Linguistiques de corpus et mathématiques du continu

Stéphanie Girault, Bernard Victorri

Histoire Epistémologie Langage, 31/1, pp. 147-170, 2009. – Version préliminaire

Linguistiques de corpus et mathématiques du continu

Stéphanie Girault & Bernard Victorri
Laboratoire Lattice (ENS-UMR 8094 du CNRS)

Introduction

Les nouvelles technologies ont transformé radicalement les rapports des linguistes avec leur objet d'étude. On peut disposer aujourd'hui très facilement d'une impressionnante masse de données sur la langue, qui couvre la plupart de ses usages (du moins dans le domaine de l'écrit), sans aucune commune mesure avec ce qui était accessible il y a à peine dix ans. De plus on commence à disposer de ressources linguistiques (comme les dictionnaires électroniques) de plus en plus exhaustives, et d'outils de traitement (analyseurs syntaxiques, etc.) de plus en plus fiables. Cet ensemble représente en fait un nouvel « instrument » qui permet de « voir » les phénomènes langagiers comme on n'avait jamais pu le faire auparavant, un peu comme la lunette de Galilée a permis de voir des phénomènes astronomiques (les ombre sur la Lune, les satellites de Jupiter, etc.) inaccessibles jusqu'alors. Il est donc assez normal que ce nouveau dispositif d'observation transforme assez radicalement les méthodes et les attendus de la modélisation informatique en linguistique et de ses applications en traitement automatique des langues. Comme le fait observer Benoît Habert (2005b) : « à côté d'une linguistique sans instrument s'impose clairement une linguistique « à l'instrument » ». C'est un tel instrument que nous allons présenter ici. Fondé sur les mathématiques du continu, ANASEM est un logiciel d'analyse sémantique sur corpus linguistique. Après avoir montré l'intérêt d'utiliser les mathématiques du continu et présenté les différents types de modélisations possibles dans ce cadre, nous présenterons une application concrète à la modélisation du sens d'un marqueur discursif : la conjonction de subordination *quand*.

1. Une nouvelle donne pour la mathématisation de la linguistique

1.1. De nouveaux instruments – un nouvel observatoire

L'accessibilité à de gros corpus et surtout les instruments et outils[1] permettant de les traiter sont en passe de conférer à la linguistique le statut d'une science à part entière. Encore assez récemment, Chomsky assurait : « Corpus linguistics does not exist. » (Entretien avec Baas Aarts, 1999 cité par Rastier 2002). D'aucuns, spécialistes de l'épistémologie, ont apporté des arguments rationnels à cette position : la linguistique ne peut accéder au statut de science tant qu'elle ne répond pas à ces deux critères essentiels dans ses méthodes :« 1) [être] en relation directe avec une hypothèse explicite à tester ; 2) [correspondre] à la production d'un phénomène » (Auroux, 1998, p.183). Milner, en introduisant la notion d'observatoire (Milner 1989) a eu une position moins tranchée : la linguistique, selon lui, est bien une science expérimentale, mais elle dépourvue d'observatoire. Développons quelque peu cette idée. Pour

[1] La ligne de partage que Simondon établit entre outil et instrument dans le cours *L'invention et le développement des techniques* (1968–1969) est la suivante : « [l'instrument] est l'inverse de l'outil, car il prolonge et adapte les organes des sens : il est un capteur, non un élément effecteur. L'instrument équipe le système sensoriel, il sert à prélever de l'information, tandis que l'outil sert à effectuer une action » [Simondon, 2001, p. 88–89].

certaines disciplines comme l'astronomie ou la linguistique, le chercheur ne dispose pas à proprement parler d'une paillasse avec des observables manipulables directement. Si l'on construit bien, comme dans toutes les sciences, des modèles théoriques que l'on élabore à partir d'observations, il n'est en revanche pas possible de modifier ces observations pour mettre le modèle à l'épreuve : les conditions initiales de l'expérimentation sont des données *a priori* qui ne supportent pas la manipulation. Par exemple, l'astronome ne peut pas changer la lune de place pour évaluer les incidences de cette nouvelle configuration du système observé. Ainsi, dans certaines disciplines, et la linguistique en fait partie, le chercheur est amené à essayer de faire des prédictions sur ce qui se passe et non sur ce qu'il peut lui-même monter comme expérience. Parfois, le chercheur est amené à faire des inductions à partir d'observations incomplètes, parce qu'il est justement très difficile de dire à quel moment toutes les conditions de l'observation sont réunies. Un exemple célèbre est la découverte de Neptune. C'est en observant Uranus que les astronomes, en constatant que son orbite ne se conformait pas aux lois de Newton firent l'hypothèse de l'existence d'une autre planète exerçant son influence sur la trajectoire d'Uranus, Neptune. La première fois que l'on observa Neptune au télescope en 1846, ce fut au point exact prédit par les calculs scientifiques quelques temps plus tôt. Ces calculs étaient fondés sur l'observable de l'époque, à partir des positions relatives de Jupiter, Saturne et Uranus. Or des observations ultérieures ont révélé que les trajectoires calculées alors divergeaient assez rapidement de l'orbite actuelle de Neptune, et que si les recherches avaient été menées quelques années avant ou après, elles n'auraient pas pu prédire la présence de Neptune. Cette anecdote révèle au moins deux points. D'abord, qu'il est possible de faire des prédictions scientifiques et de les vérifier même quand on ne peut pas construire une expérience que l'on maîtrise de bout en bout. Et ensuite, que la validation d'une théorie dans ce cadre est extrêmement délicate.

Ceci tend à plaider en faveur de la notion d'*observatoire* décrite par Milner :

> « Sa première fonction est de permettre de construire une expérience, laquelle permet de tester une théorie, c'est-à-dire de choisir entre deux propositions contradictoires (…). [Cette fonction] donne ainsi accès à l'instance de réfutation. On peut résumer cette fonction d'un nom simple : l'outillage de l'expérimentation construit l'instance de l'observatoire. Pour que cela soit possible, il convient que (…) les propositions de la théorie qui fondent l'expérimentation soient indépendantes de la proposition qu'il s'agit de tester » (op. cit., p. 127). Après avoir donné l'exemple de la lunette astronomique, il précise : « Il y a là un risque de circularité : si la science physique est un tout, comment veut-on établir la moindre proposition physique en se fondant sur une expérimentation qui elle même dépend pour partie de fragments de la théorie physique. La résolution de ce cercle est cependant possible : il faut et il suffit que l'observatoire en question soit localement indépendant. Autrement dit, les propositions dont il dépend peuvent relever de la science, elles peuvent pour autant ne pas dépendre de la proposition, entendue au sens étroit, que l'expérimentateur en question vise à tester. Ainsi, il est vrai que des propositions de la théorie astronomique dépendent de la lunette, mais l'optique dont dépend la lunette est localement indépendante de l'astronomie. » (ibidem).

Après avoir constaté que la linguistique, du moins celle de l'objet théorique « langue » tel qu'il le conçoit, est une science sans outillage donc sans observatoire, Milner conclut :

> « En bref, en linguistique, il y a des expérimentations, mais il n'y a pas d'observatoire- ou, ce qui revient au même, ce qui passe pour observatoire inclut toujours un fragment de théorie linguistique qui ne peut être rendu totalement indépendant de la donnée soumise à l'expérimentation. » (ibidem, p. 128).

S'il est vrai que la pratique des jugements d'acceptabilité sur des exemples construits que l'on peut manipuler à sa guise tombe pleinement sous le coup de la critique de Milner, on peut se demander, avec Sylvain Auroux[2], s'il n'existe pas d'autres observatoires dont pourrait

[2] Cf. Auroux (1998 : 215) : "Nous sommes parfaitement capables d'identifier de multiples observatoires de langue : l'écriture, les textes, les autres langues, les corpus d'exemples, les dictionnaires, etc. La plupart des observatoires sont des construits théoriques; la linguistique ne diffère en rien des sciences physiques sur ce point. L'origine des théories

bénéficier la linguistique. Notamment, le développement des ressources électroniques nous donne accès à des quantités de données qui changent radicalement l'échelle de ce que l'on peut observer des usages de la langue. Comme se développent aussi des outils de traitement automatique à large couverture, bien adaptés au traitement de ces données, on peut penser qu'émerge un nouveau dispositif, avec ses nouveaux *instruments*[3], qui mérite pleinement le nom d'*observatoire* de la langue au sens de Milner. En effet, la condition d'indépendance locale réclamée par Milner semble de mieux en mieux satisfaite : pour ne donner qu'un exemple, l'observation du comportement syntaxico-sémantique d'un marqueur grammatical comme *quand*, telle que nous la mettons en pratique (cf. 2ème partie de cet article), est très largement indépendante des choix théoriques qui ont présidé à la conception du programme informatique d'analyse syntaxique des énoncés que nous étudions.

1.2. Que faire des données quantitatives ?

D'une manière générale, l'utilisation de ce nouvel « observatoire de la langue » implique inévitablement de faire appel à des méthodes quantitatives, puisque son efficacité réside dans la taille des données accessibles.

Les méthodes statistiques descriptives (mesures de fréquence, calculs de cooccurrence, etc.) sont les outils de base qui permettent de « mettre de l'ordre » dans les observations, en classant les phénomènes et en mettant en évidence des corrélations entre eux. Ces méthodes sont utilisées depuis bien longtemps pour les formes de surface, mais, avec la mise au point ces dernières années d'analyseurs syntaxiques robustes, comme par exemple, pour le français, l'analyseur SYNTEX (Bourigault, 2007), ce sont les relations structurelles de la langue, aussi bien syntagmatiques que paradigmatiques, qui peuvent ainsi être observées. L'observation des relations syntagmatiques est pour ainsi dire directe, puisqu'elle repose fondamentalement sur les mesures de cooccurrence d'unités linguistiques en relation de dépendance syntaxique dans un corpus donné. Les relations paradigmatiques, en revanche, ne sont pas directement observables dans un corpus, du moins dans toute leur généralité[4]. Il faut utiliser les techniques de l'analyse distributionnelle pour les retrouver, en appliquant le principe selon lequel sont en relation paradigmatique des unités linguistiques qui partagent les mêmes contextes syntagmatiques dans un corpus.

A ces méthodes statistiques sont tout naturellement associées des représentations graphiques (analyse en composantes principales, analyse factorielle de correspondance, etc.) qui permettent de visualiser des regroupements en classes ou d'autres formes d'organisation des données observées. Souvent une distance (comme la métrique du χ^2 : cf. par exemple Ploux et Victorri 1998) peut être définie pour conférer aux représentations graphiques obtenues un véritable statut d'espace géométrique[5].

---

linguistiques ne se confond pas avec l'origine du langage; là où il y a langage, il n'y a pas nécessairement théorie linguistique".

[3]Nous suivons ici la définition proposée par Benoît Habert : "Par instrument (en anglais *tool*), on entendra un logiciel qui prend en entrée une donnée langagière (du texte, de l'oral, un lexique...) et qui permet d'obtenir en sortie une représentation transformée (annotée), soit automatiquement soit semi-automatiquement soit manuellement." (Habert 2005a)

[4] Certaines méthodes, comme l'extraction d'énumérations ou de syntagmes définitoires, permettent de déceler directement certaines relations paradigmatiques (cf. par exemple Hamon et Nazarenko, 2001), mais ces procédés restent très limités.

[5] La théorie des graphes a aussi été sollicitée ces dernières années : il se trouve que la plupart des graphes de relations syntagmatiques et paradigmatiques d'une langue ont une structure de « petit monde » (*small world graphs* cf. Watts &

Cependant, toutes ces techniques, quel que soit leur degré de sophistication, restent fondamentalement descriptives. Elles ne constituent pas un « modèle » à elles seules. Elles ne font que révéler une organisation des données dont il faut pouvoir prouver par ailleurs le bien-fondé. Il faut insister sur ce point, notamment pour les représentations graphiques : l'expérience montre qu'on leur attribue souvent un pouvoir explicatif un peu magique (le sens semble « émerger » de la figure), alors que seules nos capacités illimitées d'interprétation sont en cause (le sens est dans nos têtes).

En fait, même si la structuration des observations obtenue par ces méthodes statistiques nous semble cohérente et « parlante », rien ne prouve que ce soit la meilleure possible, ni même qu'elle ne masque un aspect essentiel du phénomène étudié, qu'une autre disposition des mêmes données aurait mis en lumière. Du point de vue du traitement automatique des langues, la réponse à ces questions est d'ordre pragmatique : on évalue les performances relatives de systèmes qui intègrent les diverses représentations possibles, et l'on juge que la meilleure d'entre elles est celle qui produit les meilleurs résultats. Mais du point de vue de la théorie linguistique, qui seule nous intéresse ici, le problème se pose autrement, puisqu'il s'agit de savoir si ces observations peuvent servir de « preuves » expérimentales validant ou invalidant des hypothèses théoriques.

Si l'on s'en tient aux théories linguistiques classiques, fondées pour l'essentiel sur des modèles discrets, on doit reconnaître que l'utilisation de données quantitatives est peu appropriée, comme le montre de manière assez approfondie Manning (2003). L'un de ses arguments, cité par Habert et Zweigenbaum (2002), concerne la notion d'acceptabilité. Prenant des exemples de sous-catégorisation verbale, il montre que l'on trouve dans les corpus de manière peu fréquente mais régulière des constructions verbales jugées non acceptables par une grammaire de référence (Pollard & Sag 1994), et que, en revanche, il existe des différences énormes de fréquence parmi les constructions jugées également acceptables par la même grammaire. D'où la conclusion que les modèles discrets sont à la fois :
(1) trop contraignants pour couvrir la totalité des données observées (« *Categorical linguistic theories claim too much. They place a hard categorical boundary of grammaticality where really there is a fuzzy edge, determined by many conflicting constraints and issues of conventionality vs. human creativity.* »)
(2) trop pauvres pour rendre compte de leur complexité (« *Categorical linguistic theories explain too little. They say nothing at all about the soft constraints which explain how people choose to say things (or how they choose to understand them).* »).

Un autre argument de Manning concerne la diachronie. Analysant un phénomène de grammaticalisation en anglais contemporain, en l'occurrence l'acquisition progressive du statut de préposition par des participes présents tels que *following*, il montre que l'on trouve dans les corpus un grand nombre d'occurrences pour lesquels les deux analyses (faisant de *following* un gérondif ou une préposition) sont tout aussi valides. Un modèle discret force à choisir, et donc empêche de comprendre la spécificité de ces cas qui sont, en fait, au cœur de toute explication du phénomène de grammaticalisation. Il propose plutôt un modèle continu de l'espace des catégories linguistiques, dans lequel les catégories traditionnelles correspondraient à des points d'accumulation (« *It seems that it would be useful to explore modeling words as moving in a continuous space of syntactic category, with dense groupings corresponding to traditional parts of speech* »).

Strogatz 1998), ce qui a conduit à utiliser les techniques de représentation de ces structures mathématiques : cf. par exemple Ferrer & Solé 2001, Gaume *et al.*2006.

Enfin un dernier argument de Manning, concernant la typologie des langues, mérite qu'on le reprenne ici plus longuement, notamment parce qu'il lui sert à introduire les modèles probabilistes que nous présenterons dans la section suivante. Le point de départ est une réflexion de Givón (1979) critiquant l'opposition compétence/performance :

> In many of the world's languages, probably in most, the subject of declarative clauses cannot be referential-indefinite. In other words, the subject position in the sentence is one in which *new information* cannot be introduced. In order to violate this *categorical* constraint, the speaker must resort to a special, *marked* sentence type, the *existential-presentative* construction. Languages of this type are, for example, Swahili, Bemba, Rwanda (Bantu), Chinese, Sherpa (Sino-Tibetan), Bikol (Austronesian), Ute (Uto-Aztecan), Krio (Creole), all Creoles, and many others.
> In a relatively small number of the world's languages, most of them with a long tradition of literacy, referential-indefinite nouns may appear as subjects of nonpresentative sentences […]. When one investigates the text frequency of [such] sentences in English, however, one finds them at an extremely low frequency: About 10% of the subjects of main-declarative-affirmative-active sentences (nonpresentative) are indefinite, as against 90% definite. Now this is presumably not a fact about the "competence" of English speakers, but only about their actual "language behavior." But are we dealing with two different kinds of facts in English and Krio? Hardly. What we are dealing with is apparently the very same *communicative tendency* – to reserve the subject position in the sentence for the *topic*, the old-information argument, the "continuity marker." In some languages, (Krio, etc.) this communicative tendency is expressed at the *categorial* level of 100%. In other languages (English, etc.) the very same communicative tendency is expressed "only" at the *noncategorial* level of 90%. And a transformational-generative linguist will then be forced to count this fact as competence in Krio and performance in English. But what is the communicative difference between a rule of 90% fidelity and one of 100% fidelity? In psychological terms, next to nothing. In communication, a system with 90% categorial fidelity is a highly efficient system (pp.26-28).[6]

De fait, même si elle n'adopte pas la distinction chomskyenne entre compétence et performance, une théorie s'appuyant sur un modèle discret ne pourra pas traiter dans un même cadre les faits en krio et en anglais. En revanche un modèle continu pourra rendre compte à la fois des différences entre les deux langues (90% est différent de 100%) et de l'universalité du phénomène, qu'il soit régi par une contrainte forte dans certaines langues ou par une contrainte plus faible dans les autres.

Il faut donc changer de cadre de modélisation, si l'on veut utiliser pleinement le nouvel observatoire que nous offrent les nouvelles technologies : les linguistiques de corpus se doivent d'être des théories s'appuyant sur les mathématiques du continu.

On peut distinguer deux grandes classes de modèles continus, suivant le type d'outils qu'ils utilisent, qui peuvent provenir soit de la théorie des probabilités, soit de la théorie des systèmes dynamiques. Nous allons présenter succinctement ces deux types de modèles, que nous appellerons de manière abrégée respectivement modèles probabilistes et modèles dynamiques.

---

[6] En écho à cette réflexion, on peut rappeler une discussion entre Culioli et Milner sur l'acceptabilité de l'énoncé *Un chien aboie* qui conduit ce dernier à écrire (Milner 1992, p.22) : *« La théorie de Culioli a défini un phénomène discriminant propre. […] A. Culioli, un jour, formula l'intuition linguistique : « 'un chien aboie' est mal formé ». […] « Soyons clair : il est indubitable que la découverte était capitale et d'une nature propre à susciter la réflexion. Premièrement, il s'agissait d'une donnée empirique, confirmée par des observations multiples et croisées ; secondement, la donnée concernait un secteur central de la langue française : les conditions d'apparition de l'article indéfini dit « spécifique » (par opposition au « générique ») ; troisièmement, elle n'avait jamais été notée : cela seul soulevait une question grave, touchant à la nature des instruments de l'observation linguistique ; s'ils étaient impropres à enregistrer une donnée aussi massive, cela ne signifiait-il pas qu'ils étaient mal réglés ? plus exactement, cela ne signifiait-il pas que les principes reglant leur système de discrimination étaient mal conçus , »*

1.3. Les modèles probabilistes

Dans un article au titre suggestif, *Régler les règles*, Habert et Zweigenbaum (2002) montrent l'actualité du programme qu'avait dessiné Harris dans ses derniers travaux (Harris, 1988 ; Harris *et al.*, 1989 ; Harris, 1991), et, en le confrontant aux développements actuels des linguistiques de corpus, ils concluent (p. 101) par cette citation de Pereira (2000, p. 1250) :

> « . . . il est bien possible que nous assistions à l'émergence d'une nouvelle version du programme harrissien, dans lequel des modèles computationnels contraints par des considérations grammaticales définissent des grandes classes [*broad classes*] de grammaires possibles, et des principes empruntés à la théorie de l'information spécifient comment ces modèles s'ajustent aux données linguistiques attestées. » .

Pour progresser dans cette voie, ils appellent de leurs vœux des approches de modélisation « permettant d'incorporer des traits linguistiques arbitraires dans un cadre probabiliste » (p. 100), ajoutant en note : « Ce serait prolonger la recherche de satisfaction simultanée de contraintes de différente nature qui a grandi dans les grammaires d'unification, particulièrement en HPSG, mais en y ajoutant des propensions, des poids distincts ».

C'est en effet une bonne façon de caractériser les modèles probabilistes : ils s'appuient généralement sur un modèle discret, qu'ils rendent continus en associant aux règles de ce modèle des variables aléatoires dont on peut ajuster la distribution en fonction des données observées.

Manning (2002) présente deux exemples simples de ces modèles, en les appliquant à la modélisation d'un même phénomène : le choix du groupe sujet dans la proposition assertive dans différentes langues. Il retient, pour simplifier, trois facteurs qui peuvent influencer ce choix : le premier, que nous avons déjà évoqué dans la section précédente avec Givón est d'ordre discursif : le sujet doit être un élément topical. Comme nous l'avons vu, il s'agit d'une règle grammaticale dans certaines langues et d'une simple préférence dans d'autres. Le deuxième facteur est d'ordre argumental : le sujet doit être l'agent de l'action évoquée par le verbe. Là encore, suivant les langues, la contrainte peut être plus ou moins forte (elle est impérative pour les langues qui ne disposent pas de construction passive). Enfin le dernier facteur porte sur la marque de personne : le choix préférentiel de la première et deuxième personne pour la fonction sujet. Manning cite le lummi (langue amérindienne de Colombie Britannique) parmi les langues qui imposent une contrainte forte à ce sujet : il est agrammatical d'avoir pour complément un pronom personnel de première ou deuxième personne si le sujet n'est pas lui-même un de ces deux pronoms.

Le premier modèle proposé par Manning est une version stochastique de la théorie de l'optimalité, développée par Boersma (Boersma et Hayes 2001). La théorie de l'optimalité (Prince & Smolensky 1993) repose à la base sur un ordonnancement des contraintes : la contrainte la plus élevée ne doit pas être violée, mais, quand elle est satisfaite, chacune des contraintes de rang inférieur peut éventuellement s'exprimer à son tour. En changeant l'ordre des contraintes, on peut modéliser des langues différentes, et, notamment, prendre en compte les diverses agrammaticalités que nous avons évoquées. Mais la théorie de l'optimalité reste une théorie discrète, avec un inconvénient majeur pour le type de phénomène qui nous intéresse ici : il ne peut y avoir dans chaque cas de figure qu'un résultat possible. La variabilité que l'on observe dans les corpus ne peut pas être modélisée. Pour circonvenir ce problème, il suffit de passer à un modèle continu, en associant à chaque contrainte un poids aléatoire distribué sur la droite réelle. Si l'on ordonne les contraintes suivant leur poids, le rang de certaines contraintes peut varier, pour peu que les distributions des poids correspondants se recouvrent partiellement (fig. 1). On peut alors ajuster le modèle aux données de telle ou telle langue, et même de tel ou tel corpus, en utilisant une méthode

d'apprentissage : le meilleur jeu de paramètres (moyenne et écart-type de chaque distribution) sera celui qui se rapproche le plus des fréquences relatives observées dans chaque cas de figure.

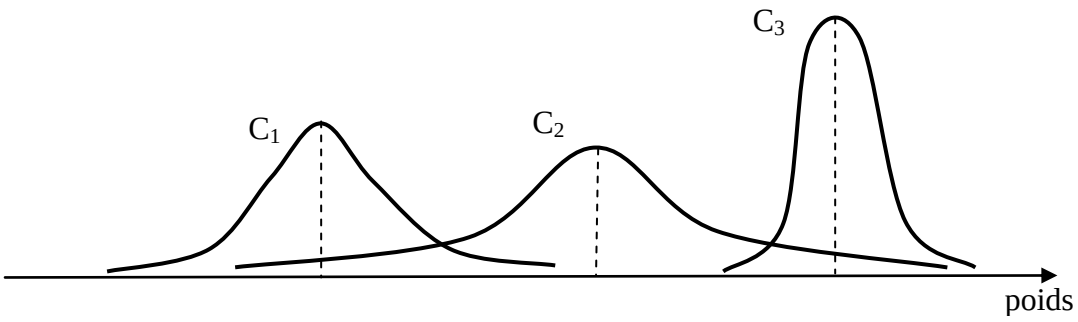


Figure 1 : Distributions probabilistes associées à des contraintes (C_1, C_2 et C_3)

Le deuxième modèle que propose Manning est plus proche de ce que prônent Habert et Zweigenbaum dans la note que nous avons citée ci-dessus. Il s'agit en effet d'un modèle continu de satisfaction de contraintes, c'est-à-dire en quelque sorte l'équivalent continu des systèmes discrets basés sur l'unification. On associe toujours à chaque contrainte un poids aléatoire, caractérisé par une distribution donnée sur les réels. Mais cette fois, la probabilité de la construction choisie d'une fonction globale sur l'ensemble des poids, et non plus simplement de leur rang. La fonction globale le plus souvent choisie est une fonction log-linéaire (le logarithme de la probabilité résultante dépend linéairement du logarithme de la probabilité d'application de chaque contrainte), fonction très utilisée, justement, en théorie de l'information (maximisation de l'entropie). Un mécanisme d'apprentissage permet, comme dans le cas de l'optimalité stochastique, d'ajuster les distributions des poids aux données d'un corpus.

Manning montre que les deux modèles ne sont pas équivalents : notamment l'optimalité stochastique ne permet pas de rendre pleinement compte d'un phénomène assez répandu, « l'effet de gang » (*ganging up*), c'est-à-dire la possibilité que plusieurs contraintes de niveau inférieur contribuent par leur nombre à contrebalancer une contrainte hiérarchiquement plus élevée. Ainsi, tous les modèles continus ne se valent pas, et, quelle que soit la méthode d'apprentissage, un modèle probabiliste ne pourra pas rendre compte des données s'il ne modélise pas correctement le phénomène étudié. Il s'agit donc là d'un champ de recherche tout à fait passionnant, comme le conclut Manning (p. 334) :

> There are many phenomena in syntax that cry out for non-categorical and probabilistic modeling and explanation. The opportunity to leave behind ill-fitting categorical assumptions, and to better model probabilities of use in syntax is exciting. The existence of 'soft' constraints within the variable output of an individual speaker, of exactly the same kind as the typological syntactic constraints found across languages, makes exploration of probabilistic grammar models compelling. We saw that one is not limited to simple surface representations: I have tried to outline how probabilistic models can be applied on top of one's favourite sophisticated linguistic representations. The frequency evidence needed for parameter estimation in probabilistic models requires a lot more data collection, and a lot more careful evaluation and model building than traditional syntax, where one example can be the basis of a new theory, but the results can enrich linguistic theory by revealing the soft constraints at work in language use. This is an area ripe for exploration by the next generation of syntacticians.

1.4. Les modèles dynamiques

Les modèles fondés sur la théorie des systèmes dynamiques s'écartent plus radicalement des modèles discrets que les modèles probabilistes. En effet, il ne s'agit plus d'appliquer des règles ou des contraintes. Tout cela est remplacé par un principe général d'interaction entre éléments de l'énoncé et de rétroaction de l'énoncé global sur ses composants. Tous les éléments sont concernés par ce processus : unités lexicales, marqueurs grammaticaux, et aussi constructions syntaxiques. Chacun contribue à sa façon à construire un sens global de l'énoncé, et en retour, ce sens global détermine le sens précis que l'élément en question prend dans cet énoncé. Mathématiquement, on peut modéliser ce processus de construction comme l'action d'une dynamique dans un espace continu multidimensionnel. Chaque dimension de l'espace correspond à une variable continue que l'on veut étudier. Un état du système est constitué par la donnée d'une valeur pour chacune des variables, donc par un point de cet espace. Se donner une dynamique sur l'espace consiste à associer à chaque point un vecteur vitesse, qui indique dans quelle direction le système évolue quand il passe par l'état correspondant. Le système décrit donc une trajectoire dans l'espace d'état, et il se stabilise en un point d'équilibre, que l'on appelle un attracteur de la dynamique. On peut, si l'on veut, interpréter ce vecteur vitesse en un point donné comme l'expression de l'ensemble des contraintes qui s'exercent quand on se trouve dans l'état en question. De cette manière, on peut dire qu'un système dynamique est un système de satisfaction de contraintes, mais avec la particularité que les contraintes ne sont pas les mêmes en tout point : elles dépendent de la valeur des variables sur lesquelles elles s'exercent.

La plupart des modèles dynamiques de phénomènes langagiers ont été conçus dans les années 90 par des tenants du connexionnisme en linguistique, ce qui fait que l'on confond souvent modèles dynamiques et modèles connexionnistes. En fait, on peut construire un système dynamique sans passer par une implémentation informatique par un réseau connexionniste, et, par ailleurs, seuls certains modèles connexionnistes, à savoir ceux qui mettent en œuvre un réseau connexionniste récurrent, sont réellement des modèles dynamiques. De nombreuses architectures de réseaux récurrents ont été proposées pour traiter divers types de phénomènes en sciences du langage (cf. entre autres St John & McClelland 1990, Miikkulainen & Dyer, 1991, Hinton & Shallice 1991, Elman 1991, Victorri & Fuchs 1996, Tabor et Tanehaus 1999). Bien que ces modèles aient suscités un intérêt certain, cette voie de recherche a été progressivement abandonnée. On peut invoquer au moins deux causes de cette désaffection. La première est technique : les réseaux récurrents sont assez difficiles à mettre en œuvre, notamment en ce qui concerne les méthodes d'apprentissage, contrairement aux réseaux plus classiques (unidirectionnels), qui continuent, eux, à être largement utilisés pour des tâches de classification. La deuxième est plus conceptuelle : chacun des modèles proposés a été construit pour traiter un phénomène particulier, ce qui donne le sentiment d'avoir à faire avec une multitude de systèmes dédiés différents, dont on ne voit pas très bien comment ils pourraient fonctionner ensemble pour nous donner une intelligibilité du fonctionnement général de la langue.

Mais comme nous l'avons dit, on peut utiliser un modèle dynamique autrement que par l'intermédiaire d'un réseau connexionniste récurrent. On peut notamment déterminer les attracteurs d'une dynamique sur un espace donné sans pour autant simuler complètement la dynamique elle-même, à condition de disposer de suffisamment d'exemples. Avec le développement de gros corpus et d'outils d'analyse capables de les traiter, ce type de méthodes devient de plus en plus efficace.

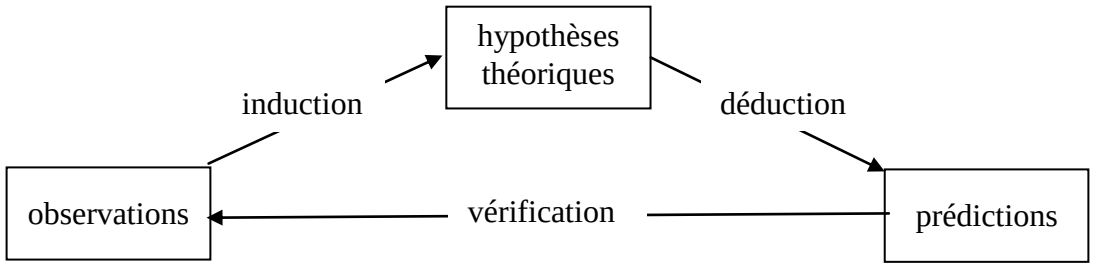
C'est ainsi que nous avons développé ces dernières années dans notre laboratoire (Jacquet *et al.* 2005) une méthode de désambiguïsation d'unités lexicales basée sur un modèle dynamique de la polysémie. Nous avons pour cela utilisé des calculs des données distributionnelles sur un gros corpus analysé par le logiciel SYNTEX (Bourigault 2007). Nous procédons en deux étapes. On construit d'abord un espace sémantique de l'unité étudiée, à l'aide des cliques du graphe de synonymie de cette unité (Ploux & Victorri 1998). On mesure ensuite, pour chaque élément du co-texte susceptible d'influer sur le sens de l'unité étudiée, la fréquence de ses cooccurrences avec chacun des synonymes de l'unité. On calcule alors un degré d'affinité de l'élément co-textuel avec toutes les cliques du graphe de synonymie, ce qui permet de situer l'influence de cet élément dans l'espace sémantique de l'unité. Du point de vue du système dynamique, cela veut dire que l'on est capable de déterminer la position et la forme des bassins d'attraction de la dynamique associée à l'élément cotextuel considéré. La méthode est limitée, notamment parce qu'elle ne peut pas prendre en compte tous les éléments cotextuels à la fois, et nous ne l'avons appliquée pour l'instant qu'à l'étude de quelques adjectifs et verbes polysémiques, mais les premiers résultats que nous avons obtenus sont encourageants.

Nous travaillons actuellement à la mise au point d'une méthode analogue pour l'étude des marqueurs grammaticaux. Deux raisons essentielles empêchent d'étendre la méthode que nous venons de présenter à ces marqueurs. D'abord il y a une multiplicité d'éléments cotextuels susceptibles d'influer sur le sens d'un marqueur grammatical, et, sauf exception, aucun d'entre eux ne joue à lui seul un rôle prépondérant. D'autre part il est difficile d'utiliser la notion de synonymie pour ces marqueurs, la plupart des paraphrases acceptables pour un tel marqueur impliquant une reformulation globale de l'énoncé plutôt que la simple substitution par un synonyme. Nous avons donc opté pour une autre approche, dont le point de départ remonte aux premières études que nous avions menées sur ce modèle dynamique de la polysémie. En étudiant le comportement d'une unité grammaticale, en l'occurrence l'adverbe *encore*, nous avions remarqué que l'étude de ses emplois sur un corpus faisait apparaître des points privilégiés de son espace sémantique, autour desquels s'accumulaient ces emplois (Victorri et Fuchs, 1996, chap. 5). Nous avons appelé *valeurs typiques* de l'unité étudiée ces points privilégiés, et nous avons tenté de donner une description systématique des éléments cotextuels qui conduisaient à ces valeurs. Mais l'absence de gros corpus et de moyens de les analyser ne nous avait pas permis à l'époque d'exploiter complètement cette propriété par des techniques informatiques. Nous avons donc repris cette idée en utilisant les ressources dont nous disposons aujourd'hui. Cela nous a conduit à mettre au point un logiciel d'étude de marqueurs grammaticaux, ANASEM, que nous allons présenter dans la deuxième partie de cet article, car il nous semble bien illustrer le type de travaux linguistiques sur corpus que l'on peut mener dans le cadre d'un modèle continu (on trouvera dans Tabor & Tanehaus 1999 une approche assez voisine, mais qui utilise, elle, des réseaux connexionnistes, avec les inconvénients dont nous avons parlé).

2. Un exemple d'instrumentation : le logiciel ANASEM

2.1. Motivation

Notre objectif est de permettre la modélisation d'un phénomène sémantique à partir de données extraites d'un corpus. Il s'agit donc d'aider à passer des données quantitatives « brutes » à un modèle, basé sur des mathématiques du continu. A priori, nous n'avons pas de prétention à la validité statistique. Il s'agit plutôt de révéler des interactions que de les évaluer quantitativement de manière précise. Nous nous situons dans la branche « induction » de la boucle ci-dessous :



Comme nous l'avons dit, notre approche consiste à nous placer dans le cadre d'un modèle dynamique, mais sans essayer de calculer explicitement la dynamique elle-même. Nous cherchons à dégager des caractéristiques qualitatives de cette dynamique en faisant l'hypothèse qu'il existe un certain nombre de valeurs sémantiques typiques pour chacun des marqueurs étudiés qui correspondent à autant d'attracteurs de cette dynamique. Ces attracteurs peuvent être mis en évidence grâce à l'étude d'un corpus. Pour chaque exemple du corpus, on note la valeur sémantique du marqueur et tous les traits co-textuels pertinents : l'hypothèse que les valeurs typiques correspondent à des attracteurs de la dynamique revient alors à prédire que les points représentatifs des énoncés vont s'accumuler autour de ces attracteurs.

2.2. Présentation

ANASEM (pour ANAlyse SEMantique) a été développé au laboratoire LATTICE par Bernard Victorri. Dans sa première version, ce logiciel offre la possibilité de calculer automatiquement des corrélations entre différents paramètres linguistiques. L'outil informatique permet ainsi de dégager des règles linguistiques à partir de calculs statistiques. Dans sa nouvelle version, ANASEM permet également de calculer l'espace géométrique associé à un marqueur linguistique à partir d'un faisceau de descriptions. Ces descriptions linguistiques peuvent relever de niveaux structurels assez divers dont les quatre panneaux ci-dessous permettent de rendre compte :

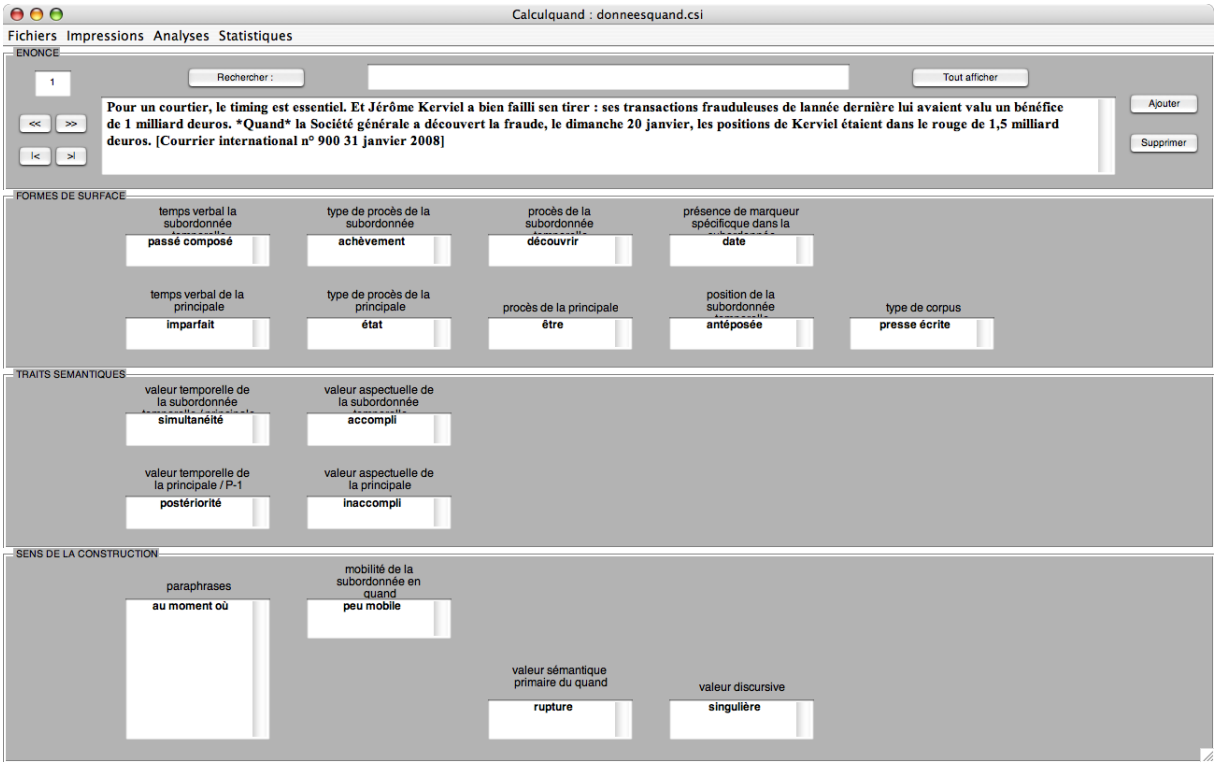


Figure 3 : Les 4 panneaux d' ANASEM

Le premier panneau donne accès à l'observable de premier niveau : l'énoncé étudié. Il est possible à tout moment d'ajouter ou de supprimer un énoncé. Une recherche d'occurrence sur l'ensemble des énoncés est rendue possible à partir de la fonction « rechercher ». Notons que dans le cas présent, nous ne travaillons pas sur des énoncés à proprement parler mais sur des séquences textuelles. Nous avons ainsi, pour chaque occurrence de *quand* dans notre corpus, élargi la fenêtre d'observation au contexte pertinent (les phrases antérieures et postérieures à la phrase d'accueil du marqueur *quand*) toutes les fois que cela était possible.

Le deuxième panneau inventorie les formes de surface. Nous désignons par « formes de surfaces » tous les faits linguistiques observables directement, c'est à dire qui ne nécessitent pas un calcul interprétatif. Ainsi nous renseignons dans cette catégorie les temps verbaux, le type de procès et le type de corpus. Bien que ce ne soit pas nécessaire au calcul sémantique, nous avons précisé ici le lemme du procès étudié pour lever toute ambigüité dans le codage des formes.

Le troisième panneau concerne les formes sémantiques. Nous avons retenu les valeurs temporelles et aspectuelles des énoncés comme critères pertinents. Il s'agit cette fois d'un véritable calcul sémantique car ces valeurs sont données pour le sens global de l'énoncé et non pas pour le verbe seul.

Enfin, le quatrième panneau fait état du sens de la construction. Nous avons indiqué ici les paraphrases possibles à la construction en *quand*: *si, au moment où, puisque*…. Nous avons également tenu compte de la mobilité de la proposition en *quand* selon un gradient : *proposition isolée, pas mobile, peu mobile, mobile*. Enfin, nous avons dégagé à chaque fois une valeur sémantique « primaire » pour l'occurrence étudiée, que nous avons complétée par une valeur « discursive ». Nous verrons un peu plus loin à quoi correspondent ces valeurs pour le marqueur *quand.*

2.3. Un exemple d'utilisation : l'étude du marqueur grammatical *quand*

Partant d'une observation sur large corpus narratif de plus d'un million de mots, nous avions dégagé dans un précédent travail (Girault 2007) trois grandes valeurs discursives de *quand*: le *quand* de rupture, le *quand* de cohésion et le *quand* introducteur.

Sans entrer dans les détails ici, précisons que ces valeurs discursives se fondent sur la notion de *Situation Narrative* que nous avions définie comme la structure conceptuelle élaborée à partir de segments linguistiques et de données extra-linguistiques. Ainsi, nous avions procédé à l'examen linguistique détaillé de plusieurs marqueurs grammaticaux pour valider cette notion de Situation. Parmi ces marqueurs, le marqueur *quand* avait attiré notre attention. En effet, d'un point de vue discursif, *quand* est un marqueur hybride, qui tantôt introduit une Situation nouvelle dans la narration (exemple 1), tantôt participe à la cohésion textuelle en présentant une transition entre deux Situations (exemple 2), et enfin, troisième possibilité, *quand* introduit une rupture de la Situation en créant un point de bifurcation narrative (exemple 3). Voici une illustration de ces trois cas de figure :

(1) **Quand** vous surveillez quelqu'un ou **quand** vous écoutez une conversation, ça se voit tout de suite. (Corpus Frantext – P. Modiano, *Les boulevards de la ceinture*, 1972, page 43)

(2) Le secrétaire s'en alla rapidement. **Quand** il revint, cinq minutes plus tard, il annonça, d'un air surpris, qu'il n'avait pas trouvé l'inspecteur Vérot. (Corpus Lupin, M. Leblanc, *Les dents du tigre*.)

(3) Gourel sonna de nouveau furieusement. Rien. Il se disposait à partir **quand**, soudain, il se baissa et appliqua vivement son oreille contre le trou de la serrure. (Corpus Lupin – M. Leblanc, *813*.)

Précisons que la valeur de *quand* introducteur se subdivise elle-même en trois :

- *quand* est introducteur d'une Situation nouvelle et inédite dans la narration
- *quand* est introducteur d'une Situation réitérée
- *quand* est introducteur d'une Situation générique

ce que nous illustrons respectivement par :

(1a) Je m'appliquai quelque temps à de petites notions qui ne me donnèrent grand éclair de presque rien. Puis je me lassai, et je fus tout à autre chose pendant six mois, **quand** un jour que je rentrais chez moi, je trouvai, assis sur une chaise et me regardant, l'ennui vêtu de son grand uniforme. (Corpus Frantext – L. Aragon, *Le paysan de Paris*)

(1b) Parfois, **quand** il marchait dans un corridor, Ismaïl voyait s'ouvrir un battant : un groom sortait, portant un plat ; il allait sur lui, l'ignorant. (Corpus Frantext – G. Perec, *La disparition*.)

(1c) **Quand** vous surveillez quelqu'un ou **quand** vous écoutez une conversation, ça se voit tout de suite. (Corpus Frantext – P. Modiano, *Les boulevards de la ceinture*, 1972, page 43).

Cette démarche d'observation linguistique sur grand corpus, bien que systématique grâce à des outils de traitement automatique que nous avions élaboré alors, reste insatisfaisante, parce que finalement assez intuitive.

L'intérêt majeur du logiciel ANASEM dans cette démarche typologique est qu'il permet de dégager des éléments de descriptions quantitatifs venant corroborer l'intuition linguistique.

Nous avons donc élaboré un nouveau corpus, en reprenant les occurrences de *quand* dans une notre corpus narratif d'origine, et en ajoutant des énoncés issus d'un corpus de presse écrite (*Le Monde* et *Le Courrier International*) afin de vérifier nos hypothèses sur un corpus plus hétérogène. Les données se répartissent à peu près en 60 % d'énoncés de registre narratif et 40 % de discours journalistique sur une centaine d'énoncés.

2.4. Les différentes caractéristiques étudiées

Pour chacun de ces énoncés en *quand*, nous avons procédé à une analyse « manuelle », renseignant les trois champs mentionnés plus haut : (1) formes de surface, (2) traits sémantiques, (3) sens de la construction.

Pour les formes de surfaces, nous indiquons pour la proposition principale et la subordonnée en *quand* le temps verbal du procès (présent, imparfait, passé simple…) et le type aspectuel auquel il renvoie selon la quadripartition de Vendler (1967) : état activité, accomplissement ou achèvement.

Pour les traits sémantiques, nous renseignons la valeur temporelle de la subordonnée par rapport à la principale et de la principale par rapport au contexte antécédent *quand* celui-ci est accessible. Nous avons distingué trois valeurs: antériorité, simultanéité et postériorité. De même les valeurs aspectuelles des propositions subordonnée et principale sont étiquetées accompli ou non accompli.

Enfin, pour le sens de la construction, outre les paraphrases possibles quand elles existent, et le degré de mobilité de la subordonnée temporelle, nous proposons une analyse en termes de valeur sémantique primaire du marqueur et sa valeur discursive. Là encore, nous discernons trois valeurs sémantiques primaires pour *quand* : introducteur, cohésif et rupture. Les trois valeurs discursives sont directement déduite du type de Situation déclenchée par *quand* : singulière, générique, réitérée.

2.5. Les résultats obtenus

Les statistiques sur les champs étudiées sont peu probantes. Pour ce corpus et cette description linguistique particulière, nous ne sommes pas parvenus à identifier des règles statistiques du type : « *la présence de tel indice* (aussi bien forme de surface que trait sémantique) *induit une valeur sémantique primaire de* quand *de type* rupture*,* cohésion *ou* introducteur ». Nous pouvons tout de même en déduire un élément d'information : dans l'analyse sémantique de *quand* en discours, aucun élément co-textuel à lui seul ou combiné à d'autres ne permet de

faire basculer la valeur sémantique d'un prototype à un autre. Ce qui plaide en faveur d'une approche continue de la construction du sens.

L'étude des corrélations entre les champs est un peu plus satisfaisante. ANASEM nous confirme par exemple que l'indice *« valeur aspectuelle du procès de la principale = accomplissement »* est fortement corrélé avec la valeur sémantique primaire de *rupture*.

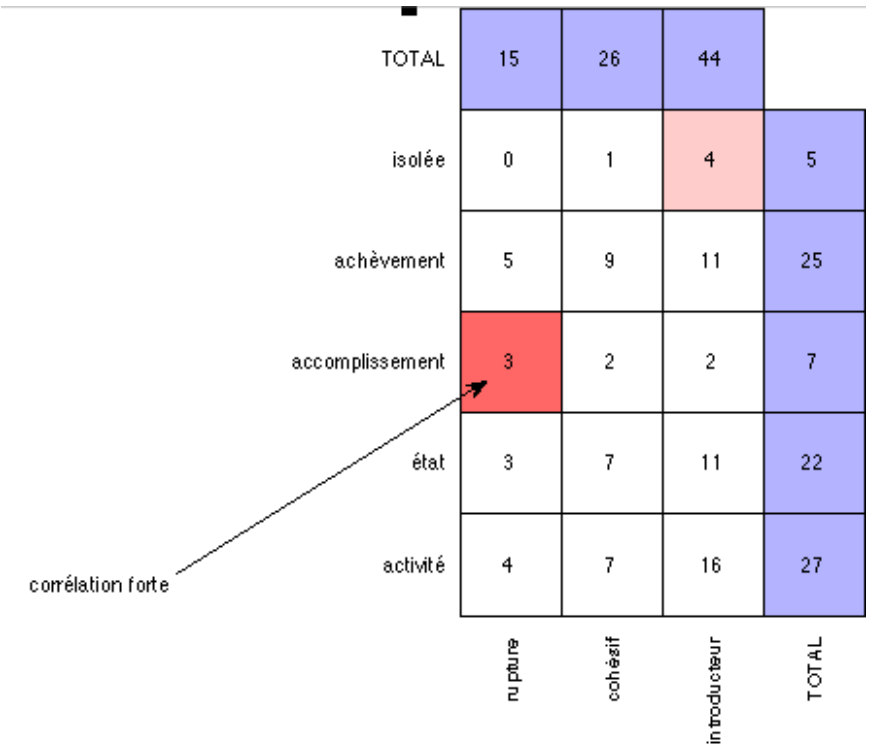


Figure 4 : Calcul des corrélations

D'autres corrélations sont ainsi mises en avant. Nous en avons relevées quelques unes que nous présentons à l'aide du tableau ci-dessous :

indice	**Valeur sémantique primaire de *quand***
Passé antérieur dans la subordonnée temporelle	cohésif
Passé simple dans la subordonnée temporelle	rupture
Procès inaccompli dans la subordonnée temporelle	introducteur
Postériorité de la principal / au contexte antérieur	cohésif
Situation réitérée	introducteur

En définitive, ces résultats viennent confirmer l'analyse linguistique en apportant, notamment, des arguments quantitatifs. Il devient possible de donner un « poids » aux corrélations des indices linguistiques, ce que l'analyse « à la main » ne permet pas de faire.

Mais c'est avec la représentation géométrique que les résultats sont réellement spectaculaires. En calculant la répartition géométrique des valeurs sémantiques primaires de *quand* à partir de tous les indices linguistiques que nous avons détaillés plus haut, on obtient la figure 1 ci-dessous :

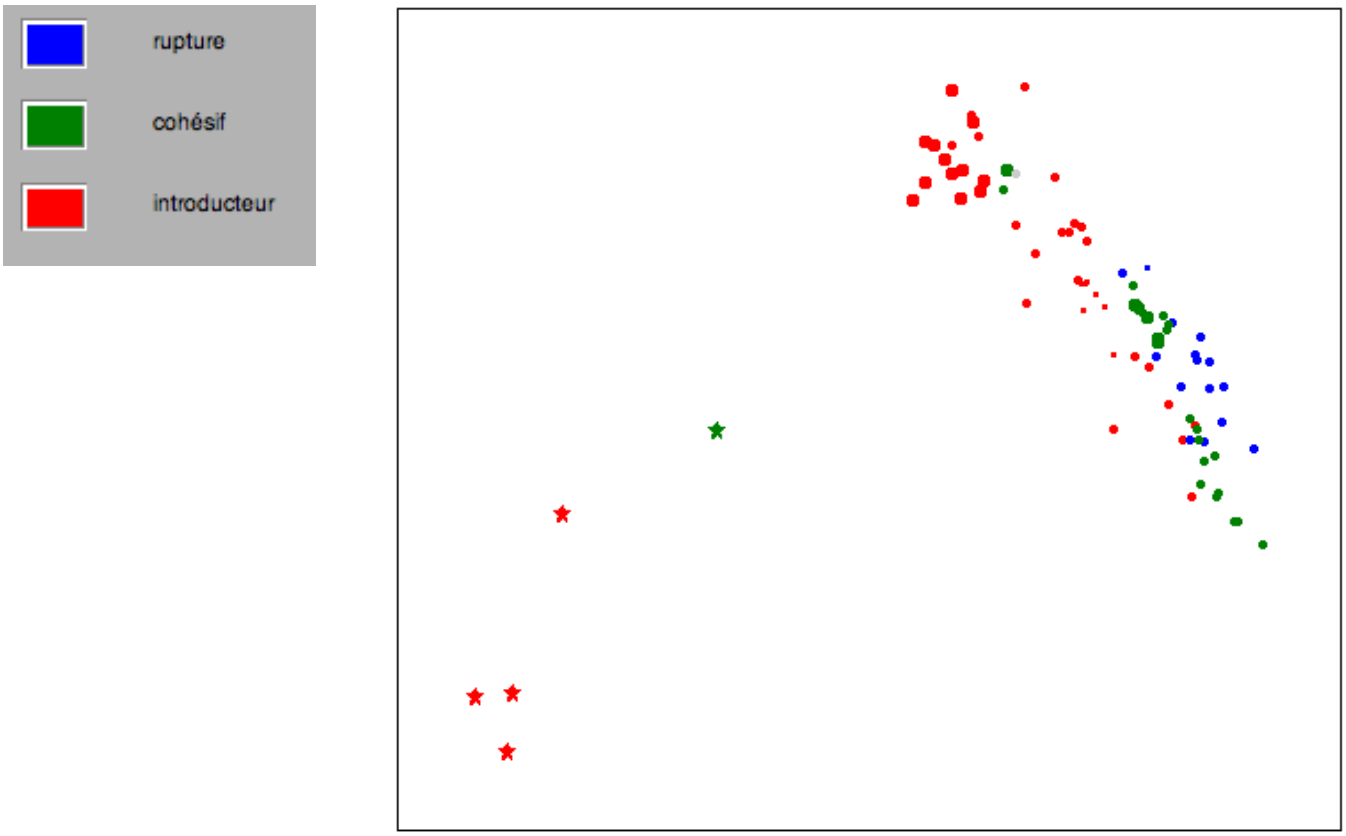


Figure 5 : Valeurs sémantiques primaires de *quand*

On distingue clairement dans cette représentation géométrique deux zones bien distinctes : d'un côté, un nuage de points assez dense, et de l'autre, en bas, à gauche de la représentation, quelques points isolés, que nous avons marqués par une étoile : quatre rouges et un vert. Comme l'indique la légende, les points bleus correspondent aux valeurs de rupture de *quand*, les verts aux *quand* cohésifs et les rouges aux introducteurs.

Les points isolés correspondent à des cas particuliers. Nous allons les supprimer du corpus provisoirement, nous y reviendrons par la suite. On obtient alors la représentation ci-dessous, avec une remarquable répartition des valeurs sémantiques primaires en trois branches qui correspondent précisément aux trois valeurs que nous avions identifiées au préalable : *rupture*, *cohésif* et *introducteur*.

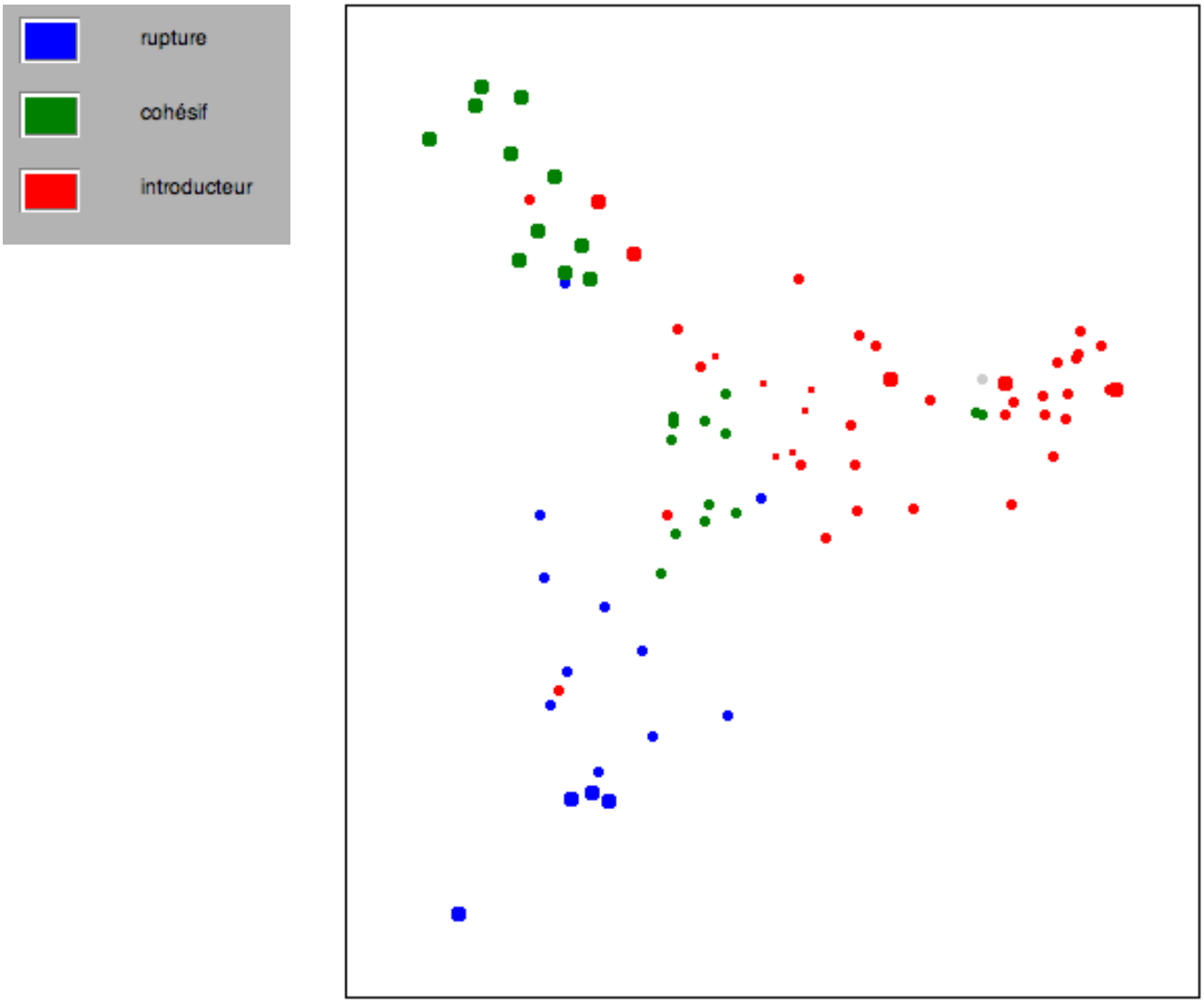


Figure 6 : Valeurs sémantiques primaires de *quand* (2)

Ce qui signifie que les hypothèses linguistiques que nous avions formulées au préalable quant aux valeurs prototypiques du marqueur *quand* en contexte discursif (cf. 2.3) se voient validées par le modèle.

Nous allons à présent entrer plus en détail dans cette représentation géométrique et montrer par quelques exemples comment cet instrument de modélisation offre de nouvelles perspectives à l'analyse linguistique sur corpus. Parmi les introducteurs en rouge, on distingue deux points verts. Ils correspondent aux énoncés :

(1) L'agent du bureau de l'immigration australien rit de bon cœur **quand** je lui dis que je vais séjourner au Menen, un hôtel cinq étoiles. [*Courrier international* n° 900 - 31 janvier 2008]

(2) A trop vénérer, on enterre trop profond. Si l'on ignore les termes entre parenthèses, l'adage javanais ci-dessus signifie littéralement : Vénérer haut et enterrer profond. Autrement dit, honorez vos aînés, d'autant plus **quand** ils sont morts. Taisez leurs fautes et placez-les à la hauteur de leurs bienfaits et de leurs dévotions. [*Courrier international* n° 900 - 31 janvier 2008]

Le *quand* de l'exemple (1) a clairement été mal analysé. Nous l'avons par erreur étiqueté comme un *quand* cohésif alors qu'il s'agit en fait d'un introducteur. ANASEM est de fait un outil d'aide à la vérification de l'analyse linguistique et permet d'identifier facilement les quelques erreurs inévitables de l'expertise humaine.

L'exemple (2) nous livre une information d'un autre ordre. Cette fois, il n'y a pas à proprement parler d'erreur dans l'étiquetage des indices linguistiques. C'est l'analyse globale qui demande à être affinée. Il semblerait qu'ici, la présence du marqueur de degré d'intensité « d'autant plus » vient modifier la valeur sémantique de *quand*. Il y a là un choix théorique à faire : peut-on considérer que *quand* ici est encore autonome ou faut-il l'envisager comme une nouvelle forme : « d'autant plus quand », qui entrerait dans le même paradigme que « surtout quand » ? Tout ce que nous pouvons dire pour le moment, c'est que cette « pseudo exception » nous invite à étendre notre corpus afin d'observer le cas échéant de nouvelles régularités qui ne sont pas perceptibles dans le contexte actuel.

Si l'on revient à la représentation globale, nous pouvons également noter la présence d'un point rouge (un *quand* introducteur) au milieu de la zone bleue des valeurs de rupture. Il correspond à l'exemple suivant :

> (3) On avait senti qu' Albertine avait cessé d' être une petite enfant **quand** un jour, pour remercier d' un cadeau qu' une étrangère lui avait fait, elle avait répondu : « je suis confuse ». [Corpus Frantext - M. Proust, 1921, *À la recherche du temps perdu*, p. 355]

Or, lorsqu' on examine cet exemple d'un peu plus près, on s'aperçoit qu'il s'agit en fait d'un emploi hybride ; *quand* exprime en effet l'introduction d'un nouvel épisode, une nouvelle Situation, mais en même temps, il exprime un point de rupture dans la narration : jusqu'ici, Albertine a été une enfant, à partir de ce jour, elle ne l'est plus.

Pour terminer, laissons de côté l'analyse des cas particuliers pour revenir aux cinq énoncés que nous avions provisoirement isolés artificiellement.

Il s'agit des énoncés suivants :

> (4) La situation des syndicats de Trondheim se dégrada rapidement, dans un secteur public de plus en plus privatisé. Difficile de recruter de nouveaux membres **quand** les emplois stables disparaissent. Sans compter l'augmentation du stress au travail et du nombre d'arrêts maladie. [*Courrier international* n° 900 - 31 janvier 2008]

> (5) Jonathan Yeo a peut-être besoin d'approbation, mais il n'a pas grand-chose de l'artiste torturé. Je m'en rends compte dès que j'entre dans le fantastique atelier de Chelsea qu'il occupe depuis cinq ans et qui semble avoir les plus hauts plafonds et les plus vastes fenêtres du Londres résidentiel. Pour mes modèles, il est plus facile de venir ici que, par exemple, d'aller au fin fond de Dalston [un quartier populaire de l'East End]. **Surtout quand**, parmi les modèles en question, figurent de grands noms du cinéma, des princes et des personnalités politiques. [*Courrier international* n° 900 - 31 janvier 2008]

(6) **Quand** portable et avion riment avec violon [Titre du *Courrier international* n° 900 - 31 janvier 2008]

(7) **Quand** la grenade remplacera le pavot [Titre du *Courrier international* n° 900 - 31 janvier 2008]

(8) **Quand** les terroristes font leur marché [Titre du *Courrier international* n° 900 - 31 janvier 2008]

Parmi les outils proposés par ANASEM (nous ne pourrons pas tous les énumérer ici), il est possible de calculer les propriétés d'une sélection de points, comme ci-dessous :

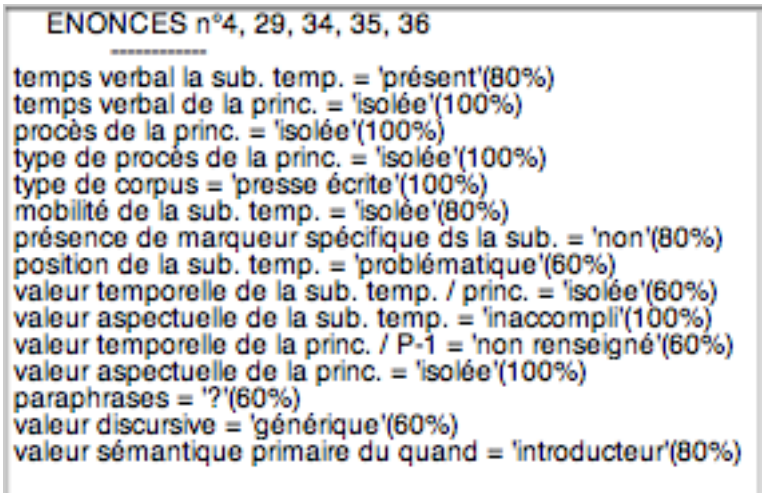


Figure 7 : Propriétés de la sélection d'énoncés

Cette fenêtre de propriété ne relève que les indices pertinents à plus de 50 % pour la sélection. Ainsi, on remarque que le point commun de tous ces énoncés est que leur proposition d'accueil est une subordonnée isolée. Le procès de la principale ou la principale toute entière sont implicites. Il est remarquable aussi que les énoncés (6), (7) et (8) correspondent à des titres d'articles. Ce qui leur confère un statut sémantique un peu particulier. Nous les avions analysés comme des introducteurs, mais d'un certain point de vue, un titre, c'est aussi un résumé de ce qui suit. Que l'on pense par exemple à l'exemple célèbre de Candide, le roman philosophique de Voltaire :

(9) Chapitre Premier
Comment candide fut élevé dans un beau château, et comment il fut chassé d'icelui.

Ceci plaide en faveur de la construction dans la théorie d'une valeur prototypique à part pour les titres. Les exemples (4) et (5) sont analysés alors comme se situant à mi –chemin sur le continuum sémantique entre deux valeurs prototypiques : le *quand* introducteur de Situation et le *quand* introducteur de titre.

Conclusion

Nous avons présenté ici un instrument de recherche pour la linguistique, le logiciel ANASEM qui permet de calculer les valeurs en usage d'un marqueur grammatical. sur un espace

sémantique modélisé, notamment, grâce aux mathématiques du continu. Les résultats que nous avons obtenus sont très encourageants et nous invitent à poursuivre dans cette direction. Ils tendent en effet à montrer que cet instrument présente un intérêt à au moins deux titres : du point de vue de la validation théorique d'une part et du point de vue de la découverte de nouveaux observables d'autre part.

En ce qui concerne la validation théorique, les trois valeurs sémantiques de *quand* dégagées dans le travail de thèse de S. Girault (Girault 2007) sont parfaitement identifiées par ANASEM qui calcule une représentation à trois branches correspondant aux trois grandes valeurs typiques du marqueur. Nous projetons à très court terme de travailler sur d'autres marqueurs grammaticaux. Une étude sur *si* est en cours et les résultats sont tout aussi satisfaisants.

Quant aux nouveaux observables linguistiques, c'est par l'intermédiaire du modèle même qu'ils deviennent accessibles. La représentation des données telles qu' ANASEM les « donne à voir » nous permet d'identifier les cas limites ainsi que les occurrences à la frontière entre deux valeurs, autrement dit, ceux-là même qui intéressent la linguistique.

Toutefois, nous modèrerons quelque peu notre enthousiasme : comme le soulignent à juste titre Marcel Cori et Sophie David (2008), le nouvel usage des corpus ne saurait en tant que tel "fonder une nouvelle linguistique", de même que, pour reprendre l'analogie que nous avons développée au début de l'article, l'utilisation de la lunette de Galilée n'a pas fondé à elle seule une nouvelle astronomie : elle a simplement fourni à Galilée des éléments qui l'ont aidé à concevoir sa théorie. Tout ce que l'on peut donc espérer en linguistique, c'est que l'usage de ces nouveaux instruments d'analyse de gros corpus jouera un rôle similaire, en permettant l'émergence dans les prochaines années de nouvelles théories linguistiques mieux à même de rendre compte de toute la complexité du langage.

Auroux S. (1998). La raison, le langage et les normes, Paris, Presses Universitaires de France.

Boersma P., Hayes B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32(1):45–86.

Bourigault D. (2007). *Syntex, analyseur syntaxique opérationnel*, Mémoire d'HDR, Université de Toulouse-Le Mirail.

Cori M., David S. (2008), Les corpus fondent-ils une nouvelle linguistique ?, *Langages*, 171, p. 111-129.

Elman J. L. (1991). Distributed representations, simple recurrent networks and grammatical structure. *Machine Learning*, 7, 195-224.

Ferrer R., Solé R. V. (2001). The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1482):2261-2265.

Gaume B., Venant F., Victorri B. (2006) Hierarchy in lexical organization of natural language, in D. Pumain (éd.), *Hierarchy in natural and social sciences*, Methodos series, vol 3, Springer, 2006, 121-142.

Girault S. (2007). Recherche sur les marques aspectuelles et temporelles dans les organisations narratives. Thèse de doctorat de l'Université de Caen.

Givón T. (1979). *On Understanding Grammar*. New York: Academic Press.

Habert B. (2005a). *Portrait de linguiste(s) à l'instrument.* *http://perso.limsi.fr/spip/IMG/pdf/BHabertPortraitDeLinguisteALInstrumentV4.pdf*

Habert B. (2005b). Instruments et ressources électroniques pour le français. Gap/Paris, Ophrys, L'essentiel français.

Habert B., Zweigenbaum P. (2002). Régler les règles. *Traitement automatique des langues*, 43(3):83-105.

Hamon T., Nazarenko A. (2001). Detection of synonymy links between terms: experiment and results, *Recent Advances in Computational Terminology*. John Benjamins.

Harris Z., Gottfried M., Ryckman T., Mattick JR P., Daladier A., Harris T.,Harris S. (1989). *The Form of Information in Science, Analysis of Immunology Sublanguage*, Kluwer.

Harris Z. (1988). *Language and information*, Columbia University Press, New York.

Harris Z. S.(1991). *A theory of language and information. A mathematical approach*, Oxford University Press.

Hinton G. E., Shallice T. (1991). Lesioning an attractor network: investigations of acquired dyslexia. *Psychol. Rev.*, 98, 74-95.

Jacquet G., Venant F., Victorri B. (2005). Polysémie lexicale, in P. Enjalbert (éd.), *Sémantique et traitement automatique des langues*, Hermès, 99-132.

Manning C. (2003). Probabilistic Syntax, in Bod, Hay and Jannedy (eds), *Probabilistic Linguistics*, MIT Press, 289-341.

Miikkulainen R., Dyer M. G. (1991). Natural language processing with modular PDP networks and distributed lexicon. *Cogn. Sci.*, 15, 343-400.

Milner J-C. (1992). De quelques aspects de la théorie d'Antoine Culioli in *La théorie d'Antoine Culioli, ouvertures et incidence*s. Ophrys, Paris.

Milner J-C. (1989). Introduction à une science du langage. Le Seuil, Paris.

Pereira F. (2000). Formal grammar and information theory : together again ?, *Philosophical Transactions : Mathematical, Physical and Engineering Sciences*, 358:1239-1253.

Ploux S., Victorri B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes, *Traitement automatique des langues*, 39(1) :161-182.

Pollard C., Sag. I. A. (1994). *Head-Driven Phrase Structure Grammar*, University of Chicago Press.

St John M. F., McClelland J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, 46, 217-56.

Prince A., Smolensky P. (1993). *Optimality theory: Constraint interaction in generative grammar*. Technical Report TR-2, Rutgers University Center for Cognitive Science.

Rastier F. (2002). Enjeux épistémologiques de la linguistique de corpus. In Williams G. (Ed.), *Actes des deuxièmes Journées de linguistique de corpus de Lorient*. Presses Universitaires de Rennes.

Simondon G. (2001 réed.) Du mode d'existence des objets techniques, Paris, Aubier Tabor W., Tanenhaus M. K. (1999). Dynamical Models of Sentence Processing, *Cognitive Science*, 23 (4):491-515

Vendler Z. (1967). *Linguistics in Philosophy*, Cornel University Press.

Victorri B., Fuchs C. : *La polysémie, construction dynamique du sens*, Hermès, 1996.

Watts D.J., Strogatz S.H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393:440-442.