 2000.

—, « Machine translation over fifty years », *Histoire, Épistémologie, Langage*, n° 12 vol. 1, 2001, p. 7-31.

Kaplan, R. et Bresnan, J., « Lexical-Functional Grammar : A formal system for grammatical représentation », in J. Bresnan (éd.), *The mental representation of grammatical relations*, Cambridge (Mass.), The MIT Press, 1982.

Kayser, D., « Une sémantique qui n'a pas de sens », *Langages*, n° 87, 1987, p. 33-45.

Lamiroy, B. et Charolles, M., « Les verbes de parole et la question de l'(in)transitivité », *Corpus*, n ° 2, 2008, (http://discours.revues.org/3232).

Lenat, D., « CYC : A Large-Scale Investment in Knowledge Infrastructure », *Communications of the ACM*, n° 38, vol. 11, 1995.

Léon, J., « De la traduction automatique à l'automatisation de la traduction. Parcours historique », *Bulag*, n° 25, 2000, p. 5-21.

—, « Claimed and unclaimed sources of corpus linguistics ». *Henry Sweet Society Bulletin*, n° 44, 2005, p. 36–50.

Lin, D. et Pantel, P., « Discovering Word Senses from Text », *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining* (KDD), Edmonton, Canada, 2002, p. 613–619.

Manning, C. D., « Probabilistic Syntax », in R. Bod, J. Hay et S. Jannedy (éd.) *Probabilistic Linguistics*, Cambridge (Mass.), The MIT Press, 2003, p 289-341.

McEnery, T., Xiao, R. Z., et Tono, Y., *Corpus-based language studies : An advanced resource book*, Londres, Routledge, 2005.

McEnery, T. et Wilson, A., *Corpus linguistics,* Edimbourg, Edinburgh University Press, 1996.

Messiant, C., Gábor, K. et Poibeau, T., « Acquisition de connaissances lexicales à partir de corpus : la sous-catégorisation verbale en français », *Traitement Automatique des Langues*, n° 51 vol. 1, 2010, (www.atala.org/Acquisition-de-connaissances).

Morin, E. et Daille, B., « Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé », *Traitement Automatique des Langues (TAL)*, Lavoisier, n° 45, vol. 3, 2004, p. 103-122.

Navigli, R., Litkowski, K. C. et Hargraves, O., « SemEval-2007 Task 07 : Coarse-Grained English All-Words Task », *Proceedings of the SemEval Workshop*, Prague, 2007.

Navigli, R. et Crisafulli, G., « Inducing Word Senses to Improve Web Search Result Clustering », *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2010), MIT Stata Center (Mass.), 2010, p. 116-126.

Och, F. J. et Ney, H., « Improved Statistical Alignment Models », *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics* (ACL'2000), Hong kong, 2000.

Poibeau, T., *Traitement automatique du contenu textuel,* Paris, Lavoisier, 2011.

—, (*à paraître*), « Pour une sémantique de l'usage. Le cas de l'instrumental sujet en français ».

Rumshisky, A., Hanks, P., Havasi, C. et Pustejovsky, J., « Constructing a Corpus-based Ontology using Model Bias », *Proceedings of the FLAIRS Conference*, Melbourne (FL), 2006.

Schmid H., « Probabilistic part-of-speech tagging using decision trees ». *Proceedings of the International Conference on New Methods in Language Processing*, Manchester (UK), 1994.

Spärck Jones, K., « Computational Linguistics : What about the Linguistics ? », *Computational Linguistics*, n° 33, vol. 3, 2007, p. 437-441.

Teubert, W., « Corpus Linguistics and Lexicography », *International Journal of Corpus Linguistics,* n ° 6, 2002, p. 125-154.

Thuilier, J., Fox, G. et Crabbé, B., « Approche quantitative en syntaxe : l'exemple de l'alternance de position de l'adjectif épithète en français ». *Actes de la conférence Traitement automatique des langues naturelles* (TALN 2010), université de Montréal et École polytechnique de Montréal, 2010.

van de Cruys, T., Poibeau, T. et Korhonen, A., « Latent Vector Weighting for Word Meaning in Context », *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2011), Edimbourg, 2011.

Wilks, Y., « On whose shoulders? », *Computational Linguistics*, n° 34, vol. 4, 2008, p. 471-486.

—, « Computational Semantics requires computations », *Proceedings of the 24th FLAIRS Conference*, Melbourne (FL), n° 34, vol. 4, 2011.

Wittgenstein, L., *Philosophische Untersuchungen*, trad. P. Klossowski, *Investigations philosophiques*, Paris, Gallimard, 1961.

Zweigenbaum, P., Rapp, R. et Sharoff, S. (éd.), *Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, Association for Computational Linguistics, Portland, 2011.

NOTES

1. Encore récemment, dans la bouche de Chomsky lui-même : « *[m]y judgment, if you like, is that we learn more about language by following the standard method of the sciences. The standard method of the sciences is not to accumulate huge masses of unanalyzed data and to try to draw some generalization from them* ». (Chomsky, 2004, p. 97)

2. Cf. Fillmore (1992), aussi cité par Habert (2002) : « *Armchair linguistics does not have a good name in some linguistic circles. The caricature of the armchair linguist is something like this. He sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, "Wow, what a neat fact !", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like. (There isn't anybody exactly like this, but there are some approximations.)*

Corpus linguistics does not have a good name in some linguistic circles. A caricature of the corpus linguist is something like this. He has all the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence. (There isn't anybody exactly like this, but there are some approximations.

These two don't speak to each other very often, but when they do, the corpus linguist says to the armchair linguist, "Why should I think that what you tell me is true ?", and the armchair linguist says to the corpus linguist, "Why should I think that what you tell me is interesting ?". »

3. La lexicographie est essentiellement concernée par l'étude des mots, pour constituer des dictionnaires par exemple. On peut la considérer comme une des branches de la linguistique.

4. On remarquera au passage que les approches statistiques ont quasiment toujours une assise symbolique, dans la mesure où le comptage repose souvent sur une première étape d'analyse symbolique (découpage en mots, analyse de relations entre mots).

5. Wordnet est un réseau sémantique de l'anglais disponible en ligne (http://wordnet.princeton.edu/) et très utilisé en traitement automatique des langues. Des équivalents de Wordnet ont été développés pour de nombreuses langues, avec toutefois une couverture souvent largement inférieure à celle de l'anglais.

6. *n* vaut généralement 2 ou 3, c'est-à-dire qu'un mot peut le plus souvent être catégorisé simplement en regardant les deux ou trois mots précédents dans la phrase, même si les systèmes les plus performants incluent aussi des mécanismes de décision plus élaborés – il faut en outre garder à l'esprit que même 3 % d'erreur peuvent avoir des conséquences importantes : ce taux signifie que plus d'une phrase sur deux a un mot mal catégorisé, ce qui peut ultérieurement empêcher l'analyse syntaxique de la phrase toute entière

7. Par exemple 72.5 % lors de Senseval-3 (Snyder et Palmer, 2004) ou 67.3 % pour la campagne *Open Mind, Word Expert annotation exercise* (Chklovski et Mihalcea, 2002).

8. Comme dans « *il a free il a tout compris* ». On peut remarquer que peu de personnes en définitive perçoivent le jeu de mots du premier coup, ce qui montre bien que naturellement nous ne percevons pas les ambiguïtés et toute la complexité de la langue, mais que certaines significations priment sur d'autres.

9. Il existe en fait plusieurs modèles ayant des degrés de complexité divers.

10. Les traductions ainsi obtenues sont généralement mauvaises, notamment sur le plan syntaxique mais les systèmes développés manuellement ne donnent guère de meilleurs résultats, même sur ce plan où ils devraient en théorie être plus performants ; une syntaxe même approximative n'est souvent pas un frein à une compréhension globalement correcte d'un énoncé.

11. C'est-à-dire que peu de règles couvrent beaucoup de données mais beaucoup de règles sont nécessaires pour couvrir toutes les données ; les statistiques permettent de repérer les règles les plus utiles pour l'analyse mais ne permettent pas toujours de « couvrir » les cas rares, même si les méthodes progressent vers le repérage des phénomènes les moins fréquents.

12. *I was not sure that [...] IBM's « fundamental equation of MT » was in fact producing the results, and suggested that something else they were doing was giving them their remarkable success rate of about 50 % of sentences correctly translated. As their general methodology has penetrated the whole of NLP/CL, I no longer stand by my early criticisms ; IBM were of course right, and had everything to teach the rest of us.* (Wilks, 2008)

13. *One thing the empirical movement has taught us is the vital importance of scale and the need to move away from toy systems and illustrative examples.* (Wilks, 2008)

14. http://www.semantique-gdr.net/dico/index.php/Donkey_sentence

15. *There ARE no donkey sentence, aside of course from the corpus of linguistic discussions. They do not exist, any more than anyone ever says, « John wants to marry a Norwegian » to mean a particular Norwegian--- you simply do not put it that way ; hence the class of sentences are not ambiguous in the way the theory requires.*

16. *This is not an issue of access to limited data, but of a whole tradition of making up and discussing examples that do not correspond to anything anyone would or could say in real life. It is as glaring an artificial construction of a subject as Russell's attributed observation that no one had ever in fact proved anything with a syllogism.*

17. *So we need to be alert. It's not just that we may find ourselves putting the cart before the horse. We can get obsessed with the wheels, and finish up with uncritically reinvented, or square, or over-refined or otherwise unsatisfactory wheels, or even just unicycles. What matters is the way the cart, its load, and the horse, together make a rational journey [...], it's about how you integrate the language and the computation on the one hand, the symbolic and the statistical on the other and to succeed in this we should not forget that mainstream linguistics may have some things to offer us, even if not as many as linguists themselves may suppose.*

18. *[...] Corpus linguistics – or « usage-based models of grammar » – has all the right rhetoric about being an objective, falsifiable empirical science interested in the totality of language use, but is failing by largely restricting itself to surface facts of language, rather than utilizing sophisticated formal models of grammar, which make extensive use of hidden structure (things like phrase structure trees, and other abstract representational levels).*

19. On peut d'ailleurs s'interroger sur l'intérêt d'ajouter l'adjectif *formel* à la dénomination du domaine. La formalisation peut-elle être une fin en soi ?

RÉSUMÉS

Cet article porte sur la façon dont l'arrivée massive de données textuelles sur support numérique a récemment changé la façon de faire des recherches en linguistique. Plusieurs branches de la linguistique travaillent à partir de grands ensembles de données attestées : nous examinerons essentiellement le cas de la linguistique de corpus et celui du traitement automatique des langues. Nous essaierons de mettre en avant les ruptures, les avancées mais aussi les limites des approches reposant sur des données massives. Notre regard sera donc avant tout épistémologique, davantage que technique ou historique.

This paper examines to what extent the massive availability of textual data in digital form has recently changed the way people carry out research in linguistics. Several subfields of the domain require large amounts of attested data : here, we primarily consider the case of corpus linguistics and natural language processing. We consider recent breakthroughs but also the main limitations of current approaches based on massive data. We will adopt an epistemological point of view, rather than a technical or historical one.

INDEX

Mots-clés : linguistique, statistiques, sémantique, traduction automatique
Keywords : linguistics, statistics, semantics, machine translation

AUTEUR

THIERRY POIBEAU

Directeur de recherche au CNRS et directeur du laboratoire LATTICE (Langues, Textes, Traitements informatiques et Cognition) CNRS-ENS-Université Paris 3.
Parmi les publications :
Avec B. Devereux, N. Pilkington et A. Korhonen, « Towards unrestricted, large-scale acquisition of feature-based conceptual representations from corpus data », *Research on Language and Computation*, 7(2-4), 2010, p. 137-170.
Traitement automatique du contenu textuel. Paris, Lavoisier, 2011.
Avec H. Saggion, J. Piskorski et R. Yangarber, *Multi-source, Multilingual Information Extraction and Summarization*, Berlin et Heidelberg, Springer-Verlag, « Theory and Applications of Natural Language Processing », XXIV, 2012.
Avec G. Col, J. Aptekman et S. Girault, « Gestalt Compositionality and Instruction-Based Meaning Construction », *Cognitive Processing*, vol. 13, n° 2, 2012, p. 151-170.
Avec A. Villavicencio, A. Korhonen et A. Alishahi, « Computational Modeling as a Methodology for Studying Human Language Learning », in *Cognitive Aspects of Computational Language Acquisition* , Springer « Theory and Applications of Natural Language Processing », 2013.

Avec E. Omodei et J.-Ph. Cointet, « The Socio-Epistemic Dynamics of Scientific Research », *Proc. European Conf. on Complex Systems*, Barcelone, 2013.