

***Parce que, perché, porque* dans les langues romanes médiévales : l'utilité des études sur corpus¹**

FAGARD Benjamin
Lattice (ENS / CNRS)

Résumé : Nous proposons ici une réflexion sur l'utilité du corpus pour les linguistes, diachroniciens en particulier. Afin de montrer que le corpus est le complément idéal des outils traditionnels de la philologie (grammaires et dictionnaires de référence), nous étudions les conjonctions de subordination *par/por ce que*, *perché* et *porque* en français, italien et espagnol médiévaux, en deux temps : d'abord à partir des outils philologiques, puis avec une étude sur corpus. Cette dernière nous permet d'émettre une hypothèse sur les différences observées dans le comportement des trois conjonctions, à savoir les emplois intersubjectifs de *perché* et *porque* et l'absence de tels emplois pour *par/por ce que*.

Mots-clés : subjectification, intersubjectification, corpus électroniques, langues romanes, cause

Abstract : In this paper, we try to find out in what ways corpora can be useful to linguists, and specialists of diachrony in particular. In order to show that corpora are in fact the ideal complement to the more traditional tools of philology (reference grammars and dictionaries), we propose an analysis of causal conjunctions as they were used in the Middle Ages in three Romance languages: *par/por ce que* in French, *perché* in Italian and *porque* in Spanish, all meaning roughly «because». We first compile the data from various philological tools, and we then compare these results with a corpus study of the same items. This corpus study enables us to put forward a hypothesis which might explain the differences observed in the

¹ Nous tenons ici à remercier les relecteurs anonymes pour leurs commentaires.

behaviour of the three conjunctions under study, that is to say the fact that *perché* (Italian) and *porque* (Spanish) have intersubjective uses, which are not available to *par/por ce que* (French).

Keywords: subjectification, intersubjectification, electronic corpora, Romance, cause

« many aspects of corpus-based research remain mysterious for readers, since research articles can rarely afford the space to discuss methodological considerations or analytical procedures in complete detail. » Biber & al. (1998 : ix)

0. Introduction

Le vingtième siècle a apporté à l'étude diachronique des langues des outils précieux, sous la forme de textes en format électronique, sélectionnés et rassemblés dans des bases de données. Pour presque toutes les langues romanes, en particulier, on dispose de corpus électroniques de textes médiévaux, du 10^{ème} au 16^{ème} siècle². Il devient possible de faire des études diachroniques et/ou comparées, sur corpus. Reste à savoir si cela en vaut vraiment la peine : nous entendons ici, à partir d'une étude de cas, proposer une réflexion sur l'utilité des corpus, à la fois d'un point de vue théorique et sur la base d'un exemple pratique.

Dans la première partie de cet article, nous cherchons ainsi à montrer que les corpus constituent un complément idéal aux outils traditionnels (grammaires et dictionnaires de référence), qui conservent tout leur intérêt. Suivant Prévost (2005), nous avançons ici que les corpus permettent dans une certaine mesure de compenser l'absence de locuteur pour les états anciens de la langue.

Dans la seconde partie, nous passons à l'étude de cas, et comparons les conjunctions *par ce que*, *perché* et *porque* en

² Nous pensons en particulier aux bases suivantes (la liste qui suit ne constitue pas, cependant, un relevé exhaustif) : BFM et DMF (français), OVI et CLPIO (italien), Corde (espagnol), Phrasis, CIPM et MedDB (portugais et galicien), RIALC (catalan), COM (occitan)... Nous utilisons dans cet article et présentons en annexe trois d'entre elles : la BFM, OVI et Corde, et renvoyons à Fagard (2006) pour les autres.

français, italien et espagnol médiévaux. Nous proposons ainsi d'étudier la naissance des conjonctions causales dans les langues romanes, à partir des prépositions latines *per* et *pro* et du connecteur *quod* (qui a abouti à *que* en français et en espagnol, à *che* en italien) : elles ont donc des origines proches³, et les comparer nous permettra d'observer les orientations différentes de leur évolution. Cela devrait en outre nous permettre de clarifier les rapports entre grammaticalisation, lexicalisation et subjectification, phénomènes désormais bien connus, mais dont les contours respectifs restent à définir : la grammaticalisation suppose-t-elle toujours une subjectification, ou inversement ? La lexicalisation est-elle une étape de la grammaticalisation ?

Nous présenterons d'abord les données accessibles à partir des ouvrages de référence, pour chacune des conjonctions choisies. Nous résumerons ensuite les résultats de l'étude sur corpus, en présentant l'évolution et en particulier le figement des différentes formes. Nous chercherons en outre à établir un lien entre le mode de formation des différentes formes et leur évolution sémantique – plus précisément, à montrer la subjectification plus importante de *perché* et *porque*.

En annexe, nous présentons les corpus utilisés pour les différentes langues romanes, à l'état ancien : il s'agit d'évaluer le *contenu* (quantité/qualité), l'*interface* et l'*accessibilité* ainsi que la *période* couverte, pour chaque langue.

³ Et une origine non identique : l'absence de pronom dans *perché* et *porque* les distingue nettement de *parce que* (l'italien connaît bien une construction avec le démonstratif, *perciò ché*, dont nous réservons l'étude pour une autre occasion). Nous avons sélectionné ces trois conjonctions d'une part en raison de leur proximité formelle, d'autre part parce qu'elles ont survécu toutes trois jusqu'à la langue moderne, ce qui permettra par la suite d'étudier leur évolution sur une diachronie large.

1. Complémentarité des ouvrages de référence et des corpus numérisés

1.1 Limites des ouvrages de référence

1.1.1 Limitations liées à leur cadre théorique

Les limites des outils philologiques traditionnels sont principalement liées à leur caractère *fini* : étant donné qu'ils ont été écrits et publiés, pour certains il y a plus d'un siècle, ils n'offrent aucune possibilité d'adaptation à l'évolution des connaissances. Lorsque, par exemple, il est fait l'hypothèse de l'existence d'une nouvelle classe de mots dans la grammaire, comme celle des marqueurs discursifs – que cette hypothèse soit valable ou non – les grammaires comme les dictionnaires ne fournissent aucune information à ce sujet. En outre, on peut dire que jusqu'à la fin du siècle dernier le partage des tâches entre grammaires et dictionnaires était tel que les éléments situés à mi-chemin entre grammaire et lexique, comme par exemple les locutions de tous types (verbales, adverbiales, prépositionnelles), n'étaient traités de manière satisfaisante ni par les premières, ni par les seconds. Les grammairiens n'en rendaient pas compte de manière exhaustive parce qu'ils les considéraient comme des éléments du lexique, et les lexicologues en rendaient compte, mais pas de manière systématique, entre autres pour des raisons de place. D'autre part, même dans les parties du discours « traditionnelles », on remarquera que les classes « mineures », grammaticales – prépositions, adverbes, conjonctions ... –, sont généralement peu traitées par les grammaires traditionnelles.

Cette limite des outils traditionnels se retrouve, de manière plus générale, pour toutes les évolutions théoriques ou terminologiques. Comme le note de Haan (2002 : 91), « l'un des buts de la recherche sur corpus est de tester les propositions théoriques que l'on trouve dans la littérature »⁴. Les outils

⁴ « One of the purposes of corpus-based research is the **testing of theoretical claims made in the literature** » (nous soulignons ; notre traduction).

traditionnels ne permettent pas toujours de le faire, tout simplement parce qu'ils sont limités par leur propre cadre théorique. Par exemple, dans le cas d'espèce que nous avons choisi d'étudier ici – celui des connecteurs de causalité – il semble que le concept de subjectivité soit important. Or ce concept est apparu il y a longtemps (Bréal 1900 : chapitre 25), mais il n'est guère utilisé par les grammaires et dictionnaires de référence. Si l'on veut avoir une vision précise du rôle de ce concept dans l'évolution des connecteurs de causalité, le recours aux corpus est donc inévitable.

1.1.2 Analyse quantitative et subjectivité

L'avantage principal des grammaires et dictionnaires de référence est qu'ils ont été écrits par des spécialistes, dont les connaissances dans leur domaine ne sont pas à remettre en question. Cependant, c'est là également un des défauts ou au moins une des limites principales de ces outils : ils sont nécessairement subjectifs. Les corpus permettent justement de compenser en partie cette subjectivité inévitable du chercheur, comme le notent Oakey (2002 : 111) « Il nous a semblé qu'une recherche sur corpus pouvait apporter une vision des choses plus claire, moins intuitive »⁵ ou encore Mair (1995 : 260), précisant que **l'analyste de corpus est en position de décrire les tendances statistiques avec précision**, ce qui lui permet avant tout de séparer l'usuel et le normal de l'exceptionnel. Cela est lié à la possibilité de mettre en œuvre des éléments objectifs, comme le repérage systématique de certains éléments contextuels, ou le calcul statistique (voir entre autres Prévost 2005).

L'intérêt du corpus pour compenser la subjectivité du chercheur est particulièrement net lorsque l'auteur précise suffisamment sa méthode et son corpus pour que son étude soit reproductible (cf. Oakey 2002 : 115) :

Le corpus utilisé pour cette étude a été sélectionné afin de garantir la reproductibilité de

⁵ « It was felt that **corpus research may provide a clearer, less intuitive insight** » (nous soulignons ; notre traduction).

l'étude. **Assigner des valeurs fonctionnelles à des unités linguistiques est une tâche très subjective**, et les chercheurs, dans ce domaine, travaillent sur leurs propres sous-corpus, qui ne sont pas accessibles aux autres chercheurs. Le résultat est que ces études **ne peuvent être facilement reproduites**, et leur degré de subjectivité peut s'en trouver difficile à évaluer.⁶

Il ne suffit pas cependant d'avoir des outils statistiques pour échapper à la subjectivité et surtout à l'erreur ; de plus, il faut utiliser les données statistiques avec précautions, comme le rappellent Biber, Conrad & Reppen (1998 : 268)

Lorsqu'on recourt aux mesures statistiques, il faut prendre en compte certains aspect, et avant tout la fréquence des éléments étudiés. Les calculs statistiques ne sont généralement pas valables avec des fréquences très basses et, pour quelque test statistique que ce soit, il est important de s'assurer que l'on a suffisamment d'éléments pour l'analyse. En outre, il faut utiliser une approche qui convienne à la question à l'étude.⁷

1.1.3 Phénomènes rares et structure du langage

Une autre limite des outils traditionnels est qu'ils ne recensent pas toujours les phénomènes rares ou très rares, ou bien que, lorsqu'ils les recensent, ils ne donnent pas systématiquement suffisamment d'informations à leur sujet. Un

⁶ « The corpus data used in this study was chosen to allow replicability. **Applying functional labels to linguistic units is a highly subjective activity**, and researchers in the field often compile their own subcorpora which are not available to other researchers. This means that **previous studies can not easily be replicated**, and thus the consistency of subjectivity can be hard to gauge » (nous soulignons ; notre traduction).

⁷ « For the use of all statistical measures, certain considerations are important. Prime among these is the frequency of the items that are studied. Statistical calculations are generally not reliable with very low frequencies, and for any statistical test, it is important that you make sure you have enough tokens for the analysis. In addition, it is important that you use a procedure that fits your research question » (notre traduction).

corpus bien fourni permet de remédier à ce problème, et d'étudier y compris les faits de langue les plus rares. Il permet même d'aller plus loin, en évaluant non plus – comme le font les grammaires normatives – la grammaticalité ou non de telle ou telle structure, mais sa fréquence relative, autrement dit son degré de (proto-)typicité (voir Biber & al., 1998 : 3). Il ne s'agit plus de distinguer simplement ce qui fait partie du langage et ce qui en est exclu, mais de manière plus fine ce qui est courant ou non, en étudiant non seulement la structure linguistique mais l'usage, comme le proposent Biber & al. (ibid. : 1) : « Plutôt que de chercher ce qui est théoriquement possible dans une langue, nous étudions la langue telle qu'elle est effectivement utilisée dans des contextes naturels »⁸.

Un corpus permet ainsi – idéalement – de donner une image fidèle de la structure du langage (ou plutôt de la structure de la langue *du corpus*), comme le note Kennedy (2002 : 73) :

Les corpus électroniques modernes nous permettent maintenant d'explorer la nature et l'emploi de phénomènes linguistiques dans des textes de natures bien plus variées. Grâce à ces descriptions, **on va bien au-delà de l'étude de ce qui est grammaticalement** ou sémantiquement **possible**, en ajoutant une dimension distributionnelle qui caractérise les traits linguistiques, en termes de probabilité d'emploi (nous soulignons)⁹

Pour clore cette section, on pourra reprendre Prévost (2005 : 147) : pour les linguistes travaillant sur des états de

⁸ « Rather than looking at what is theoretically possible in a language, we study the actual language used in naturally occurring contexts » (notre traduction).

⁹ « Modern electronic corpora now make it possible to explore the nature and use of linguistic phenomena in a much wider variety of texts. **Such descriptions go beyond exploring what is grammatically** and semantically **possible**, and add a distributional dimension which characterizes linguistic features in terms of probability of occurrence » (notre traduction).

Benjamin FAGARD

langue disparus, « le corpus est indispensable¹⁰, puisqu'il conditionne l'existence même de l'objet à décrire » – ceci pour une raison principalement, à savoir que les médiévistes ne possèdent pas de compétence de production (Marchello-Nizia 1995 : 22) ; ou encore, comme le notent Habert, Nazarenko & Salem (1997 : 132-133) : « L'érudit contemporain ne saurait affirmer : cet énoncé n'est pas acceptable ». Reste à voir comment le constituer et l'utiliser – et, autre aspect fondamental, comment échapper au caractère circulaire de son utilisation (pour Prévost *ibid.* : 149, le linguiste médiéviste est contraint d' « étudier des textes avec une connaissance de la langue exclusivement fondée sur ces mêmes textes »).

1.2 Problèmes liés à l'étude sur corpus et précautions à prendre

Bien sûr, tout cela suppose que l'on dispose d'un corpus 'idéal'. Dans les faits, les défauts du corpus (en particulier pour les langues anciennes) peuvent également limiter la validité des résultats. Il y a cependant souvent moyen de compenser en partie ces limites ; il faut en tout cas rester conscient du fait que les résultats d'une étude sur corpus sont valables non pour la langue dans sa généralité, mais pour l'état de langue représenté par le corpus.

1.2.1 Constitution du corpus

Un avantage majeur du corpus est donc son caractère *dynamique*, à savoir qu'il peut (et doit) s'adapter à chaque fois à l'objet d'étude. Ceci nous amène cependant à ajouter une contrainte supplémentaire à celles qui régissent déjà la constitution de tout corpus (voir Biber & al. 1998 : 251) « Lorsque l'on rassemble un corpus, qu'il soit synchronique ou diachronique, la première étape est de décider pour quelle

¹⁰ On notera à ce propos que des corpus électroniques sont utilisés de plus en plus fréquemment pour la constitution d'outils philologiques (grammaires et dictionnaires), et ce, depuis un certain temps déjà.

recherche on désire l'utiliser »¹¹). Nous ne rappelons ici que les principales contraintes.

Si l'on part de l'idée qu'un corpus doit représenter la langue ou, du moins, une partie de la langue (Biber & al. 1998 : 246, Habert 2000), il reste à voir comment obtenir un corpus représentatif. Un des éléments fondamentaux à cet égard est la diversité des textes inclus dans le corpus, qui doit reproduire la variation ou plutôt les variations propres à la langue (Biber & al. 1998 : 247-248, Biber 1990, Prévost 2005). Cette diversité doit donc être comprise de plusieurs manières : le corpus doit en effet reproduire autant que possible la variation régionale (dialectes), stylistique (genres textuels), de niveau de langue... La constitution d'un corpus diachronique est plus complexe encore, comme cela a déjà été noté (Biber & al. 1998 : 251, Prévost 2005), puisque à ces dimensions de variation s'ajoute l'axe diachronique.

Enfin, selon Habert, Fabre & Issac (1998 : 33), « accumuler les données textuelles ne suffit pas : elles doivent être réunies en un tout cohérent, **caractérisable**, si l'on veut qu'elles deviennent le support d'une analyse dont les conclusions soient généralisables » (nous soulignons) – on peut en conclure avec Lebel (2003) que « la disponibilité des données ne doit pas être le principal critère de regroupement ». Le but est de parvenir à la constitution d'un corpus tel que le définit Sinclair (1996 : 4, cité in Habert et al., *op. cit.*) : « une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques **explicites** pour servir d'échantillon du langage ».

En fonction de l'objet d'étude, on choisira donc un corpus englobant autant que possible l'ensemble de ces variantes, ou bien une partie seulement. La taille du corpus doit également être adaptée à l'objet d'étude, afin d'assurer la représentativité des occurrences qu'il contient. Il ne semble pas possible cependant de donner *a priori* une taille minimale nécessaire pour chaque type d'étude.

¹¹ « When designing either a synchronic or a diachronic corpus, the first step is to determine the intended research purposes » (notre traduction).

1.2.2 Utilisation du corpus

Une fois le corpus constitué, il reste un certain nombre de choix à effectuer. Il convient ainsi de déterminer le nombre d'occurrences de la structure étudiée, et le nombre d'occurrences que l'on compte analyser. Il arrive souvent en effet que la quantité de données soit trop importante pour permettre au chercheur de les prendre en compte dans leur totalité. Il faut alors se décider, soit, en amont, à limiter le corpus, soit, en aval, à opérer une sélection aléatoire. Dans les deux cas, il faut s'assurer que cela ne compromet pas la validité de l'étude, et vérifier notamment que le corpus retenu en fin de compte reste représentatif. Il semble indispensable, en tout état de cause, de rendre compte de la fréquence du phénomène étudié, et pas seulement des occurrences traitées. On pourra à cette fin associer une étude semi-automatisée sur l'ensemble des occurrences, et une étude de détail sur une partie de ces dernières, sélectionnées aléatoirement, chronologiquement ou sur un autre critère (pour autant qu'il soit justifié, et surtout explicité ; voir Biber & al. 1998 : 4). La plupart des bases facilitent ce volet quantitatif de l'étude avec des fonctions comme l'index (BFM) ou le calcul de fréquence (OVI, Corde). L'analyse quantitative permet en outre, au prix de quelques efforts supplémentaires, de faire appel à la statistique mathématique pour déterminer dans quelle mesure les observations effectuées sont significatives.

Un autre problème est la lisibilité des occurrences : il faut s'assurer en effet d'avoir des résultats utilisables. Certaines bases de données ne fournissent pas un contexte suffisant pour certains types d'analyse – les études sur le discours, notamment, nécessitent la prise en compte d'un contexte très large¹². Il va de soi en effet, comme le note Oakey (2002 : 116), que d'avoir

¹² Ainsi de la base Frantext et du DMF pour le français ; de la COM pour l'occitan... la base OVI pour l'italien et la base CORDE pour l'espagnol proposent deux options d'affichage des résultats, avec un contexte très limité ou bien nettement plus large. Le meilleur système selon nous est celui de la BFM, qui laisse le chercheur choisir la taille du contexte qu'il désire.

un corpus à sa disposition n'implique pas que l'on puisse se limiter à des études 'automatisées' :

L'analyse sur corpus permet d'identifier plus rapidement des séquences dans un grand ensemble de textes, mais le jugement et l'intuition du locuteur doivent garder la part belle lorsque le linguiste veut déterminer si une séquence donnée remplit une fonction donnée.¹³

Il faut enfin prendre en compte le bruit et le silence : le bruit, inévitable, consiste à avoir dans les résultats autre chose que la structure recherchée, tandis que le silence, qu'il convient au contraire d'éviter, résulte de l'oubli de certaines variantes, par exemple graphiques, de cette même structure. C'est là qu'intervient la complémentarité des deux approches, sur corpus et traditionnelle.

1.3 Combiner corpus numérique et instruments traditionnels

C'est en effet l'utilisation préalable des outils traditionnels tels que grammaires et dictionnaires qui permet au linguiste d'éviter par la suite, dans son étude sur corpus, de passer à côté de certaines formes ou constructions. Nous fournirons dans la section suivante quelques exemples de cette complémentarité des deux approches, notée déjà par Biber & al. (1998 : 9-10 : « Cependant, il convient de voir l'analyse sur corpus comme une approche complémentaire des approches traditionnelles, plutôt que comme la seule bonne approche »)¹⁴.

L'utilisation du corpus ne s'oppose donc pas nécessairement à la méthode philologique traditionnelle : elle doit plutôt être un moyen de compléter et de mettre à jour les

¹³ « Corpus analysis can speed the identification of word strings in a large corpus, but traditional intuitive judgement must still play a large part in deciding whether a particular string performs a particular function » (notre traduction).

¹⁴ « However, corpus-based analysis should be seen as **a complementary approach to more traditional approaches**, rather than as the single correct approach » (nous soulignons ; notre traduction). Cette complémentarité se voit par ailleurs dans la publication de grammaires ou dictionnaires « d'usage ».

outils de cette dernière, en tenant compte des dernières ‘avancées’ épistémologiques et en même temps en portant sur ces mêmes ‘avancées’ un regard critique et juste – parce que fondé justement à la fois sur une vision de la langue ‘statique’ des outils traditionnels et sur celle, plus dynamique, des corpus.

En s’appuyant sur les données que lui fournissent les outils de la philologie (grammaires et dictionnaires de référence), le linguiste de corpus pourra analyser de manière plus pertinente le fonctionnement et/ou l’évolution de structures ou morphèmes donnés.

2. De la théorie à la pratique : une étude de cas

2.1 Choix de *par ce que*, *per ché*, *por que*

Le choix des locutions conjonctives à sens causal s’imposait assez naturellement pour cette étude de cas, pour plusieurs raisons. En premier lieu, parce qu’il s’agit de constructions complexes, à mi-chemin entre lexique et grammaire. Elles ont donc été relativement peu étudiées jusqu’ici, et sont relativement peu prises en compte par les grammaires et dictionnaires de référence. Elles présentent de plus, de ce fait, un intérêt tout particulier pour les études sur la grammaticalisation et plus précisément du renouvellement des paradigmes de formes grammaticales : il s’agit en l’occurrence d’étudier le remplacement progressif des conjonctions subordinatives synthétiques du latin (*ut, quod, quia, quam, ...*) par les conjonctions analytiques romanes (*pour que, par/por ce que > parce que, puis que > puisque, à fin que > afin que, étant donné que, etc.*).

En second lieu, ce caractère non figé des structures en question rend leur étude sur corpus assez complexe, ce qui nous permet de tester la souplesse des bases de données utilisées, et de déterminer plus précisément quels sont les outils nécessaires au linguiste de corpus ; nous verrons (en annexe) que c’est sur ce point que les bases de données varient le plus, en fonction de l’interface qu’elles proposent.

En dernier lieu, nous participons actuellement à un projet¹⁵ sur la subjectification (telle que définie par Sweetser 1990 et surtout Traugott 2003 ; y compris l'intersubjectification, telle que définie par Traugott & Dasher 2002 : 23-24). Le but du projet est d'étudier le rôle de la subjectification dans le processus de grammaticalisation : ce critère n'étant généralement pas pris en compte par les outils traditionnels, seule une étude sur corpus peut nous apporter des informations sur le degré de subjectification de ces structures, avec la prise en compte systématique de critères objectifs : présence de telle ou telle marque dans le contexte, position dans la phrase, TAM (temps, aspect, mode) et personne du verbe principal, etc. (pour une liste précise de ces critères, voir entre autres Pander Maat & Degand 2001, Pander Maat & Sanders 2001).

2.2 Outils traditionnels

Dans cette section et la suivante, nous comparons les informations que nous apportent respectivement les outils traditionnels (dictionnaires et grammaires de référence, voir bibliographie) et l'étude sur corpus (2.3). Nous avons ainsi l'occasion de constater les avantages de chaque outil : d'un côté, les ouvrages de référence – du moins nous l'espérons – tirent leurs données des manuscrits, et donnent pour les locutions retenues les dates d'apparition, les formes et graphies rencontrées dans les textes, les sens principaux ainsi que, pour les grammaires, un paradigme des locutions conjonctives causales, pour chaque langue (voir 2.2.1 à 2.2.3). De l'autre, le corpus permet d'affiner l'étude fonctionnelle, éventuellement de préciser les datations, de proposer une analyse quantitative et enfin de comparer les trois conjonctions avec une même approche (voir 2.3.1 à 2.3.3).

Nous présentons ici ces données, pour chaque locution ; on peut déjà dire qu'elles apparaissent dans l'ensemble dès les

¹⁵ Projet Gramis (IAP P6/44), financé par le gouvernement fédéral de Belgique : http://webh01.ua.ac.be/gramis/personnel_UCL.html.

premiers textes littéraires (soit 10^{ème} siècle en ancien français, 12^{ème}-13^{ème} siècles en ancien espagnol et italien), avec une certaine variabilité formelle (formes attachées ou non, graphies variées) et fonctionnelle (toutes ont plusieurs emplois).

2.2.1 Données pour le français Parce que

On trouve dès le 10^{ème} la forme *por cio que* (Saint Léger), dès 1200 la forme *par ce que*, enfin vers 1375 la forme soudée *parce que*.

Les sens repérés comprennent la cause, la condition, et (uniquement avec un verbe au subjonctif) le but ; les dictionnaires et grammaires ne notent pas (du moins pas tous) la manière, dont on verra pourtant qu'elle est assez fréquente dans notre corpus.

Les variantes sont principalement liées à la préposition utilisée (*por* ou *par*) et à la forme du démonstratif : *cio*, *ço*, *ce*, *ice*, *o* ; il peut également être fusionné à la préposition (*poruec*, *paruec*, *pruec*) ou même absent.

Le paradigme des conjonctions à valeur causale est assez large en ancien français : *que*, *par/por ce que*, *puisque*, *ainz/ainçois que*, *tant que*, *si que*, *tel que*, *car (quer)*, *com(e)*, *quant*. Elles se différencient les unes des autres principalement d'un point de vue sémantique, selon qu'elles ont outre les emplois spatiaux des emplois temporels, hypothétiques, complétifs, de but, d'accompagnement de l'impératif, etc. ; on notera en outre que seul *car* est une conjonction de coordination.

2.2.2 Données pour l'italien Perché

Perché apparaît également dès les premiers textes littéraires, soit au début du 13^{ème} siècle. On trouve également dès le 13^{ème} siècle la forme avec le démonstratif (*perciocchè*), qui n'est en général pas employée comme conjonction de subordination, à la différence de son équivalent français : il est plutôt relatif ou coordonnant, et on pourrait le gloser par « et pour cette raison » ou « raison pour laquelle ».

Le morphème *perché* semble plus polyvalent encore que *parce que*, du fait de ses emplois comme particule interrogative (avec le sens de « pourquoi ? » ; y compris dans l'interrogation indirecte), comme relatif (avec le sens de « raison pour laquelle »), comme simple conjonction introductrice de complétives (il équivaut alors à *che* « que »), et enfin comme nom (*il perché* « le pourquoi »).

Comme conjonction, *perché* peut prendre un sens final (*afin que*, avec verbe au subjonctif) ou concessif (*pour autant que*), outre ses emplois comme conjonction causale, que nous détaillerons plus bas.

Les variantes graphiques relevées sont les suivantes : *per ché*, *perché*.

Le paradigme des conjonctions à valeur causale est également assez large en italien médiéval, surtout bien évidemment si l'on retient les formes dialectales. Les principales sont *che*, *ca*, *perchè* *perocchè*, *perciocchè*, auxquelles il faut ajouter les conjonctions temporelles en emploi causal : *poichè*, *poi*, *giacchè*, *come*, *quando*, *alora che*, *mo che*, *adés che* ; et quelques autres qui peuvent également prendre un sens causal : *pe'via che*, *dacchè*, *onde*, *con ciò sia (cosa) che (conciòsiacosacché)*, sans compter quelques autres formes régionales.

2.2.3 Données pour l'espagnol Porque

Porque apparaît au 12^{ème} siècle, soit encore une fois dès les premiers textes littéraires (mais plus tard que *ca*, présent dès les *Glosas Silenses*, deuxième moitié du 10^{ème} siècle).

On trouve des emplois comme interrogatif et relatif, comme pour *perché* en italien. Pour les emplois comme conjonction, le premier sens observé est le sens causal de « parce que », un sens final apparaissant au 14^{ème} siècle, ainsi que le sens causal de « puisque ».

Les variantes notées sont *por que* et *porqué*.

Le paradigme des conjonctions à sens causal est également assez large en espagnol médiéval, comprenant au

moins les formes suivantes : *ca, que, pues que, pues (pus), porque, quando, como*.

2.3 Etude sur corpus

Pour l'étude sur corpus, nous avons sélectionné aléatoirement 150 occurrences de chaque construction, au 13^{ème} siècle, dans trois bases de données contenant des textes principalement littéraires : la BFM pour le français, la base OVI pour l'italien, la base Corde pour l'espagnol. Le recours aux textes littéraires, pour l'étude de phénomènes de subjectification, se justifie dans la mesure où la distinction discours/récit joue un rôle important dans la subjectification. On peut postuler que cette dernière est plus forte dans les contextes discursifs¹⁶, or les textes sélectionnés sont assez riches en dialogues. Il faudra sans doute, cependant, répéter l'expérience sur des corpus plus diversifiés.

Nous avons ensuite étudié les occurrences ainsi sélectionnées en fonction des critères indiqués dans l'annexe 1. L'échelle de subjectivité décrite en annexe a été développée, à partir de la notion de subjectification, par plusieurs auteurs (Pander Maat & Degand 2001, Pander Maat & Sanders 2001, Pit 2003). C'est la possibilité d'utiliser de tels outils – la liberté de choix concernant la manière d'aborder le corpus, et donc, encore une fois, le caractère *dynamique* de cette méthode – qui fait tout l'intérêt de l'étude sur corpus : pour chaque nouvelle étude, le linguiste peut adopter l'approche la plus idoine, et ne doit pas s'en tenir aux descriptions nécessairement plus statiques des ouvrages de référence.

2.3.1 Données pour le français Parce que

Notre corpus montre pour *parce que* un figement très progressif, qui ne semble pas achevé même en moyen français –

¹⁶ Parce que la subjectification est liée au degré d'implication du locuteur, et que le locuteur se met naturellement en avant dans les dialogues plus que dans le récit (voir Spooren et al., à paraître).

et pour cause, puisqu'il continue en français moderne¹⁷. Les graphies observées dans le corpus sont en effet très variées¹⁸, plus encore que ce que notent les grammaires : *par, por, pour + ce, ceo, ceu, che, cio, ice, içou + que, ke (qu', q', k'), parce, porce + que (qu')* (et même, avec un emploi légèrement différent, les formes *poruec* et *peruec*).

Dans notre corpus, pas plus que dans les ouvrages de référence, on ne trouve d'autres emplois pour *parce que* que celui de conjonction de subordination, en français médiéval. Ceci différencie clairement, comme on le verra plus bas, *parce que* de *perché* et *porque* en italien et espagnol. Ce qui les rapproche, en revanche, est leur sémantisme : en effet, *parce que* présente des emplois non seulement causaux, mais encore finaux et de manière.

Les emplois non causaux sont illustrés par les exemples 1 et 2 ci-dessous.

(1) *Sunent mil grailles por ço que plus bel seit.*

Ils font sonner de nombreuses cloches, pour que ce soit plus joli (*Chanson de Roland*, v. 1004, 11^{ème} siècle)

(2) *En grant peine nus mist Par ço quë il mangat Ço que Eve li dunat Sur le defens de Dé, Ulte sa volenté.*

Il nous a mis dans une situation bien douloureuse, en mangeant ce qu'Eve lui avait donné (*Comput*, v. 532-6, 12^{ème} siècle)

Les emplois causaux présentent divers degrés de subjectivité. Plus précisément, on trouve des emplois où *parce que* introduit la cause objective (la cause décrite correspond à la

¹⁷ Notamment avec la graphie *paske*, que l'on trouve désormais régulièrement sur internet : dans les pages en langue française, on trouve plus d'un million d'occurrences des graphies *pask(e)*, *parsk(e)* (pour 44 millions de *parce qu(e)* – recherche effectuée sur Google le 27 juin 2008 ; avec le moteur de recherche Yahoo, sur les pages en français, on trouve près de 750 000 occurrences des mêmes graphies, pour environ 100 millions de *parce qu(e)*). La graphie n'est pas le meilleur indice de figement, bien sûr, mais on voit bien que la construction s'est contractée progressivement, de *por ce que* à *paske*.

¹⁸ Ceci malgré le travail d'uniformisation des éditeurs : la BFM est en effet composée d'éditions critiques.

cause « réelle », voir annexe 1), comme dans l'exemple 3. La cause introduite peut être plus subjective, comme dans l'exemple 4 : il s'agit de la cause « volitive », où le locuteur indique la raison pour laquelle il a agi ; on s'éloigne de la cause « réelle » pour se rapprocher de la cause « subjective ». Enfin, on trouve des occurrences où la cause introduite est de nature épistémique : le locuteur y indique, comme dans l'exemple 5, pourquoi il pense ce qu'il énonce ; on est ici en plein dans la cause « subjective ». On ne trouve cependant pas d'emplois intersubjectifs au sens de Traugott & Dasher (ibid.), qu'on trouvera pour *perché* et *porque*.

(3) *Li anfes ploroit de grant fin por ce que n'avoit que mengier*

L'enfant pleurait de faim, parce qu'il n'avait rien eu à manger (*Roman de Renart*, branche XI, v. 11482-3, 13^{ème} siècle)

(4) *mort me fis en mi la voie por ce que trop grant fain avoie*

Je fis le mort sur la route car j'avais terriblement faim (ibid., branche X, v. 9769-70)

(5) Le monde despit et confoule **Par ce qu'ele** voit bien qu'il boule Et a boulez toz ses amis Et en enfer toz les a mis (*Les Miracles de Nostre Dame*, Gautier de Coincy, v. 3017-20, 13^{ème} siècle)

« Elle méprise et écrase le monde, car elle voit bien qu'il trompe et a trompé tous ses amis et les a tous envoyés en enfer »

2.3.2 Données pour l'italien *Perché*

Pour la conjonction *perché*, on trouve dans notre corpus les graphies suivantes : *perché*, *perche*, *perchè* ; *per/pèr che/chè/ché*.

En italien médiéval, *perché* est employé comme introducteur d'interrogation directe et indirecte ; on trouve également des emplois comme relatif (« la raison *pour laquelle* ») ou même nominalisés (*il perché*), comme l'illustrent les exemples 6-8 ci-dessous.

(6) Morte, **perché** m'hai fatta sì gran guerra, che m'hai tolta madonna, ond'io mi doglio? (Giacomo Pugliese, *Morte*, 13^{ème} siècle, p. 146)

« O Mort, **pourquoi** as-tu été si cruelle à mon égard, me prenant ma dame, pour ma plus grande douleur? »

(7) [chella] ... lascialo al tutto, non diciendoli le cagione perchè tu ti voli partire da lui

« [celle-là (= son amitié)] ... laisse-la tout à fait, sans lui dire les raisons pour lesquelles tu veux te séparer de lui »

(8) Li cavalieri li fecero cerchio intorno domandando il **perché** (Novellino, 13^{ème}, p. 312)

« Les chevaliers l'entourèrent, lui demandant le *pourquoi* (de son comportement) »

Lorsqu'il est employé comme conjonction de subordination, ce morphème peut introduire des propositions finales (ex. 9), ou concessives (ex. 10).

(9) Anche de' combattere co le mani, perchè non ti sia fatto forza (Andrea da Grosseto, 1268, livre II, chap. 49, p. 157)

« Tu dois aussi te battre à mains nues, pour éviter qu'on ne te fasse violence »

(10) Secondo che avvenne a santo Iob, lo quale, **perché** perdesse tutti gli sui figliuoli e tutti gli ben sui, [e] sostenne in sé molte tribulazione e pene nel corpo suo, sempre stette dritto e sempre ne rendé grazie a Dio (Andrea da Grosseto, 1268 L. 2, chap. 3, p. 207)

« Comme ce fut le cas pour Saint Job, qui, bien qu'il ait perdu tous ses fils et ses biens, et subisse lui-même, dans sa chair, de nombreux tourments et peines, resta toujours honnête et rendit toujours grâce à Dieu ».

Enfin, employé comme conjonction causale, *perché* introduit des causes diverses : cause objective (11), volitionnelle (12), épistémique (13) ; on trouve en outre des emplois intersubjectifs, où la clause qu'introduit *perché* justifie l'énonciation précédente, notamment lorsque cette dernière contient un ordre (14).

(11) ...eloquenzia avea più grande bisogno per lo male che faceano i folli arditi nelle cittadi, e **perché**

guastavano la cosa onestissima e dirittissima, cioè eloquenzia che ssi pertiene alle cose oneste e diritte (Brunetto Latini, Rettorica, 1260-61, p. 34)

« L'éloquence était en grand péril à cause du tort que lui causaient les fous téméraires dans les villes, et parce qu'ils corrompaient sa nature des plus honnêtes et franches, c'est-à-dire l'éloquence, qui appartient aux choses honnêtes et franches »

(12) Et imperciò ti dissi dagli amici, **perchè** Salamon disse : el corpo si diletta di buoni unguenti e diversi odori, et l'anima si ralegra delettosamente di buoni consigli dell'amico (Andrea da Grosseto, 1268, L. 2, chap. 21, p. 85)

« Et c'est pour cela que je t'ai parlé des amis, parce que Salomon a dit : comme le corps aime les bons onguents et les parfums, ainsi l'âme se réjouit des bons conseils de l'ami »

(13) Madonna sapientissima, tu ci ài prevenuti in benedizione di dolcezza : imperciò che quello que tu ài detto a' nnoi, noi lo dovavamo inprima dire a te. **Perchè** lo cominciamento de la discordia venne de la nostra stoltezza ; et così lo cominciamento de la concordia dovea venire da noi (Andrea da Grosseto, 1268, L. 2, chap. 49, p. 161)

« Ma dame pleine de sagesse, tu nous as précédés dans la bénédiction de ta douceur, car ce que tu nous as dit, c'était à nous de te le dire les premiers. Parce que c'est de notre sottise qu'est née la discorde, et le retour de la concorde devait de même venir de nous »

(14) Et non ti fidar troppo nel lor consiglio, **perchè** un savio disse : le femine vincono gli uomini ne'ma'consigli (Andrea da Grosseto, 1268, L. 3, chap. 23, p. 275)

« Et ne te fie pas trop à leur avis, parce qu'un sage a dit : les femmes sont supérieures aux hommes, dans les mauvais conseils »

2.3.3 Données pour l'espagnol Porque

La conjonction *porque* semble assez figée, dès les premiers textes. Les graphies rencontrées dans le corpus sont en effet peu variées : *por que*, *porque*, *por qué*, *porqué* (noter cependant que les limitations de l'interface sont peut-être en cause – voir annexe 2).

Comme *perché* en italien, *porque* sert entre autres à introduire les interrogations directes et indirectes (ex. 15), et connaît également des emplois comme relatif (ex. 16) et des emplois comme introducteur de complétives (ex. 17).

(15) ¿E **porqué** son vuestras casas iguales ? (1250
Anónimo Bocados de oro)

« Et **pourquoi** vos maisons sont-elles identiques ? »

(16) mas tengo por seso et por consejo, si me tú quisieres creer et los que contigo son, una cosa **porque** fio en Dios que vençeremos nuestro enemigo et tornaremos al mejor estado que nunca fuemos (*Calila e Dimna*, 1251, 343)

« le mieux à mon avis est que toi et tes compagnons me croient quant à la **raison pour laquelle** je garde foi en Dieu et pense que nous vaincrons notre ennemi et reviendrons au meilleur état dans lequel nous avons jamais été »

(17) mas es acaesçido tanto de mal, que me non plaze **porque** estás así ; et non es ninguno que mejor me pueda librar desto en que esté et deste tan grant peligro en que esté salvo tú (*Calila e Dimna*, 1251, 268)

« Mais il y eut tant de malheurs qu'il ne me plaît pas **que** tu restes ainsi ; et il n'y a personne qui puisse mieux que toi me délivrer de cette situation et du terrible danger dans lequel je me trouve »

Les emplois comme conjonction causale sont, encore une fois, assez variés, comprenant la cause objective (ex. 18), la cause volitionnelle (ex. 19), la cause épistémique (ex. 20) et même des emplois intersubjectifs, comme *perché* (ex. 21).

(18) Dixo la madre del león : - Esto que tú vees estar al león triste et cuidadoso non es sinon **porque** te ha dexado sano et salvo fasta oy, faziéndole tú engaño et

enridándole con tu mestura et con tu falsedad para matar a Senseba (*Calila e Dimna*, 1251, 182)

« La mère du lion dit : - Si tu vois le lion triste et peiné ce n'est que **parce qu'**il t'a laissé sain et sauf jusqu'ici, alors que toi tu l'as trompé et que tu t'es moqué de lui avec tes tours et ta fausseté, pour tuer Senseba »

(19) Et el que non faze bien sinon por aver bien et por ganar alguna alegría deste siglo et algund pro es tal en esto commo el paxarero que echa los granos a las aves non por les fazer ayuda sinon **porque** quiere ganar (*Calila e Dimna*, 1251, 208)

« Et celui qui ne fait du bien que pour recevoir des bienfaits et tirer quelque joie et profit de ce monde est comme l'oiseleur qui donne des graines aux oiseaux, non pour leur venir en aide, mais **parce qu'**il veut en tirer bénéfice »

(20) Dixo el rey : - ¿Et qué viste dese por que entendiste que era de buen seso ? Dixo : - Por dos cosas : la una **porque** consejava mi muerte et la otra **porque** consejava lealment[e] a su señor et le non çelava nada (*Calila e Dimna*, 1251, 252)

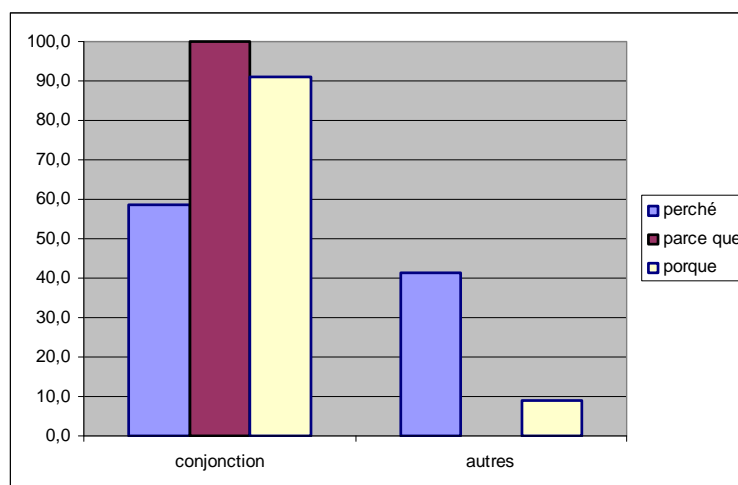
« Le roi dit : - Et que l'avez-vous vu faire, qui vous a fait penser qu'il était malin ? Il répondit : - Pour deux raisons : d'abord **parce qu'**il conseillait ma mort, ensuite **parce qu'**il conseillait loyalement son seigneur, sans rien lui cacher »

(21) Ya oí lo que dexiste muy bien, enpero véote estar así commo triste et remiénbraste de cosas que tienes en el coraçón; et **porque** aquí eres connusco en ageno lugar non seas de tal acuerdo, et déxate ende [...] (*Calila e Dimna*, 1251, 217-218)

« J'ai très bien entendu ce que tu m'as dit, mais je vois que tu es triste et que tu te remémore des choses que tu caches en ton cœur; et **puisque** tu es ici avec nous, dans un lieu qui t'est étranger, ne sois pas de cette humeur, mais libère-t'en »

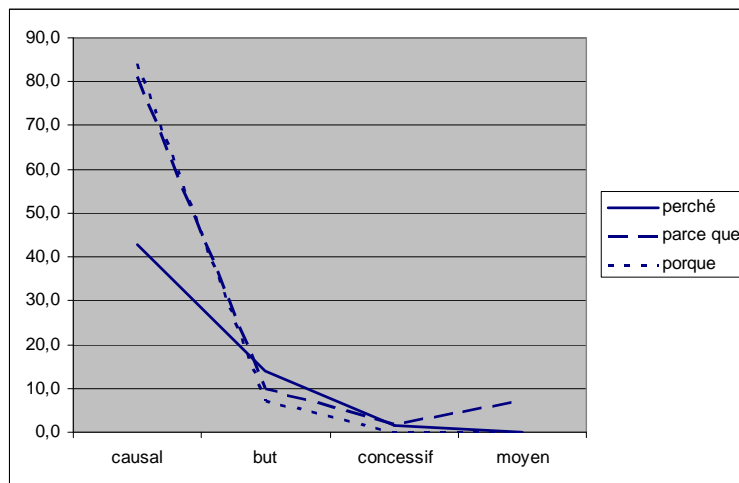
2.4 Résultats de l'étude comparative

Si l'on compare dans un premier temps les emplois des trois constructions d'un point de vue syntaxique, on peut observer une gradation de *parce que* (le plus souvent employé comme conjonction) à *perché* (le plus souvent employé autrement), avec *porque* entre les deux, comme l'illustre le graphique 1 ci-dessous.



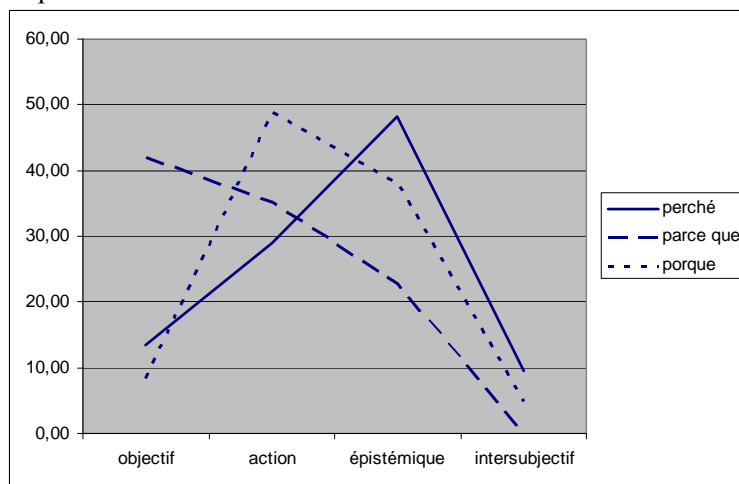
Graphique 1 : emploi des trois conjonctions dans nos corpus.

Les sens de ces trois constructions lorsqu'elles sont employées comme conjonctions diffèrent également assez nettement, comme le montre le graphique 2 ci-dessous : outre les emplois causaux, *perché* est assez souvent final, tandis que *parce que* indique souvent le moyen ou la manière ('moyen' dans le graphique 2) ; c'est *porque* qui est le plus souvent causal.



Graphique 2 : sémantisme des trois conjonctions dans nos corpus.

Enfin, si l'on observe en détail les emplois causaux, il y a encore une fois un écart assez grand. En particulier, on peut remarquer que la gradation est encore une fois dans le même sens, avec *parce que* plus « objectif », *perché* plus « intersubjectif », et *porque* entre les deux, comme le montre le graphique 3 ci-dessous.



Graphique 3 : emplois causaux des trois conjonctions dans nos corpus.

On pourra faire l'hypothèse d'un lien entre la subjectification de *perché* et la fréquence de ses emplois comme interrogatif, donc en situation énonciative marquée justement par l'interaction. Le fait que *porque* soit également plus subjectif que *parce que*, alors qu'il est lui aussi employé comme interrogatif, nous semble confirmer cette hypothèse. Il faudra cependant, pour s'en assurer, répéter l'expérience sur un corpus plus étendu.

4. Conclusion

Le but de cet article était de plaider pour la complémentarité des approches traditionnelle (philologique) et 'assistée' (corpus électroniques). La première étape est nécessairement la consultation des ouvrages de référence, sans lesquels la recherche sur corpus risque de laisser de côté des variantes (graphiques ou autres) importantes. Nous avons montré qu'une fois passée cette première étape, l'apport du corpus numérisé – même avec un échantillonnage limité – est d'offrir une première évaluation de la fréquence d'un morphème donné, de la part respective de ses différents emplois, et de faciliter ainsi la comparaison entre plusieurs morphèmes. L'étude sur corpus permet ainsi d'étudier de manière plus fine le renouvellement des conjonctions de subordination entre le latin et les langues romanes, comme nous l'avons vu pour *parce que*, *perché* et *porque*, et de suggérer que l'origine des deux dernières (*perché* et *porque* sont à l'origine, semble-t-il, des adverbies interrogatifs) pourrait expliquer leurs emplois plus subjectifs.

Il faut noter enfin que les exigences formulées en 1.2, sur la constitution et l'utilisation du corpus, ne sont pas toujours faciles à satisfaire. Il nous semble avoir rempli la plupart de ces exigences : corpus adapté à l'étude (qualité et taille), critères linguistiques explicités, analyse quantitative, lisibilité des occurrences, gestion du bruit et du silence. Pour d'autres, la méthode reste à affiner : diversité des textes, reproductibilité de l'étude.

Références bibliographiques

- Biber B., Conrad S. & Reppen R., (1998). *Corpus linguistics – Investigating Language Structure and Use*. Cambridge University Press : Cambridge.
- Degand L. & Bestgen Y. (2004). « Connecteurs et analyses de corpus : de l'analyse manuelle à l'analyse automatisée », in S. Porhiel & D. Klingler (éds.) *L'Unité texte*. Pleyben : Perspectives, 49-73.
- Degand L. & Pander Maat H. (2003). « A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale », in A. Verhagen & J. van de Weijer (éds.) *Usage based approaches to Dutch*. Utrecht : LOT, 175-199.
- de Haan F. (2002). « Strong modality and negation in Russian », in R. Reppen, S. Fitzmaurice & D. Biber, *Using Corpora to Explore Linguistic Variation*. Amsterdam / Philadelphie : John Benjamins, 91-110.
- Fagard B. (2006). *Evolution sémantique des prépositions dans les langues romanes : illustrations ou contre-exemples de la primauté du spatial ?* Thèse de doctorat, Université Paris 7 & Università Roma 3 (disponible sur le site du RISC).
- Habert B. (2000). « Des corpus représentatifs : de quoi, pour quoi, comment ? », in M. Bilger (éd.), *Cahiers de l'université de Perpignan*, 31, 'Linguistiques sur corpus. Etudes et réflexions'. Perpignan : Presses universitaires de Perpignan, 11-58.
- Kennedy G. (2002). « Variation in the distribution of modal verbs in the British National Corpus », in R. Reppen, S. Fitzmaurice & D. Biber, *Using Corpora to Explore Linguistic Variation*. Amsterdam / Philadelphie : John Benjamins, 73-90.
- Marchello-Nizia C. (1995). *L'évolution du français : ordre des mots, démonstratifs, accent tonique*. Paris : Colin.
- Oakey D. (2002). « Formulaic language in English academic writing », in R. Reppen, S. Fitzmaurice & D. Biber, *Using Corpora to Explore Linguistic Variation*. Amsterdam / Philadelphie : John Benjamins, 111-129.

- Pander Maat H. & Degand L. (2001). « Scaling causal relations and connectives in terms of Speaker Involvement », *Cognitive Linguistics* 12-3, 211-245.
- Pander Maat H. & Sanders T. (2001). « Subjectivity in causal connectives : An empirical study of language in use », *Cognitive Linguistics*, 12/3, 247-273.
- Pit M. (2003). *How to express yourself with a causal connective. Subjectivity and causal connectives in Dutch, German and French*. Amsterdam / New York : Rodopi.
- Prévost S. (2005). « Exploitation d'un corpus de français médiéval : enjeux, spécificités et apports », in A. Condamines (éd.) *Sémantique et corpus*. Paris : Hermès / Lavoisier, 147-176
- Rossari C. & Jayez J. (1996). « *Donc* et les consécutifs. Des systèmes de contraintes différentiels », *Linguisticae Investigationes*, XX-1, 117-143.
- Simon A.C & Degand L. (2007). « Connecteurs de causalité, implication du locuteur et profils prosodiques. Le cas de *car* et de *parce que* », *Journal of French Language Studies*, 17, 323-341.
- Spooren W., Sanders T., Huiskes M. & L. Degand. (à paraître). « Subjectivity and Causality: A Corpus Study of Spoken Language », in J. Newman & S. Rice (éds.), *Conceptual Structure in Discourse and Language*.
- Sweetser E. (1990). *From etymology to pragmatics. Metaphorical and cultural aspects of semantic structure*. Cambridge : Cambridge University Press.
- Traugott E. (2003). « From subjectification to intersubjectification », in R. Hickey (éd.), *Motives for language change*. Cambridge : Cambridge University Press, 124-139.
- Traugott E. & Dasher R. (2002). *Regularity in Semantic Change*. Cambridge : C.U.P.
- Vincent D. (1993). *Les ponctuels de la langue et autres mots du discours*. Québec : Nuit Blanche.
- Bases de données*

Base de Français Médiéval (Laboratoire ICAR, ENS-LSH / CNRS).

Base textuelle du moyen français (Laboratoire ATILF, CNRS).

Base Frantext (Laboratoire ATILF, CNRS).

Base Corde (*CORpus Diacrónico del Español*, Real Academia Española).

Base OVI (*Opera del Vocabolario Italiano*, consortium ItalNet).

Ouvrages de référence

Alonso M. (1986). *Diccionario medieval español (desde las Glosas Emilianenses y Silenses (s. X) hasta el siglo XV)*. Universidad Pontificia de Salamanca.

Buridant C. (2000). *Grammaire nouvelle de l'ancien français*. Paris : Sedes.

Corominas J. & Antonio Pascual J. (1980-1991). *Diccionario crítico etimológico castellano e hispánico*. Madrid : Gredos.

Diccionario histórico de la lengua española. Casares J. (dir.) & Ramirez S. (red.). (1960). Madrid : Real academia española.

Penny R. (2005). *Gramática histórica del español*. Barcelone : Ariel Lingüística.

Rohlf G. (1954). *Historische Grammatik der italienischen Sprache und ihrer Mundarten*. Verlag Bern.

Tobler A. & Lommatzsch E. (1925). *Altfranzösisches Wörterbuch*. Berlin / Wiesbaden : Weidmannsche Buchhandlung.

Tesoro della Lingua Italiana delle Origini. (Nuovo TLIO : <http://tlio.ovi.cnr.it/TLIO/>)

Trésor de la langue française. (1971-1994). Sous la direction de Paul Imbs puis de Bernard Quemada. Paris : Éd. du CNRS puis Gallimard.

Vocabolario de gli academici della Crusca. (1612). Venezia : Giovanni Alberti.

Zauner A. 1921. *Altspanisches Elementarbuch*. Heidelberg : Carl Winter.

ANNEXE 1 DEGRÉS DE SUBJECTIFICATION

On peut mettre en évidence le degré de subjectification d'un terme à l'aide de différents marqueurs ou indices, principalement l'emploi du temps présent et de la première ou deuxième personne du singulier (voir Pander Maat & Degand 2001, Pander Maat & Sanders 2001, Pit 2003). On peut ainsi établir une échelle de subjectivité pour différents emplois d'un même terme, ou plus généralement pour différents types de causalité. Nous avons ainsi défini quatre degrés principaux de subjectivité dans la causalité, de la causalité objective aux actes de langage. Le schéma 1 ci-dessous et les exemples (a-e) correspondants illustrent et précisent cette approche.

cause **objective** : 1. met en relation de manière factuelle deux « états des choses » (a)

> cause **volitionnelle** : 2. explique/justifie la décision / l'action du locuteur (b-c)

> cause **épistémique** : 3. explique/justifie l'état d'esprit du locuteur (d-e)

> **intersubjective** : 4. explique/justifie un acte de discours (f)

Schéma 1 : échelle de subjectivité pour différents types de causalité

a. En la teste ad e dular e grant mal : Rumput est li temples, por ço que il cornat.

« Sa tête lui fait horriblement mal : il a la tempe rompue, parce qu'il a sonné le cor ». (*Chanson de Roland*, v. 2102, 11^{ème} siècle)

b. Li duze per, Li paien controverent Les nuns que as jurz dunerent Li premier, que apelum Diemeine par num, Al soleil le dunerent, E sun num li poserent Pur ço que enluminout Le mund e nuit chazot.

« Il lui donnèrent son nom parce qu'il enluminait le monde » (*Comput*, v.429, 12^{ème} siècle)

c. Reposed sei quar lassét sunt.

« Ils se reposent, car ils sont las » (*Saint Brendan*, v. 638, 12^{ème} siècle)

d. Sire, dist Guenes, ço ad tut fait Rollant ! Ne l'amerai a trestut mun vivant, Ne Oliver, por ço qu'il est si cumpainz.

« Je le haïrai jusqu'à ma mort, et Olivier aussi, parce qu'il lui est proche » (*Chanson de Roland*, v. 324, 11^{ème} siècle)

e. Forment m'en poise, quar mout l'avoie amé ; Mes par mon chief ja sera comparé.

« Cela me pèse fort, car je l'aimais tendrement ». (*Aliscans*, laisse XLVI, 12^{ème} siècle)

f. Ensembl'od lui i ferrunt veirement. De ço qui calt ? car ne lur valt nient. Demurent trop, n'i poedent estre a tens.

« Mais qu'importe ? Car cela ne leur sert à rien » (*Chanson de Roland*, v. 1840, 11^{ème} siècle)

ANNEXE 2 Présentation des bases de données

1 Base « BFM »

La BFM (Base du Français Médiéval) a été constituée par l'ex-ELI (Equipe linguistique et informatique), co-dirigée par Christiane Marchello-Nizia et Benoît Habert, aujourd'hui UMR 5191 ICAR (CNRS – ENS-LSH Lyon). Elle a été créée afin de servir de base de données aux chercheurs travaillant sur cette période (linguistes, littéraires, historiens), en complément du corpus du DMF (Dictionnaire du moyen français), qui contient des textes à partir de 1330. Elle est donc formée principalement de textes d'ancien français, mais contient également des textes plus tardifs, du 14^{ème} au 16^{ème} siècle. Pour l'ancien français, les 9^{ème}-10^{ème} siècles sont peu fournis, parce que peu de textes nous sont parvenus de cette époque ; il y a davantage de textes pour les 11^{ème}-13^{ème} siècles.

Les genres présents dans la base sont variés : il y a principalement des textes littéraires, qui forment en effet la majorité des textes disponibles pour cette période, mais aussi des textes plus techniques, qu'il s'agisse de textes historiques, de textes juridiques (chartes) ou de recueils de coutumes ; théâtre et poésie sont peu représentés. Ces textes ont été obtenus pour la plupart par saisie et scannage systématique à partir des

éditions de référence¹⁹, élaborées selon des principes peu interventionnistes et sans vérification supplémentaire sur les manuscrits. Elle contient près de quatre-vingts textes, pour environ 3 millions de mots. Une partie de la base est catégorisée, mais il n'y a pas de lemmatisation.

On peut l'exploiter à l'aide de deux logiciels : SATO, conçu par François Daoust (centre ATO, UQAM, Montréal), moteur de filtrage qui permet des recherches fines en contexte s'appuyant sur l'utilisation d'expressions régulières, et Weblex, conçu par Serge Heiden, accessible en ligne sur internet, qui met en œuvre l'ensemble de la méthodologie lexicométrique, et permet la génération rapide de concordances, le calcul de cooccurrences, des analyses statistiques dans des environnements hypertextuels, ainsi que des indexations ou relevés d'occurrences. Associé aux « requêtes CQP » (*corpus query processor*, langage formel d'interrogation de la base), il constitue un outil très performant, rendant possible la recherche de formes complexes à l'aide de caractères de choix (*,., ?, [], |, etc.) ; mais nécessite un apprentissage/une prise en main préalable. Il permet d'autre part de rechercher, outre des formes simples, des expressions complexes. Toutes ces requêtes peuvent être effectuées aussi bien sur les textes étiquetés que sur les autres. Enfin, tous les renseignements paratextuels nécessaires sont fournis en en-tête.

Il est possible de plus d'effectuer des recherches sur une partie du corpus seulement, en sélectionnant un genre, un siècle, une œuvre, etc.

Enfin, élément d'importance majeure pour la recherche en linguistique, la taille du contexte est modulable (bien que limitée), ce qui est rarement le cas dans les bases de données, comme nous le verrons. On notera enfin que la BFM est

¹⁹ Ce qui signifie qu'ils ont pu être *corrigés* par un éditeur moderne : cela pose le problème de la méthode de création du corpus, à partir d'éditions nouvelles – ce qui permet d'être plus sûr du texte mais suppose un travail énorme d'édition et de révisions – ; ou bien à partir d'éditions déjà faites, ce qui pose un problème de fidélité (on ne peut pas connaître avec précision le degré de fiabilité d'une édition sans avoir recours au manuscrit) mais permet d'étendre plus rapidement son corpus.

accessible à tous, après inscription (gratuite) et signature d'une 'charte'.

2 Base « OVI »

La base OVI (Opera del vocabolario italiano) est issue d'une collaboration du consortium ItalNet, fondé par le *Centro di Studi Opera del Vocabolario Italiano* (centre de recherches du CNR italien) établi à Florence, le *projet ARTFL* du département de langues romanes de l'université de Chicago, le *William and Katherine Devers Program in Dante Studies* de l'université de Notre Dame (Indiana, Etats-Unis) et le *département d'études italiennes* de l'université de Reading (Royaume-Uni). Elle a été créée pour contribuer à la compilation du dictionnaire historique *Tesoro della lingua italiana delle origini* (TLIO).

Les textes la constituant sont au nombre de 1960 ; ce sont des textes en vers et en prose, datant pour la majorité d'avant 1375. Y sont compris aussi bien les monuments de la littérature italienne (Dante, Pétrarque, Boccace) que des textes littéraires moins connus et des textes plus techniques, notamment historiques ou juridiques. La taille de la base est de 22,3 millions de mots.

L'interface de recherche en ligne, PhiloLogic, a été élaborée par Marc Olsen dans le cadre du projet ARTFL, à l'université de Chicago. Elle permet une série d'opérations qui nous intéressent : partitionnement du corpus pour limiter la recherche à une période donnée, à un ou plusieurs auteurs ou textes ; recherche en contexte et index des occurrences ; extraction de textes et établissement de la bibliographie.

Pour la recherche d'une forme précise ou d'une série de formes, on a recours à un ensemble d'expressions spécialisées relativement standards et faciles d'utilisation, qui permettent notamment la neutralisation des diacritiques (toute majuscule équivaut à n'importe quelle forme de la lettre en question : E = E, È, É, Ê, Ë, é, e, è, ë, ê ...), le choix entre plusieurs caractères ([a-d] = a, b, c ou d), la troncation (. * = n'importe quelle série de caractères, x.* = n'importe quelle série de caractères

commençant par « x »), etc. On peut rechercher des suites de mots, ou bien plusieurs expressions différentes (x|y = « rechercher x ou y »), ce qui est également utile.

On peut noter, enfin, deux fonctionnalités intéressantes : la possibilité de choisir le format des résultats, avec une ligne de texte par occurrence (format KWIC ou Key Word in Context) ou bien davantage (format « concordance », avec contexte plus large) ; celle d'obtenir la fréquence d'une forme par œuvre ou par auteur.

3 Base « CORDE »

La Real Academia a lancé dès 1993 un projet de corpus électronique, ayant pour but de servir à la constitution d'un dictionnaire historique, ainsi que de servir d'outil de connaissance de la langue, outil philologique accessible à tous. Il contient une section synchronique, le CREA (*Corpus de referencia del español actual*), que nous ne présenterons pas en détail ici, et une section diachronique, le CORDE ou Corpus diacrónico del español, dont nous présentons ici les caractéristiques principales.

Le Corde est formé de textes écrits en espagnol (et en latin pour les textes les plus anciens), depuis les origines de la langue jusqu'en 1975 (avec la partition suivante : origines-1491, 1492-1712, 1713-1974), dans l'ensemble des pays hispanophones. Ces textes sont diversifiés du point de vue du genre : littéraires et techniques, en prose et en vers ; textes narratifs, lyriques, dramatiques, scientifiques, historiques, juridiques, religieux, journalistiques. La partition la plus ancienne, des origines à 1492, contient quelque 26 millions de mots. Les textes ont été obtenus par scannage de livres ou bien par édition directe sous format électronique.

L'accès à la base est libre et se fait par l'intermédiaire d'une page internet, accessible sans inscription.

L'interface mise en place permet d'effectuer différents types de recherches, en incluant des critères spécifiques : délimitation d'une période donnée (avant *x*, après *y*, entre *x* et *y*), choix d'un ou plusieurs genres, d'un ou plusieurs auteurs ou

Benjamin FAGARD

œuvres. La lemmatisation de la base est en cours, et elle n'est pas catégorisée morpho-syntaxiquement : aucune recherche par lemme ou par catégorie syntaxique n'est donc possible. On pourra déplorer par ailleurs l'absence de fonctionnalités proprement « linguistiques », qui permettraient d'effectuer des requêtes plus fines : il n'est possible en effet que de rechercher une expression donnée, sans les options de troncation, d'alternative et autres habituellement disponibles pour ce genre de base.

On notera, enfin, la possibilité d'obtenir les résultats sous deux formes : l'occurrence dans une ligne de texte, ou bien dans le paragraphe – voire la page – où elle se situe.