Autour des prépositions en, dans, dedans. Vers une approche diachronique sur corpus outillé

Denis Vigier

UNIVERSITÉ TOULOUSE JEAN JAURÈS

Volume 1

MÉMOIRE

présenté pour l'obtention de l'

HABILITATION À DIRIGER DES RECHERCHES

Discipline : Sciences du langage
Soutenu publiquement le 13 novembre 2017

par

Denis Vigier
Maître de Conférences à l'Université Lumière Lyon 2

Autour des prépositions *en, dans, dedans*. Vers une approche diachronique sur corpus outillé

JURY

M. Michel AURNAGUE,	Directeur de Recherche CNRS – Garant
M. Peter BLUMENTHAL,	Professeur émérite, Université de Cologne - Rapporteur
M. Michel CHAROLLES,	Professeur émérite, Université Sorbonne-Nouvelle, Paris 3
Mme Cécile FABRE,	Professeur, Université Toulouse Jean Jaurès - Rapporteur
Mme Nathalie FOURNIER,	Professeur émérite, Université Lumière Lyon 2 - Rapporteur
M. Gilles SIOUFFI,	Professeur, Université Paris-Sorbonne, Paris 4

La linguistique cumulative pense que l'on peut et que l'on doit progresser à partir des acquis antérieurs. Modestement sans doute, à petits pas le plus souvent.
G. Kleiber (1994, 19).

Toute ma gratitude va à Michel Aurnague qui m'a fait l'amitié d'accompagner de ses conseils la rédaction de ce mémoire. Son souci de rigueur et de précision, les perspectives qu'il m'a ouvertes sur la composante fonctionnelle du sens des prépositions dans le cadre de la linguistique cognitive, sa très grande modestie enfin, m'ont énormément apporté.

Je remercie aussi vivement les rapporteurs, Peter Blumenthal, Cécile Fabre et Nathalie Fournier ainsi que les autres membres du jury, Michel Charolles et Gilles Siouffi.

J'éprouve enfin beaucoup de reconnaissance envers celles et ceux - collègues et ami-e-s - qui, depuis ma thèse, m'accompagnent de leur estime. Puissent-ils se reconnaître dans ces lignes.

TABLE DES MATIÈRES

PREMIÈRE PARTIE
Construire l'identité sémantique d'une préposition en synchronie du français contemporain. Le cas de « *en* »

TROISIÈME PARTIE
Les prépositions *en, dans, dedans* du XVI^{e} s. au XX^{e} s. Approche statistique en corpus

PREMIÈRE PARTIE

Construire l'identité sémantique d'une préposition en synchronie du français contemporain. Le cas de « *en* »

RAPPEL DE LA TABLE DES MATIÈRES DE LA PREMIÈRE PARTIE

Introduction

On évoquera d'abord (§1) deux points de portée générale : le premier concerne l'unité de la classe des prépositions et leur statut de relateur ; le second, d'ordre méthodologique, touche à la notion d'identité sémantique d'une préposition. On s'attachera ensuite (§ 2) à définir l'identité de *en* à la lumière des travaux déjà conduits sur cette préposition et des réflexions développées dans la section précédente. Cette partie s'achèvera (§ 3) par l'examen de deux emplois de cette préposition. Le premier conduira à argumenter *contra* D. Leeman & C. Vaguer (2014) qui proposent d'intégrer le trait aspectuel perfectif dans l'identité sémantique de *en* ; le second permettra de cerner les contours d'un usage encore peu étudié de cette préposition.

1. Repères et propositions méthodologiques

On propose de s'arrêter, dans les lignes qui suivent, sur quelques notions à nos yeux décisives pour l'étude de la sémantique des prépositions : celles de relateur, d'identité et d'invariance sémantiques des polysèmes grammaticaux, d'approches « verticale » *versus* « horizontale » de la sémantique prépositionnelle, ...

1.1. La préposition comme classe grammaticale

Les paliers d'analyse que sont la morphologie, la sémantique et la syntaxe permettent de définir l'unité et l'unicité de la classe des prépositions en français. Au premier palier[1] s'impose leur invariabilité qu'elles possèdent en partage avec la classe des conjonctions et des adverbes[2]. Au niveau sémantique s'impose la notion de « relateur » traditionnellement affectée aux adpositions[3] et aux formes casuelles marquées (voir par ex. C. Hagège, 1982, [2013] ; B. Pottier 1974). Quant à la syntaxe, elle fournit le faisceau le plus étoffé de traits constitutifs de la classe.

1.1.1. Unité de la classe des prépositions : critères syntaxiques

1.1.1.1. Le cas du français contemporain

Pour le français moderne, on dispose de travaux récents (entre autres D. Leeman (éd.) (2006-2007, 2008), L. Melis (2003)) qui discutent les critères syntaxiques mobilisés pour définir la catégorie des prépositions en français et s'interrogent sur leur pertinence lorsqu'on les passe au crible des emplois en discours. On examinera les quatre critères suivants :

1. La préposition constitue la tête syntaxique du syntagme prépositionnel.
2. Elle est pourvue d'une valence propre.

---

[1] On soulignera après B. Fagard & W. de Mulder (op.cit.: 14-15) que C. Di Meola (2000 : 146), outre la propriété morphologique d'invariablilité, propose aussi pour les prépositions prototypiques celle de *brièveté*.

[2] Pour les adverbes, se pose cependant la question de la variabilité de *tout* notamment.

[3] Comme D. Creissels (2014 : 3) nous distinguons la « morphologie casuelle » de « l'adposition » qui tous deux permettent un *marquage* des rôles syntaxiques : « D'une langue à l'autre le codage des rôles syntaxiques peut faire appel au *rangement linéaire* des constituants nominaux, au *marquage* des constituants nominaux au moyen de la morphologie casuelle ou d'adpositions, ou à *l'indexation* des participants qu'ils représentent. »

3. Son complément n'est pas nécessairement un SN (ou l'un de ses équivalents fonctionnels) mais toute préposition possède dans sa valence la possibilité de construire un SN.
4. En français, la préposition précède linéairement son complément.

Ces quatre critères se vérifient pour la plupart des prépositions mais il existe à chaque fois des « exceptions » qui conduisent à adopter une conception flexible de la notion de catégorie.

Soit le critère 1 - désormais, [C1] : on peut, à la suite de L. Melis (*op. cit.* : 139), définir la tête syntaxique d'un groupe comme

« *(i)*[4] détermin[ant] en principe la nature des unités qui peuvent apparaître dans le groupe, *(ii)* le mode de construction de ces unités et *(iii)* les propriétés sémantiques auxquelles elles doivent répondre. En outre, *(iv)* la tête donne au groupe dans son ensemble sa catégorie (…) et *(v)* l'intégration du groupe dans une structure plus vaste se fait par référence aux propriétés de la tête qui est sélectionnée et construite par la tête du groupe supérieur dans lequel elle s'intègre. »

On peut illustrer cet ensemble de propriétés au moyen de la préposition *en* qui constituera notre fil rouge dans cette première partie. Cette préposition détermine en effet :

(i) la nature des unités susceptibles d'apparaître dans sa complémentation : par ex., l'usage de l'infinitif y est prohibé ;
(ii) le mode de construction des unités qu'elle régit : par ex., elle accepte de régir une complétive à condition que celle-ci soit construite avec le relais pronominal *ce* (voir H. Bat-Zeev Shyldkrot, 2008) et non directement.
(iii) Sur le plan sémantique, les propriétés des unités qu'elle régit ont donné lieu à de nombreuses études (entre autres, J.-J. Franckel & D. Lebaud, 1991 ; D. Leeman 1995 ; B. Martinie & D. Vigier 2013).
(iv) *En* confère au groupe entier le statut de syntagme prépositionnel (d'où sa capacité par ex. à compléter un nom).
(v) Ce SP peut à son tour être sélectionné par une tête externe – par ex. un verbe qui en fait son complément argumental : *la cargaison consiste en bananes* (ex. emprunté à D. Leeman, 1998 ; sur la complémentation verbale au moyen de *en*, voir I. Khammari, 2008).

En constitue bien la tête syntaxique des SP qu'elle construit.

Pour la vérification des critères [C2] et [C3] relatifs à la valence, on se reportera pour l'essentiel à C. Vaguer (2008) qui propose un tableau récapitulatif de la valence des prépositions simples du français[5], et aux analyses de L. Melis (*op. cit.*, 17-18) qui note que la combinatoire de chaque préposition ne devrait être, en bonne méthode, définie qu'au niveau de ses emplois[6].

[4] C'est nous qui ajoutons les items (i), (ii) etc. de manière à rendre plus clair la suite de notre propos.
[5] Tableau incomplet cependant à nos yeux : la combinatoire avec les adjectifs n'y est pas envisagée. Par ex. : *« Pour basse, la Loire, elle l'est »* (P. Cadiot, 1991 : 113), « *Il passe pour intelligent / laissé pour mort »* (D. Leeman, 2006 : 2).
[6] On peut par ex. observer que la possibilité qu'a *pour* d'être suivi ou non d'un nom nu est liée à certains des emplois syntactico-sémantiques du SP: on comparera ainsi : *il a été exclu pour (*son* + *Ø) dopage* [ajout de sens causal] et *il travaille pour (son* + ** Ø) avenir* [ajout de sens final].

Le critère 4 [C4] enfin (« en français, la préposition précède linéairement son complément ») est massivement vérifié en français pour les prépositions et notamment par *en*[7] mais peut être remis en cause pour certaines structures du français qui font figure d'exceptions. Par ex. :

(1) *Il a reçu* ***coup sur coup*** *deux appels de Pierre.*
(2) *Il a glosé le texte* ***mot à mot****.*
(3) *Je suis d'accord avec vous,* ***à une nuance près****.*
(4) *Il a neigé* ***plusieurs jours durant****.*

L. Melis (*op. cit.,* 22-23) montre que les structures en interposition illustrées[8] dans (1)(2) forment un seul constituant, la préposition assurant la liaison entre les deux noms et l'intégration de l'ensemble dans la phrase. Quant à (3), il illustrerait un cas de circumposition (*à… près)*, rarissime en français. L'énoncé (4) enfin, emprunté à D. Leeman (2006 : 10), signale la coexistence en français moderne de deux étapes (par comparaison avec « *durant plusieurs années* ») dans le processus de grammaticalisation au sein de la chaîne allant des verbes aux prépositions (voit B. Fagard, 2006 : 110-111 & *passim*).

L'existence d'exceptions pour le quatrième critère syntaxique, illustrée par (1) à (4), se vérifie tout autant pour les trois autres critères. Pour [C1], L. Melis (2001 : 13-14 ; 2003 : 33-34) et D. Leeman (*op. cit.*) signalent le cas des prépositions a-sélectives *sauf* et *excepté* qui ne gouvernent pas le mode de construction de leur complément (cf. critère [C1], item (*ii*)). Ce caractère a-sélectif a pour conséquence d'ôter à ces prépositions toute valence propre dans leurs emplois exceptifs, invalidant du même coup les critères [C2] et [C3].

Ainsi, dans les exemples suivants

(5) *Il a tout lu sauf* ***ø*** *ce document.*
(6) *Il a pensé à tout sauf* ***à*** *ce document.*
(7) *Il se souvient de tout sauf* ***de*** *ce document.*

la sélection d'une construction directe du SN *ce document* derrière *sauf* ou le choix de la préposition *à* ou *de* dans les SP *à/de ce document* sont gouvernés non par la préposition mais par la construction des verbes qui la précèdent (*lire, penser, se souvenir*).

Autrement dit, pour chaque critère syntaxique énoncé *supra*, on peut trouver des prépositions qui ne s'y conforment pas. Faut-il pour autant révoquer en doute la catégorie entière ? Une solution[9] à cette difficulté consiste à adopter une conception scalaire de la « prépositionnalité » qui rend possible une approche flexible de la catégorie. Dans le cadre théorique qu'offre la sémantique du prototype (dans sa version standard si l'on suit G. Kleiber, 1990[10]), nous distinguerons un « centre organisateur » de la catégorie réunissant

[7] Que *en* vérifie l'ensemble des traits morphologiques, sémantiques et syntaxiques cités ici – et d'autres : voir C. Di Meola 2000 : 146 – ne peut guère étonner : si l'on en croit cet auteur, *en* constitue l'une des prépositions les plus prototypiques du français (voir aussi B. Fagard & W. de Mulder, 2007).
[8] Pour les exemples donnés, on observera qu'il s'agit de structures (semi-)figées.
[9] Solution choisie par J.-P. Lagarde (1988 : 105-106) et plus récemment par C. Guimier (2007), L. Melis (*op. cit.*) ou B. Fagard (2006).
[10] La version standard de la sémantique du prototype « postule une organisation interne des catégories dans laquelle le prototype, conçu comme le meilleur représentant de la catégorie, joue un rôle prédominant. Il fournit directement le principe d'organisation et de représentation des catégories : les catégories sont structurées selon une échelle de prototypicalité qui mène des meilleurs représentants, placés au centre de la catégorie, aux moins

« tous les traits caractéristiques de la préposition centrale prototypique » pour envisager ensuite « chaque préposition en termes de distance relativement à ce centre » (C. Guimier, *op. cit.*: 98). La question décisive ne consiste plus alors à déterminer si tel ou tel groupe d'unités linguistiques appartient ou non à telle ou telle catégorie[11], mais à quelle distance de tel ou tel noyau catégoriel prototypique ce groupe se trouve situé. Conséquemment, les prépositions situées à distance du centre organisateur - prépositions a-sélectives, prépositions à deux compléments (qu'il s'agisse d'emplois de quasi-coordonnant, d'interposition voire de circumposition), etc. - peuvent être simultanément repérées par rapport aux pôles organisateurs d'autres catégories grammaticales dont elles se rapprocheraient. L. Melis (*op. cit.:* 41-43) propose ainsi de dénombrer quatre pôles (en dehors de la préposition) tous situés parmi les catégories traditionnelles invariables, et qui pourraient du même coup être réunies en une classe subsumante des « particules » : les marqueurs casuels[12], l'interposition, le coordonnant et l'adverbe.

1.1.1.2. Le cas du français du XVI^{e} s.

Evoquons sans nous y attarder la question de la « catégorisation » des prépositions dans l'état de langue que constituait le français au sortir du moyen français.[13]

Sur un plan général d'abord, on conviendra qu'une catégorie – ou son noyau prototypique - est en dernière analyse réductible à la somme des propriétés qui la – ou le – constitue. Parler de *prépositions* au XVI^{e} s. et donc recourir à la même étiquette catégorielle que celle utilisée pour le français moderne exige qu'on reconduise pour les morphèmes qu'on catégorise comme tels les mêmes propriétés que celles du français moderne. Et de fait, les critères morphologiques, sémantiques et syntaxiques que nous avons détaillés plus haut d'une part, la démarche d'analyse en termes de gradient de prépositionnalité d'autre part, s'appliquent aux morphèmes prépositionnels du français pré-classique[14].

Le point sur lequel nous voudrions nous attarder concerne le caractère pluricatégoriel des morphèmes qui assuraient le rôle de préposition jusqu'à l'aube du XVI^{e} s. Certes, ce caractère est une constante des unités linguistiques dans toutes les langues, mais on est en droit de considérer que la période allant du IX^{e} s. au XV^{e} s. constitue une étape particulièrement marquante dans l'histoire du français dans la mesure où « certains morphèmes étaient largement pluri-fonctionnels ou même pluricatégoriels » (C. Marchello-Nizia, 2002 : 207). Or aucune autre catégorie ne semble avoir été aussi « peu exclusive » que

---

bons exemplaires situés à la périphérie. Le prototype fournit en même temps le principe de catégorisation : les entités sont rangées dans une catégorie selon leur degré de ressemblance avec le prototype ». (G. Kleiber, *op. cit.* : 185-186).

[11] Approche mise en jeu par une étude raisonnant à partir de conditions qu'on juge « nécessaires et suffisantes » (CNS). En ce cas, une unité est rangée dans une catégorie si et seulement si elle satisfait à la totalité des critères explicites considérés comme nécessaires (l'absence de l'un d'entre eux excluant d'intégrer l'unité dans la catégorie) et suffisants. Or on peut se demander si une telle démarche n'est pas génératrice de problèmes insolubles. Le cas des adverbes est à cet égard révélateur : outre qu'il apparaît vain de chercher à déterminer des critères positifs que pourraient partager l'ensemble des « adverbes » de la grammaire traditionnelle, l'application d'un classement par CNS fait surgir un nombre de classes exagérément étendu, dont certaines pourraient être réduites à très peu d'éléments. Pour une approche des classes grammaticales en syntaxe générale selon le modèle de la prototypie, voir D. Creissels (1995, 2006).

[12] Cas des groupes introduits par *à* et alternant avec les pronoms *lui/leur* qu'on qualifie de datifs. Voir entre autres R.-S. Kayne, 1977, M. Van Peteghem, 2006.

[13] Concernant la coupure du XVI^{e} s. par rapport au moyen français, nous suivons C. Marchello-Nizia (1997 : 3-6). Sur ce point, voir aussi J. Ducos & O. Soutet (2012 : 3-10).

[14] Pour une application du modèle de la « prototypicalisation » aux prépositions de l'ancien et du moyen français, voir par ex. B. Fagard & W. de Mulder (2007 : 9-29).

les prépositions à cette période[15]. Parmi les catégories dans lesquelles pouvaient entrer les morphèmes prépositionnels à cette époque figure celle des *particules séparées*, que C. Buridant (2000 : 438) définit comme suit :

« Une particule séparée est un adverbe fondamentalement locatif jouant par rapport au verbe le rôle de vecteur sémantique en complétant son sémantisme, comme le font les particules séparables de l'allemand contemporain (*hinaus-gehen « aller dehors » → sortir*) ou les particules des *phrasal verbs* de l'anglais, appelés aussi verbes discontinus (*to go out « aller dehors » → sortir*).

Il est intéressant ici d'observer qu'un lien peut être établi entre cette définition et la distinction typologique de L. Talmy (1985, 2000)[16] entre les langues à cadre verbal (*verb-framed languages*) et langues à satellites[17] (*satellite-framed languages*). Les premières encodent préférentiellement la composante sémantique « trajectoire » (PATH) dans le verbe (par ex. en français : *le chien est sorti de la maison*), les secondes encodent préférentiellement cette composante dans un satellite du verbe (particules du verbe en anglais : *up, down, over,* …, préfixes verbaux dans les langues slaves et en allemand, etc. Par ex., en anglais : *The dog ran out of the house*). Les particules en ancien français telles que définies par C. Buridant illustreraient donc un cas d'encodage de la trajectoire au moyen d'un satellite.

On peut illustrer la pluricatégorialité de certains morphèmes de l'ancien français avec le mot *contreval* qui, dans (8) et (9), est successivement préposition et particule séparée. Quant à (10), il présente un cas d'ambiguïté entre ces deux catégories :

(8) *Si jeta ses mains a ses cheveus … Elle amena sa main contreval sa face* (SageP , 5, 3)
(*Elle s'en prit à ses cheveux … Elle porta ses mains plus bas sur sa figure)*

(9) *Si se leva et devala conme ainz pot contreval* (*op. cit.,* 21, 12)
(*Il se leva et descendit le plus vite qu'il le put)*

(10) *Les norrices descendirent contreval les degrez du mur*[18] (*op. cit.*, 10, 15)
(*Les nourrices descendirent les escaliers du mur)*

Selon l'auteur de la *Grammaire nouvelle de l'ancien français*, la disparition des particules séparées s'explique pour des raisons d'ordre typologique, le français étant passé d'une langue OV à une langue VO.

A l'aube du XVI^{e} s., donc, les morphèmes relevant du paradigme catégoriel des prépositions étaient en outre susceptibles d'appartenir à la classe des adverbes, des préfixes et

[15] Si l'on examine le cas de *en*, ce mot pouvait être préposition, préfixe verbal lexicalisé et préfixe séparable (*op. cit.*: 218).

[16] Merci à M. Aurnague de m'avoir suggéré ce rapprochement. On pourra se reporter par ailleurs à D. Stosic (2009b) pour un focus sur la composante « manière » dans l'analyse que fait L. Talmy de ces deux familles typologiques de langues.

[17] « *It is the grammatical category of any constituent other than a noun-phrase or prepositional phrase complement that is in a sister relation to the verb root. It relates to the verb root as a dependent to the head. The satellite, which can be either a bound affix or a free word, is thus intended to encompass all of the following grammatical forms, which traditionally have been largely treated independently of each other: English verb particles, German separable and inseparable verb prefixes, Latin or Russian verb prefixes, Chinese verb complements…*» (L. Talmy, 2000, vol. II : 102)

[18] Dans cet exemple, commente C. Buridant (*op. cit.*), « *contreval* peut être analysé comme préposition, formant alors le syntagme *contreval les degrez du mur*, ou comme la particule séparée de *descendre* ».

des particules séparées (en cours de disparition).

1.1.2. La relation *X R Y*

1.1.2.1. Une relation syntaxique ou sémantique ?

La notion de « relateur », presque systématiquement mobilisée dans les études sur les prépositions, reçoit selon les auteurs une définition syntaxique et/ou sémantique.

C. Hagège (1982, [2013][19]) illustre la première position : « *Le relateur relie un complément au prédicat et se distingue ainsi d'un nominant, incident au seul nom et supprimable* » (*op. cit.*: 45). On observera que dans cette définition (qui s'applique, outre aux prépositions, aux postpositions et aux désinences casuelles), l'auteur est conduit à exclure du statut de relateur les emplois de *de* internes au SN comme dans *L'ami de Paul*, lui affectant le statut de *joncteur* (*op. cit.,* 74).

> « Je ne comprends pas pourquoi un certain nombre de linguistes appellent préposition le *de* qui est un joncteur. Or autant le *de* de *mourir de faim*, *de honte, de froid* est évidemment une préposition, autant le *de* interne au syntagme nominal ne peut être traité (sauf tradition didactique de l'enseignement primaire en France depuis trois siècles!) comme une préposition. Il ne l'est pas! Je l'appelle joncteur, car pour moi une préposition est un élément qui a pour fonction de mettre dans la dépendance d'un prédicat verbal un lexème ou un groupe nominal. *De* en français est ambigu : il est relateur et donc préposition dans *mourir de faim*, alors qu'il est joncteur dans *le jardin de mon père*. » (C. Hagège, http://fdl.univ-lemans.fr/fr/liste-des-numeros/n9/n9_presentation.html)

A la lecture de ces lignes, on conçoit qu'il puisse exister entre les auteurs adoptant une définition syntaxique de la notion de *relateur* de notables différences. Ainsi, P. Cadiot (1997) qui affecte à la formule A-(PREP-B) le rôle d'exprimer non un simple rapport mais une relation syntaxique de « subordination », fait de *de* en emploi adnominal une préposition. J. Cervoni (1991), adoptant une perspective guillaumienne, reconnaît le rôle de relateur syntaxique des prépositions (plan de la « syntaxe résultative », *op. cit*. : 125) mais le juge inapte à rendre compte de la spécificité de la classe dont seule la notion d'incidence « diastématique[20] » peut rendre compte.

La deuxième position, suivant laquelle *relateur* doit s'entendre dans un sens strictement sémantique, peut être illustrée par B. Pottier qui dès 1974 désignait par ce terme « l'ensemble des signes établissant une relation (ayant donc une double incidence) entre deux termes (du simple lexème aux propositions). Cela incluait les prépositions (simples et complexes), les postpositions, les préfixes et préverbes, les conjonctions, les déictiques, les marques casuelles. » (1997 : 29). Certains auteurs, favorables à une spécialisation sémantique de l'acception du terme *relateur*, arguent de la nécessité de distinguer soigneusement relation syntaxique et sémantique dans le cas notamment des SP régis par le verbe. C'est le cas de J.-J. Franckel & D. Paillard (2007) qui soulignent que dans de telles constructions - comme par ex.

[19] Voir aussi C. Hagège (1997).
[20] La psycho-mécanique analyse les prépositions comme des mots (de langue) non prédicatifs destinés à échoir en discours à un intervalle entre deux termes prédicatifs : incidence dite diastématique.

(11) *Le chasseur a tiré sur le lapin* (*op. cit.*: 107)

le terme X de la relation prépositionnelle *X R Y* n'est pas le verbe (comme il serait dit souvent) mais le projectile tiré, référent non réalisé par une expression linguistique à la surface de l'énoncé. Autrement dit, dans une perspective syntaxique, le terme X du relateur prépositionnel dans (11) serait le régissant du SP tandis que dans une perspective sémantique, ce terme X est un SN non réalisé linguistiquement.

On observe enfin que certains auteurs se dispensent de préciser si la fonction de relateur qu'ils affectent à la préposition désigne une relation d'ordre syntaxique ou sémantique. On tombe alors dans ce travers dénoncé par C. Guimier (2007) chez R. Quirk & al. (1985) : « Ce caractère relationnel [de la préposition] (…) fait rarement l'objet d'une réflexion théorique : (…) S'agit-il d'une relation syntaxique ou d'une relation sémantique ? ».

Dans les lignes qui suivent, nous adopterons une conception strictement sémantique de la notion de relateur, le volet syntaxique des relations entre la préposition, son complément et sa possible tête externe étant traité *via* les propriétés évoquées *supra* (préposition comme tête du SP et valence). Ce faisant, nous affranchissons l'identification du terme X de toute perspective a priori syntaxique.

Ajoutons pour finir que dans ses nombreux travaux relatifs à la préposition, P. Cadiot (1997a : 35-36, 1997b : 134) fait observer que la relation qu'instaure la préposition entre ses deux termes relève pour partie du « codage » - dimension instructionnelle du relateur - et pour partie de « l'inférence » - sa dimension « catalysante » -, inférence et instruction entrant dans des dosages qui varient.

> « [T]out mot – et d'abord la préposition – catalyse une part essentielle de sa valeur dans son environnement (il prend du sens), mais le fait selon des instructions qui lui appartiennent en propre (il alloue du sens). Mais d'une préposition à l'autre, d'un type à l'autre, les dosages sont très différents. On peut estimer que moins un segment a de sens interne, plus il est en mesure de « prendre » des sens variés. A l'inverse une préposition peut avoir un sens relationnel fort. Elle est porteuse de ce sens qu'elle code. » (P. Cadiot, (1997a : 35-36).

Cette distinction est intéressante en ce qu'elle invite à raffiner la notion de *relation* (instanciée par la préposition) en distinguant les prépositions où la (relation de) catalyse l'emporte sur la (relation de) codage (Inférence^{+} Codage^{-}) – on songe aux prépositions « incolores » (*de, à,* peut-être *en*) -, et celles où c'est l'inverse : les prépositions les plus colores (Inférence^{-} Codage^{+}).

1.1.2.2. Identification des termes X et Y

Sur le plan de l'analyse, l'identification du terme Y ne pose en général pas de problème : placé à la suite de la préposition, il est linguistiquement réalisé par son complément si celui-ci n'est pas nul. Sinon[21], Y est récupérable en contexte :

(11) *Le chasseur a tiré sur le lapin*.
(12) A- *Et concernant le référendum?* B- *Je suis contre*[22].

[21] Moyennant éventuellement la sélection d'une variante allomorphique de la préposition employée devant régime (*dans -> dedans ; sur -> dessus ; …*)

On observera que l'identification du terme Y nécessite la prise en compte du complément de la préposition dans sa dimension énonciative et textuelle, et non simplement dans sa réalisation linguistique de surface. En voici une illustration :

(13) *[Vivre chacun de son côté, avoir des aventures sentimentales et sexuelles : leur seule promesse était de tout se raconter, de ne jamais se mentir]$_{PX}$. En résumé, [une liberté totale dans une transparence parfaite]$_{PY}$.* (B. Lamblin, *Mémoires d'une jeune fille dérangée*, 1993).

Le SP figé (*en résumé)* est ici disjonctif de style (C. Molinier & F. Levrier, 2000) puisqu'il concerne la relation du locuteur à l'énoncé qu'il formule – plus précisément, à la *forme* de ce dernier. Le prédicat P_Y « *une liberté totale dans une transparence parfaite* » est présenté par le locuteur, au moyen du disjonctif, comme exprimant sous une *forme résumée* le même état de choses[23] que celui exprimé par les prédicats P_x antérieurs. Le N syncatégorématique[24] *résumé* quant à lui, pourvu d'une valence, implique un complément non exprimé en surface mais qui énonciativement et textuellement équivaut à quelque chose comme : *(en résumé) de l'état de choses formulé en amont par P_X.*

L'identification du terme X de la relation peut être plus épineux, comme illustré *supra* par l'examen de l'énoncé (11) dans lequel, si l'on suit D. Paillard (2001 : 117), X n'est pas le procès dénoté par le verbe mais un référent auquel pourrait référer le SN « une balle / un projectile ».

1.1.3. Conclusion

De notre propos antérieur, nous soulignerons deux points particulièrement saillants.

En premier lieu, le choix que nous avons fait en faveur d'une conception strictement sémantique de la fonction de *relateur* traditionnellement affectée à la préposition dans la structure *X R Y* afin d'affranchir le terme X de la relation syntaxique qui lie le SP à sa tête externe.

En second lieu, la préférence que nous accordons à une approche scalaire de l'appartenance des unités à la catégorie des *prépositions*. Deux arguments au moins plaident pour une telle option. La série d'exceptions observables chez certaines prépositions (cf. entre autres, L. Melis, 2003) pour telle ou telle propriété syntaxique considérée traditionnellement comme critère d'appartenance à la catégorie, et qui mène à adopter de cette dernière une vision plus flexible. La perspective diachronique enfin qui conduit, *via* le cadre de la théorie de la grammaticalisation (cf. entre autres B. Fagard 2006, W. de Mulder 2001), à appréhender la dynamique permettant à certaines unités entrantes, d'abord marginales dans la catégorie, d'acquérir progressivement des propriétés qui les rapprochent du noyau prototypique, voire qui leur permet d'y entrer.

---

[22] Il y a ici ellipse du complément de la préposition employée absolument : *contre* en français accepte dans sa valence un régime nul, ce qui n'est par ex. pas le cas de *en.*

[23] Dans la typologie proposée par C. Rossari (1997 :17), *en résumé* est à rapprocher d'autres marqueurs de reformulation non paraphrastique (MNRP) comme *en un mot, bref...* : il opère « une rétrointerprétation du point de vue auquel il renvoie selon une nouvelle perspective énonciative annoncée par les instructions sémantico-pragmatiques du connecteur ». En l'occurrence, cette nouvelle perspective est celle de la « récapitulation » : « La prise de distance due au changement de perspective énonciative est donc peu accentuée, car le locuteur ne remet pas en question le point de vue exprimé dans la première formulation en ce qui concerne son contenu, mais se contente d'en donner une expression plus condensée. » (*op. cit.* : 18).

[24] G. Kleiber (1981 : 39-40),

1.2. Construire l'identité sémantique d'une préposition

Construire l'identité sémantique d'une préposition nécessite qu'on résolve la difficulté que pose tout polysème grammatical (C. Fuchs, 1997) et plus largement toute unité polysémique : comment rendre compte de manière synthétique de la variété plus ou moins étendue des sens que le polysème peut prendre en discours[25] ? A cet égard, *en* n'est pas en reste puisque dans le *Trésor de la Langue Française*, son traitement lexicographique ne nécessite pas moins de cinq pages (999-1003). Formuler l'identité sémantique d'une préposition, en particulier abstraite, nécessite donc une *construction* qui implique de prendre des décisions méthodologiques à plusieurs niveaux[26] .

1.2.1. Polysémie « verticale » *versus* « horizontale »

Le « pari» de la polysémie verticale coïncide avec le postulat selon lequel à toute préposition, il est possible de faire correspondre une identité sémantique « *à l'œuvre dans tous [s]es emplois* » (J.-J. Franckel & D. Paillard, 2007) et qui soit

- suffisamment *abstraite* pour se dégager de la multiplicité des effets de sens en emploi ;
- suffisamment *différencié* [27] pour se distinguer de l'identité des autres prépositions avec lesquelles elle entre en concurrence ;
- suffisamment *puissante* pour engendrer les effets de sens attendus ;
- suffisamment *sous-déterminée* pour permettre au co(n)texte d'interagir avec elle[28].

La préposition se voit ainsi attribuer « une valeur de base abstraite, générique (hyperonyme) (…), les sens en emploi (hyponymes) étant attribués par la spécification sensible aux contextes. » (P. Cadiot*,* 1997a : 10).

Le choix de la polysémie « horizontale » fait en revanche l'hypothèse qu'il est impossible de rendre compte de tous les sens que la préposition exprime en discours au

[25] A de nombreux égards, les réflexions que nous allons développer dans les sections suivantes croisent la question posée par D. Stosic & B. Fagard (2012 : 3-24) : « le signifié des polysèmes est-il monolithe, composite... ou extrêmement malléable ? ».

[26] Pour P. Cadiot (1997a), « le pari monosémique » ouvre deux pistes à l'analyse : « (i) Une version verticale (*polysémie verticale*) où telle préposition se voit attribuer une valeur de base abstraite, générique (hyperonyme) et non représentationnelle, les sens en empli (hyponymes) étant attribués par la spécification sensible aux contextes. (ii) Une version horizontale (*polysémie horizontale*) qui consiste à valoriser un des sens en emploi en le promouvant au statut de prototype. (…) La version (i) repose sur un pari. » » (*op. cit.* : 10-11).

[27] Comme l'observe P. Cadiot (1997a : 38), « il n'y a guère de fléchage peu ou prou biunivoque des valeurs sémantiques sur les prépositions, et tout particulièrement bien sûr sur les incolores et mixtes. De la même façon, la notion de « direction », même en la cantonnant dans l'espace, passe par plusieurs prépositions (*à, pour, vers, sur*, sans même évoquer les locutions spécialisées). Celle, plus indirecte ou de second degré, « d'approximation » passe par *environ* … mais aussi par *dans* (il mesure dans les 1, 80 mètres) et *sur* (habiter sur Paris). » D'où un travail nécessaire de réduction, de distillation en quelque sorte, qui permette de se hisser jusqu'à une formulation distinctive.

[28] On trouvera ces deux derniers traits formulés dans L. Melis (2003 : 99). On observera que le dernier renoue, à travers le concept d'interaction, avec l'idée de catalyse (voir *supra*) avancée par P. Cadiot à propos de la « relation » prépositionnelle.

moyen d'une identité sémantique fixe, conçue comme un ensemble de traits sémantiques (représentationnels ou instructionnels : cf. *infra*) tous présents dans chacun de ses emplois. En d'autres termes, la polysémie des acceptions en discours ne peut être ramenée à l'unité d'une valeur sémantique stable en langue qui serait conçue comme « une conjonction suffisante de traits nécessaires». (G. Kleiber, 1990 : 23). Le cadre conceptuel qui s'offre alors au linguiste pour développer une telle approche « horizontale » est celui de la sémantique du prototype dans sa seconde version (« étendue » au sens de G. Kleiber, *op. cit.*) qui s'appuie sur la notion wittgensteinienne[29] de « ressemblance de famille ». Une préposition peut ainsi se voir attribuer une identité sémantique structurée comme une ressemblance de famille : y sont regroupés des traits qui n'ont pas besoin d'être tous vérifiés dans chacun des emplois observés. On illustrera cette approche par deux exemples.

D'abord, par la préposition *dans* telle qu'elle est analysée par C. Vandeloise[30] : ce mot décrit selon lui la relation C/c (contenant / contenu) structurée comme une ressemblance de famille et réunissant trois[31] traits déterminants[32] : (i) *Le contenant contrôle la position du contenu* ; (ii) *Le contenu se déplace vers le contenant ;* (iii) *le contenu est inclus, au moins partiellement, dans le contenant ou dans la fermeture convexe de sa partie contenante.* (1993 : 33)

Les couples d'énoncés suivants (L. Sarda, 2010 : 10) permettent d'illustrer chacun de ces traits, le second énoncé du couple (b) ne vérifiant pas le trait actualisé dans le premier (a):

(14) [Trait *i)*] (a) *La lampe est dans la douille /* (b) *Le doigt est dans la bague.*
(15) [Trait *ii)*] (a) *La mouche est dans le bol /* (b) *La soupe est dans la louche.*
(16) [Trait *iii)*] (a) *Le vin est dans le verre /* (b) *Les fleurs sont dans le vase.*

La seconde illustration que nous donnerons d'une approche « horizontale » de la polysémie compare deux prépositions sémantiquement très proches dans deux langues différentes : *à travers* en français et *kroz* en serbe. Comme le montre D. Stosic (2009a), toutes deux convoient le concept de *guidage* qui se définit comme une ressemblance de famille réunissant sept traits[33]: *dynamicité, intériorité, opposition au mouvement, orientation latérale, unicité du site, minimum de parcours, focalisation sur le parcours du site*. Là encore aucun de ces traits « ne doit être considéré comme une condition nécessaire et suffisante pour la description des usages spatiaux de *à travers* » (24-25). Ainsi, pour l'exemple suivant proposé par l'auteur :

(17) « *Eh! bien, qu'en dites-vous ? lui demanda mademoiselle des Touches en jetant la lettre à travers la table à Vignon* » (Balzac H., *Scènes de la vie privée*).

[29] L. Wittgenstein (1953, § 65-67)
[30] Merci à M. Aurnague qui me fait remarquer que dans la « ressemblance de famille » telle que l'entend L. Wittgenstein, tous les traits ne peuvent pas être réalisés ensemble alors que chez C. Vandeloise, cette réunion est possible. On retrouve là, me semble-t-il, les deux types de regroupements de traits dans la version *étendue* de la sémantique du prototype telle que l'expose et l'illustre G. Kleiber (1990 : 160).
[31] Le nombre de traits affectés à la relation C/c et leur formulation a évolué dans le temps : on comparera à cet égard C. Vandeloise 1986 et 1993 par ex. (trois traits, mais formulations qui diffèrent) et C. Vandeloise (2001) où sont dénombrés cinq traits (voir L. Sarda, 2010), le français ne faisant pas appel au cinquième trait pour étendre l'usage de la préposition *dans*.
[32] Il s'agit de « traits nécessaires pour justifier au moins un usage [de la préposition] qui ne pourrait pas être motivé par les traits déjà existants » (1993 : 36)
[33] Nous nous abstenons de les présenter ici, préférant renvoyer le lecteur à l'article.

le trait « intériorité » n'est pas vérifié en ceci que « le déplacement ne se fait aucunement par l'intérieur du site » (*op. cit.*, 25). Une des originalités de l'analyse présentée consiste à montrer que les différences sémantiques observables entre les deux prépositions (*à travers* / *kroz)* ne sont pas dues à des différences situées au niveau du concept qu'elles expriment mais à des différences relatives aux degrés de saillance de chacun des traits qui composent ce concept, ce qui fait entrer en jeu la dimension culturelle[34].

Polysémies « verticale » et « horizontale » constituent donc deux voies possibles pour la construction de l'identité sémantique d'une préposition. Concernant *en,* l'ensemble des études passées et présentes sur son identité sémantique ont uniquement adopté, à notre connaissance (voir D. Vigier, 2013) la voie de la polysémie verticale.

Dans les lignes qui suivent et dans toute notre première partie, nous nous en tiendrons exclusivement à une approche verticale de la sémantique des prépositions. Nous croyons en effet être à même - dans le *seul* cadre de la polysémie verticale - de déployer une réflexion méthodologique assez ample pour revêtir un certain intérêt, en vue de répondre à cette délicate question : comment rendre compte de l'unité *et* de la variabilité du sens de ce polysème grammatical « incolore » qu'est *en*?

Nous ne manquerons pourtant pas de revenir dans la conclusion de cette partie à l'approche horizontale de la sémantique prépositionnelle car nous comptons en faire un de nos axes de recherche dans un avenir proche.

1.2.2. La notion d'identité sémantique présuppose-t-elle un invariant ?

Par *invariance*, on entend *l'identité de soi à soi* de la valeur sémantique affectée à un polysème dans tous ses emplois en discours.

Une telle hypothèse exclut de considérer que certaines prépositions (généralement, les plus incolores) puissent être considérées comme sémantiquement vides dans leurs emplois les plus fonctionnels. Cette question de la vacuité possible du signifié prépositionnel a été amplement discutée et nous la considérons comme tranchée[35] : le sens d'une préposition, si ténu soit-il, n'est jamais vide.

La possibilité de l'invariance exclut en outre l'idée que les prépositions incolores, dans leurs emplois les plus fonctionnels, puissent voir leur sens se modifier. Pour illustrer notre propos, on peut examiner après L. Melis (2003) les constructions où *de* i) joue le rôle de quantificateur (18)-(19), ii) entre dans la construction d'un prédicat second (20), iii) joue (21) le rôle de « complémenteur » (terminologie issue de la grammaire générative) ou d' « indice » (terme souvent employé en grammaire : entre autres, P. le Goffic, 1993):

(18) *J'ai de la monnaie*
(19) *Je n'ai pas de monnaie*
(20) *Il y a une place de libre*
(21) *Il regrette de partir*

Relativement à ces emplois, notre position est dans son principe analogue à celle argumentée par L. Melis (*op. cit.*: 53-54, 83-84, 125-131). L'auteur montre d'une part que

[34] « Puisqu'on peut formuler l'hypothèse selon laquelle la saillance des différents éléments est déterminée au moins partiellement par la culture, cela pourrait indiquer que ce qui est au prime abord le même concept est quand même défini de façon différente par chaque communauté linguistique. » (W. de Mulder & D. Stosic, 2009 : 4-5).

[35] Pour une discussion, nous renvoyons à J. Cervoni 1991, L. Melis (2003), D. Leeman (2006, 2008), …

dans ses emplois illustrés de (18) à (21), le mot *de* partage une partie seulement de ses caractéristiques sémantiques et syntaxiques avec la préposition homonyme, d'autre part que ce mot a acquis dans ces mêmes emplois des propriétés sémantiques et syntaxiques qui conduisent à le faire sortir de la classe des prépositions. Il resterait cependant à définir (ce que l'auteur à notre connaissance ne fait pas) à partir de quand on considère que le mot (qu'il s'agisse de *de, à,* voire *en*[36]) ne partage pas suffisamment de propriétés avec son homonyme prépositionnel et doit être exclu de la classe.

Ces observations formulées, on peut aisément convenir que l'hypothèse de l'invariance de l'identité sémantique domine parmi les auteurs qui optent pour la « polysémie verticale ». Concernant *en*, on citera l'exemple de G. Guillaume (1919), de G. Gougenheim (1951), de C. Guimier (1978), de L. Waugh (1976), de J.-J. Franckel & D. Lebaud (1991), …

Pour dominante qu'elle soit, cette position n'est cependant pas systématique. Témoins les travaux de P. Cadiot & Y.-M. Visetti (2001) et de P. Cadiot (2002) qui proposent d'attribuer à chaque préposition un *motif* à entendre comme « [un] « germe [...] » instable [...], apte [...] à se stabiliser en syntagme par reprise au sein de dynamiques de « profilage » qui ne [lui est] pas immanente [...]. » (P. Cadiot, 2002 : 21). A propos de l'invariance de ces motifs, P. Cadiot écrit : « Il est (…) difficile de les qualifier d'invariants[37] : le terme renvoie à une problématique dont nous voulons justement dépasser les apories. Ce sont des unités de couplage, hautement instables, entre des dimensions qui n'apparaissent comme hétérogènes qu'à d'autres niveaux de stabilisation ». (*op. cit.:* 21). C'est donc à une remise en cause de l'invariance qu'on assiste, l'auteur cherchant à situer les *motifs* à un niveau non encore stabilisé du sens qui lui permettrait de coupler des dimensions (*quantité / qualité, ego / alter, intérieur / extérieur,* ...) perçues comme hétérogènes à des degrés plus aboutis de stabilisation.

1.2.3. Cadre théorique et formulation de l'identité

Toute formulation d'identité sémantique pour un polysème s'inscrit dans un cadre théorique plus vaste qui confère à cette identité un statut au sein de l'économie conceptuelle mise en jeu par la théorie. On illustrera ce point par deux exemples tirés des études sur *en*.

Pour G. Guillaume (1919), l'identité sémantique affectée à *en* possède en psycho-mécanique le statut de *signifié de puissance*, c'est-à-dire de valeur en système à laquelle on peut remonter à partir des *signifiés d'effets*[38] de la préposition. « Toutes les valeurs d'emploi, si diverses soient-elles, sont réductibles à la valeur constante, invariante, qu'a la forme en système : c'est-à-dire à la valeur de langue. » (R. Valin, 1997 : 213).

Chez J.-J. Franckel & D. Paillard (2007 : 21), cette identité a le statut de *forme schématique* tel que défini par A. Culioli :

> « Les phénomènes linguistiques forment des systèmes dynamiques qui sont réguliers, mais avec une marge de variation due à des facteurs d'une grande diversité : on a affaire à des phénomènes qui sont à la fois stables et plastiques. (…) Pour qu'il y ait

[36] Pour ce qui regarde *en*, un emploi justifiant qu'on ne le considère pas comme une préposition mais comme le composant d'un morphème verbal discontinu (position adoptée par G. Kleiber, 2007 : 107 ; voir aussi W. de Mulder & D. Amiot, 2013 : 33-36) est celui où il entre dans la constitution du traditionnel gérondif.

[37] Dans P. Cadiot & Y.-M. Visetti 2002, cette invariance est clairement écartée. Par ex. « Caractère 'figural' des motifs (…) à l'opposé d'une logique […] fondée sur la délimitation d'une couche de sens homogène (i.e. accueillant des 'invariants' qui sont foncteurs d'homogénéité) » (45).

[38] « En tant qu'unité de langue, chaque mot est pourvu d'un invariant cognitif, ou signifié de puissance, prévoyant l'ensemble des réalisations contextuelles, ou signifié d'effet. » (D. Bottineau, 2005 : 42).

> déformabilité, il faut que l'on ait affaire à une forme schématique (telle qu'il puisse y avoir à la fois modification et invariance), que l'on ait des facteurs de déformation et que l'on ait une marge de jeu, un espace d'ajustement muni de propriétés topologiques. » (A. Culioli, 1990 : 129-130)

1.2.4. Modèle du sens et formulation de l'identité

Toute formulation d'identité sémantique engage à réfléchir au « modèle de sens » que l'on adopte pour une telle formulation. Comme le montre G. Kleiber[39] (1997, 1999), vouloir rendre compte du *sens* d'un mot ou d'un segment linguistique (syntagme, phrase, …) peut conduire à s'interroger sur la relation qu'on établit entre sens et existence d'une part, entre sens et référence d'autre part.

Nous ne développerons pas le premier couple (*sens-existence*), étant en parfait accord avec l'auteur qui défend une thèse objectiviste tempérée (ou critique) et conclut : « le langage en tant que système de signes est tourné vers le dehors » (*op. cit.*, 16), c'est-à-dire vers l'extralinguistique. *Exit* donc toute idée d'une référence purement interne au langage. Nous nous intéresserons en revanche au second couple *sens-référence*. Si l'on adopte le point de vue d'un réalisme objectiviste modéré, la question suivante se pose: comment préparer la « sortie » du langage « sur l'extralinguistique » (*op. cit.*, 19), c'est-à-dire la référence ? Sans reprendre en détail son argumentation, on rappellera que G. Kleiber regroupe les conceptions du couple sens-référence en sémantique dans deux grandes paradigmes (*op. cit.*, 20-21) : le *paradigme du sens référentiel* et le *paradigme du sens aréférentiel*. Dans le premier, le sens d'une expression linguistique est conçu comme « un programme référentiel » : il « est en prise avec la référence par le biais de ses conditions d'applicabilité référentielle » (*op. cit.*, 22). Pour que le nom « cheval » puisse être appliqué à un segment de la réalité extralinguistique, il faut qu'il vérifie un faisceau de conditions que le sens (ou la *référence virtuelle* de l'expression linguistique, dans la perspective de J.-C. Milner (1978 : 332-333[40])) a pour rôle de stipuler. Dans le second cas (*paradigme du sens référentiel*), le lien entre sens et référence est brisé de diverses manières (indétermination fondamentale du sens des énoncés, option constructiviste radicale, etc.). Le point à retenir est le suivant : plusieurs des approches du paradigme dit « aréférentiel » ont été développées pour construire[41] l'identité sémantique de morphèmes différents des unité lexicales puisqu'il s'agit des connecteurs, des temps verbaux, des symboles indexicaux, des affixes, ou encore des polysèmes grammaticaux tels que certains adverbes (par ex. *encore* ; voir B. Victorri & C. Fuchs, 1996 ; C. Fuchs 1997) ou les prépositions (en particulier incolores). Rendre compte de l'identité sémantique de ces dernières nécessite qu'on opte pour une conception du sens moins prédicative qu'instructionnelle, qui fasse la part belle non pas aux conditions d'applicabilité référentielle de l'expression mais aux *procédures* à suivre pour accéder au sens qu'elle revêt en contexte. Selon qu'on cherche à construire l'identité de la préposition *de, à* ou *en*, ou bien du nom *bicyclette* ou *linguiste*, on a donc intérêt à se doter d'un modèle du sens différent mais non hétérogène en ce que, comme y insiste G. Kleiber, ces modèles s'avèrent tous « référentiels » puisqu'ils se branchent *in fine* sur la référence :

> « L'hypothèse que nous suggérons est que le sens obéit à deux modèles référentiels différents : le modèle descriptif, celui qui indique quelles sont les conditions

[39] A qui nous avons emprunté l'expression de « modèle de sens »

[40] « Le sens établit les conditions de possibilités générales d'une désignation : on peut le considérer comme une relation référentielle en puissance ou virtuelle ». (J.-C. Milner, 1978 : 333)

[41] Du moins dans un premier temps. « Ce n'est que récemment que [ces approches] se trouve[nt] également testée[s] sur des unités lexicales comme *arbre, boîte, lit, cendrier,* etc. (*op. cit.* : 29)

> (nécessaires et suffisantes ou prototypiques) auxquelles doit satisfaire une entité pour pouvoir être désignée ainsi, et le modèle instructionnel, qui marque le moyen d'accéder au, ou de construire le référent. ». G. Kleiber (1997 : 32 ; 1999 : 50)[42]

Conformément à la perspective de l'auteur, nous considérons que les prépositions relèvent des unités « mixtes » en termes de sens, la part des ingrédients instructionnel et référentiel dans la formulation de leur identité sémantique étant liée au caractère plus ou moins abstrait de celle-ci. Plus une préposition est incolore, plus sa part de sens instructionnel domine. Ce que l'on peut illustrer par l'exemple de la locution *au creux de* : dans ses acceptions spatiales, elle exige que le référent désigné par son complément soit doté d'une concavité. D'où le caractère étrange voire inacceptable de séquences comme : **au creux de la boule, *au creux du plan, *au creux de la ligne*[43], etc. Construire l'identité de *au creux de* nécessite qu'on fasse appel à des formes telles que courbure ou concavité, c'est-à-dire à des critères d'applicabilité référentielle. En revanche, plus la préposition devient abstraite, plus la composante instructionnelle de son identité sémantique constitue un ingrédient majeur. Dans le cas de *en*, on verra que la formulation que nous adoptons (§ 2.2.1) fait appel à des concepts (coalescence, restriction au cadre extensionnel, …) invitant à accomplir des opérations sur les signifiés mis en jeu par les termes X et Y (et au-delà, par le co(n)texte) de la relation prépositionnelle sans que ne soient mobilisés des traits « prédicatifs » visant à décrire des référents.

1.2.5. Modéliser une structure de déploiement des interactions contextuelles

Peut-on et/ou doit-on ménager, entre une identité stable nécessairement abstraite et épurée (cadre d'une approche verticale) et la variété (particulièrement étendue pour les prépositions incolores) de ses acceptions possibles, un niveau intermédiaire dont la structure propre (ou « structure de déploiement des interactions contextuelles») reflèterait des hypothèses quant à la manière dont le sens prépositionnel se déploie en discours? L'existence d'une telle structure suppose qu'aient été mis au jour un ou plusieurs critères qui règleraient ces interactions, la question étant ouverte de savoir si cette structure est « générique » en ce qu'elle s'appliquerait à toute préposition (voire à toute unité du lexique) ou bien si elle est (pour partie ? complètement ?) spécifique, chaque préposition possédant sa propre structure de déploiement.

Pour « cadrer » notre propos, nous examinerons cette question à travers le filtre de notre préposition-fil rouge qu'est *en.* Parmi les travaux qui ont jalonné les études sémantiques sur cette dernière, on peut distinguer deux types de positionnement des auteurs :

=> Une première attitude consiste à ne pas thématiser ni *a fortiori* modéliser une telle structure. L'auteur recourt bel et bien un niveau intermédiaire pour organiser la variété des effets de sens pris par la préposition en contexte, mais ce niveau apparaît comme un dispositif rhétorique de « classement » : bref, un plan. On peut regrouper les plans rencontrés selon la préférence qu'ils donnent à l'un des deux critères de structuration thématique suivants :

- un critère « générique notionnel » qui mobilise les grandes catégories que sont le *lieu,* le *temps*, la *manière* etc. Ces catégories sont recrutées pour organiser les sens possibles

[42] L'auteur ajoutant qu'une même expression peut « relever de ces deux modèles de sens » (1997 : 33 ; 1999 : 51). De fait, pour M. Aurnague, les prépositions spatiales relèvent de ce type d'expression : « *Comme d'autres marqueurs, les relations spatiales semblent donc présenter une double nature intructionnelle et descriptive*. » (2004 : 216-217).

[43] Concavité implique tridimensionnaité.

pris par le terme Y, suggérant du même coup que la variation des effets de sens produits par la préposition en contexte est indexée sur un tel découpage.

- Un critère « configurationnel »: le classement est alors opéré à partir de propriétés syntaxiques et sémantiques stables affectables à une structure *X R Y* distinguée par l'auteur : par ex. pour *en,* les structures avec verbe trivalent pouvant exprimer un changement « essentiel [44] » opéré sur le référent de l'objet syntaxique N1 (N0 *transformer / métamorphoser / changer / ... N1 en N2*) ou encore celles à attribut essentiel ou accessoire où le SP exprime une manière d'agir, de se comporter (*il se comporte en goujat, il est revenu en vainqueur, il a conclu cette affaire en expert, ...*).

Illustrons ce propos par quelques exemples tirées d'études sur *en*. La présentation de G. Guillaume (1919) adopte une structure uniquement fondée sur un critère « notionnel générique » : sont successivement envisagées l'*idée de lieu* (*op. cit.*, § 158), l'*idée de situation* (*op. cit.*, § 159) les *états moraux* (*op. cit.*, § 160) et enfin l'*idée de proximité ou d'éloignement d'une norme* (*op. cit.*, § 161) (par ex. « être en faute »). D. Leeman (1995) adopte elle aussi un même critère de classement puisque après avoir envisagé les domaines que sont la manière (*op. cit.*, 56), le temps (*op. cit.*, 61), le lieu (*op. cit.*, 62), elle s'intéresse à celui des sentiments. G. Gougenheim (1950) recourt quant à lui à un plan plus hybride: recensant les divers sens de *en* en français moderne, il rappelle d'abord sa valeur de *lieu* puis de *temps* [critère notionnel générique] ; puis il enchaîne sur l'examen de structures syntaxiques à sens stable (critère « configurationnel ») : les verbes exprimant la transformation, ceux signifiant « *croire, penser* », enfin les structures où *en* sert « à marquer la façon dont quelqu'un s'est comporté dans une action donnée ou dont on a traité quelqu'un : *il a agi en roi, il est mort en brave, on le traite en esclave* » (60-61). Quant à I. Khammari (2007), elle indexe son étude sémantique de *en* sur une approche uniquement configurationnelle[45].

=> La seconde attitude consiste à thématiser pour elle-même la structure de déploiement des interactions de *en* avec ses contextes. Plusieurs études sur *en* témoignent d'une telle démarche. Parmi les plus anciennes, on peut citer celle de C. Guimier (1978) qui puise dans le schéma tensif binaire guillaumien un principe de régulation des effets de sens en discours. Plus proches de nous, P. Cadiot & Y.-M. Visetti (2001) adoptent une démarche qui, à de nombreux égards, s'approche de la structure dont il est ici question à travers la triade *motif, schéma, thème*. Chacun de ces trois termes désigne une phase au sein d'un parcours dynamique de construction du sens. Mais le caractère complexe (et pour tout dire parfois abscons[46]) de cette représentation de la dynamique interprétative nous a conduit à ne pas intégrer leurs analyses au sein de notre propre réflexion sur la représentation du sens des prépositions. Nous nous arrêterons donc sur deux autres études.

La première est celle de J.-J. Franckel & D. Paillard (2007) : convaincus que « l'interaction de la préposition avec son co-texte relève, pour partie de ces variations

[44] Voir C. Vandeloise, 2000 pour la différence entre les verbes *devenir* et *se transformer*.

[45] SP ayant pour tête *en* uniquement employés comme compléments argumentaux d'un verbe, les occurrences relevées dans Frantext de cette structure syntaxique étant réparties dans des classes sémantiques « sur une base intuitive » (*op. cit.* : 58).

[46] A de nombreux égard, certaines des analyses développées nous semblent s'exposer à la critique de G. Kleiber (1999, 47) qui parle du caractère « incontrôlable » de certaines analyses sémantiques adescriptivistes : « Par *incontrôlables*, nous voulons dire deux choses : premièrement qu'il est parfois difficile d'interpréter la formule définitoire abstraite proposée et, deuxièmement, que, du coup, il est aussi difficile de la falsifier, c'est-à-dire de vérifier sa pertinence applicative sur des données discursives ».

[observables], de principes réguliers » (*op. cit.*, 13), les auteurs proposent pour chacune des prépositions qu'ils étudient un « format de description » systématique :

« 1. Proposition, pour chacune des prépositions, d'une FS [*forme schématique*] fondée sur un premier parcours descriptif de ses emplois.
2. Déploiement de la variation fondée sur l'organisation de l'interaction entre la préposition et X-Y. »

Deux grands ordres de « variation » sont distingués : une variation dite « interne » car considérée comme propre à la préposition, une autre dite « externe » liée aux propriétés lexicales de X et Y et aux rapports que ces unités entretiennent avec le verbe.

On peut tenter d'illustrer rapidement[47] une telle démarche avec l'exemple de *dans*. Sa forme schématique serait la suivante : « *Dans* marque que le repérage de X par Y correspond au rattachement de X à Y de telle sorte que la zone de rattachement de X à Y est indifférenciée. Le domaine associé à Y est indifférencié pour ce qui est de la zone où X s'y rattache. » (*op. cit.* : 151). En adoptant cette FS, les auteurs se distinguent des analyses de la sémantique de *dans* en termes d'intériorité (comme C. Guimier : 1978 ou B. Victorri : 2003, par ex.) puisque « *dans* rend indistincte la zone de rattachement de X à Y » (*op. cit.* : 151).

Concernant la variation et la déformabilité propre à la sémantique de *dans* (variation interne), les auteurs distinguent trois cas : (i) « Le point de rattachement de X à Y n'est pas singularisable du fait qu'il a partout un voisinage » (*op. cit.* : 153). Par ex. dans « X se fond dans le lointain » ; (ii) « Le point de rattachement est un point quelconque de Y, indiscernable relativement à tout autre » (*ibid.*). Par ex. « Il est dans le vrai » ; (iii) « Y se réduit au point de rattachement de X » (*op. cit.* : 154). Par ex. « Il arrive dans 28 minutes exactement ».

Trois cas sont également distingués pour la variation externe (liée aux propriétés lexicales de X et Y): (i) « l'homogénéisation de Y n'est due qu'à *dans* » (*op. cit.* : 165). Par ex. « Dans un juron, il sauta sur ses pistolets » ; (ii) « Y fait par ailleurs l'objet d'une homogénéisation » (*ibid*.). Par ex. « J'aime travailler dans le calme » ; (iii) « Y correspond à un terme dont l'identité incorpore celle d'homogénéité » (*ibid.*). « Il s'agit de termes tels que *fil, continuité, lignée, cheminement, cours, courant, prolongement, durée* etc. » introduits par *dans* (*op. cit.* : 156).

Le plan de variation lié aux rapports que *dans* entretient avec le verbe (ou : *configurations*), présenté dans les pages 167 à 172 de l'ouvrage, ne sera pas repris ici pour ne pas (trop) alourdir notre propos.

La seconde étude est celle de L. Melis (2003) plusieurs fois évoquée ici. L'auteur propose d'abord un examen critique des « domaines d'emplois » des prépositions (*lieu, temps*, ...) souvent convoqués par les dictionnaires (mais aussi les grammairiens et les linguistes) lorsqu'ils examinent le sens d'une préposition. A l'issue d'une discussion relative à la délimitation des domaines pertinents pour la description du sens d'une préposition, il est conduit à ne retenir que le *lieu*, le *temps* et les *relations argumentatives* qui possèdent chacun une structure propre influant de manière réglée sur l'interprétation de la préposition. « La notion de domaine peut être utile pour rendre compte des interprétations des prépositions dans la mesure où le domaine est un dispositif d'interprétation indépendant de la préposition et qui interfère avec les composantes sémantiques que celle-ci a en propre » (*op. cit.*, 75). Ce ne sont

[47] Ce qui constitue immanquablement une gageure lorsqu'on a à l'esprit que l'analyse de *dans* occupe trente-quatre pages dans l'ouvrage. Notre propos ne peut donc qu'être allusif et tronqué : nous renvoyons le lecteur au développement intégral.

cependant pas là les seuls modes d'organisation des sens d'une préposition, l'auteur en identifiant trois autres : le premier serait gouverné par des conditions d'ordre syntaxique ; le deuxième serait directement lié à la sémantique propre à la préposition. Resterait enfin une forme d'unité située « dans la dépendance, par figure ou par figement, d'autres emplois de la même préposition » (*op. cit.*, 75).

Dans la suite de cette première partie (cf. § 2.3.), nous reviendrons sur cette question du déploiement du sens en discours dans le cadre restreint de la sémantique de *en* pour articuler l'exigence d'abstraction associé au pari de la polysémie verticale avec la couverture la plus exhaustive possible des effets de sens en discours. En d'autres termes, nous ne souscrivons pas vraiment au diagnostic formulé par D. Stosic & B. Fagard (2012 : 7-8) :

> « Si, en dépit de la multiplicité des sens, on maintient l'idée d'un signifié monolithe, comment éviter le piège d'un sens schématique, i.e. d'un « invariant supérieur » unificateur qui, ayant pour vocation d'expliquer toute la variabilité du mot en discours, est nécessairement sous-déterminé, très général et donc trop puissant (trop puissant du point de vue explicatif, mais généralement peu efficace du point de vue interprétatif, cf. Kleiber 1999 ; Kleiber, 2008 : 89).

L'idée d'un invariant unificateur ne constitue pas, croyons-nous, un « piège » et la question de la puissance explicative, judicieusement observée, peut être selon nous « contrôlée » comme nous nous proposons maintenant de le montrer.

1.2.6. Contrôler la puissance de l'identité sémantique de la préposition

G. Kleiber (1999 : 47-48), à propos des formules définitoires très abstraites (schématiques et/ou instructionnelles) visant à établir l'identité sémantique d'une unité linguistique, note :

> « Elles courent [le] danger (…) d'être trop puissantes. (…) Par *trop puissantes,* nous entendons souligner le fait qu'elles peuvent convenir également à des entités qui ne se trouvent pas désignées par l'expression en question ».

Dans l'esprit de l'auteur, il s'agit de mettre au jour le fait que certaines définitions adescriptivistes de polysèmes lexicaux se révèlent inadéquates en ceci qu'elles peuvent s'appliquer aussi à d'autres entités : ainsi en irait-il selon lui de la définition avancée par P. Cadiot (1997, 201-214) pour le nom *boîte,* qui pourrait convenir pour désigner un cartable ou une serviette. Le reproche d'excès de puissance que nous avons en tête diffère de celui visé G. Kleiber. Il concerne les restrictions sélectionnelles que la préposition fait peser sur ses régimes, restrictions dont l'identité sémantique formulée est le plus souvent incapable de rendre compte. Le cas de *en* est à cet égard pertinent : quelle que soit l'identité sémantique qu'on lui affecte, la question des restrictions sélectionnelles qu'elle fait peser sur ses régimes demeure dans les études souvent éludée voire impensée. Or ces restrictions s'avèrent particulièrement complexes :

> « Les emplois de *en* révèlent des contraintes particulièrement foisonnantes. En particulier, la distribution des termes qui peuvent ou non suivre *en* se présente de façon apparemment anarchique. » (J.-J. Franckel & D. Lebaud, 1991 : 57)

Pourquoi peut-on dire par ex. des prisonniers, mais non des robinets, qu'ils sont *en fuite* ? Pourquoi peut-on circuler *en train (de nuit)*, mais non **en train bondé* ? S'il nous arrive d'être *en émoi*, *en joie, en dépression ou en colère*, pourquoi ne peut-on pas être **en émotion, *en jovialité, *en chagrin ou *en irritation ...* ?

C'est à D. Leeman que l'on doit le plus grand nombre de travaux mettant en jeu cette démarche « prudentielle » visant à explorer patiemment les restrictions sélectionnelles que fait peser la préposition sur ses régimes, en vue de mieux contrôler ensuite la puissance de l'identité sémantique proposée. Comment procède-t-elle ? Après avoir rassemblé de manière intuitive ou sur une base syntaxique et distributionnelle un ensemble de noms associable à une même notion[48] (par ex., les noms de vêtements, les noms de moyens de déplacement, les noms de sentiment...), sont listées les possibilités et impossibilités distributionnelles observées pour la combinaison de la préposition *en* avec ses arguments. Des hypothèses d'invariant sémantique sont alors avancées, mises en perspective et comparées avec celles proposées pour d'autres ensembles lexicaux (voir en particulier D. Leeman 1994, 1995, 1996, 1997, 1998). Cette méthode permet de contrôler la « puissance » explicative de l'identité sémantique énoncée en lui adjoignant certaines restrictions d'application mise au jour par l'étude systématique des contextes. Pour notre part, dans B. Martinie & D. Vigier (2013), nous avons cherché à mettre au jour certaines des contraintes exercées par *en* sur ses régimes nominaux abstraits dans la structure *être en N*, en mobilisant pour cela les catégories aspectuelles forgées initialement pour le lexique verbal (Z. Vendler, 1967). L'une de nos conclusions est que cette préposition « exige de son régime nominal abstrait qu'il entretienne un rapport avec le temps, ce qui exclut de la construction *être en N* les noms d'achèvement » (*op. cit.*, 78).

1.2.7. Conclusion

Dans cette section 1.2., nous avons voulu poser les jalons méthodologiques pour la construction de l'identité sémantique d'une préposition fortement polysémique (car peu colore), construction i) qui fasse le pari d'une approche « verticale » de la polysémie et de l'invariance de l'identité sémantique, ii) qui cherche à contrôler la puissance explicative de cette dernière, iii) qui propose un modèle de déploiement du sens articulant l'unicité abstraite de l'invariant sémantique avec l'extrême variété des acceptions observables en discours. Pour reprendre le titre de D. Stosic & B. Fagard (*op. cit.*) : qui parvienne à conjoindre *unicité* et *variabilité*.

2. Quelle identité sémantique pour *en* en français contemporain?

Dans une première étape, nous nous arrêterons sur le schème (ou « forme schématique ») que G. Guillaume avait affecté à *en* dans son étude de 1919 (§ 2.1.), étude séminale en ce qu'elle a fécondé (que cette dette soit ou non explicitement reconnue) un grand nombre d'études qui ont suivi. Dans une deuxième étape (§ 2.2.), nous présenterons ce qui constitue à nos yeux la meilleure formulation de l'identité sémantique de *en*. Nous tirerons profit des réflexions méthodologiques conduites dans la section précédente pour distinguer dans cette identité deux modules : un « module sémantique invariant », un module « restrictionnel » visant à contrôler la puissance du premier. Nous ferons enfin (§ 2.3.) une proposition de modélisation des variations du sens de *en* en discours.

[48] D'où le choix de structures de présentation des inventaires de sens de *en* de type « notionnel-générique » comme indiqué *supra*.

2.1. Le « schème » guillaumien : notions d'intériorisation réciproque de X et de Y, et de réversion de l'idée nominale sur le sujet

Il a souvent été reproché à G. Guillaume (voir en part. J .-J. Franckel et D. Lebaud, 1991 ; P. Cadiot, 1997) d'avoir sur-valorisé le couple *en-dans* au détriment d'autres couples tels que *en-chez* (*il y a en lui / chez lui beaucoup de haine*)*, en-de* (*un cartable en / de cuir*) *, en-à* (*être à la cuisine / en cuisine*), *en-comme* (*travailler en maçon / travailler comme maçon*) etc. Ce reproche est fondé et l'on sait aujourd'hui l'importance méthodologique que revêt l'exploration la plus systématique possible des micro-systèmes prépositionnels mis en jeu par les diverses configurations syntactico-sémantiques dans lesquelles *en* est susceptible d'entrer. Il n'en demeure pas moins que G. Guillaume a réussi à mettre magistralement en lumière ce qu'il nomme le *schème* propre à *en* : *l'intériorisation réciproque* des référents des termes X et Y mis en relation par la préposition. Cette facette de la sémantique de *en* n'a jamais été démentie depuis, même si souvent les auteurs ont affecté à ce schème un autre nom ou lui ont conféré une autre place dans l'économie générale de leur analyse.

2.1.1. Présentation du schème

En marche de pair avec *dans,* l'une étant dans la langue *la valeur déformée* de l'autre (1919 : 266). Le signe visible de cette déformation consiste en l'absence très régulière (quoique non systématique) de l'article devant le régime de *en*. Ainsi, tandis que *dans* met en relation deux entités nettement distinctes (*jeter un livre* dans *le feu* donne à voir séparément un livre et du feu), *en* offre à l'esprit l'image de deux entités d'abord distinctes qui prennent position si intimement l'une dans l'autre qu'elles se confondent en une seule (*jeter un livre* en *feu* montre un seul objet jeté). Dans ses emplois les plus visibles, cette intériorisation réciproque permet d'exprimer le résultat de processus concrets tels que la fabrication des artéfacts (*une table en bois*), la transformation (réelle ou magique : *Max a transformé son sous-sol en bowling* ; *la sorcière a transformé le prince en souris*), la destruction (*un livre en feu*, *réduire quelque chose en miettes*), etc. Le plus souvent cependant, *en* exprime un mouvement qui n'a pas de support matériel et qui s'opère de façon plus subtile par une « réversion de l'idée nominale » convoyée par le régime nominal de la préposition en « mode sur le sujet ». Par ex., le syntagme *un homme en prison* donne à voir « l'idée d'un homme sur qui s'est appesanti tout ce que le mot prison, interprété moralement, enferme de douloureux : bref, le prisonnier. » (*op. cit.*, 268). Réversion et intériorisation réciproque constituent l'avers et l'envers d'un même schème, la première étant la réalisation plus subtile et plus abstraite de la seconde.

Une telle analyse permet de rendre compte de nombreux contrastes observables dans l'usage de *en* et de *dans* et que mettent particulièrement en lumière des paires minimales. Par ex. :

(22) *Je suis*[49] *dans le fauteuil.* [le locuteur est assis dans son fauteuil au moment où il parle.]

[49] Le présent de l'indicatif dans les énoncés doit être entendu avec sa pleine valeur déictique

(23) *Je suis en fauteuil.* [le locuteur est handicapé et ne peut se déplacer sur ses jambes ; au moment où l'énoncé est formulé, « je » peut ne pas être dans son fauteuil roulant mais dans son lit par ex.]

(24) *Je suis dans le collège.*
(25) *Je suis en collège.* [le locuteur est élève ou enseignant dans un collège. « Je » n'est pas nécessairement situé dans son collège au moment où il s'exprime.]

(26) *Je suis dans le village.*
(27) *Dans mon cas, je suis en village, et il n'y a pas de transports en commun, c'est une calamité.* (Extrait d'un forum sur la Toile.) [Le locuteur veut spécifier avant tout un mode d'existence lié au fait qu'il vit en village ; là encore, une localisation spatiale de « je » n'est pas nécessairement opérée au moment où il formule l'énoncé.]

Tandis que les prédicats *être dans le (fauteuil + collège + village*) opèrent une localisation de leur argument externe, les prédicats *être en (fauteuil + collège + village*) opèrent une qualification de ce dernier : « je» est (respectivement) handicapé, personnel de l'enseignement ou élève, habitant d'un village. Le prédicat n'est plus nécessairement localisant puisque « je» peut se déclarer *être en (fauteuil + collège + village*) tout en se trouvant physiquement, au moment de l'énonciation, ailleurs que dans un *(fauteuil + collège + village*). Pour reprendre les mots de P. Cadiot (1997 : 192), *fauteuil, village, collège* « n'existent qu'à travers les activités du sujet et en viennent à construire de simples états sans localisation ».

Il convient à ce stade d'établir un lien avec les travaux de M. Aurnague (2010, 2012a) consacrés aux emplois de la préposition *à* de type « routine sociale[50] » comme dans *être au piano, être à l'hôpital, être au coin* etc. Sur un plan général d'abord, ce type de configuration dont la valeur sémantique met en avant une relation « fonctionnelle » entre cible et site – au même titre que celle instanciée par *en* dans des exemples comme (23)(25)(27) – dévoile l'intérêt d'une approche *fonctionnelle* et non simplement *géométrique* de la préposition. Comme le soulignent M. Aurnague & L. Vieu (2013), l'approche géométrique tend à attribuer « le rôle principal […] au complément de l'adposition et, par conséquent, au site » (*op. cit.*, 24) alors que l'approche fonctionnelle ouvre à une réelle analyse relationnelle « attribuant une égale importance à la cible et au site pour le calcul du sens » (*op. cit.*, 11). Sur un plan plus particulier, maintenant, les emplois de type « routine sociale » étudiés par M. Aurnague impliquent i) que les Nsite « n'ont pas pour fonction de localiser la cible », ii) que « leur interprétation est souvent non spécifique ou 'générique' », iii) qu'ils sont associés, « dans la connaissance de sens commun (et dans le lexique), à une activité ou un état – auquel prend part la cible » (2012, 190-191). Il est très frappant de constater les similitudes (mais aussi certaines différences) entre les emplois de « routine sociale » auxquels donnent lieu *à* et ceux que construit *en* (si l'on adopte la même terminologie).

Du coté des similarités, on rangera par exemple le fait que les « profils de routines » dégagés par M. Aurnague pour *à* (sans prétention d'exhaustivité précise-t-il) s'observent aussi pour *en*[51]. Ce point peut être illustré pour les quatre profils suivants : expression d'une « activité » ou d'un « état » de la cible mettant en jeu un site de type « objet » ou de type « lieu »:

[50] Ces travaux s'ancrent dans la distinction initialement proposée par C. Vandeloise (1988) pour les deux usages de la préposition *à* dans des configurations statiques : localisation et routines.
[51] Nous ne distinguons ici, pour des raisons de longueur, que 4 types (activité de la cible + site-objet etc.) sur les six proposés par l'auteur et pour tous lesquels *en* offre des exemples.

(28) *Pierre est au piano / Pierre est en moto* [activité de la cible + site-objet : *Pierre fait du piano / de la moto*]
(29) *Pierre est au balcon / Pierre est en cuisine* [activité de la cible + site-lieu : *Pierre regarde la rue depuis le balcon, … / fait la cuisine*]
(30) *L'eau est au congélateur / Pierre est en fauteuil* [état de la cible + site-objet : *l'eau est congelée / Pierre est handicapé*]
(31) *Pierre est au coin / Pierre est en prison* [état de la cible + site-lieu : *Pierre est puni / Pierre est prisonnier*]

Autre similitude : la prédication opérée n'implique pas nécessairement la présence d'un site du type attendu. Ainsi, un professeur peut déclarer dans sa cuisine *être en collège*, ou une infirmière dans son jardin *être à l'hôpital* (par contraste avec *être en libéral*).

Du côté des différences, on rangera le fait que si *à* peut exprimer une routine avec un régime nominal déterminé ou non (*être à l'hôpital / être à table*), *en* exige un régime nominal nu. *En* sélectionne l'interprétation d'activité ou d'état routinier *via* l'exploitation *intensionnelle* (car non déterminé) du contenu du Nsite, alors que pour *à* se pose la question de l'interprétation intensionnelle (voir N. Furukawa, 2010) *versus* définie ('para-intensionnelle') (voir F. Corblin, 2011) du défini. M. Aurnague (*op. cit.*) quant à lui argumente pour *à* en faveur d'une solution intermédiaire entre celles avancées par ces deux auteurs.

Il reste qu'une étude comparative fouillée de telles constructions avec *à* et *en* serait, comme l'observe l'auteur (2012 : 212), d'un très grand intérêt, notamment pour examiner certaines « paires minimales » exprimant toutes deux une routine. Par ex.

(32) *Pierre est à la mer / Pierre est en mer*
(33) *Pierre est au collège / Pierre est en collège*

2.1.2. Le schème guillaumien s'applique-t-il à tous les emplois ? Le cas de *en* suivi des noms de pays

Une des difficultés du schème guillaumien pointées dans D. Vigier (2013 : 6-7) concerne le caractère *systématique* de la réversion. On illustrera ce point en nous focalisant sur les emplois de *en* suivi d'un nom de pays (N_{PAYS}). Par ex.:

(34) *Max est en Espagne.*
(35) *Marie est en Irak.*

A suivre G. Guillaume, on devrait observer dans (34)(35) une réversion sur le sujet qui serait en revanche absente des deux énoncés suivants :

(36) *Max est au Portugal.*
(37) *Max est aux Etats-Unis.*

Adopter la thèse guillaumienne de la réversion dans les emplois de *en* suivi d'un N_{PAYS} pose des difficultés insurmontables. En premier lieu, on peine à contraster (34)(35) avec (36)(37) sous l'angle d'une réversion qu'opèrerait *en* mais non *à*. Autant celle-ci était très perceptible dans les paires minimales présentées *supra* ((22)-(23), (24)-(25), …), autant elle semble insaisissable ici. Parviendrait-on à la mettre en lumière qu'il faudrait alors expliquer

pourquoi la sémantique de *en* ne s'accommoderait pas des N_{PAYS} masculin singulier à initiale consonantique (36) ou de nombre pluriel (37).

D. Leeman (2015) a récemment travaillé sur cette distribution et, mobilisant le schéma tensif binaire de G. Guillaume, développe l'hypothèse suivant laquelle la préposition *en* d'une part, le genre grammatical féminin d'autre part, appartiendraient à une première tension tandis que *à* et le genre grammatical masculin relèveraient d'une seconde tension :

> « Le réseau explicatif réside donc dans le fait que (1) *en*, le nom sans article, le genre féminin relèvent de «l'avant» (c'est-à-dire du «donné», de «la nature», pour reprendre les termes de Eskénazi), à partir de quoi (2) les prépositions suivies de l'article (*à* en l'occurrence), le nom déterminé, le genre masculin relèvent de «l'après», de l'actuel (construit sur «l'avant») (*op. cit.:* 197)

Elle défend conséquemment l'idée suivant laquelle, dans un énoncé comme

(38) *Il est en Suisse.*

> « la lecture (française) charge immédiatement le sujet d'un certain statut, qu'il s'agisse d'un compte en banque, d'un enfant mis en pension, d'un artiste ou d'un sportif préférant ne pas acquitter ses impôts en France *»* (*op. cit.* 192).

Autrement dit, *en* sélectionnerait certains stéréotypes attachés à l'évocation de la *Suisse* qui seraient reversés en mode d'être sur le sujet (« il » veut dissimuler son argent au fisc, ou bien se débarrasser de ses enfants en les envoyant dans une prison dorée etc.).

A l'exemple proposé, on pourrait opposer le suivant :

(39) *Cette société est au Panama.*

où l'inférence selon laquelle le propriétaire de ladite société cherche à échapper au fisc serait tout aussi prégnante….

En tout état de cause, cette thèse énoncée dans D. Leeman (2015, 2016) et reprise dans D. Leeman & A. Falaise (2017) suivant laquelle la préposition *en*, le genre féminin et la voyelle relèveraient « d'une saisie d'avant, première, relativement floue, intuitive et subjective, de la notion à verbaliser » alors que la préposition *à*, le genre masculin et la consonne relèveraient « d'une saisie d'après, donc élaborée, abstraite, conceptuellement construite » nous semble très discutable.

D'abord, il conviendrait d'expliquer pourquoi on trouve aussi la corrélation *en* + *Nom masculin de pays* + *voyelle* : *en Iran, en Irak*, ….[52] Il conviendrait aussi d'expliquer pourquoi, dans des domaines d'emplois autres que la localisation au moyen d'un nom propre géographique, *en* s'accommode sans difficulté de noms masculins à initiale consonantique. Ainsi, pour les Nabstraits : *être en (déplacement* + *voyage* + *progrès* + *développement* + ... ; pour les Nconcrets : *être en (train* + *bus* + *plateau* + *rayon* + ...). Enfin, comme en

[52] Sur ce point, D. Leeman n'apporte à notre connaissance pas de réponse : « Une hypothèse possible, pour expliquer que *en* se trouve devant les noms masculins commençant par une voyelle, est que l'article, élidé, ne permet pas de savoir si l'on a affaire à un genre ou à un autre (voir l'Iran, l'Afghanistan, l'Ouganda, l'Uruguay, comme l'Europe, l'Irlande, l'Afrique, l'Egypte...) : pour une raison qui reste à trouver, ces associations auraient été versées au profit du féminin. » (D. Leeman, 1995 : 63-64). (C'est nous qui soulignons)

conviennent les auteurs (D. Leeman & A. Falaise, *op. cit.* : 59) « de plus en plus de noms masculins de région ou de département sont introduits par *en*, qu'ils commencent par une voyelle ou une consonne, et […] aucun n'est construit avec *à*. » Ainsi *en (Gers /Finistère / Calvados, ...) vs *au (Limousin / Vaucluse / Nord Pas-de-Calais, ...)*. » Comment rendre compatibles ces emplois en infraction avec la thèse tensive présentée *supra* ?

Pour ce qui concerne la distribution de *en* devant des noms de région et de département, D. Leeman & A. Falaise (*op. cit.*) écrivent:

> « Le nom géographique introduit par *en* représente (…) non une entité objective, abstraite, rationnellement élaborée, mais, au contraire, une réalité affective appréhendée à travers un vécu concret (…). Dans cette même perspective, l'emploi des noms de région (et de département) masculins avec la préposition *en* s'interprète comme l'écho, dans les perceptions des territoires, de la réforme de décentralisation entreprise dans les années cinquante, qui, conférant aux régions une autonomie et des prérogatives nouvelles, leur permet du même coup de se prévaloir d'une identité spécifique, de même que les départements qui les constituent prétendent désormais au statut de « terroir »».

Sans entrer dans le rôle accordé à la décentralisation pour expliquer ce type d'emploi, nous peinons à être convaincu par l'ensemble du raisonnement. Récapitulons-en rapidement les grandes lignes : on souhaite montrer que la réversion guillaumienne de l'idée nominale en mode sur le sujet se produit dans les cas où *en* est suivi d'un N_{PAYS} (*être en (France /Allemagne /...)*). Or la distribution de ces noms derrière *en* semble obéir à des règles morpho-phonologiques étrangères à la sémantique de la préposition. Comment expliquer en ce cas que la langue choisirait de mettre en jeu une réversion sur le sujet pour un prédicat comme *être en France* mais non pour *être en Portugal ?* On fait alors l'hypothèse que la distribution est en réalité soumise à des critères d'ordre sémantique plus profonds que les caractéristiques morpho-phonologiques observées : les noms féminins[53] et la voyelle[54] relèveraient d'une saisie tensive « d'avant, relativement floue, intuitive et subjective, de la notion à verbaliser ». Inversement, *à,* les noms masculins et la consonne relèveraient d'une saisie d'après. Autrement dit, les contraintes morpho-phonologiques observées affectant la distribution de *en* et de *à* devant les N_{PAYS}, tout en étant conservées, sont posées comme les manifestations en langue d'une saisie en *pensée* qui les fonde. A partir de quoi, on peut remotiver l'emploi de la préposition *en* et argumenter en faveur d'une réversion qu'il convient alors rendre sensible. Il demeure cependant à expliquer aussi la possibilité de trouver *en* devant des noms de région ou de département masculins à initiale consonantique – distribution en totale infraction avec l'hypothèse tensive[55]. On fait alors une seconde hypothèse suivant laquelle ces régions et département seraient saisis de manière affective et subjective, c'est-à-dire de manière précoce et cela malgré leur genre et leur initiale consonantique, tombant du même coup dans l'escarcelle de *en*. Mais ce faisant, on viole le principe précédent sur lequel on a fondé la distribution de *en* et de à devant N_{PAYS}. Comment expliquer que la saisie tardive associée au genre masculin et à l'initiale consonantique s'accommoderait désormais de la préposition *en* (saisie précoce) alors qu'une telle configuration était exclue précédemment (*être en *Pérou /*

[53] *«* Le féminin est de l'ordre d'une perception directe, spontanée, donc du côté de la préposition *en*. On peut expliquer par là que les noms de pays féminins soient introduits par la préposition *en*. » (D. Leeman, 2016 : 120)

[54] *«* Il subsiste toutefois un problème, c'est le statut des noms masculins qui demandent *à* s'ils commencent par une consonne mais *en* s'ils commencent par une voyelle (…) on voit de quel côté chercher à résoudre la difficulté : c'est de ranger la consonne du côté de l'après (donc du masculin et de la préposition *à*) mais la voyelle du côté de l'avant (donc du féminin et de la préposition *en*). » (D. Leeman, *op. cit.*: 120-121).

[55] Ici, aucune élision ne peut être invoquée (voir notre note précédente pour les cas de *en* devant N_{PAYS} masculins à initiale vocalique).

*en *Portugal / …*?). Comment justifier qu'on abandonne une mise en corrélation entre des formes et des saisies qui fondait l'hypothèse précédente sans pour autant abandonner aussi cette hypothèse ? On l'aura compris, l'argumentation développée par D. Leeman (2016), D. Leeman & A. Falaise (2017), A. Falaise & D. Leeman (2017) nous paraît présenter des difficultés insolubles.

Si l'on se détourne de l'hypothèse tensive présentée *supra*, quel parti prendre alors pour expliquer l'usage de *en* non seulement devant les N_{PAYS} mais aussi de régions et de département? En tout état de cause, il convient d'abandonner la thèse de la réversion guillaumienne qui, pour ces emplois, conduit à des apories. Comme nous l'avons dit dans D. Vigier (2013),

> *«* Vouloir à tout prix – et notamment d'explications souvent alambiquées – mettre au jour un mouvement de réversion dans tous les emplois de *en* conduit à conférer à ce mouvement mille visages qui finissent par en brouiller les contours » (*op. cit.*, 6)

Quant aux restrictions sélectionnelles observées avec les N_{PAYS}, nous considérons que seule la piste diachronique jouit d'une véritable consistance en la matière. On ne peut passer sous silence que la distribution de *en* devant N_{PAYS} recouvre exactement les restrictions sélectionnelles qu'a héritées cette préposition au cours du XVI^{e} s. à la suite de la disparition des amalgames issus de *en* + *le* et de *en* + *les* (cf. *infra,* partie 3). C'est la thèse que défend C. Molinier (1990 : 47) et à laquelle nous adhérons.

2.2. Quelle identité sémantique pour *en* ?

Nous proposons, à la lumière de nos réflexions méthodologiques antérieures, de distinguer dans l'identité sémantique de *en* deux « modules » : un module « sémantique » visant à établir la valeur sémantique invariante que nous affectons à cette préposition. Un module « restrictionnel » visant à contrôler la puissance du module précédent.

2.2.1. Module sémantique invariant

Nous considérons que l'« instruction de saturation référentielle réciproque » (ISRR) proposée par P. Cadiot (1997a) pour rendre compte de la sémantique de *en* permet à ce jour la meilleure couverture des emplois de cette préposition en discours, ainsi que la meilleure « économie » définitionnelle dans la mesure où (voir § 3.1.) elle évite d'introduire dans son identité des caractéristiques plus spécifiques comme celle de l'aspect.

Cette instruction se décompose en deux éléments:

1- coalescence de X avec les dimensions de Y ;
2- restriction de Y au cadre extensionnel fixé par X.

Dans la première composante instructionnelle, l'opération de coalescence est orientée de X vers Y : ainsi, dans

(40) *Rousseau en poète.*[56]

[56] Les exemples (40)(41)(42) sont repris de P. Cadiot (*op. cit.*)

l'être désigné par le N propre *Rousseau* est « projeté » dans les dimensions d'un poète. Pour le dire autrement, il y a « codimensionnalisation » et « fusion » du référent du Np *Rousseau* avec l'idée nominale abstraite de *poète*. Le même processus opère dans

(41) *Pays en guerre.*
(42) *Arbre en fleurs.*

respectivement entre les réalités dénotées par les N *pays, arbre* (X) et les N *guerre, fleurs* (Y).

La seconde composante instructionnelle est symétriquement orientée (de Y à X) et correspond au processus d'« application référentielle » (*op. cit.*, 191) de Y à X. Elle permet d'expliquer certains effets de sens propres à *en* tels qu'observables dans les paires minimales suivantes :

(43) *Il mourut en général d'armée.*
(44) *Il mourut général d'armée.*

(45) *Il a rédigé sa thèse en trois ans.*
(46) *Il a rédigé sa thèse pendant trois ans.*

Nous allons étudier le premier de ces couples d'exemples afin d'illustrer en détail les deux composantes de l'ISRR. Nous reviendrons plus loin (§ 3.1) sur la seconde paire d'énoncés qui met en jeu un complément prépositionnel de type *DétQuant NTps*.

L'exemple (43) met en jeu un prédicat second (*en général d'armée)* greffé sur une prédication première : nous parlerons d'attribut accessoire du sujet[57]. (44) permet de mieux saisir par contraste le rôle joué par *en* :

(44) *Il mourut général d'armée*

Dans cet énoncé, le SN prédicat second *général d'armée* stipule qu'au moment de sa mort, « il » possédait le grade de général d'armée.

Dans (43) en revanche, la présence de la préposition en tête de l'attribut ne conduit pas à la même interprétation. *Stricto sensu*, « il » n'est pas identifié comme possédant le grade de général d'armée de sorte que l'énoncé pourrait parfaitement s'appliquer à un civil prenant la tête d'une rébellion par ex. « Il » a eu simplement le comportement d'un général d'armée dans sa mort. Autrement dit, le locuteur qui énonce (43) produit un jugement de conformité qui consiste à attribuer au comportement adopté par « il » dans la réalisation du procès dénoté par le verbe l'ensemble des qualités stéréotypiquement associées au statut dénoté par Y.

> « Dans ce type d'emploi, le GP *en GN* exprime le résultat d'un jugement de conformité opéré par un locuteur/énonciateur[58] à propos du comportement d'un actant. Par *conformité*, nous entendons (comme D. Leeman (1995 : 60)) que ce comportement est évalué par rapport à une norme. Ainsi, déclarer par ex. que *X a parlé en tyran*, c'est considérer que dans son comportement, X s'est montré (plus ou moins) identique à une norme que l'on a en tête et qui réunit l'ensemble des traits typiques que l'on affecte à un tyran ». (Vigier, 2013 : 10)

Pour en revenir au rôle que joue la préposition *en* dans (43), nous dirons que

[57] Sur ce point, D. Vigier (2008, 2013)
[58] Sur cette distinction d'ordre énonciatif (locuteur *versus* énonciateur), voir D. Vigier (2008: § 47-51).

i) elle reprofile (ou travaille…) le signifié de Y en y sélectionnant un complexe de traits sémiques (désormais [C]) stéréotypiquement associés dans les circonstances fixées par le co(n)texte[59],

ii) elle projette par *coalescence* le référent de X dans les dimensions de Y reprofilé. Au terme de cette première composante instructionnelle (item *ii*)[60], il y a coalescence de X avec les dimensions de Y.

Le faisceau de traits qui forme le complexe [C] varie selon le contexte (large) de l'énoncé, comme l'illustre les exemples suivants :

(45) *Dans les discussions avec le gouvernement relatives au budget de la nation, il agit en général d'armée*
[suite 1] *en réclamant plus de crédits pour la défense.*
[suite 2] ?? *en prenant la tête des opérations.*

(46) *Dans cette ultime bataille à la baïonnette qui devait décider du sort de la victoire, il agit en général d'armée*
[suite 1] *en prenant la tête des opérations.*
[suite 2] ?? *en réclamant plus de crédits pour la défense.*

Selon le contexte, le complexe de traits activés par *en* pour le terme Y diffère, ce que met au jour le caractère pragmatiquement acceptable *versus* peu acceptable de la suite proposée.

Concernant la seconde composante instructionnelle, la comparaison de (43) et (44) est aussi éclairante :

(43) *Il mourut en général d'armée.*
(44) *Il mourut général d'armée.*

L'interprétation de (44) conduit à distinguer l'intervalle temporel de validité de la qualité attribuée (*être général d'armée* - au sens d'en posséder le grade) et l'intervalle temporel de validité du procès (*mourut*). Cette non-coïncidence d'intervalles apparaît lorsqu'on recourt à la glose (a) qui fait de la prédication première dans (44) une subordonnée temporelle, et du prédicat second le prédicat premier appliqué au sujet au moyen de la copule *être*.

(44a) *Quand il mourut, il était général d'armée.*

L'alternance passé simple et imparfait dévoile obliquement la disparité des intervalles.

En revanche, l'interprétation de (43) conduit à considérer que la qualification opérée sur le sujet par le prédicat second ne vaut que pour le temps de la mort de « il » : ce dernier s'est montré *général d'armée* dans sa mort seulement. La meilleure glose possible serait alors :

[59] Comme y insiste P. Cadiot (1997a : 25) « la préposition relie non des mots, mais des représentations ».
[60] Comme nous allons le préciser plus loin en effet, l'item i) ne relève pas de la composante instructionnelle de *en* mais doit être pris en charge par un autre module.

(43a) *Quand il mourut, il fut général d'armée.*

Cette co-extension de l'intervalle de validité de Y à celui de X s'explique par la seconde composante instructionnelle (« restriction de Y au cadre extensionnel fixé par X »), le terme X devant être identifié non pas au référent de « il » mais à l'état de chose construit par la prédication première « il mourut ».

Récapitulons : dans (43), les termes X et Y de la relation prépositionnelle sont respectivement :

- l'état de chose dénoté par la prédication première « *il mourut* »,
- un complexe de trait sémiques [C] associé au SN *général d'armée* et qui caractérise ce que serait son comportement stéréotypique dans les circonstances définies par le co(n)texte. Le « profilage » de Y est assuré par *en* dans le cadre d'une opération qui ne relève pas du noyau instructionnel (invariant) de la préposition mais qui est pris en charge par un autre module (voir *infra*).

Les deux composantes intructionnelles de *en* opèrent comme suit :

i) elle projette par *coalescence* le référent de X (« il » saisi dans la phase précédant sa mort) dans les dimensions de Y (reprofilé) : au terme de cette opération, le comportement du référent de « il » tel qu'il fut juste avant sa mort est jugé identique au comportement stéréotypiquement attaché à un général d'armée plongé dans des circonstances identiques ;

ii) elle restreint l'extension de Y au cadre extensionnel fixé par X : au terme de cette opération, la caractérisation de X par Y est conçue comme ne valant *que* pour l'intervalle temporel de validité fixé par X.

Cette analyse rejoint, par d'autres voies, celle de G. Gougenheim ([1950], 1970) qui contestait par ailleurs le statut attributif[61] du SP dans ce type de construction :

> « *Il a agi en roi, il est mort en brave, on le traite en esclave.* (…) On pourrait être tenté de considérer le groupe ainsi constitué comme une sorte d'attribut du sujet ou de l'objet. Mais à la différence de l'attribut, il n'est pas une qualification, même transitoire, du sujet ou de l'objet. Il n'est pas non plus une comparaison : *comme un roi, comme un esclave*. Il implique une sorte d'identité temporaire, limitée à l'action exprimée, entre le sujet ou l'objet et le nom précédé de *en*. » (*op. cit.*, 60)

L'évocation de *comme* par G. Gougenheim invite à s'arrêter sur la triade d'énoncés suivante :

(47) *Il travaille en maçon.*
(48) *Il travaille comme maçon.*
(49) *Il travaille comme un maçon.*

Selon C. Fuchs (1999), l'interprétation du tour *qualifiant* (48) induirait l'interprétation « *il est maçon* » : « il est maçon (ou du moins présenté comme tel), et travaille ès qualités » (*op. cit.*, 77)). Ce serait l'inverse (*« il n'est pas maçon »*) pour le tour *échantillant* (49): « « il » a seulement la manière de travailler d'un ou du maçon » (*ibid.*). Cette différence d'interprétation ne serait pas due à *comme* mais à la présence / absence de l'article.

[61] Il reste, comme l'écrit plaisamment M. Riegel (1985 : 12), que les divergences que suscite l'attribut « *se résolvent habituellement par la formule consacrée : « tout dépend de ce que vous entendez par attribut* » ».

La précision de l'auteur « *du moins présenté comme tel* » est cruciale pour distinguer le rôle de *comme* qualifiant de celui joué par *en*. Ainsi, lorsque quelqu'un déclare :

(50) *Je travaille comme maçon depuis une semaine.*

cela n'implique pas nécessairement qu'il a pour métier d'être maçon et qu'il est employé conformément à son métier (« ès qualités » dit C. Fuchs). L'énoncé suivant serait donc parfaitement acceptable :

(51) *J'ai pour métier d'être cuisinier et pourtant je travaille comme maçon depuis une semaine*.

Ce qui importe, c'est la manière dont le prédicat *présente* « il » dans le tour considéré. Dans (47)(50), « il » est présenté comme ayant le *statut* d'un maçon, qu'il le soit ou non en réalité : ainsi peut-on travailler comme maçon tout en ayant eu la formation d'un cuisinier (51). Dans (48) en revanche, « il » est présenté comme ayant le comportement stéréotypique d'un maçon dans son travail, qu'il le soit ou non en réalité. Voilà pourquoi (dans un tout autre registre) on peut *mourir en héros* tout en ayant vécu *en lâche*…

2.2.2. Module distributionnel-restrictionnel

Ce module consiste à contrôler la *puissance* du module sémantique d'invariance par la mise au jour des restrictions sélectionnelles qu'impose la préposition à ses régimes.

Il s'agit du module le plus « ardu » pour les études linguistiques portant sur les prépositions. L'expérience a montré combien il est difficile de mettre au jour les caractéristiques sémantiques susceptibles d'expliquer les restrictions que manifeste l'usage de telle ou telle préposition. Pour l'instant, en ce qui concerne *en,* la récolte demeure maigre. D'autant plus que s'avèrent inappropriées certaines des restrictions sélectionnelles jusqu'ici identifiées. Ainsi :

- Concernant les noms de matière (*une montre en or, un mur en pierre, ...*), D. Leeman (1994 : 114) parvient à la conclusion que le N régi par *en* ne peut pas dénoter la matière « naturelle » de X (*une maison en bois versus* ??*un arbre en bois*). Est-ce là une restriction sélectionnelle imposée par *en* ? C'est la thèse de l'auteur. On peut néanmoins s'interroger : ne s'agirait-il pas plutôt, comme nous l'a suggéré M. Aurnague, d'un principe pragmatique plus vaste s'imposant à diverses constructions de type *prép. N* lorsqu'elles expriment une composante d'un référent ? Ainsi A. Borillo (1996) observe-t-elle que dans les constructions binominales *N1 à N2* de type partie-tout (*verre à pied, fauteuil à roulettes, ...*), les cas où N2 « fait partie de l'objet de manière définitoire » (*op. cit.*, 113) rend la construction peu acceptable. Ainsi *« ??une voiture à volant, ??un vélo à pédales, ... »*. Cela tiendrait à ce que ces expressions « sont perçues comme redondantes, non-informatives »[62]. On peut se demander si les constructions binominales *N1 en N2* où N2 exprime une partie – en l'occurrence, la matière – de N1, ne tomberaient pas sous le coup du même principe pragmatique, le caractère peu acceptable de la séquence ??*un arbre en bois* n'ayant alors aucun rapport avec une restriction sélectionnelle imposée par *en*. Ainsi pourrait-on opposer la faible acceptabilité de ??*un diamant en carbone* à l'acceptabilité de *un*

[62] Il suffit d'ajouter à N2 une qualification pour que la construction redevienne acceptable : *une voiture à volant réglable, un vélo à pédales escamotables, ...* Le SP redevient alors informatif.

diamant en carbone synthétique/de synthèse, la caractérisation de N2 rétablissant l'informativité du tour.

- Nous sommes aussi en désaccord avec les conclusions de D. Leeman (1994) selon qui (*op. cit.*: 104-107) les noms dénotant un moyen de déplacement nécessiteraient d'avoir affaire à des entités *construites* (venir *en voiture / à pied* versus **à voiture / *en pied*). A cela on peut opposer que les locuteurs emploient sans difficulté des séquences telles que « *safari / excursion /… en chameau* ».

- Nous peinons en outre à considérer que pour les noms de sentiment, les restrictions de sélection en jeu (par ex. **être en peur* versus *être en colère*) coïncideraient avec le statut d'état manifesté (*colère*) *versus* « la concrétisation d'une qualité naturelle de l'individu » (D. Leeman, 1995 : 67). Si une telle opposition paraît s'appliquer aux noms ayant un sens proche de celui signifié par colère – *être en (rogne + rage + pétard + fureur)* –, elle ne permet pas de rendre compte de la difficulté de **être en mécontentement, *être en peur*, etc., puisque le mécontentement ou la peur sont a priori des états que l'on peut manifester ostensiblement, qui peuvent se lire sur un visage, etc. :

  (52) *Il a manifesté (son mécontentement + sa peur).*
  (53) *(Le mécontentement + la peur) se lisait sur son visage.*

- Concernant enfin les noms abstraits, nous espérons avoir montré (B. Martinie & D. Vigier : 2013) que dans les constructions *N0 être en N1*, *en* exige de son régime nominal abstrait qu'il entretienne un rapport avec le temps, ce qui exclut de la construction *être en N* les noms d'achèvement (**être en explosion, …*). En outre, *être en* N_{STATIF} implique une lecture stative-résultative au sens où l'entend D. Creissels (1999 : 95) : « représentation d'un état comme découlant d'un événement antérieur au repère temporel relativement auquel cet état est envisagé ». Dès lors, l'impossibilité pour *en* de sélectionner des $N_{STAT.QUAL}$ trouve une explication évidente, puisque les qualités ne peuvent pas, par définition, être associées à des processus antérieurs dont elles constitueraient le résultat. D'où l'impossibilité d'énoncés tels que **être en maigreur, *être en intelligence, ….*

Comme on le voit, le champ immense ouvert par ce module est pour l'essentiel encore en friches…

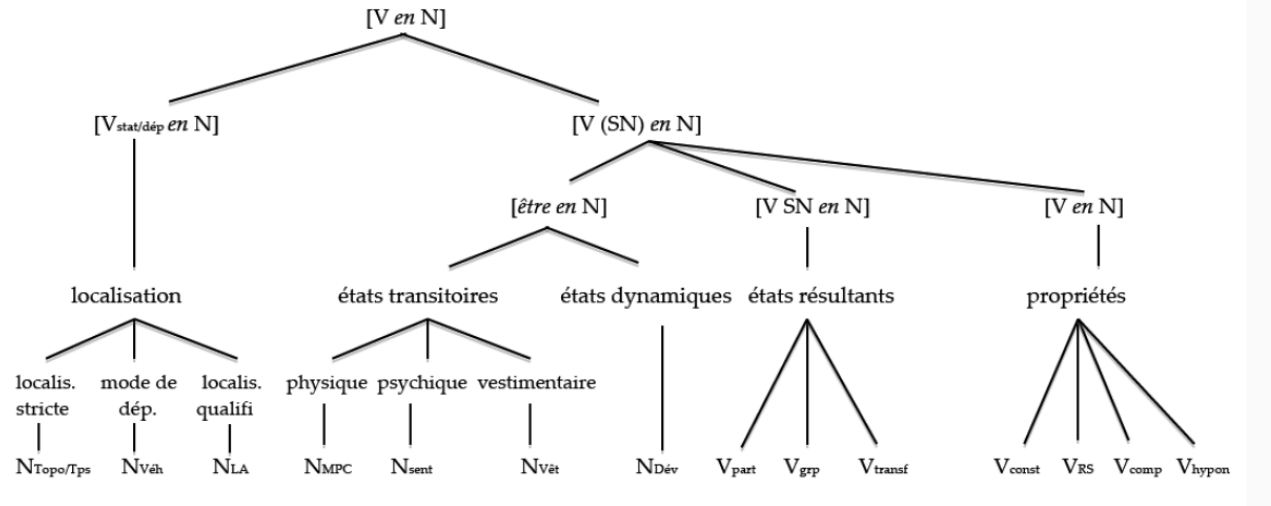
2.3. Modélisation des variations de sens en discours. Approche « constructionnelle »

Il est nécessaire de ménager, entre le niveau très abstrait de l'identité sémantique de la préposition et celui de l'extrême variété de ses acceptions en discours, un niveau intermédiaire dont la structure reflète nos hypothèses quant à la manière dont le sens prépositionnel se déploie en discours. Nous avons cursivement évoqué *supra* quelques propositions de structure de déploiement des interactions contextuelles (C. Guimier (1976), J.-J. Franckel & D. Paillard (2007), L. Melis (2003)). Nous voudrions dans les lignes qui suivent présenter une autre modélisation possible fondée sur une approche syntaxique et constructionnelle. L'essentiel de notre propos se fondera sur W. de Mulder & D. Amiot (2013) et, plus ponctuellement, sur I. Khammari (2008).

Comme le font observer W. de Mulder & D. Amiot (*op. cit.:* 22), selon la grammaire de

construction (notamment A. E. Goldberg 1995, 2006 ; W. Croft & A. Cruse, 2004) « le lexique-grammaire est un inventaire structuré de constructions, c'est-à-dire d'unités de forme et de sens, dotées de règles d'interprétation sémantiques spécifiques et suffisamment fréquentes pour être stockées telles quelles dans l'esprit des locuteurs ». Qu'il s'agisse de « substantive constructions », de « schematic constructions » ou de « semi-schematic constructions » (voir W. de Mulder & D. Amiot (*op. cit.:* 23)), les constructions envisagées ont l'immense avantage pour nous (en vue de proposer un inventaire stable et structuré des effets de sens de *en* en discours) de présenter des unités stabilisées de forme et de sens.

Voici – pour ce qui regarde du moins les configurations de type [V (SN) en N] où *en N* est dans la dépendance d'un verbe – l'arbre[63] auquel aboutissent les auteurs et que nous illustrerons par des exemples tirés de leur article :



(54)	(55)	(56)	(57)	(58)	(59)	(60)	(61)	(62)	(63)	(64)	(65)	(66)	(67)

(54)*Séjourner en Bretagne / Espagne / …*
(55)*Venir en voiture / bus / chameau / …*
(56)*Etre en ville / prison / clinique / …*
(57)*Etre en sueur / sang / …*
(58)*Etre en colère / extase / panique / …*
(59)*Etre en pyjama / slip / costume / clown / chevalier / …*
(60)*Etre en vadrouille / pleurs / fuite / …*
(61)*Diviser/répartir / ... en trois parts égales / …*
(62)*Regrouper / réunir / ... en un seul groupe / …*
(63)*Transformer / changer / métamorphoser / ... le prince en grenouille / …*
(64)*Le dîner consistait en une soupe froide et un morceau de fromage sec / …*

[63] Le sens des troncations opérées par les auteurs pour désigner le sémantisme du nom ou du verbe impliqué dans la construction sont les suivants : « **Vstat/dép** »= verbes statifs ou de déplacement. Feuilles de l'arbre : « **Ntopos/tps** » : toponymes/noms de temps ; « **Nvéh** » = noms dénotant une entité permettant de se déplacer ; « **NLA** » = noms de lieu d'activité ; « **NMPC** » : noms qui dénotent une manifestation physique concrète ; « **Nsent** » = noms de sentiment ; « **Nvêt** » = noms qui dénotent un vêtement ; « **Ndév** »= noms déverbaux ; « **Vpart** » = verbe exprimant une partition ; « **Vgrp** » = verbe exprimant un regroupement ; « **Vtransf** » = verbe exprimant une transformation ; « **Vconst** » = verbe exprimant une constituance ; « **VRS** » = les auteurs ont ici commis une erreur dans leur tableau ; il faut lire « Vcomp » : verbe de comportement suivi d'un nom dénotant un rôle social ; « **Vcomp** » : seconde erreur des auteurs ; il faudrait lire « Vquant » = Verbe de comparaison de valence 3 ; « **Vhypon** » : verbes sémantiquement différents mais avec des objets « hyponymiques » : dans tous les cas le nom introduit par *en* explicite une des composantes internes du verbe.

(65)*Pierre se comporte en automate / se conduit en adulte / ...*
(66)*Marie égale X en gentillesse / dépasse X en intelligence / ...*
(67) *Pierre parle / s'exprime en anglais // Les pertes se chiffrent en tonnes / ...*

On observera en premier lieu que les auteurs discriminent entre elles des sous-constructions

- sur le critère des **noms régimes** de *en* : ainsi, à l'intérieur de la construction [*V stat/dép en N*] sont distinguées sur le critère du nom régime les *localisations* : stricte (54) *versus* qualifiante (56) *versus* mode de déplacement (55). A l'intérieur de la construction [*V être en N*] sont distingués sur le critère du nom régime les états transitoires d'une part – qu'ils soient physique (57) *versus* psychique (58) *versus* vestimentaire (59), les états dynamiques d'autre part (60).
- Sur le critère des **verbes** ayant dans leur rection un complément à tête *en* : dans la structure [*V SN en N*] à interprétation « état résultant » sont ainsi distinguées les sous-constructions à verbes de : partition (61) *versus* regroupement (62) *versus* transformation (63). Enfin, dans la structure [*V en N*] à interprétation « propriétés » sont discriminées sur le critère du sémantisme verbal ou de la relation sémantique objet-verbe les sous-constructions où *en N* entretient une relation de constituance avec le sujet (64) *versus* les constructions que nous avons nommées *supra* à « jugement de conformité comportementale » (65) *versus* d'évaluation quantitative (66), *versus* à « objets hyponymiques » (67).

On peut bien entendu discuter certains des choix opérés dans cette classification. Par ex., les « états transitoires vestimentaires » (59) possèdent aussi un statut d'état « résultant » : « *Pierre est en culotte courte* » s'interprète comme le résultat d'un procès du type « *Sa mère habille Pierre en culotte courte / Pierre s'habille en culotte courte* ». Certaines constructions nous semblent en outre absentes : par ex. celles du type [*N0 V N1 en N2*], où N1 exprime une forme et N2 une matière : *construire une maison en briques*, qui relèverait des prédicats à état résultant, mais distincts des prédicats transformateurs en ceci que le verbe ne change pas une entité existante en une autre, mais la porte à l'existence.

Il n'en reste pas moins que cette voie nous semble particulièrement fructueuse dans la mesure où elle permet de conjoindre des effets de sens stabilisés en discours et des configurations syntaxiques « dédiées ».

Il nous paraît enfin qu'à ces constructions peuvent être, dans un second temps, associées d'autres constructions de type [*N0 V (N1) en N2]* où *en N2* assume une fonction syntaxique non de complément argumental du verbe (cas envisagé par les auteurs, à la suite I. Khammari, 2008) mais d'ajout verbal. Nous ne développerons pas ici ce point, nous proposant simplement de l'illustrer par quelques pistes de travail qui s'attachent à la construction catégorisée par W. de Mulder & D. Amiot comme : « *propriétés > verbes de comportement* (Vcomp) » comme dans :

(65) *Pierre se comporte en automate / se conduit en adulte / ….*

A cette construction, nous proposons d'associer les structures déjà étudiées *supra* où i) le verbe ne ressortit plus à la catégorie lexicale spécifique du comportement, ii) où le SP n'est plus un complément syntaxiquement argumental du verbe iii) et peut s'analyser comme un attribut accessoire du sujet. Par ex.

(68) *Pierre agit / revient / parle /… en tyran.*

L'extension à des compléments adnominaux nous semble également possible, étant entendu qu'à chaque construction verbale ne s'associe pas nécessairement un (seul) type de structure *N1 en N2*. Ainsi, à la sous catégorie « *état transitoire* > *vestimentaire* » on pourrait faire correspondre des séquences à complément adnominal comme

(69) *Un homme en costume.*
(70) *Un enfant en maillot de bain.*
(71) *Louis Jouvet en Jean-Jacques Rousseau.*
…

Hors contexte, on observera en revanche que la séquence

(72) *Le roi en grenouille*

est ambiguë : *état transitoire* > *vestimentaire* (déguisement : état résultant d'un déguisement vestimentaire ?) ou *état résultant* d'une transformation magique?

Nous ne voyons pas, en revanche, quelle séquence à complément adnominal pourrait être affectée à la catégorie « *propriétés* > *verbes de comportement* » qui paraît exiger la présence d'un verbe pour être actualisée.

Pour ce qui regarde les structures de type [*ADJ en N*], sous réserve d'inventaire plus systématique, elles relèveraient plutôt du type *localisation qualificative*, pourvu qu'on ouvre cette catégorie aux localisations « abstraites » dénotant des domaines d'activité et de connaissance (*expert en médecine/linguistique/chèvres/…*). Cette ouverture devrait concerner aussi les constructions avec complément argumentaux (*je suis en linguistique*[64]) ou ajouts (*je m'ennuie en médecine*) ainsi que les compléments adnominaux (*un étudiant en médecine*).

Evoquons pour finir le cas des « incidents » (B. Lavieu, 2006), compléments syntaxiquement affranchis de toute relation avec le verbe et qui sont i) aptes à figurer détachés en tête de phrase négative et ii) inaptes à être focalisés par *c'est… que*. (Voir aussi C. Molinier & F. Levrier, 2000 cités *supra*). Tel est le cas par ex. de

(73) *En résumé / En outre / En toute sincérité, cette affaire a été un fiasco*.

Les résultats des travaux conduits sur les adverbes (entre autres, C. Guimier (1996), Le Goffic (1993), L. Melis (1986), C. Molinier & F. Levrier (2000),…) offrent aujourd'hui la possibilité de dresser un tableau raisonné des grandes catégories de sens que convoient ces incidents. On peut illustrer ce point par la sous-catégorie des disjonctifs de style qui disent quelque chose « de la relation de l'énonciateur à l'interlocuteur ou à l'énoncé qu'il formule » (C. Molinier & F. Levrier, *op. cit.* : 49). La plupart de ces compléments peuvent être considérés comme portant sur un performatif de type « dire » placé dans une phrase supérieure. Ainsi, pour les énoncés suivants

(74) *En clair,*
(75) *En deux mots,*

[64] Énoncé par ailleurs ambigu, la discipline pouvant dénoter *via* une connexion pragmatique le lieu où elle est enseignée (= *je suis en cours de linguistique*). Sur ce point, voir D. Vigier (2003).

(76) *En bref,*
(77) *En toute honnêteté, cette affaire est un fiasco.*
(78) *En toute franchise,*
(79) *En toute confidence,*

le SP incident peut constituer le modifieur du verbe *dire* dans une phrase matrice :

(80) *Je dis* { *en clair,* / *en deux mots,* / *en bref,* / *en toute honnêteté* / *en toute franchise,* / *en toute confidence,* } *que cette affaire est un fiasco.*

Or on peut distinguer sur des critères sémantiques les énoncés (74) à (76) d'une part, (77) à (79) d'autre part. Les premiers disent quelque chose de la relation du locuteur avec son énoncé tandis que les seconds disent quelque chose de la relation du locuteur avec son interlocuteur.

Bref, les sous-catégorisations disponibles des adverbes du français fournissent – pourvu qu'on les étende aux SP adverbiaux – un cadre de modélisation possible des grandes catégories d'effets de sens associables aux incidents de type *en N*.

Récapitulons : pour ce qui concerne la préposition *en* (mais le raisonnement devrait pouvoir être étendu à d'autres prépositions), les cadres qu'offrent d'une part les grammaires de construction, d'autre part les travaux récents sur les adverbes et les adverbiaux permettent d'élaborer une structure susceptible de rendre compte des principales « zones » de stabilité sémantique discernables parmi l'ensemble des interactions possibles de la préposition avec ses contextes d'emploi.

2.4. Conclusion

Notre ambition dans cette première partie ne consistait pas à présenter une nouvelle identité sémantique pour *en* mais, oserons-nous dire, à tirer le meilleur partie de l'offre disponible sur le marché des études déjà publiées. La préoccupation principale qui nous a animé a consisté à ressaisir dans un point de vue surplombant, moins[65] le « territoire » des prépositions que la « carte » qui en a été dressée par les linguistes. Cela afin de revenir sur certains choix qu'offrent des carrefours « stratégiques » dans l'analyse : choix entre approches verticale *versus* horizontale de la polysémie, choix d'approche de l'identité sémantique (en termes de « signifié de puissance », de « forme schématique » etc.), choix ou non de l'invariance sémantique, etc.

Dans les lignes qui suivent, nous voudrions revenir sur l'instruction de *en* et montrer que l'ISRR permet de faire l'économie du trait aspectuel de perfectivité que plusieurs linguistes ont été tentés d'affecter à *en* (outre D. Leeman & C. Vaguer 2014, discuté ci-dessous, voir aussi J.-M. Merle, 2008). Nous présenterons enfin cursivement les pièces d'un dossier que

[65] « Moins » car celui-ci est resté dans le champ de l'examen, tout en occupant une place plus périphérique.

nous comptons instruire dans un proche avenir : celui des relations entretenues par *en* avec les termes de couleur.

3. Dialogue critique autour de l'identité de *en*

3.1. Réfutation de la thèse suivant laquelle *en* possèderait une valeur aspectuelle perfective[66]

Nous allons traiter des adverbiaux de type *en DétQuant Ntps*[67] occupant une position postverbale liée comme dans (81) (adverbial intra-prédicatif endophrastique (C. Guimier 1996 & *infra*)). Ces constituants ont été récemment étudiés dans D. Leeman et C. Vaguer (2014)[68], article qui servira de fil conducteur à notre propre analyse.

Nous ne traiterons pas les emplois où le SP figure en position liée dans un prédicat verbal dénotant une itération (82) ou modifié par une négation totale (83),[69] ni ceux où il apparaît en position détachée préverbale (84). Nous les avions déjà étudiés de près dans notre thèse (2004 : 117-127, 211-222) et même si des mises à jour devraient être apportées à ces analyses, elles demeurent pour l'essentiel valides sur le fond.

(81) *Je ferai le trajet en trois jours.* (exemple forgé par G. Gougenheim (1950) et souvent cité dans les travaux sur la sémantique de *en.*[70])
(82) *Max a fumé deux cigarettes en un an.*
(83) *Je tiens à préciser que je n'ai jamais trompé ma femme en douze ans.* (Forum sur la Toile.)
(84) *En trente ans d'indépendance, l'armée française a manifesté concrètement sa présence durant vingt-et-un ans*. (*Le Monde Diplomatique,* janvier 1991.)

Nous reviendrons d'abord rapidement sur le fait bien connu que les SP étudiés imposent aux énoncés au sein desquels ils figurent des restrictions sélectionnelles touchant à l'aspect. Nous enchaînerons sur le rappel d'une hypothèse présentée par D. Leeman et C. Vaguer (*op. cit.*) touchant à la valeur aspectuelle perfective attribuable à *en* et que nous voudrions discuter. Nous montrerons que la notion d'*instruction de saturation référentielle réciproque* (P. Cadiot, 1997) présentée et discutée *supra* suffit à rendre compte des restrictions sélectionnelles que

---

[66] L'argumentation développée dans cette section a fourni une partie d'un article à paraître (D. Vigier, *à par.* 2017).

[67] Cette notation est empruntée à A. Borillo (1998), qui reprend à M. Gross (1986 : 207) la notion de « *nom de temps notée Ntps* ». Tous les noms listés par M. Gross dans son ouvrage ne sont pas susceptibles d'apparaître dans ce type de complément. Parmi les noms dénotant les « *divisions habituelles du temps* » par ex., les N *seconde, minute, heure* semblent échapper à toute restriction (*En (deux + quelques + plusieurs + ...) (secondes + minutes + heures)*) tandis que d'autres Ntps apparaissent moins attendus dans ce type de syntagme : *En (deux + quelques + plusieurs + ...) * (aubes + aurores + lendemains +...)*. En outre, on peut trouver dans des séquences plus ou moins figées outre les *Ntemps*, des noms qui « incorporent » une notion de durée sans être des Ntps à proprement parler : *en (deux + trois) coups de (cuillère à pot + louche)*, *en un clin d'œil*, e*n cinq sec*, etc.

[68] Les deux contributrices adoptent la notation suivante : $<en + Dét_{Indéf.} + N_{temps}>$.

[69] Nous avons pour partie traité ces configurations dans D. Vigier (2004 : 117-127, 211-222).

[70] L'auteur déclare à propos de cet énoncé qu'il « signifie que le trajet s'effectuera (et s'effectuera rapidement) en occupant ces trois jours. Il (…) y a (…) une espèce d'identité entre le trajet et les trois jours qu'il durera, une prise de possession de ces trois jours par le trajet » (189). A cette formulation, nous n'ajouterions pour notre part pas grand chose, la remodelant et la précisant simplement dans le cadre de l'ISRR de P. Cadiot (1997) – voir *supra*. Quant à la « rapidité » associée à cette quantification temporelle, nous la contestons dans notre article (à par. 2017).

ces SP imposent à la valeur aspectuelle de la situation construite par leur prédication d'accueil, et qu'il est en conséquence inutile d'intégrer dans l'invariant de *en* un trait relatif à l'aspect.

3.1.1. Restrictions de sélection imposées par les SP *en DétQuant Ntps* en emploi intra-prédicatif sur l'aspect des situations dénotées par le reste de la prédication

Dans la perspective de C. Guimier (1996 : 5-7), sont syntaxiquement intra-prédicatifs les constituants intégrés au prédicat verbal : ils ne sont en général pas séparés du verbe par une virgule à l'écrit, ni intonativement à l'oral. Sur le plan sémantique, ils sont endophrastiques c'est-à-dire qu'ils participent à la construction du sens référentiel. Si l'on considère le cas des SP *en DétQuant Ntps* comme illustrés sous (81), ils dénotent le temps nécessaire à l'effectuation intégrale de la situation construite par l'ensemble de la prédication[71], ce qui n'est plus le cas si le syntagme verbal (SV) exprime une itération[72] comme dans (82) ou que le prédicat est nié (83).

(81') *Faire intégralement le trajet me réclamera trois jours.*
(82') ♯ *Fumer intégralement deux cigarettes a réclamé un an à Max.*[73]
(83') ♯ *Ne jamais tromper ma femme m'a réclamé douze ans.*

N'importe quel énoncé n'est pas compatible avec un SP *en DétQuant Ntps* en position intraprédicative liée. Les restrictions sélectionnelles que cet adverbial impose à la situation dénotée par sa prédication d'accueil ont été utilisées comme test, généralement en parallèle avec le test de la compatibilité avec le complément aspectuel *pendant DétQuant Ntps*. On peut ainsi distinguer deux grandes classes de *situations* : les accomplissements et les activités pour reprendre la terminologie (francisée) de Z. Vendler (1957). Seules les premières sont compatibles avec une quantification temporelle au moyen de *en DétQuant Ntps* ((85) *versus* (86)). Les secondes acceptent une quantification de la durée avec *pendant DétQuant Ntps* (87), possible aussi (sous certaines conditions[74]) pour les accomplissements (88).

(85) *Mon voisin repeindra sa cuisine en deux heures.* (situation d'accomplissement : [+durée][+dynamique][+télique].)
(86) **Mon voisin nagera en deux heures.* (situation d'activité: [+durée][+dynamique] [-télique].)
(87) *Mon voisin nagera pendant deux heures.*

---

[71] Suivant en cela A. Borillo (1991), nous adoptons le point de vue selon lequel l'aspect ne résulte pas seulement de la sémantique du verbe voire du SV, mais inclut la prise en compte d'un ensemble de morphèmes présents dans la prédication et influant sur la valeur aspectuelle de la « situation » construite par la totalité de la prédication (semi-auxiliaires, déterminants, adverbiaux aspectuo-temporels, …).

[72] Sauf dans le cas où la somme des procès itérés occupe tout l'intervalle de temps dénoté par l'adverbial. Par ex. « Son cœur bat à 50 pulsations par minute. Autrement dit, il bat 36000 fois en une journée. » Glose possible : « Battre 36000 fois réclame à son cœur une journée ».

[73] Le symbole ♯ signale que la glose proposée, même si elle est interprétable (Max pourrait faire un concours de lenteur pour fumer deux cigarettes, l'accès à la fidélité conjugale pourrait avoir réclamé douze ans au locuteur), ne rend pas compte du sens de l'énoncé source.

[74] Il convient que le contexte favorise une interprétation du procès d'accomplissement comme provisoirement interrompu avant de pouvoir reprendre jusqu'à atteindre sa borne finale inhérente. J.-P. Desclés (1991 : 183) parle en ce cas d' « événement non achevé » : « Jean a écrit sa thèse pendant deux mois cet été puis, il est parti au Canada. » (*versus* « événement achevé » : « Jean a écrit sa thèse en deux ans avant de partir dans un laboratoire étranger. »)

(88) *Mon voisin repeindra sa cuisine pendant deux heures, s'arrêtera pour déjeuner, puis continuera l'après-midi.*

Les situations dénotant des états et des achèvements peuvent être compatibles avec une quantification de la durée au moyen d'un SP *en DétQuant Ntps* : ce dernier exprime alors la durée nécessaire à l'effectuation de la phase préparatoire à l'entrée dans l'état ou dans l'activité.

(89)*Marie fut prête à sortir en à peine dix minutes.*
(89') *Marie mit à peine dix minutes avant d'entrer dans l'état « être prête à sortir »* [glose possible].
(90)*L'avion supersonique a franchi le mur du son en quelques secondes.*
(90') *L'avion a mis quelques secondes avant de franchir le mur du son* [glose possible].

3.1.2. Adverbiaux aspectuels *en DétQuant Ntps* et valeur sémantique de *en* : l'hypothèse aspectuelle de D. Leeman & C. Vaguer (2014)

On est immanquablement conduit à chercher quel lien établir entre l'identité sémantique de *en* (quel que soit son contexte d'emploi) et la valeur qu'elle confère aux SP aspectuels dont elle constitue la tête. D. Leeman et C. Vaguer (*op. cit.*) s'y sont employées et nous renvoyons donc prioritairement à leur étude. Nous voudrions cependant présenter ici un point sur lequel notre analyse diffère de la leur : la préposition *en*, nous disent les auteurs, convoierait une valeur aspectuelle de type perfectif. Leur hypothèse tire son origine du fait « *que ce type de constituant* [= <*en* + $Dét_{Indéf.}$ + N_{temps} >] sert *précisément de test pour repérer un emploi verbal de type « accomplissement »* » (*op. cit. :* 407). Ainsi, dans l'exemple

(91) *Balthazar fait ses devoirs en un quart d'heure*

« l'ajout en un quart d'heure spécifie la durée de « *faire ses devoirs* » en la montrant « *bornée* », i.e. dotée d'un point de départ et d'une limite de fin (le procès est télique : le résultat est atteint), ce qui constitue l'apport de la préposition (par opposition, par ex., à *pendant (un quart d'heure), pour (une semaine)*, susceptibles également de marquer la durée. » (*op. cit.*, 403)

Soucieuses d'examiner si leur hypothèse élaborée à partir des emplois de *en* tête des compléments aspectuels peuvent s'étendre aux autres emplois de la préposition, D. Leeman et C. Vaguer (*op. cit. :* 412-414) se tournent successivement vers ses usages « spatial » et « notionnel ». Dans le premier cas, sauf erreur de notre part, nous n'avons trouvé aucune justification du caractère présumé borné de *en.* Pour le second, les auteurs envisagent des énoncés comme :

(92) *Un élève en difficulté.*
(93) *Un colis en attente.*

L'aspect clairement statif du nom régi par *en* dans ces deux exemples, observent-elles, semble contredire la valeur aspectuelle de perfectivité attribuée à *en* : « l'hypothèse que *en* renferme une valeur aspectuelle de l'ordre du borné, donc du perfectif, voire de l'instantané,

n'est-elle pas contredite par le fait que cette préposition sélectionne des noms d'aspect statif[75] – l'état étant, justement, aspectuellement non borné *?* » (*op. cit.*, 414).

Selon les auteurs, la valeur perfective de *en* se manifesterait, en ce cas, sur un plan plus conceptuel : le pouvoir de discrétisation propre à la perfectivité s'exercerait en ceci que le SP délimite un sous-type, une catégorie (d'élève (92), de colis (93)). « L'expression *en attente* indique certes l'état présent du colis (et en cela n'est pas d'ordre perfectif, encore moins ponctuel), mais en même temps définit le statut du colis – et c'est en cela que le syntagme est perfectif (il délimite un type, une catégorie). (*op. cit.*, 414)

3.1.3. Discussion et conclusion

Il est possible de rendre compte des restrictions sélectionnelles imposées par le SP *en DétQuant Ntps* à la situation dénotée par sa prédication d'accueil sans recourir à l'hypothèse d'une valeur perfective propre à *en*.

Si l'on cherche à appliquer l'ISRR aux adverbiaux aspectuels *en DétQuant Ntps*, on dira que *en* opère

1. une coalescence entre la situation dénotée par la prédication (=X) et (les dimensions de) l'intervalle temporel quantifié dénoté par le SN (=Y).
2. Cet intervalle se trouve aussitôt restreint au cadre extensionnel fixé par la situation.

Autrement dit, la durée exprimée par le SN (Y) dénote la durée *intégrale* de la situation (X), X et Y s'impliquant l'un l'autre. Pour reprendre une formulation empruntée à J.-J. Franckel & D. Lebaud (1991), l'intervalle temporel (Y) a le statut de « *condition de manifestation* » de la situation (X), la borne finale de l'intervalle temporel impliquant (donc) « *l'achèvement* » (au sens de J.P. Desclés *op. cit.*) de la situation.

L'une des conséquences de ce processus de saturation référentielle réciproque est que la borne finale de la durée devient la borne inhérente de la situation. C'est ici, selon nous, la clef de la restriction de sélection que manifestent ces compléments aspectuels vis-à-vis de l'aspect des situations d'accomplissement dont ils quantifient la durée.

Si l'on nous accorde le bénéfice d'une telle analyse, on nous accordera aussi qu'elle dispense de toute hypothèse sur la perfectivité de *en*. Ce faisant, on évite d'avoir à chercher à étendre cette valeur aspectuelle à d'autres emplois de la préposition. Extension qui conduit à forger des hypothèses parfois hasardeuses, soit par leur contenu (considérations sur l'instantanéité des jugements opérés par les locuteurs, possibilité d'une discrétisation d'essence aspectuo-temporelle qui s'opérerait sur un plan autre que l'aspect) soit par leur méthode (passage d'une valeur sémantique de *en* qui dirait quelque chose de la situation référentielle associée à l'énoncé (vitesse de réalisation du procès) à une valeur sémantique qui dirait quelque chose des processus cognitifs mis en jeu lors de l'énonciation (vitesse d'accomplissement d'un jugement par le locuteur, discrétisation de sous-type[76])).

---

[75] On pourrait préciser : N statifs comptables (*versus* N statifs de qualité). La question de savoir comment s'actualise le trait perfectif propre à *en* dans les cas où la préposition est suivie d'un nom abstrait d'activité - qu'il soit [-comptable] (*Max est en linguistique*), [+comptable] (*Max est en réunion*) ou [±comptable] (*Max est en randonnée*) – n'est pas traité dans l'article. Pour toutes ces catégories et sous-catégories aspectuelles des noms, voir B. Martinie & D. Vigier (2013).

[76] *Mutatis mutandis*, nous serions ici tenté de faire, en toute amitié, à D. Leeman et C. Vaguer, le reproche qu'adressait J.-C. Anscombre (2001 : 185) à la première à propos de la notion de résultat : « En utilisant le même mot *résultat* pour désigner le résultat d'un procès désigné par un verbe et le résultat du jugement du locuteur, D.L. s'expose au risque d'assimiler langue et métalangue. Il n'y a a priori aucune raison pour que ces deux notions de 'résultat' soient identiques, et si tel était le cas, il conviendrait de le justifier. »

3.2. La construction[77] *N_0 être en $X^{Couleur}$*

La notation « *$X^{COULEUR}$* » désigne non seulement les termes catégorisables parmi les adjectifs de couleur - adjectifs catégorisateurs (*rouge, noir, lie-de-vin…*) et adjectifs de caractérisation générale des couleurs (*sombre*[78]*, clair, mat…*) chez C. Molinier (2006) -, mais aussi le nom hyperonyme *couleur* et la séquence figée *noir et blanc*.

Soient les énoncés :

(94) *La mariée est (blanche + noire).*
(95) *La mariée est (en + *ø) (blanc + noir + bleu +…).*

Dans (94), l'adjectif dénote la couleur de peau de la mariée tandis que dans (95), le prédicat *en $X^{COULEUR}$* dénote la couleur des habits qu'elle porte. L'énoncé (94) est par ailleurs un pur *statif* : il construit une « représentation d'un état abstraction faite de toute prise en considération du processus par lequel il a pu être instauré. » (D. Creissels, 1999 : 185). De fait, la pigmentation de la peau de la mariée ne peut pas être associée à un processus antérieur dont elle constituerait le résultat. L'énoncé (95) est en revanche *statif-résultatif* : il construit une « représentation d'un état comme découlant d'un événement antérieur au repère temporel relativement auquel cet état est envisagé » (*ibid.*). Dire de quelqu'un qu'il est *en (blanc + noir + bleu +…)* présuppose que la personne a revêtu des habits de couleur blanche, noire, bleue etc. On peut par conséquent associer (95) — qui exprime un résultat — à un événement antérieur dont (96 a+b) constitue une formulation possible :

(96)(a) *La mariée s'est (habillée + nippée + fringuée +…) en (blanc + noir + bleu+ …).*
(b) *On a (habillé + nippé + fringué +…) la mariée en (blanc + noir + bleu +…).*

Comme l'illustrent les énoncés (96 a+b) et (95) ci-dessus, il est possible d'associer à certains énoncés de type *N_0 $V^{TRANSFORMATION}$ N_1 en $X^{COULEUR}$* un énoncé statif-résultatif de type *N_0 être en $X^{COULEUR}$*.

Il s'agit-là d'une propriété partagée par un bon nombre d'énoncés (mais non tous : cf. *infra*) qui incluent des verbes de transformation / de création / de destruction + SP *en X* dénotant le résultat (cf. Leeman, 1995). Voici un autre exemple :

(97) (a) *Max s'est (déguisé + travesti) en pompier.*
(b) *On a (déguisé + travesti) Max en pompier.*
(98) *Max est en pompier*.
(99) *Max est pompier*.

[77] Cette construction a fait l'objet d'un travail en commun entamé par B. Martinie et moi-même qui devait initialement être intégré dans notre article « Le régime nominal de la préposition *en* dans la construction *être en + N abstrait* : une étude aspectuelle » (voir bibliographie). Cette sous-section rend compte des conclusions auxquelles nous étions parvenus.

[78] « Je n'ai pas fait attention à ses vêtements. Il me semble qu'il était en sombre, probablement en bleu marine. » (G. Simenon).

Être en pompier présuppose un événement antérieur par rapport auquel l'état *être en pompier* est envisagé. En revanche, (99) comme (94) est - hors contexte, du moins – un pur statif.

Soient maintenant les couples d'énoncés suivants :

(100) *Max a peint la façade de sa maison en rouge*.
(101) *Maintenant, la façade de sa maison est (?en + ø) rouge*.

(102) *Max a teint les rideaux de sa cuisine en rouge*.
(103) *Maintenant les rideaux de sa cuisine sont (?en + ø) rouge(s).*

(104) *Le potier a coloré l'argile en rouge.*
(105) *Maintenant, l'argile est (?en + ø) rouge.*

Il semble que la langue traite *similairement* la couleur de la peau des entités vivantes (94) et la couleur obtenue par peinture (101), par teinture (103) ou par coloration (105) des entités inertes, et *différemment* la couleur de l'habillement et le déguisement (95)(98). Faut-il voir dans ce traitement qu'elle réserve à ces deux procès le signe qu'ils aboutiraient à un état plus passager, plus « circonstanciel » (au sens de J.-J. Franckel & D. Lebaud (1991: 59)) des individus — le prédicat attributif réclamant alors *en* ? Telle est globalement la position de A. Dagnac (2010 : 77) — « *l'habillage est superficiel, la teinture affecte l'objet dans sa totalité* » — laquelle reprend peu ou prou J.-J. Franckel & D. Lebaud (*op. cit.*: 64) pour qui la séquence « *être en* $X^{COULEUR}$ dénoterait *une propriété circonstancielle, nettement associée à une actualisation* ».

Il reste que dans deux cas de figure au moins, une telle hypothèse semble sinon prise en défaut, du moins devoir être précisée voire amendée.

Tout d'abord, c'est à la construction N_0 *être en* $X^{COULEUR}$ qu'on recourt lorsqu'il s'agit de dénoter la couleur des films, des photographies, des planches de bandes dessinées, … En quoi s'agirait-il là d'états plus « circonstanciels » que la peinture ou la teinture par ex. ? A moins que l'emploi des termes seuls autorisés (*couleurs + noir et blanc)* ne modifie la donne ?

(106) *Max a tourné un film en (noir et blanc + couleurs)*
(107) *Le film de Max est (en + ?ø) (noir et blanc + couleurs)*

(108) *Max a dessiné des planches de BD en (couleurs + noir et blanc)*
(109) *Les planches de la BD sont (en + *ø) (couleurs + noir et blanc)*

En second lieu, dans certaines situations d'énonciation qu'il conviendrait d'étudier, l'expression de la couleur obtenue par peinture, par teinture ou par coloration s'accommode sans difficulté de la séquence *être en* $X^{COULEUR}$. Voici quelques exemples :

(110) *La miniature est minutieusement exécutée ; les couleurs variées. Le manteau de la Vierge est bleu ; (…) Les murs sont en blanc ; le toit de l'église, au fond, en rose*. (…) [*Bulletin de la Société Française de Reproductions de Manuscrits à Peintures, Vol. 19, 1938*].
(111) *Dans une autre boutique où je suis allée, cette même robe était en noir* (*énoncé forgé*).
(112) *Sur les véhicules appartenant aux services publics, les caractères sont en rouge sur fond blanc*. (*Les marchés tropicaux et méditerranéens, 1970*).

(113) *Toutes les granulations des leucocytes sont colorées (…) les noyaux des leucocytes et des globules rouges sont en bleu.*[79] (*Traité d'hygiène pratique, 1908*).

L'examen des énoncés présentés ci-dessus ainsi que d'autres devrait conduire à affiner voire à réviser sur certains points les analyses déjà proposées dans la littérature.

Conclusion de la première partie

Nous voici parvenu au terme de cette première partie. Notre objectif a consisté à tenter d'y formuler en un seul « mouvement argumentatif » l'état actuel de notre réflexion sur la préposition *en* envisagée en synchronie pour le français contemporain.

Notre première section avait pour objectif d'opérer une forme de « ressaisie » des options méthodologiques sur lesquelles nous avions eu à réfléchir ces dix dernières années, sans articuler cependant cette réflexion au sein d'un développement suivi.

Nous avons en premier lieu opté pour une approche scalaire de la « prépositionnalité » qui permette de distinguer dans la classe des prépositions, outre les unités linguistiques réunissant l'ensemble des traits morphologiques, syntaxiques et sémantiques caractéristiques de la catégorie (son « centre organisateur »), celles qui sont situées à une distance plus ou moins grande de ce centre et dont le caractère plus ou moins prépositionnel peut le disputer avec leur caractère plus ou moins coordonnant, casuel, etc. On peut à cet égard faire observer que certains emplois de *en* relèvent - comme l'avait jadis remarqué G. Gougenheim (1950, [1970][80]) – du cas sémantique *translatif* pris en charge par des marques casuelles dans d'autres langues. L'exemple de certains verbes transformateurs illustre parfaitement ce point : en français, la préposition *en* (*into* en anglais) exprime le passage d'un état à un autre, qui est marqué en finnois au moyen du suffixe casuel *ksi* :

(114) *Hän muuttu-i (touka-sta) perhose-**ksi**.*[81]
*S/he changed (from a caterpillar) **into** a butterfly.*
*Il/elle s'est transformé(e) (de chenille) **en** papillon.*

(115) *Taikuri muutt-i perhose-n touka-**ksi**.*
*The magician changed a/the butterfly **into** a caterpillar.*
*Le magicien a changé un/le papillon **en** chenille.*

Dans la suite de la première section, nous avons choisi d'affecter à la relation *X R Y* mise en jeu par la préposition un statut strictement sémantique afin de désintriquer dans l'analyse la relation syntaxique entre terme recteur et Y (dans le cas des SP compléments placés dans la rection du verbe ou du GV en particulier) du rapport sémantique proprement dit (voir l'ex. *Le chasseur a tiré sur le lapin* analysé par J.-J. Franckel & D. Paillard, 2007 : 107 & *passim*). Après quoi nous nous sommes employé à dresser une forme de « cartographie »

[79] Il s'agit ici de couleurs obtenues par réaction chimique sur des éléments organiques.
[80] « « Claire peut se transformer en d'autres femmes encore aimables » (J. Chardonne, *Claire*, Select Collection, p. 8). Comment classer ces sortes de compléments ? Sont-ce des objets ou des objets secondaires ? Cela est bien douteux. Nous y verrions plutôt une fonction toute spéciale que la langue finnoise exprime par un cas particulier, le « translatif » : *talo* signifiant « *la maison, la ferme* », le translatif *taloksi* veut dire « *changé en maison, en ferme* ». Cette idée de « translation », de passage d'un état à un autre, est marqué en français par la préposition *en*. » (G. Gougenheim, *op. cit.*, 59-60).
[81] Ces deux exemples sont tirés de V. Fong (2003 : 202).

(non exhaustive) des choix stratégiques que le linguiste a à accomplir lorsqu'il veut rendre compte de l'identité sémantique d'une préposition. Choix du traitement de la polysémie d'abord : opte-t-il pour une approche horizontale ou verticale ? C'est cette seconde voie que nous avons proposé d'explorer en suivant notre fil rouge qu'est la préposition *en*, ce « grand petit mot » sur lequel nous avons souvent travaillé durant ces dix dernières années. Nous avons ensuite discuté des rapports entre invariance et identité sémantique de la préposition, ainsi que de la formulation de cette identité. La mise en perspective de questions relatives à la « forme » du sens prépositionnel avec les analyses de G. Kleiber (1997) sur la triade *sens-référence-existence* nous a permis d'expliciter un parallèle de scalarité entre la « couleur » de la préposition et le caractère plus ou moins instructionnel de son identité sémantique. Les paragraphes suivants ont été consacrés à l'élaboration du triptyque méthodologique :

Noyau d'invariance sémantique – modèle de déploiement des interactions contextuelles – module distributionnel-restrictionnel visant à contrôler la puissance du noyau.

Ce triptyque constitue à nos yeux un cadre possible de formulation du sens d'une préposition incolore lorsqu'on opte pour un traitement vertical de la polysémie.

Dans la deuxième section, nous avons d'abord discuté du schème guillaumien de la réversion, fondateur dans les études sur *en* mais selon nous (et *contra* G. Guillaume) d'occurrence non-systématique dans les emplois en discours de cette préposition. L'analyse critique de la position défendue par D. Leeman sur les N_{PAYS} nous a permis de le montrer. La réversion guillaumienne « de l'idée nominale en mode sur le sujet » n'appartient donc pas au noyau instructionnel de *en* mais doit être affectée à certaines *constructions* seulement. Elle rejoint donc le module de « déploiement du sens » qui recense les régions de stabilité sémantique des effets de discours.

La dernière section de cette première partie nous a permis de présenter des éléments d'une étude à paraître consacrée aux SP aspectuels intraprédicatifs de type *en DétQuant Ntps* lorsqu'ils expriment la durée intégrale du procès exprimé par le verbe ou le GV. Nous avons cherché à montrer, *contra* D. Leeman[82] & C. Vaguer (2014), qu'il n'est pas nécessaire d'inclure dans l'invariant sémantique de la préposition *en* un trait aspectuel perfectif. Selon nous, la seconde composante instructionnelle de l'ISRR (*instruction de saturation référentielle réciproque*, P. Cadiot, 1997a) permet à elle seule d'expliquer la valeur aspectuelle de perfectivité que véhiculent les compléments intraprédicatifs examinés, cette valeur relevant donc (comme la réversion) de certaines régions de sens en discours.

L'ensemble de cette première partie constitue, dans l'économie générale de ce mémoire, une sorte de photographie de notre conception actuelle de la sémantique de *en* en français contemporain. Comme nous l'avons dit *supra*, une des directions de notre travail de recherche à venir sur cette préposition consistera à explorer une nouvelle formulation possible de son identité dans la perspective cognitive et fonctionnelle développée par C. Vandeloise.

[82] Et en toute amitié vis-à-vis de cette éminente linguiste dont les recherches sur *en* ont permis des progrès majeurs dans la compréhension de l'identité sémantique de cette préposition.

DEUXIÈME PARTIE

Constituer un corpus historique du français pour une exploration automatisée

RAPPEL DE LA TABLE DES MATIÈRES DE LA DEUXIÈME PARTIE

Introduction

Le contenu de cette deuxième partie est étroitement lié aux objectifs poursuivis, en termes de constitution et d'exploration de corpus dans le programme Presto[83]. Au cours des chapitres suivants, on s'efforcera :

- de présenter le corpus Presto constitué dans le cadre de ce programme, et qui sera utilisé dans la troisième partie ;
- d'exposer les critères qui ont présidé aux choix opérés lors de la constitution du corpus et lors de l'annotation de ce dernier ;
- de présenter les outils et les fonctionnalités auxquels nous avons recouru pour son exploration.

On souhaite ainsi mettre en lumière combien il est important (chaque fois que cela est possible) que le linguiste

i) participe étroitement aux étapes de constitution et d'annotation des corpus sur lesquels il compte travailler ;
ii) prenne en main les outils d'exploration et de calcul automatisés qu'il projette d'utiliser ;
iii) s'approprie l'esprit des fonctionnalités (notamment statistiques) que ces outils mettent à sa disposition.

Corpus numérisés et outils automatiques constituent de fascinantes *fenêtres* d'accès aux textes, mais ils forment aussi des *écrans* redoutables qui peuvent (plus ou moins) s'interposer entre le linguiste et ces textes. Qu'il s'agisse des choix opérés à chaque étape de la constitution du corpus, des décisions arrêtées lors de la mise en œuvre de la chaîne de traitement, des options prises dans le champ des calculs statistiques automatisés, sans compter les éventuelles erreurs commises et les simplifications ou approximations opérées[84], tout cela a une incidence plus ou moins forte sur l' « image » que se construit le linguiste du corpus qu'il étudie. Il serait donc naïf de croire que les outils informatiques donnent à « voir » le corpus sur lequel on travaille : ils en construisent plutôt une « image » nécessairement déformée. Une des conditions, donc, pour que le linguiste domine au mieux la contradiction - intrinsèque à toute linguistique *sur*[85] corpus outillée - entre « fenêtre » et « écran » est qu'il

[83] Le programme franco-allemand Presto (http://www.agence-nationale-recherche.fr/?Projet=ANR-12-FRAL-0010; http://presto.ens-lyon.fr) a été financé par l'ANR et la DFG pour une durée initiale de trois ans (avril 2013 - avril 2016) portée ensuite à quatre (avril 2017). Ce projet avait pour but l'étude diachronique de l'emploi, des valeurs sémantiques et discursives des prépositions françaises *à, en, par, contre, dès, devant, entre, pour, sans, sur, sous, vers, dans*, de l'ancienne langue jusqu'au français contemporain. Instrumentée, adossée à une approche statistique et distributionnelle, cette étude se proposait de porter sur les variations du comportement combinatoire des prépositions suivant des critères de dates (évolution diachronique), de genres et d'auteurs.

[84] Simplification que présupposent en particulier les analyses statistiques multivariées.

[85] Nous nous astreindrons dans ce mémoire à parler de linguistique « sur » corpus et non de linguistique « de » corpus. Nous considérons en effet que cette dernière appellation suggère qu'on aurait affaire à une sous-catégorie de la linguistique dotée de ses méthodes et de ses raisonnements propres - en d'autres termes à une sous-discipline de la linguistique. Or il convient selon nous (et comme le souligne M. Cori (2008) qui s'appuie sur K.R. Popper (1956, [1985])) de soigneusement distinguer l'*instrument* de la *science* qui y recourt, et d'éviter ainsi tout « brouillage des enjeux épistémologiques» (M. Cori, *op. cit.* : 109). L'expression « linguistique sur corpus » évite – du moins le pensons-nous - un tel brouillage. (Voir aussi M. Cori & S. David, 2008). Notre distinguo entre *de* et *sur* ne croise donc que très partiellement celui avancé par S. Azzopardi (2010) qui, à la suite de M.-P. Jacques (2005), rabat la *linguistique « de » corpus* sur l'approche *corpus-driven* et la *linguistique « sur » corpus* sur l'approche *corpus-based.* Notre propos est plus épistémologique.

connaisse avec un certain degré de profondeur les opérations ayant permis de construire les données linguistiques qu'il manipule, ainsi que les outils et les méthodes grâce auxquels il accède aux résultats qu'il interprète. *In fine*, le linguiste sur corpus qui travaille dans un environnement numérique n'examine jamais des « données » mais toujours déjà des « construits » (sur ce point, voir par ex. M. Cori & S. David, 2008 : 120).

Nous traiterons successivement - dans les lignes qui suivent - du corpus, de l'annotation et du traitement automatisé des données.

1. Construire un Corpus. De la perspective « théorique » à la réalisation pratique (le corpus Presto)

Dans cette première section, nous proposons de partir d'une définition de la notion de corpus en linguistique pour étudier ensuite tout particulièrement la question relativement polémique de la représentativité. Un corpus peut-il être représentatif ? La population langagière nous est-elle suffisamment connue pour parvenir à une telle représentativité? Suivant quelles voies ? Faut-il, dans un corpus de référence, recourir à des textes échantillonnés ou intégraux ? Quelle relation établir entre représentativité (*representativeness*) et équilibre (*balance*)? Les corpus historiques peuvent-ils eux aussi prétendre à la représentativité ? … Ces diverses questions - ainsi que d'autres - seront examinées, et certaines caractéristiques spécifiques aux corpus historiques seront mises en lumière.

1.1. Qu'est-ce qu'un corpus ?

Le programme Presto présentait comme un réquisit la mise au point d'un corpus diachronique couvrant la période allant de 1500[86] à 2000. C'est en grande partie grâce à l'appui constant et amical de certains chercheurs français et étrangers travaillant à l'évolution et à l'enrichissement de bases de données textuelles de premier rang sur les plan national et international – nous songeons en particulier à V. Montémont et G. Souvay de l'équipe « Ressources, normalisation, exploitation et annotation » du laboratoire ATILF qui gère la base Frantext et à R. Morrissey de l'Université de Chicago, directeur de l'ARTFL (*American and French Research on the Treasury of the French Language*) – que ce corpus a pu être mis au point. Quelle que soit la qualité de ce dernier – la communauté des chercheurs en jugera – sa constitution nous a montré combien est juste la formule de S. Hunston (2008 : 156): « *All corpora are a compromise between what is desirable (...) and what is possible* ».

1.1.1. Définir la notion de corpus

On peut, comme il est fait souvent (voir par ex. B. Habert, A. Nazarenko & A. Salem, 1997 ; B. Habert 2000, A. Geyken 2008) partir de la définition de J. Sinclair (1996 : 4) :

« *A corpus is a collection of pieces of language that are selected according to explicit linguistic criteria in order to be used as a sample of the language* »

et de ses éléments constitutifs qui fixent le cap à suivre pour la construction d'un corpus :

[86] Initialement, le projet intégrait l'équipe de la Base de Français Médiéval (http://bfm.ens-lyon.fr) et le corpus devait donc aussi comprendre des textes allant du IX^{e} au XV^{e} s. Des difficultés de collaboration avec cette équipe, et plus particulièrement avec l'un de ses membres, nous ont conduits à sélectionner la date de 1500 comme borne initiale de notre corpus (et de nos recherches).

i) réunir une collection de données langagières,
ii) résultat d'une sélection,
iii) opérée à partir de critères linguistiques explicites,
iv) et réglée par un objectif : servir d'échantillon du langage.

Il existe aujourd'hui un consensus parmi les linguistes pour considérer qu'un *corpus* doit être pensé et construit en fonction d'objectifs scientifiques préalablement définis. Comme l'écrit F. Rastier (2011 : 80) « un corpus reflète le point de vue qui a présidé à sa constitution ». Cette conception nous conduira, comme cela est souvent fait, à distinguer les *corpus* des *archives* et autres *bases / banques de données*.

Il reste que nombre des éléments qui forment la définition de Sinclair prêtent à discussion.

La notion de « données langagières ». J. Pearson (1998, 42) fait judicieusement observer que dans une publication antérieure, J. Sinclair (1991 : 171) avait défini le *corpus* comme « *a collection of naturally-occurring language text* (…) ». D'une définition à l'autre, la notion de *text* a donc été remplacée par celle de « *pieces of language* » (traduite par B. Habert, A. Nazarenko & A. Salem, *op. cit.*: « données langagières ») : l'auteur craignait en effet que l'on n'interprète le mot *text* dans le sens de *texte intégral.* Bref, dans la définition de 1996, « c'est à dessein que le mot « texte » n'est pas employé » (B. Habert, A. Nazarenko & A. Salem, *op. cit.* : 146). J. Sinclair ménage ainsi dans sa définition une place à l'échantillonnage des textes[87]. Le caractère bien-fondé ou non de l'échantillonnage qui brise « la séquentialité des textes » (B. Habert, A. Nazarenko & A. Salem, *ibid*.) dans un corpus est une question très discutée. F. Rastier (*op. cit.* : 33-34), dont on sait qu'il fait du texte et non de la phrase l'unité première (en droit) d'analyse des études linguistiques[88], définit quant à lui la notion de corpus en fermant la porte à tout échantillonnage des textes : « *Un corpus est un regroupement structuré de* ***textes intégraux***[89]*, documentés, éventuellement enrichis par des étiquetages, et rassemblés : i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications* ». M.-P. Péry-Woodley (2000 :150) propose une voie moyenne à laquelle irait notre préférence : « Comment concilier une bonne pratique d'échantillonnage et le respect de l'unité texte ? Les corpus pour l'anglais, y compris le récent *British National Corpus*, sont faits d'échantillons pris à divers moments du texte pour ne pas privilégier certaines formes, associées par exemple aux débuts de texte. La conséquence est bien évidemment que l'organisation textuelle n'y est plus accessible. Pour les linguistes du texte, un corpus idéal serait peut-être fait d'échantillons correctement sélectionnés et calibrés, mais dont certains tout au moins donneraient accès au texte intégral. » Pouvoir *en même temps*… disposer d'un échantillonnage des textes permettant une pondération en taille des sous-corpus et recourir aux textes intégraux lorsqu'on le souhaite : telle a été une de nos ambitions dans le corpus Presto où nous avons systématiquement

[87] De fait, il est spécifié au-dessous de la définition : « *Note that the non-committal word `pieces' is used above, and not `texts'. This is because of the question of sampling techniques used. If samples are to be all the same size, then they cannot all be texts. Most of them will be fragments of texts, arbitrarily detached from their contents*. »

[88] Voir par ex. F. Rastier, *op. cit.* : 24-25 : « Les textes demeurent les seuls *objets empiriques* de la linguistique. (…) Si le morphème est bien l'unité linguistique élémentaire, le texte demeure l'unité minimale d'analyse, car le global détermine le local. »

[89] C'est nous qui soulignons.

ménagé, pour chaque « niveau » du corpus (voir *infra*), une « version » échantillonnée et une version intégrale[90].

Les critères de *sélection* des textes. B. Habert (*op. cit*.) propose d'ajouter aux critères linguistiques explicites évoqués par Sinclair des critères *extralinguistiques*. A. Geyken (*op. cit*.) commente cette proposition comme suit : « Il est important de noter ici [chez B. Habert] que la sélection des données langagières repose non seulement sur des critères linguistiques comme la richesse du vocabulaire ou la variabilité syntaxique, mais aussi sur des critères extralinguistiques, par exemple le choix de types de textes constituant un corpus. » De fait, B. Habert parcourt la question des types de textes et de leur classification, en opposant les démarches *a priori* (fondées sur une typologie des conditions de production des textes etc.) et les démarches *inductives* qui consistent « à faire émerger les types de textes – considérés comme des agglomérats de traits linguistiques – grâce à un traitement statistique de textes » (*op. cit.* : 7). Nous reviendrons *infra* sur la question des genres, types et registres[91] de textes, en lien avec celle de la représentativité (lien que font B. Habert et M.-P. Péry-Woodley) ainsi que des critères extralinguistiques.

Un « **échantillon du langage** » (*a sample of the language*). On touche à la question de la population cible que le corpus vise à représenter. Si l'on suit J. Pearson (*op. cit.* : 42), il faut entendre dans la définition de J. Sinclair « *a sample of the language or some subset of the language »*. Il existe un consensus aujourd'hui pour considérer que toute langue est régie par un principe de variation, de sorte que tout échantillon visant à représenter la langue, un (ou plusieurs) langage spécialisé ou tout autre « pratique langagière effective » suppose *a minima* qu'on se dote d'un « modèle de la variation » entendu comme un ensemble de paramètres qui déterminent ces pratiques (voir M.-P. Péry-Woodley (*op. cit.* : 158 & sq.)). Comme nous y reviendrons, la question de l'adéquation entre les données du corpus et la population qu'on vise à représenter n'est pas pour autant réglée. D'une part parce que la notion même de « langue » peut poser problème selon le sens auquel elle est entendue ; d'autre part, parce que nous sommes encore loin de posséder un « modèle de variation » de l'ensemble des pratiques discursives, qu'elles soient contemporaines (pour l'oral par exemple, on dispose encore de travaux limités relatifs aux genres, types et registres), ou *a fortiori* anciennes (Moyen Âge, Renaissance, …). Comme l'observe B. Habert (*op. cit.* : 1, note 5) « notre connaissance de la « population » des données langagières est encore extrêmement fragmentaire. »

1.1.2. Quelques axes d'opposition

Les points de discussion ou de consensus signalés plus haut dans la définition de ce qu'est un *corpus* concourent à dresser quelques couples d'opposition.

***Corpus* versus *archive, base, réservoir*.**

On peut opposer la notion de corpus à celles d'archive, de base et de réservoir sur le critère de la sélection des données qui y figurent. Une **archive** est une collection de textes i) qui ne répond pas à un critère de sélection des textes (voire de segments de textes) en fonction d'un but de recherche spécifique, ii) et qui a été amassée par un même opérateur (qu'il

[90] Suivant en cela l'exemple de l'initiative prise dans le cadre du corpus de la *Grande Grammaire Historique du Français* (GGHF) qui propose de distinguer un corpus « noyau » et un corpus « complémentaire » (voir par ex. https://bcl.cnrs.fr/rubrique137).

[91] Le terme de *registre* est beaucoup utilisé par D. Biber (voir *infra*) au côté de celui de *genre*, le premier terme renvoyant à une « conception élargie des genres » et dont l'inventaire « est destiné à rester ouvert ». (B. Habert, 2000 : 6).

s'agisse d'un particulier ou d'un opérateur public (archives parlementaires) ou privé (archives du *Monde*, …). Une **base textuelle** est une collection de textes réunie en vue de mettre à disposition des chercheurs des textes susceptibles d'entrer dans des corpus que ces derniers façonneront en fonction de leur but de recherche[92]. Tel est le cas par ex. de la *Base de Français Médiéval* (BFM http://bfm.ens-lyon.fr)), de Frantext (http://www.frantext.fr), de l'ARTFL (*American and French Research on the Treasury of the French Language*, http://artfl-project.uchicago.edu), des BVH (*Bibilothèques Virtuelles Humanistes*, http://www.bvh.univ-tours.fr), de CLAPI (*Corpus de LAngue Parlée en Interaction*, http://clapi.ish-lyon.cnrs.fr), etc. ***Un réservoir de corpus***, enfin, est une base mettant à la disposition des chercheurs des corpus et non simplement des textes. On songera par ex. à Ortolang (*Outils et Ressources pour un Traitement Optimisé de la LANGue*, https://www.ortolang.fr).

Corpus *exhaustif* vs *représentatifs*

On peut opposer ces deux notions sur le critère de la population visée. Une population textuelle ou langagière *finie* peut donner lieu à un corpus exhaustif, où la question de la représentativité ne se pose pas en ceci que *le corpus, c'est la population elle-même*. Ainsi les œuvres complètes d'un auteur, la totalité des discours d'un homme politique décédé etc. Certains, comme D. Mayaffre (2002) posent néanmoins la question de la clôture et de l'exhaustivité d'un corpus[93] à partir de la notion d'intertextualité : l'œuvre de J.-P. Sartre, par ex., peut-elle s'interpréter à partir du corpus intégral de ses œuvres seulement ? Ne faut-il pas convoquer aussi, à un moment donné du processus interprétatif, les œuvres de S. de Beauvoir, A. Camus, E. Husserl, M. Heidegger etc. avec lesquels l'auteur entre – parfois explicitement (on songera à la correspondance avec le Castor) - en dialogue ? Si cette observation est parfaitement recevable lorsqu'on adopte une perspective *interprétative* des textes[94], elle nous semble plus discutable sur un plan strictement linguistique. Si l'on adopte l'idée que la définition de la population (langagière) cible constitue le principe régulateur de toute constitution de corpus, alors se fixer pour population cible l'ensemble des œuvres publiées d'un auteur par ex. autorise, lorsque cette collection de données langagière est réunie, de parler de corpus exhaustif sans considération pour les processus nécessairement intertextuels que leur interprétation met en jeu.

Corpus clos *vs* corpus évolutif

Un corpus clos est figé, vitrifié : il n'évolue plus. C'est le cas par ex. de certains corpus « historiques » comme le *Brown Corpus*. Un corpus évolutif voit les données qu'il réunit évoluer par étapes successives. Tel est le cas par ex d'ARCHER qui a connu quatre étapes (Archer 1, Archer 2, Archer 3.1, Archer 3.2 [95]) depuis sa création. Nous avons conçu de même le corpus Presto comme évolutif.

Corpus unique *vs* corpus multi-niveaux et multi-versions

Nous introduisons cette distinction pour rendre compte d'une spécificité du corpus Presto (qu'il partage partiellement avec le corpus de la *Grande Grammaire Historique du Français*). Nous avons ménagé plusieurs « niveaux » et plusieurs « versions » dans notre corpus de manière à l'adapter à des recherches diverses. Nous reviendrons sur ce point en détail au § 1.4.1.

---

[92] B. Habert (*op. cit.* : 1) a raison d'y voir des « réservoirs à corpus ».
[93] Voir aussi S. Mellet (2002, § 3)
[94] Ainsi l'auteur plaide-t-il pour des corpus qu'il nomme « réflexifs », i.e. qui intègrent en leur sein le « co-texte » des textes qu'ils réunissent.
[95] http://www.projects.alc.manchester.ac.uk/archer/archer-versions

1.1.3. « Ne pas choisir, (…) c'est choisir de ne pas choisir »[96]

Une suite de choix raisonnés constitue le fondement de tout corpus. A l'inverse, le recours au Web non filtré – encore trop souvent confondu par certains linguistes avec un corpus[97] - constitue le degré zéro de la linguistique sur corpus. « L'utilisation du Web comme un terrain de collecte de données ne présuppose aucune démarche de construction du corpus » écrivent M. Cori, S. David & J. Léon (2008 : 6).

Choisir, c'est non seulement additionner, c'est aussi soustraire[98] c'est-à-dire écarter tel texte au profit d'un autre.

Ce peut être enfin choisir … d'être choisi. Les très fortes contraintes exercées sur l'accessibilité des textes en linguistique historique – qu'il s'agisse des droits d'édition, de l'impossibilité de consulter certains textes en dehors du fonds où ils sont conservés ou encore la mutilation des manuscrits originaux quand on s'éloigne dans le passé (un bel exemple : les romans de *Tristan*, presque tous perdus ou mutilés ; G. Paris (1900 : 7), dans sa préface à l'ouvrage de J. Bédier *Le roman de Tristan et Iseut*, parle d' « amas de décombres ») conduit à accepter non de *ne pas choisir*, mais d'*être choisi* par les rares textes accessibles qui remplissent les contraintes d'échantillonnage que l'on s'est fixées .

On peut enfin faire état de la petite typologie des « objectifs » susceptibles d'être poursuivis dans la constitution d'un corpus, selon M. Cori, S. David & J. Léon (*op. cit.*:6-7) :

1. Mettre à disposition des corpus pour la communauté (exemples du LOB (Lancaster-Oslo-Bergen http://www.helsinki.fi/varieng/CoRD/corpora/LOB), du LUND (London-Lund http://www.helsinki.fi/varieng/CoRD/corpora/LLC), du British National Corpus (BNC http://www.natcorp.ox.ac.uk/index.xml).
2. Confectionner des outils linguistiques élaborés à l'aide de corpus : bases de données, dictionnaires, grammaires etc. (ex. du COBUILD « Collins Birmingham University International Language Database » (http://www.lt-world.org/kb/resources-and-tools/language-data/ltw_x3alanguage_x5fdata_.2010-09-23.5678579368)
3. Proposer des descriptions linguistiques.
4. Se donner de nouveaux moyens de travailler certaines questions ayant trait à la variation : les genres, le changement linguistique, auxquelles on peut ajouter la productivité en morphologie.
5. Construire des outils de traitement automatique des langues à base de corpus, en particulier à base de corpus d'entraînement qui servent à paramétrer des algorithmes probabilistes.

Selon nous, les objectifs 3 et 4 sont en relation d'imbrication, 4 constituant une spécification de 3. Quant aux objectifs de Presto, ils relèvent de 4 (étude du changement linguistique dans le système des prépositions du français). Signalons que, toujours dans le cadre de Presto, un logiciel d'annotation (lemmes et catégories morphosyntaxiques) a été mis au point, en recourant au corpus, pour le français pré-classique et classique.

---

[96] J.-P. Sartre, *l'Etre et le Néant* (1943).

[97] Je ne m'exempt pas de reproches : mes premiers pas en corpus ont fait appel à la Toile. Voir par ex. Vigier (2003).

[98] Selon la jolie formule de D. Mayaffre (*op. cit.*) : « additionner en matière de corpus signifie de manière problématique avant tout soustraire. Décider de rassembler deux textes, c'est avant tout décider d'écarter tous les autres. »

1.1.4. « Hygiène » des corpus

Il est impératif de documenter de manière précise les décisions prises (et leurs fondements) à toutes les grandes étapes de la construction du corpus et de son annotation : modèle des paramètres de la variation, choix d'échantillonnage, calibrage des tailles de mots et de textes, etc. (voir B. Habert, A. Nazarenko & A. Salem, *op. cit.* :15-17 ; M. Cori & S. David, *op. cit.*: 126 ; C. Guillot, S. Heiden, A. Lavrentiev, C. Marchello-Nizia, 2008). J. Sinclair (2004 :13) souligne le caractère crucial d'une telle documentation et énonce le principe suivant : « *The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.* » L'objectif est de permettre qu'à n'importe quel moment il soit possible pour l'utilisateur de remonter aux principes de constitution du corpus, de vérifier si tel ou tel choix n'a pas été trop teinté de subjectivité voire d'irréflexion, et de l'améliorer si nécessaire. C'est aussi un renseignement précieux pour le chercheur s'étonnant d'un résultat en corpus qui contredit son intuition linguistique. « *So many of our decisions are subjective that it is essential that a user can inspect not only the contents of a corpus but the reasons that the contents are as they are*. (…) *Also at any time a researcher may get strange results, counter-intuitive and conflicting with established descriptions. (...) one of the researcher's first moves on encountering unexpected results will be to check that there is not something in the corpus architecture or the selection of texts that might account for it*. » (*op. cit.* : 7).

Signalons pour finir que la bonne documentation d'un corpus permet sa réutilisabilité dans d'autres corpus qui l'intègrent, très importante si l'on souhaite rationnaliser les efforts et les moyens investis dans la constitution des corpus. L'initiative TEI (*Text Encoding Initiative* http://www.tei-c.org/index.xml) va naturellement (et très loin) dans ce sens. De manière plus modeste, les plateformes de stockage (ou « réservoirs ») de corpus (Cocoon « *COllections de COrpus Oraux Numériques* », Ortolang, …) développent des standards d'annotation minimale[99] des corpus en vue de leur réutilisation. Dans le cas de Presto, une partie du corpus constitué sera mise à la disposition des chercheurs en accès libre téléchargeable gratuitement. Notre objectif est de pousser le plus loin que nos forces le permettront leur encodage TEI.

1.2. La notion de représentativité

Les quelques généralités formulées ci-dessus ont fait ressortir, à l'occasion de la question de l'échantillonnage, l'importance majeure que revêt la notion de représentativité.

1.2.1. Qu'est-ce que la « représentativité » ?

Selon Y. Dodge (2007 : 159), la notion de *représentativité d'un échantillon* en statistique est récente. Si le principe d'échantillonnage peut être daté du début de la seconde moitié du XVII^{e} s. (*De ratiociniis in Aleae Ludo,* 1657, C. Huygens), ce serait en 1895, lors d'un congrès international de statistique à Berne, que le directeur du *Bureau Central de Statistique* du royaume de Norvège, A. N. Kaier, introduisit la notion de représentativité dans un mémoire intitulé « *Observations et expériences concernant des dénombrements*

[99] « Dans tous les cas le dépôt d'une ressource sur ORTOLANG devra être accompagné d'un jeu minimum de métadonnées descriptives au format Dublin Core, et administratives spécifiques à ORTOLANG. Pour aider les utilisateurs à construire ces métadonnées, un éditeur interactif de ces métadonnées est proposé par ORTOLANG dans l'espace de travail lors du dépôt d'une ressource. Il permet de préciser en particulier des renseignements généraux de type descriptif de la ressource, les droits y afférant et les divers contributeurs à cette ressource. » (https://www.ortolang.fr/information/policy)

représentatifs ». C'est en 1925 que l'*Institut International de Statistique* définit dans son rapport la notion de représentativité en lien avec deux méthodes d'échantillonnage : par choix aléatoire (*random sampling)* versus par choix judicieux (*purposive sampling)*. Voici un court extrait de ce rapport, tiré d'A. Desriosières (1988 : 111-114) :

« (…) pour que les résultats d'une enquête partielle puissent être légitimement généralisés, la fraction retenue comme un spécimen de l'ensemble dont il fait partie doit être suffisamment représentative de cet ensemble. »

La *représentativité* apparaît ainsi étroitement liée aux notions *d'échantillon* (« *fraction/spécimen »*), de *population* (« *ensemble »*) et de *généralisation*. L'objectif que poursuit le chercheur lorsqu'il constitue un échantillon représentatif d'une population « cible » est de pouvoir, à partir de cet échantillon, *généraliser* avec une certaine confiance à la population représentée les résultats obtenus sur l'échantillon. Les gains d'une telle représentativité sont nombreux, comme le rappelle Y. Dodge (*op. cit.*) : gain de temps, de coût, augmentation des possibilités, ...

Dans le domaine spécifique de la linguistique sur corpus (historique ou non), la notion de *représentativité* d'un corpus convoque donc *a minima* celle de *population* qu'on souhaite représenter et celle *d'échantillonnage* entendu comme « ensemble des opérations destinées à former un échantillon à partir d'une population donnée » (Y. Dodge, *op. cit.* : 159).

Voici trois définitions de la représentativité formulée dans le cadre de la linguistique sur corpus :

La première est tirée de G. Leech (1991 : 27) : « *A corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety.* »

La deuxième est tirée de D. Biber (1993a : 243) : « *A corpus must be 'representative' in order to be appropriately used as the basis for generalizations concerning a language as a whole. (...) Representativeness refers to the extent to which a sample includes the full range of variability in a population.* »

La troisième est proposée par S. Hunston (*op. cit.*: 160) : *« Representativeness is the relationship between the corpus and the body of language it is being used to represent. A corpus is usually intended to be a microcosm of a larger phenomenon. »*

Dans ces trois définitions se trouvent exprimées les notions de *généralisation* (*generalized* D1*, generalizations* D2) ou de *représentation* (*to represent* D1, D3), S. Hunston faisant en outre allusion à l'idée plus vague de *microcosme*. On est ici, selon A. Desrosières, au cœur de la notion de représentativité statistique telle qu'elle émerge à la fin du XIX^{e} s. avec A. N. Kiaer : « L'essentiel de l'idée de représentativité était bien là : *la partie peut remplacer le tout* ». La question de la population visée est aussi présente dans les trois extraits, et spécifiée comme suit : *the language* (D1), *a language as a whole* (D2), *the body of language* (D3). Si l'on en croit D. Biber, une définition explicite de la population visée constitue même le préalable à toute constitution de corpus : « *the issue of population definition is the first concern in corpus design* » (D. Biber, *op. cit*., 244). C'est donc par cette question de la population que nous allons commencer.

1.2.2. Représentativité et population ; représentativité et taille de l'échantillon

Nous n'évoquerons pas pour le moment le cas des corpus historiques que nous traiterons plus loin.

Il convient d'abord d'écarter le cas des corpus *clos et exhaustifs* (S. Mellet 2002, §3) évoqués plus haut qui réunissent exhaustivement tous les textes constitutifs de la population cible. Tel est le cas des « monographies », des corpus d'œuvres complètes d'auteurs, etc. Une illustration en ligne pourrait être le corpus des « dossiers de Bouvard et Pécuchet » (http://www.dossiers-flaubert.fr) qui rassemble 2 400 feuillets conservés à la bibliothèque municipale de Rouen et forme un ensemble clos sur lui-même, ayant « servi à rédiger le premier volume de l'œuvre et [qui] aurait dû être réutilisé pour la composition d'un second volume, jamais écrit en raison de la mort soudaine du romancier ». La notion d'échantillon ici n'a pas droit de cité puisque le corpus est la population à décrire.

Hormis les corpus clos, on est confronté (sous réserve d'irréflexion méthodologique[100]) à la question de l'échantillonnage de la population que l'on veut décrire. Cette population peut être un sous-domaine spécialisé du langage (*corpus spécialisés* : cas évoqué par M.-P. Péry-Woodley, *op cit*[101] ; voir aussi B. Habert, A. Nazarenko & A. Salem, *op. cit.*: 148 ; D. Biber 1993b : 220[102]) ou la langue prise dans son ensemble. Dans ce dernier cas, on a affaire aux corpus dits de « référence » comme le LOB (Lancaster-Oslo-Bergen) ou le BNC.

« *A reference corpus is one that is designed to provide comprehensive information about a language. It aims to be large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials*. » (J. Sinclair, 1996: 10).

Or l'idée d'un échantillonnage représentatif d'une langue naturelle saisie dans sa globalité comme population cible – par ex . en vue de la rédaction d'un dictionnaire[103] - fait surgir bien des difficultés. Voici les principales :

- **Difficulté 1**. **Définir la population cible.** Que faut-il exactement entendre par « langue » ? Est-ce la langue telle qu'elle est *produite* par les locuteurs ? En ce cas, il faut préalablement constituer un échantillon représentatif des locuteurs natifs d'une langue donnée[104] puis l'interroger afin de modéliser la variété des genres et des registres de langue que cette

---

[100] M.-P. Péry-Woodley (*op. cit*. : 157) envisage ainsi les recherches sur la « langue générale » qui ne se posent pas la question de l'échantillonnage, sélectionnant « une population supposée non-marquée » censée représenter cette langue (souvent, la presse nationale, ou Frantext). Là encore, nous plaidons coupable, ayant eu jadis en tête (comme doctorant) ce genre de naïveté.

[101] L'auteur semble ne reconnaître la possibilité à une linguistique de corpus de pouvoir modéliser avec une confiance suffisante la population qu'elle vise que dans le cas où « le corpus doit représenter une population de textes spécifiques ». En revanche, la possibilité d'échantillonner de manière satisfaisante « la langue générale » est considérée comme une chimère. Voir aussi M.-P. Péry Woodley (1998 : 219). A. Condamines (2000 : 15-16) rappelle cette position de prudence, soulignant que D. Biber (1988 : 200) considère quant à lui qu'un corpus représentatif de la langue prise dans sa totalité est possible. Nous allons revenir sur ce point.

[102] « *Such as journal articles on lipoprotein kinetics (Sager 1986), Navy telegraphic messages (Fitzpatrick et al. 1986), weather reports (Lehrberger 1982), and aviation maintenance manuals (Kittredge 1982)* ».

[103] On songera au corpus COBUILD **«** *Collins Birmingham University International Language Database* », pour la création du *Collins COBUILD English Language Dictionary***.**

[104] Voir D. Biber (1993a : 247): « *a simple demogaphically based sample of language use would be proportional by definition — the resulting corpus would contain the registers that people typically use in the actual proportions in which they are used. A corpus with this design might contain roughly 90% conversation and 3% letters and notes, with the remaining 7% divided among registers such as press reportage, popular magazines, academic prose, fiction, lectures, news broadcasts, and unpublished writing*. ».

population produit dans un empan temporel donné (une journée ? une semaine[105] ? un mois ? ... une vie ?). A moins qu'il ne faille entendre la langue à laquelle nous sommes *exposés* à travers les productions écrites et orales que nous interprétons. En ce cas, c'est un autre travail de modélisation qui est nécessaire, fondé aussi sur un échantillon représentatif de locuteurs et accompli en vue de recenser la variété des genres et des registres auxquels ils sont exposés. Comme le signale S. Hunston (161-162) : *« the composition of a corpus will be very different depending on whether it is based on the amount of each kind of language that is produced or on the amount of each kind of language that most people come into contact with* ». A moins enfin que par « langue » on entende l'ensemble des genres et des sous-genres qui structurent la variation des usages langagiers à un moment donné sans considération pour leurs proportions respectives d'apparition dans ces usages. L'échantillonnage représentatif vise alors à donner une image la plus fidèle possible de cette variation sans chercher à l'équilibrer. Telle est la position de D. Biber (1993a : 247) : « *Researchers require language samples that are representative in the sense that they include the full range of linguistic variation existing in a language*. »

- Difficulté 2. **Connaître la (structure de la) population cible**. Si l'on en croit B. Habert, A. Nazarenko & A. Salem, « notre connaissance de la « population » des données langagières est (…) extrêmement fragmentaire » (*op. cit.* : 150)[106]. Sachant que l'un des principes qui sous-tend toute langue est celui de la variation, on ne dispose actuellement pas d'inventaire exhaustif et hiérarchisé des genres et des registres qui structurent cette variation. Comme l'écrit S. Hunston (*op. cit.* :161) : « *it is not possible to identify a complete list of « categories » that would exhaustively account for all the texts produced in a given language ».* On retrouve ici l'enjeu que constitue le modèle multidimensionnel de la variation langagière soulevé par M.-P. Péry-Woodley (*op. cit*. : 158 & sq.).

- Difficulté 3 : **Connaître la proportion entre les classes formant la structure de la population cible**. On se trouve confronté à la question de la ***proportion*** ou de **l'*équilibre* (*balance*)**. Un échantillon représentatif de la langue devrait à priori refléter les proportions des usages réels de celle-ci. Pour atteindre ce but, il s'agit de « mêl[er] dans des proportions jugées représentatives des usages des extraits textuels empruntés aux différents genres et sous-genres des discours répertoriés (oral familier, oral académique, presse, roman, théâtre, poésie, discours politique, ouvrages scientifiques, etc.) » (S. Mellet, 2008, § 4). Or la possibilité d'une telle proportionnalité suppose[107] qu'on ait préalablement résolu les difficultés 1 et 2 formulées *supra*. Le seraient-elles, il demeure la difficulté que constitue le calibrage de la *taille* de l'échantillon, comme on va le voir.

Difficulté 4 : **Définir les tailles de l'échantillon et du corpus total**. La question de la taille recouvre à la fois la taille en nombre de *mots* du corpus et sa taille en nombre de *textes* (d'où notre pluriel). Dans le cas des corpus représentatifs stratifiés[108] mais non équilibrés, la

[105] Choix du *British National Corpus* pour sa partie orale.
[106] Dans B. Habert 2000, l'auteur parle de « notre ignorance de la population d'événements que constitue un langage dans son ensemble ».
[107] D. Biber (*op. cit*. : 247-248) soulève une difficulté contre l'idée de proportionnalité, consistant à souligner qu'un texte n'a pas seulement une importance quantitativement mesurable, mais aussi une importance symbolique (qualitative) beaucoup plus difficile à mesurer : « *Proportional samples are representative only in that they accurately reflect the relative numerical frequencies of registers in a language - they provide no representation of relative importance that is not numerical. Registers such as books, newspapers, and news broadcasts are much more influential than their relative frequencies indicate* ».
[108] i.e. représentatifs des différentes dimensions/strates de la variation de la population cible

question de la taille en nombre de mots et de textes affectée à chaque strate et sous-strate[109] reste largement ouverte (voir A. Geyken, *op. cit.*: 93).

D. Biber lie la question de la taille des strates et du corpus total à la fréquence des phénomènes linguistiques qu'on cherche à observer : plus ils sont rares, plus le corpus doit être de taille importante. Ainsi, après avoir comparé la distribution de 10 traits linguistiques à l'intérieur de 481 textes appartenant à 23 genres distincts, il conclut à propos de la taille en nombre de mots des échantillons pour chaque texte[110] : « [F]*requency counts for common linguistic features are relatively stable across 1,000 word samples, while frequency counts for rare features (such as conditional subordination and WH relative clauses) are less stable and require longer text samples to be reliably represented.* » (1993a : 249).

J. Sinclair (2004 : 15) quant à lui lie la taille (minimale) du corpus au type de recherche projeté et à la méthodologie qu'on compte adopter : « *The minimum size of a corpus depends on two main factors: (1) the kind of query that is anticipated from users, (2) the methodology they use to study the data.* » Il propose ainsi d'établir une liste des « objets » qu'on veut étudier (mots, lemmes, motifs, …) et d'observer sur des corpus existants faciles d'accès quelle est leur fréquence. Après s'être fixé une fréquence minimale au-dessous de laquelle on ne veut pas descendre, on peut faire une première estimation de la taille minimale à atteindre pour le corpus à construire. Quant à la méthodologie, elle concerne le recours ou non aux outils statistiques. Si tel est le cas (calcul des cooccurrences par exemple), le nombre d'occurrence disponibles pour les objets de la recherche doivent être considérablement accru. « *If you intend to continue examining the first results using the computer, you will probably need several hundred instances of the simplest objects, so that the programs can penetrate below the surface variation and isolate the generalities. The more you can gather, the clearer and more accurate will be the picture that you get of the language.* » (*op. cit.* : 18)

Un corpus équilibré pose plus crûment encore le problème de la présence des événements langagiers rares dans la mesure où ces derniers sont susceptibles d'apparaître dans des genres/sous-genres peu représentés du fait du respect de la proportionnalité. « Les plus grands corpus équilibrés disponibles actuellement ne contiennent pas suffisamment d'attestations pour servir de base lexicographique unique, notamment pour des phénomènes tels que les expressions figées, les emplois rares. » (A. Geyken, *op. cit*. : 92)[111].

Une dernière difficulté, enfin, externe mais réelle, touche aux réalités matérielles dans lesquelles est impliquée toute recherche :

Difficulté 5 : « What time and resources are available for corpus construction » ? (D. Biber, S. Joahnsson, G. Leech, S. Conrad & E. Finegan (1999 : 27)). Le principe de réalité énoncé ici touche à des volets divers de la recherche empirique, ô combien souvent éprouvés au cours du programme Presto - qu'il s'agisse des limites de temps imparties à tout programme de recherche financé (et qui peuvent par ex. conduire à limiter certaines « expériences » destinées à calibrer le corpus) ou du caractère peu accessible de certains textes conduisant à renoncer à verser certains d'entre eux dans le corpus, pour des raisons sur lesquelles nous reviendrons *infra*.

---

[109] Strates génériques notamment, mais pas seulement : on peut aussi songer (sans prétendre à l'exhaustivité) à la question des « périodes » dans les corpus historiques à large couverture temporelle (voir *infra*).

[110] Il traite ensuite la question de la taille de l'échantillon en nombre de textes pour chaque (sous-)genre et pour le corpus total : « *how linguistic features are distributed across texts and across registers, and how many texts must be collected for the total corpus and for each register to represent those distributions*? » (1993a : 252)

[111] C'est aussi le point de vue de D. Biber : « *A proportional corpus would be useful for assessments that a word or syntactic construction is 'common' or 'rare' (as in lexicographic applications). Unfortunately, most rare words would not appear at all in a proportional (i.e. primarily conversational) corpus, making the database ill-suited for lexicographic research.* » (*op. cit.* : 256)

Ces cinq difficultés majeures pointées, les linguistes se positionnent de manière variable vis-à-vis de la possibilité de représenter la langue dans son ensemble, positions que l'on se propose de répartir suivant quatre pôles.

Les « optimistes lucides » pour qui l'objectif de constituer un échantillon représentatif d'une langue prise dans son ensemble à partir d'une modélisation qualitative multidimensionnelle de la variation langagière reste un objectif pertinent, une sorte de point d'horizon qui règle l'activité du chercheur tout en se dérobant sans cesse. Ainsi D. Biber (1993b : 220) déclare-t-il : « *I argue here that analyses must be based on a diversified corpus representing a wide range of registers in order to be appropriately generalized to the language as a whole, as in a dictionary or grammar of English, or a general purpose tagging program for English.* ». Probablement l'auteur a-t-il en tête[112], lorsqu'il évoque le cas d'une grammaire de l'anglais, le projet de la *Longman grammar of Spoken and Written English* (LSWE) auquel il a participé et qui a connu sa première édition en 1999. Dans les premières pages de cette grammaire, les auteurs déclarent à propos de la « représentativité » du « LSWE Corpus » : « *No corpus provides a perfect representation of a language, and the LSWE Corpus is no exception to this rule* » (27). Optimisme lucide donc : parler de « représentation de la langue générale » revêt un sens, mais cette représentation demeure intrinsèquement imparfaite *quoique toujours perfectible*.

Les « sceptiques constructifs » (parmi lesquels nous nous rangerions volontiers) qui, ayant renoncé à toute velléité de constitution d'un échantillon représentatif de *la langue*, visent une représentativité plus tempérée[113]. C'est dans cette catégorie que semble se ranger B. Habert (2000) : « Notre ignorance de la population d'événements que constitue un langage dans son ensemble m'amène à vouloir des objectifs plus limités et plus spécifiques » (*op. cit.* : 1). M.-P. Péry-Woodley appartiendrait aussi selon nous à cette catégorie de chercheurs[114] qui, tout en intégrant la dimension de la variation au cœur de la constitution du corpus, exclut la représentativité de la langue prise dans son ensemble comme un objectif susceptible de régler une recherche en linguistique sur corpus. Dans sa présentation du corpus *ANNODIS*, elle déclare avec ses co-auteurs : « *La ressource ANNODIS est composée de brèves et d'articles à visée informative représentant une certaine diversité en terme de genre textuel, de type dominant et de structure de document. Notre objectif a été d'intégrer d'entrée de jeu l'hypothèse de la variation dans les réalisations discursives en fonction de variations extralinguistiques (...). La diversification des textes du corpus n'a donc pas été envisagée dans l'optique de fournir un corpus de référence des genres écrits du français, mais dans l'idée de* ***constituer des données autorisant des comparaisons intergenres***[115]. » (M.-P. Péry-Woodley, S. D. Afantenos, L.-M. Ho-Dac & N. Asher, 2011 : 77)

Les « alternativo-quantitatifs » enfin, dont la devise serait « *more data is better data* ». Cette stratégie qui ne renonce pas à la représentativité, contourne cependant la voie qualitative que représente la modélisation de la variation langagière (toujours fragile, partielle et critiquable), au profit d'une voie quantitative : « engranger le maximum de données, le

[112] « *Its planning began in 1992* » déclarent les auteurs dans leur préface de la LSWE.
[113] Il est intéressant à cet égard de relever la formulation prudente adoptée dans la présentation en ligne du BNC (c'est nous qui soulignons le segment concerné) : « *The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent* ***a wide cross-section of British English*** *from the later part of the 20th century, both spoken and written.* » (http://www.natcorp.ox.ac.uk/corpus/index.xml)
[114] Voir aussi A. Condamines (2000).
[115] C'est nous qui soulignons.

poids total devant être garant de la richesse amassée. » (B. Pincemin, 1999 : 417).

On ne peut enfin exclure les « **cyniques** » qui, considérant toute velléité de représentativité en linguistique comme chimérique, y voient l'échec de toute linguistique sur/de corpus.

1.2.3. Un échantillonnage représentatif : comment ?

Avant cependant d'aborder la question plus particulière des corpus historiques, nous voudrions détailler quelques éléments trop rapidement évoqués jusqu'ici, concernant l'échantillonnage représentatif d'une population langagière quelconque.

1.2.3.1. Un échantillonnage stratifié

Si l'on suit D. Biber (1993a), le caractère représentatif d'un échantillonnage dépend d'une bonne définition préalable de la population cible (*target population)*, c'est-à-dire de la définition de ses frontières (*boundaries*) et de sa structure hiérarchique interne (*hierarchical organization within the population*). Autrement dit, l'approche préconisée consiste en un échantillonnage stratifié (*stratified sampling*[116]) : « Dans un échantillonnage stratifié, on divise dans un premier temps la population en sous-populations appelées strates. Ces strates ne doivent pas s'interpénétrer et l'ensemble de ces strates doit constituer l'ensemble de la population. Une fois que les strates ont été déterminées, on tire un échantillon aléatoire (pas forcément de même taille) de chacune des strates, cet échantillonnage étant fait indépendamment dans les différentes strates. » (Y. Dodge, op. cit. : 166).

Comment modélise-t-on ces strates, c'est-à-dire la structure interne de la population langagière visée – quelle qu'elle soit? Les critères pour y parvenir peuvent être de deux natures : des « critères externes » « *that are derived from an examination of the communicative function* » et des « critères internes », « *that reflect details of the language of the text* ». (J. Sinclair, 2004 : 6). Sur ce point, J. Sinclair et D. Biber diffèrent quant à la méthode. Si tous deux préconisent le recours aux critères externes, seul le second retient en outre l'usage de critères internes.

Les critères externes visent à modéliser, dans le processus de constitution d'un corpus, les paramètres situationnels de l'acte de communication : *« In corpus design, variability can be considered from situationnal (…) perspective[.] ».* D. Biber, *op. cit.* : 243). L'étude de ces critères a donné lieu à des typologies diverses mais relativement proches. J. Sinclair (*op. cit.* : 8) propose d'en distinguer six : « *(1) the mode of the text; whether the language originates in speech or writing, or perhaps nowadays in electronic mode; (2) the type of text; for example if written, whether a book, a journal, a notice or a letter; (3) the domain of the text; for example whether academic or popular; (4) the language or languages or language varieties of the corpus; (5) the location of the texts; for example (the English of) UK or Australia; (6) the date of the texts.* ». D. Biber propose quant à lui (1993a : 245) d'en distinguer huit:

---

[116] *« In this method, subgroups are identified within the target population (in this case, the genres), and then each of those 'strata' are sampled using random techniques. This approach has the advantage of guaranteeing that all strata are adequately represented while at the same time selecting a non-biased sample within each stratum. (…) stratified samples are almost always more representative than non-stratified samples (and they are never less representative). »*(244)

Table 1 Situational parameters listed as hierarchical sampling strata

1.	*Primary channel.* Written/spoken/scripted speech
2.	*Format.* Published/not published (+ various formats within 'published')
3.	*Setting.* Institutional/other public/private-personal
4.	*Addressee.* (*a*) Plurality. Unenumerated/plural/individual/self (*b*) Presence (place and time). Present/absent (*c*) Interactiveness. None/little/extensive (*d*) Shared knowledge. General/specialized/personal
5.	*Addressor.* (*a*) *Demographic variation.* Sex, age, occupation, etc. (*b*) *Acknowledgement.* Acknowledged individual/institution
6.	*Factuality.* Factual-informational/intermediate or indeterminate/imaginative
7.	*Purposes.* Persuade, entertain, edify, inform, instruct, explain, narrate, describe, keep records, reveal self, express attitudes, opinions, or emotions, enhance interpersonal relationship, . . .
8.	*Topics. . . .*

Dans la table ci-dessus, les « topics » figurent parmi les critères externes alors que J. Sinclair les en exclut au motif qu'ils relèvent des critères linguistiques internes : « *The most obvious manifestation of topic is certainly found in the vocabulary (...). But vocabulary choice is clearly an internal criterion.* » (*op. cit.* : 15). Ajoutons que la typologie de Biber *supra* a été revue et amendée dans D. Biber & S. Conrad (2009), et que D. Biber, S. Joahnsson, G. Leech, S. Conrad & E. Finegan (*op. cit.*) proposent une grille présentant la distribution de cinq traits situationnels dans quatre registres principaux qu'ils distinguent dans le corpus de la *Longman grammar of Spoken and Written English* : *conversation, fiction, news, academic prose.* Bien entendu, d'autres typologies de critères externes ont été envisagées.

Pour D. Biber enfin, le paramétrage situationnel est branché directement sur la question des registres et des genres : « *I use the terms genre or register to refer to situationally defined text categories (such as fiction, sports broadcasts, psychology articles).* » (*op. cit.* : 244). L'intérêt d'un tel modèle paramétrique, comme le souligne M.-P. Péry-Woodley (2000 : 162), est de « rendre compte du fait que s'il existe bien des prototypes nets, les frontières entre registres ou genres sont inévitablement floues». De fait, pour D. Biber (*op. cit.* : 245), « *'[R]egister' should be specified as a continuous (rather than discrete) notion.* » Il convient de souligner pour finir que chez ce dernier, les *registres/genres* de textes doivent être soigneusement distingués des *types* de textes[117], dont les catégories sont fondées sur des critères linguistiques (internes donc) : « *I use the terms genre or register to refer to situationally defined text categories (such as fiction, sports broadcasts, psychology articles), and text type to refer to linguistically defined text categories.* (D. Biber, *op. cit.* : 244-245).

Les critères internes sont relatifs aux dimensions linguistiques du texte (voir définition de J. Sinclair, *supra*). Pour cet auteur, cette famille de critère n'a pas à être pris en compte dans la constitution du corpus, sous peine de circularité : « *The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise. Obviously if it is already*

[117] Nous n'évoquerons pas ici la typologie « émergente » de cet auteur (voir, outre D. Biber (1989), M.-P. Péry-Woodley : *op. cit.* : 158-160)

known that certain text types contain large numbers of a microlinguistic feature such as proper nouns or passive verb phrases, it becomes a futile activity to "discover" this by assembling a corpus of such texts. » (*op. cit.* : 5). La thèse de la circularité est aussi mise en avant par T. McEnery, R. Xia & Y. Tono (2006 : 14) : « *In our view, it is problematic, indeed it is circular, to use internal criteria like the distribution of words or grammatical features as the primary parameters for the selection of corpus data. A corpus is typically designed to study linguistic distribution. If the distribution of linguistics features is predetermined when the corpus is designed, there is no point in analysing such a corpus to discover naturally occurring linguistic feature distributions* ».

A l'inverse, D. Biber plaide pour la prise en compte de critères internes pour la constitution du corpus. Pourquoi ?

1.2.3.2. Etude par D. Biber (1993a) de la distribution de dix traits linguistiques dans une population cible ; définition des tailles de l'échantillon et des strates représentées

Dans son approche de la représentativité, D. Biber lie étroitement la distribution des traits linguistiques dans la population cible (critères internes) à l'optimisation de la taille[118] du corpus. Ainsi, concernant l'étude de la distribution des traits linguistiques à l'intérieur des textes : *« I consider first the distribution of linguistic features within texts, as a basis for addressing the issue of optimal text length* ». (*op. cit.* : 248). Chaque mot ne peut pas être traité comme un individu atomisé séparé de son cotexte: si l'on agissait ainsi, on détruirait la structure des syntagmes ou de la phrase dans lesquels il est impliqué. Il s'agit donc, pour ce qui concerne la taille en nombre de mots d'un corpus, de déterminer le nombre de mots contigus (*contiguous words)* à sélectionner.[119]

Les traits linguistiques dont on examine la distribution sont au nombre de dix : *Nouns, Prepositions, Present tense, Past tense, Third person pronouns, First person pronouns, Contractions, Passives, WH relative clauses, Conditional clauses*/ Ils possèdent pour caractéristiques que chacun « *potentially represents a different statistical distribution across text categories* ». Leur variation distributionnelle est alors étudiée à l'intérieur des textes, à l'intérieur des genres, puis suivant les textes et suivant les genres (*across texts / across registers*). On ne développera pas ici ces points, se contentant de rappeler quelques conclusions de l'auteur qui ont paru intéressantes pour la constitution du corpus Presto.

Concernant la variation distributionnelle des traits à l'intérieur des textes, il apparaît après comparaison d'échantillons pris deux à deux[120] que les traits linguistiques dotés d'une haute fréquence (par ex. les catégories morphosyntaxiques *noms* et *préposition*) possèdent une répartition relativement stable au fil des textes. Pour ces traits, il est donc possible de se contenter d'échantillons de textes d'une longueur de 1000 mots. Pour les traits rares en revanche (par ex. *WH relative clauses, conditional subordination*), cette longueur doit être significativement accrue.

---

[118] Par taille, rappelons-le, il faut entendre non seulement le nombre de *mots* dans le corpus mais aussi le nombre de *textes*.

[119] L'auteur ne semble en revanche pas sensible à la structure textuelle que met en péril l'échantillonnage des textes pour lequel il plaide : ce qu'il nomme « *the overall structure of the text* » correspond davantage à une unité structurale du niveau de la phrase (complexe) plutôt qu'à une structure sémantico-pragmatique du niveau du texte au sens où l'entendent M.A.K Halliday & R. Hasan (1976 : 1-2) : « *A text is a unit of language in use. It is not a grammatical unit, like a clause or a sentence ; (…) A text is best regarded as a SEMANTIC unit : a unit not of form but of meaning* ».

[120] « *The distributions of these linguistic features were analysed in 110 1,000-word text samples (i.e. fifty-five pairs of samples), taken from seven text categories: conversations, broadcasts, speeches, official documents, academic prose, general fiction, and romance fiction.* » (249)

Concernant la variation distributionnelle des traits dans les registres et suivant les registres, D. Biber observe aussi que ce qu'il nomme la « variance normalisée » (*normalized variances*) est nettement moins prononcée pour les traits linguistiques à haute fréquence (par ex. noms et prépositions).

Le nombre de textes total requis dans le corpus total, enfin, est calculé à partir d'une équation discutée et affinée dans les § 3.2 et § 4.2.2. de l'article auxquels nous renvoyons le lecteur. D. Biber là encore observe que pour les traits à haute fréquence comme les noms et les prépositions, la déviation standard est relativement faible et autorise un nombre de textes dans le corpus complet relativement petit (« *quite small* [121]») ; à l'inverse, pour les traits rares, la déviation est nettement plus élevée et le nombre total de textes dans l'échantillon doit être significativement plus élevé pour accéder à la représentativité[122].

En conclusion, constituer un corpus visant à la représentativité pour l'étude des prépositions en anglais contemporain autoriserait, du fait de la haute fréquence d'usage de cette catégorie morphosyntaxique dans les productions langagières, i) un corpus de taille totale (nombre de textes) relativement petite, ii) un nombre de textes relativement restreint dans chaque strate représentée, iii) un nombre de mots contigus par texte échantillonnés d'environ mille. Dans quelle mesure ces observations sont-elles transférables pour réfléchir à la constitution du corpus Presto ? C'est évidemment très difficile à dire et seules des études rigoureuses - que nous n'avons pas menées - permettraient de répondre avec quelque certitude. Mais il y a selon nous une certaine similarité à établir en ceci que, quelle que soit la période représentée (du XVI^{e} s. au XX^{e} s.), nous avons toujours observé dans Presto (mais aussi dans la base Frantext) que la catégorie des prépositions possède une haute fréquence[123]. On peut donc faire l'hypothèse avec une certaine vraisemblance que cette catégorie morphosyntaxique possède une distribution linéaire stable à l'intérieur des textes (« *stable linear distribution* », *op. cit.* : 251) et que la variation de sa distribution entre les textes, entre les genre discursifs et à l'intérieur de ces derniers devrait être moindre que pour des traits linguistiques plus rares. Si tel est le cas, les choix de tailles (voir plus loin) opérés dans la version actuelle du corpus Presto possèderaient une certaine pertinence.

Il convient maintenant de nous tourner vers les corpus historiques afin de mieux cerner les contraintes spécifiques que leur constitution engendre.

1.3. Corpus historiques et représentativité

Comme le souligne C. Marchello-Nizia (2004 : 58-60) « *la linguistique diachronique [est] habituée depuis longtemps à fonder ses analyses sur des corpus* ». En d'autres termes, la linguistique sur corpus numérisés constitue plus un prolongement qu'une rupture dans les pratiques des diachroniciens, du moins pour ce qui regarde le contact avec les textes et le recours à la quantification des faits langagiers[124].

---

[121] Ainsi, pour ce qui regarde l'usage des prépositions en anglais contemporain, il faudrait constituer un échantillon stratifié d'environ 82 textes pour pouvoir accéder à la représentativité des résultats, si l'on en croit le tableau 4 (*Estimates of required sample sizes (number of texts) for the total corpus*) présenté par l'auteur.

[122] Pour ce qui regarde l'usage des *conditional clauses* en anglais contemporain, il faudrait constituer un échantillon stratifié d'environ 1,190 textes pour pouvoir accéder à la représentativité des résultats, selon le même tableau.

[123] Ceci dit sans considération ici pour le rang occupé par chacune des prépositions au sein de cette fréquence totale : le cas de *dans* (voir notre troisième partie) est à cet égard éloquent.

[124] La question de l'outillage informatique et du traitement automatique qui mène à la possibilité d'explorer automatiquement les textes est différente : à cet égard, diachroniciens et synchroniciens sont sur un pied

De fait (et cela a été fréquemment souligné), lorsqu'on s'intéresse à un état ancien de la langue, hors des corpus point de salut … qu'il s'agisse du français médiéval (voir S. Prévost 2008 : § 3-4) ou de toute autre période révolue pour lesquelles on ne dispose pas de locuteur vivant. Tout accès à une « compétence » (au sens de Chomsky) s'avère exclue : « *introspection and native-speaker competence cannot be relied on in the study of the language of previous centuries and millennia ».* (M. Rissanen, *op. cit.*: 53). Il s'agit là d'une contrainte spécifique qui a des répercussions importantes sur la constitution et la nature du corpus.

1.3.1. Définition d'un corpus historique

C. Claridge (2008 : 242) déclare :

A 'historical' corpus is one which is intentionally created to represent and investigate past stages of a language and/or to study language change. In all other respects, the defining characteristics of a corpus apply: it is a finite electronic collection of texts or part of texts by various authors which is based on well-defined and linguistically relevant sampling criteria and aims for some degree of representativeness.

C'est donc la visée présidant à la collection des données langagières - représenter et étudier les états anciens de langue et/ou le changement linguistique - qui distingue le corpus historique. Pour le reste, et nous en sommes d'accord, les critères « définitoires » du corpus restent valables. En d'autres termes, la constitution d'un corpus historique pour une langue donnée renvoie (comme tout corpus) aux enjeux discutés précédemment, à savoir :

- la définition explicite et opérationnelle de la population visée ;
- la question de la représentativité ;
- la question de l'échantillonnage (son utilité voire son éventuelle nécessité, sa mise en œuvre) et de sa taille.

Sur un point cependant, nous différons de la définition de C. Claridge : stipuler le caractère « fini » de la collection des textes écarte - au moins implicitement nous semble-t-il - la possibilité d'un corpus « évolutif » à laquelle nous tenons pour Presto (voir *supra* § 1.1.2.)

Ajoutons enfin que l'auteur introduit plus loin dans son article une spécificité des corpus historiques couvrant de larges périodes temporelles (comme c'est le cas pour Presto) :

Long corpora crucially also have an internal temporal structure, i.e. sub-periods which have a parallel or closely comparable composition: in the case of CONCE[125]*, for example, there are three periods of 20-30 years each containing an identical range of registers.* (*op. cit.*: 243)

Cette question de la « structure temporelle interne » est intéressante en ceci qu'elle ouvre à la notion de *comparabilité* dans les corpus historiques à large couverture temporelle. Cette notion doit être prise ici en un sens bien précis : elle indique que chaque tranche temporelle (d'empan à définir préalablement) qui constitue le corpus historique a été élaborée sur des critères identiques (choix de taille en nombre de mots/textes, choix de genres/sous-genres

---

d'égalité pourrait-on dire, sauf peut-être pour l'usage de certaines fonctionnalités documentaires automatisées telles la concordance.

[125] En fait, le CNNE : *Corpus of Nineteenth-century Newspaper English* (http://www.helsinki.fi/varieng/CoRD/corpora/CNNE)

discursifs, choix des domaines représentés etc.). Cette identité de construction rendant chaque tranche comparable aux autres.

On observera qu'il s'agit là d'une déclinaison possible de la notion de « comparabilité » telle que la définit B. Habert (2000 : 1) pour ce qui regarde des corpus :

« Corpus comparable : les textes, dans des langues ou des états de langue différents[126], sont rassemblés selon des critères similaires, en ce qui concerne le domaine, le genre, … ».

Une question demeure en suspens, que nous aborderons *infra* (§ 1.3.2.5) : elle concerne la relation qu'on peut ou non établir entre le découpage d'un corpus historique long en tranches chronologiques de même empan et la « périodisation » des phénomènes linguistiques qu'on étudie.

1.3.2. Caractéristiques et contraintes propres aux corpus historiques

1.3.2.1. Définir la population cible

Comme dit précédemment, définir explicitement la population cible constitue la première étape de toute constitution de corpus.

Les frontières temporelles pour commencer : pour ce qui concerne Presto, la borne temporelle initiale définie est 1500 (borne « arithmétique » du début du XVI^{e} s.), la borne finale 1944 - celle-ci ayant été définie en fonction de la période de soixante-dix ans pendant laquelle courent les droits d'auteurs avant que l'œuvre ne monte dans le domaine public[127].

Quelle population cible un corpus historique vise-t-il ? L'idée de représenter la langue prise dans sa globalité serait une douce et folle utopie… Elle se heurte de plein fouet à une difficulté considérable que tous les diachroniciens[128] connaissent bien : plus on s'éloigne de la période contemporaine, plus les données (ou les traces) linguistiques accessibles se font rares. L'oral n'a laissé quasiment aucune trace (sinon par sa « représentation » dans les écrits, essentiellement littéraires), les productions langagières des locuteurs analphabètes ou peu lettrés nous sont parfaitement inconnues, la variété des genres accessibles décroît (et l'on sait que, pour les périodes les plus reculées, tous ne nous ont pas été transmis), etc. Comme l'écrit R.-A. Lodge, « les données linguistiques parvenues jusqu'à nous des époques révolues sont rarement celles que le linguiste aurait choisies, laissé à lui-même : elles survivent de manière fortuite, elles sont fragmentaires et loin d'être représentatives de tous les registres de la langue, et, surtout, elles sont écrites et non orales » (2009 : 211). C. Claridge (op. cit. 247) fait le même constat: « *the texts transmitted to the present represent a random subsample of the whole population, due to largely extra-linguistic accidents. Thus, historical corpora can never even remotely capture the full variety of language* ». Autrement dit, il apparaît exclu que les concepteurs d'un corpus historique décidés à échantillonner des textes relevant d'un état de langue ancien puissent se fixer comme population cible la langue *telle qu'elle était pratiquée par les locuteurs du temps*. La population des événements langagiers visés par un corpus historique se réduit nécessairement (et « asymptotiquement » pourrait-on dire) à la

[126] En l'occurrence, dans Presto, il s'agit de textes relevant d' « états de langue différents », chaque tranche chronologique étant considérée comme un cous-corpus comparable aux autres sous-corpus (c'est-à-dire aux autres tranches).
[127] voir par ex. http://cahier.hypotheses.org/notions-juridiques#droit_d_auteur
[128] Et la sociolinguistique historique : voir P. Blumenthal & D. Vigier, à par. 2017.

somme des « traces » linguistiques qui nous en ont été transmises. Etudier par ex. le genre de la conversation en français préclassique équivaudrait, en terme de population, à cibler l'ensemble des textes qui nous disent aujourd'hui quelque chose des conversations de cette époque. Cette limite formulée, la question de la représentativité se pose sous un jour différent que précédemment nous semble-t-il. Ce que l'on cherche à représenter en linguistique historique, ce n'est pas une population d'événements langagiers par essence infinie comme c'est le cas pour une langue contemporaine[129] mais seulement les traces (en nombre nécessairement fini) d'une population d'événements langagiers donnés. Bref, s'il y a représentativité, ce ne peut être que des traces… Nous sommes donc réservé vis-à-vis des propos de S. Prévost (2015) pour qui représenter de la manière aussi proche que possible « des états de langue successifs que nous savons avoir existé » demeure en linguistique historique « un *idéal* vers lequel il faut tendre » (§ 29). Cet idéal nous semble plutôt devoir être de représenter le plus fidèlement les *traces* que nous avons conservées de ces états de langue. Bref, nous serions plus kantien que platonicien dans cette affaire, arguant du caractère inconnaissable des noumènes (la langue) pour nous tourner vers les seuls phénomènes (ses traces).

La définition opérationnelle de la population cible, quant à elle, met en jeu (voir *supra*) pour chaque strate identifiée une liste des textes candidats à figurer dans l'échantillon, liste que l'on soumet à un tirage aléatoire – du moins lorsque le nombre de textes disponibles le permet… La confection d'une telle liste engage, dans le cas des corpus historiques, la question de la datation du texte. « *The date of the sampled text is of great importance with regard to the time frame and the internal sub-periods. »* (C. Claridge, *op. cit*. : 244). Choisit-on la date de la première édition ou bien la date (approximative souvent, et sous réserve qu'elle soit connue) du manuscrit ? Les pratiques des bases textuelles diffèrent à cet égard. La BFM (http://bfm.ens-lyon.fr) par exemple a choisi d'affecter aux œuvres qu'elle met en ligne la date de composition du manuscrit (connue ou estimée)[130]. Les bibliothèques virtuelles humanistes (BVH http://www.bvh.univ-tours.fr) ainsi que Frantext (http://www.frantext.fr) – bases avec lesquelles nous avons le plus souvent travaillé pour Presto – utilisent le plus souvent la date d'édition originale de l'œuvre. Dans le corpus Presto, c'est ce principe que nous avons choisi[131]. Il demeure cependant que faute de temps, nous avons repris sans systématiquement les vérifier les dates d'édition originales fournies par ces bases pour les textes qu'ils nous ont communiqués. Ce travail reste donc à accomplir.

Un dernier point, qui touche à la **qualité philologique des textes** doit être ici évoqué. Dans Presto – comme dans toute entreprise de constitution d'un corpus historique – la qualité des éditions utilisées est primordiale. En particulier, lorsque nous ne disposions pas de l'édition originale pour une œuvre donnée, nous avons cherché à sélectionner un exemplaire dont l'édition s'approchait le plus de cette dernière, notamment pour ce qui regardait la ponctuation, l'orthographe et la graphie. Comme l'écrivent C. Guillot, S. Heiden, A. Lavrentiev & C. Marchello-Nizia (*op. cit.*) à propos des textes médiévaux :

> « il existe (…) des critères qui président à une sélection raisonnée des textes à numériser pour les inclure dans un corpus. L'un des critères pris en compte presque

[129] Le linguiste intéressé par le genre de la conversation aujourd'hui peut toujours, comme l'ont fait D. Biber, S. Joahnsson, G. Leech, S. Conrad & E. Finegan (*op. cit.*) par ex., demander à des locuteurs d'enregistrer un nombre d'heures donné de conversations dans lesquelles ils sont impliqués.
[130] La procédure suivie est détaillée dans http://ccfm.ens-lyon.fr/IMG/pdf/Manuel_Descripteurs_BFM.pdf)
[131] C'est aussi le choix qui a été fait pour le COPC (*The Century of Prose Corpus)* par ex., si l'on en croit C. Claridge (*op. cit*. : 243).

systématiquement est la « qualité philologique » des textes « papier » : qualité des éditions choisies (fondées sur un manuscrit lui-même bien choisi, que l'éditeur suit le plus fidèlement possible sans trop le « corriger » ou l' « amender »), mais aussi finesse de leur description (…) »

1.3.2.2. Rareté et caractère parcellaire des « traces » linguistiques transmises jusqu'à nous pour les états les plus anciens de la langue

Nous avons pointé *supra* le caractère parcellaire des données linguistiques accessibles pour des états anciens de la langue, qui interdit toute ambition de représentativité de la langue telle qu'elle était pratiquée. Précisons notre propos. D'abord, ces données sont écrites (lorsqu'on travaille sur des états de la langue antérieurs aux procédés techniques d'enregistrement sonore), les traces sonores ne pouvant pas être fixées sur un support (sinon cas de transcription). D'où toute une cascade de biais[132], puisque cette absence d'oral oblitère non seulement l'accès aux variétés populaires mais aussi à des couches de la société marginalisées pour des raisons de genre sexué (les femmes), d'âge (les enfants) etc. De surcroît, plus on s'éloigne dans le passé, plus la conservation de ces traces écrites a été soumise au hasard. Et même si, au cours du temps se sont maintenus quelques processus de conservation intentionnels (archivage institutionnel, conservation de documents dans le cadre familial, …), ce qui a été conservé n'a souvent rien à voir avec les objectifs qui animent le chercheur. En d'autres termes, comme le dit non sans humour W. Labov (1994 :11) : « *Historical linguistics can then be thought of as the art of making the best use of bad data* ».

1.3.2.3. Question des genres discursifs

La question des genres discursifs est primordiale. Elle se formule à de nombreux égards suivant les mêmes catégories que celles auxquelles nous avons recouru *supra* (recours à paramètres non-linguistiques situationnels pour les modéliser ; problème de leur formulation, de leur inventaire et de leur combinaison, mise au point d'une typologie, …). En outre se posent des difficultés propres à la diachronie. Elles sont essentiellement au nombre de trois.

La première, nous l'avons déjà mentionné, est liée au fait que bien des genres qui informaient les pratiques discursives – en particulier à l'oral – dans des états anciens de la langue ne nous ont pas été transmis. « Peut-être même ignorons-nous l'existence de certains genres, disparus sans avoir laissé de textes témoins » (S. Prévost, 2008 : § 9)

La deuxième difficulté à trait au fait qu'en cinq siècles (dans le cas de Presto), ces genres se sont profondément modifiés. *La Princesse de Clèves* publié comme roman n'a guère à voir - sur le plan linguistique, notamment - avec *Femmes* de Philippe Sollers ou un roman de Céline. Comme l'écrit C. Claridge (*op. cit*. : 248) : *« Some registers or genres are present throughout history, but with different functions and thus with partly different linguistic realisations (e.g. history writing) - in other words, while the genre remains constant, the linguistics text type undergoes change.* » Un tel changement peut aussi se faire sentir plus particulièrement quant au topic / thème d'un (sous-)genre donné[133]. On observe ainsi que dans le genre des « traités » (qu'il conviendrait certainement de raffiner) proposé par Frantext, la thématique religieuse est massivement représentée au XVI^{e} s. alors qu'elle est réduite à la portion congrue au XX^{e} s. Et que dire de l'astrologie ou de l'agriculture[134] ?

[132] Pour une revue de ces biais, voir Auer A., Peersman C., Pickl S., Rutten G. & Vosters R. (2015 : 6-7)
[133] Rappelons que D. Biber (1993a : 245) intègre le *topic* dans sa table des paramètres situationnels constitutifs des registres/genres.
[134] On rejoint là la question des langues / domaines de spécialité.

La troisième difficulté est relative à la disparition ou à l'apparition de genres ou de sous-genres sur la longue durée. Par exemple, les Mystères religieux de la première moitié du XVI[e] s. ont disparu du paysage du théâtre dès le XVII[e] s. si l'on en croit C. Mazouer (2010). Inversement, on ne sait presque rien du genre de la « conversation » avant les méthodes modernes de recueil des données orales. Quant aux tweets et autres tchats[135]…

Ces deux dernières difficultés ont une incidence évidente sur la comparabilité entre les tranches temporelles ménagées à l'intérieur du corpus historique long (voir *supra*). Disposer de tranches synchroniques comparables implique une similitude de (sous-)genres qui y sont représentés. Privilégier la comparabilité conduit ainsi le concepteur du corpus à se centrer sur quelques (sous-)genres pour lesquels on trouve des textes tout au long de la période temporelle à couvrir, nonobstant le problème pointé *supra* (et non résolu) que représente leur évolution parfois très significative dans le temps (en termes de traits linguistiques ou non linguistiques). C'est le choix qui a été fait par le CNNE[136] (voir C. Claridge, *op. cit*. : 243) et que nous avons retenu pour Presto.

Une autre option pourrait consister à choisir pour chaque tranche la totalité (idéalement) des (sous-) genres accessibles. Le corpus serait alors apte à refléter l'apparition et/ou la disparition de tel(s) ou tel(s) d'entre eux, la mise à disposition de métadonnées relatives à ces (sous-) genres permettant à l'utilisateur de sélectionner ceux qu'il souhaite étudier, qu'ils soient « continus[137] » ou non. Un autre avantage de cette seconde option est qu'elle permettrait de conduire des observations dans les différents genres/registres disponibles. Or, qui s'intéresse au changement linguistique sait que tel ou tel phénomène apparaît d'abord le plus souvent dans un (sous-)genre donné puis gagne ensuite d'autres pratiques discursives. Par ex., comme l'écrit M. Rissanen (*op. cit*. : 57): « *it is easy to see that such connectives as* except, provided (that), *and* notwithstanding *were in common use in legal and documentary texts as early as the fifteenth century, and that this use in prestigious genres may have contribute to their establishment in the emerging standard* ». S. Prévost (2015: 26) fait des observations similaires. L'inconvénient néanmoins d'une telle politique est qu'elle met en péril la comparabilité entre un certain nombre des tranches du corpus.

Probablement la solution est-elle dans l'usage que fait l'utilisateur des métadonnées qu'il a à disposition lors de ses requêtes. Grâce à elles, il doit pouvoir naviguer entre des recherches portant sur tels ou tels (sous-)genres à l'intérieur desquels il veut traquer un phénomène linguistique émergent, et des recherches portant sur d'autres (sous-)genres présents continûment dans la période temporelle couverte par le corpus.

1.3.2.4. « *Internal temporal structure* » des corpus diachroniques longs

Nous reprenons ici les termes de C. Claridge (*op. cit*. : 243) cités plus haut. Deux questions se posent pour la définition de cette structure temporelle interne. D'abord, celle de l'empan temporel de chaque tranche synchronique. Ensuite, leur taille (nombre de mots et textes dans chaque strate et dans la tranche totale).

Sans prétendre à l'exhaustivité, la définition de l'empan temporel des tranches nous semble pour partie liée à une question d'ordre pratique (de plus en plus aiguë au fur et à

[135] De même, Claridge (247) : « *Press and natural science writing (in the modern sense) are two examples of late emerging registers, which are simply not present before the late 17th century or even later.* »
[136] *Corpus of Nineteenth-century Newspaper English,* http://www.helsinki.fi/varieng/CoRD/corpora/CNNE
[137] Par continuité, nous entendons que tel ou tel genre/registre est présent dans toutes les tranches temporelles présentes dans le corpus global.

mesure qu'on s'éloigne de la période contemporaine) : combien de textes numérisés, relevant de (sous-)genres différents, sont-ils raisonnablement accessibles (étant connues les contraintes de temps et d'argent) pour les tranches de 10/20/30… ans les plus anciennes du corpus à construire ? Dans Presto nous avons fait le choix de tranches temporelles de dix ans. Ce choix n'a pas été complètement arbitraire car nous souhaitions une granularité fine – plus fine que celle d'ARCHER[138] par ex. qui a fait le choix de tranches de 50 ans, ou de 20-30 ans dans le CNNE[139] - et nous considérions que le fait de ne pas remonter au-delà de 1500 nous permettait de rester confiants quant aux ressources de textes disponibles *via* les BVH et Frantext.

Le calibrage de la taille (nombre de textes et de mots) de chaque tranche temporelle et du corpus total est aussi une affaire délicate. Il existe par ailleurs une relation étroite entre l'amplitude de la période temporelle qu'un corpus veut couvrir et sa taille, l'accroissement de l'une allant de pair avec celle de l'autre. Selon nous, cette question de la taille devrait pouvoir recevoir une réponse étayée, en s'appuyant par ex. sur les préconisation de D. Biber (*supra*) quant à l'étude de la distribution de traits linguistiques sélectionnés en fonction de leur plus ou moins grande rareté/banalité dans les textes. C. Claridge (*op. cit*.) propose quant à elle de tester certains acquis concernant le changement linguistique sur des corpus de plus ou moins grande taille pour déterminer à partir de laquelle on retrouve toutes les étapes du changement recherché. Pour ce qui concerne Presto, nous avons fait les choix suivants : chaque texte dont la longueur est supérieure à 5000 mots a été échantillonné pour ne pas excéder cette taille. Cet échantillonnage a obéi à un algorithme qu'on peut décrire grosso-modo comme suit : cinq extraits de 1000 mots contigus sont choisis aléatoirement dans 5 parties du texte réparties entre le début et la fin de ce dernier. Pour ce qui concerne le nombre de textes échantillonnés dans chaque genre et dans chaque tranche temporelle ainsi que le nombre de mots et de textes dans le corpus total, nous renvoyons au § 1.4 de cette partie.

1.3.2.5. Tranches temporelles et périodisation de la langue

Le français ne peut évidemment pas être considéré, sur le plan linguistique, comme homogène et identique à soi entre le IX^{e} s. et le XXI^{e} s. Vouloir le « représenter » dans un corpus nécessite donc qu'on se donne des périodes au sein desquelles on peut le considérer comme (relativement) stable, chacune de ces périodes constituant une strate qui structure la variation (diachronique) de la population langagière visée au même titre que les (sous-)genres. Malheureusement, on ne dispose pas actuellement d'une périodisation consensuelle des changements linguistiques du français, et la question de sa possibilité voire de sa pertinence reste ouverte (voir par ex. B. Combettes & C. Marchello Nizia 2010, B. Combettes 2012 ; R. de Dardel, M. Banniard & B. Combettes (éds.) (2011)). Il est intéressant d'observer à cet égard que, pour ce qui regarde le projet de la *Grande Grammaire Historique du Français*, une périodisation possible de la langue est envisagée par C. Marchello-Nizia (2011 : § 4.2.) non comme un point de départ mais comme un (plus ou moins hypothétique) point d'arrivée : « La partie conclusive, celle qui en fait motive chacun des participants à cette entreprise, aura pour fin de présenter une vue d'ensemble des changements qu'a connus le français en douze siècles, *avec la tentative d'une nouvelle périodisation sur critères linguistiques*[140] ».

D'où cette question : quelle relation possible entre l'empan des tranches temporelles sélectionnées dans le corpus historique à couverture longue et la périodisation des

[138] *A Representative Corpus of Historical English Registers*, http://www.manchester.ac.uk/archer
[139] *Corpus of Nineteenth-century Newspaper English, http://www.helsinki.fi/varieng/CoRD/corpora/CNNE*
[140] C'est nous qui soulignons.

phénomènes langagiers visés? Selon nous, il n'y en a aucune et les deux notions sont à découpler. En effet, le découpage en tranche chronologiques successives – qui sont autant de micro-tranches synchroniques – ne constitue finalement qu'une sorte de « quadrillage » à priori de la trame du temps (comme on quadrille une feuille) destiné à faire apparaître, sur ce fond réglé, des variations quantitatives relatives à tel ou tel phénomène linguistique. Ce sont ces variations qui pourront ensuite ouvrir à la question d'éventuelles périodes discernables au sein du français, et à cet égard, ce serait un acquis considérable si Presto concourait à l'identification d'une périodisation possible dans l'évolution du système des prépositions entre 1500 et 1944.

1.3.2.6. La question des droits

Les droits d'auteur (pour les textes postérieurs à 1944) et les droits d'édition susceptibles de peser sur les œuvres du passé constituent souvent une contrainte significative qui réduit l'accès aux œuvres (déjà numérisées ou pouvant être numérisées) lorsqu'on cherche à construire un corpus historique du français. On rappellera pour mémoire que la BFM, par ex. a dû retirer de son site en 2014 plusieurs textes à la demande des éditeurs concernés. Depuis 2013 (date de mise en route de Presto), les consortiums CAHIER (http://cahier.hypotheses.org) et CORLI (https://corli.huma-num.fr) ont travaillé cette question en vue d'informer au mieux la communauté des chercheurs, par le biais notamment de guides juridiques (http://cahier.hypotheses.org/guides-juridiques). En ce qui concerne le corpus Presto, nous n'avons sollicité (d'un commun accord) Frantext que pour le prêt de textes relevant du domaine public, ou dont l'éditeur avait probablement fait faillite ou avait cessé ses activités. Ce choix a considérablement réduit le potentiel de nos ressources, sans cependant l'assécher. Pour ce qui regarde la mise à disposition d'une partie du corpus sous licence libre, la solution adoptée (comme souvent dans ce type de recherche : voir C. Claridge, *op. cit*. : 245-246, « *copyright* ») consistera à ne proposer au téléchargement et à la consultation en ligne que la partie « domaine public » du corpus.

1.3.2.7. Echantillons ou textes intégraux ?

Sur ce plan, les corpus historiques se trouvent quasiment sur le même pied que les corpus contemporains, à ceci près que pour les périodes les plus reculées, nous ne disposons parfois que de fragments de textes. Le cas de *Tristan* évoqué *supra* en est un bon exemple. Dans Presto, nous avons choisi de « doubler » le corpus échantillonné par un corpus de textes intégraux qui permet à l'utilisateur de naviguer comme il le souhaite entre les deux versions.

1.4. Le corpus Presto : présentation et évaluation

1.4.1. « Niveaux » et « Versions » du corpus Presto

Nous avons distingué quatre **niveaux** dans le corpus : « noyau, contrôlé, étendu, spécialisé » (voir *infra*). Les trois premiers niveaux sont de taille croissante. Le corpus « étendu », dont la taille est la plus grande, contient tous les textes du corpus noyau (taille la plus petite) et du corpus contrôlé (taille intermédiaire) - ainsi que d'autres textes. Le corpus contrôlé contient une partie du corpus noyau, et d'autres textes.

La figure suivante représente ces relations d'inclusion totale ou partielle.

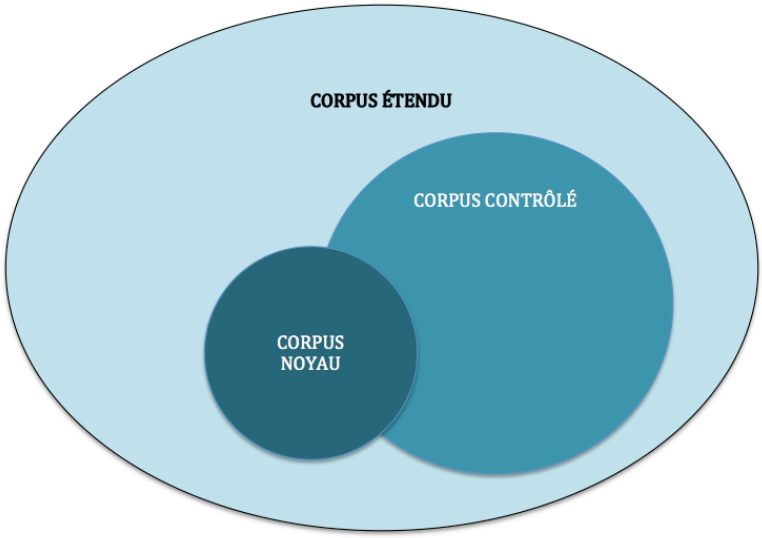


Figure 1
Rapports d'inclusion (partielle ou totale) entre les trois premiers niveaux du corpus dans Presto

Chaque niveau se présente en outre sous deux **versions** : l'une regroupe les textes dans leur intégralité ; l'autre dans une version échantillonnée suivant les principes exposés *supra* (§ 1.3.2.4).

La liste des textes des trois premiers niveaux du corpus, associés à leur principales métadonnées documentaires, est donnée dans l'ANNEXE 4.

1.4.1.1. **Corpus Noyau** (= Presto$_{-NOYAU}$)

Le corpus noyau réunit l'ensemble des œuvres mises à disposition de la communauté des chercheurs, soit 53 textes pour un nombre total de 6.820.161 mots (version intégrale) / 1.924.532 mots (version échantillonnée). Ces textes, sous licence libre, seront téléchargeables en version annotée ou non. Ils pourront être explorés (*via* des fonctionnalités documentaires (concordance, index, …) et statistiques (spécificités, AFC, …)) au moyen de la version en ligne de la plateforme PrimeStat.

1.4.1.2. **Corpus contrôlé** (= Presto$_{-CONTROLE}$)

Le corpus « contrôlé », qui inclut le corpus noyau, est le niveau sur lequel a porté l'essentiel de nos efforts en matière de construction « raisonnée » du corpus (choix de la population cible, mise œuvre du critère de comparabilité, …). Il réunit actuellement 162 textes pour un nombre total de 11.636.573 mots (version intégrale) / 5.358.382 mots (version échantillonnée). Ces textes se répartissent par décennie comme suit :

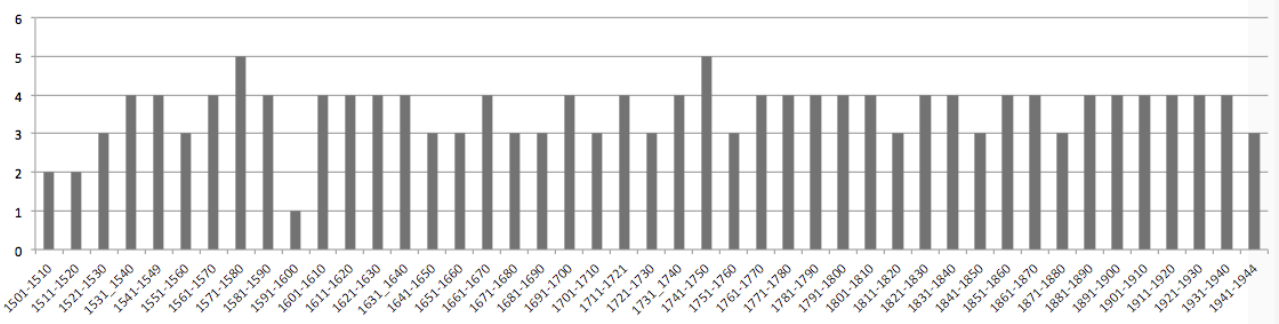


Diagramme 1
Répartition des textes par décennies dans le corpus Presto, niveau contrôlé (Presto–^{CONTROLE})

Cette répartition est relativement régulière ; les irrégularités les plus nombreuses figurent - comme on le devine aisément - dans la période la plus ancienne : 1501-1600.

1.4.1.3. **Corpus étendu** (= Presto–^{ETENDU})

Ce niveau du corpus réunit 315 textes pour un nombre total de 28.309.240 mots (version intégrale) / 11.002.199 mots (version échantillonnée). Y ont été agrégés selon les opportunités i) des textes relevant d'autres genres discursifs ii) des textes dont le statut juridique était indécis, l'ensemble permettant d'étoffer le corpus pour permettre des études plus précises sur certaines occurrences (mots, lemmes… motifs) dont on peut penser qu'elles sont peu nombreuses.

Voici la répartition des textes qu'on y observe pour la période 1501-1944 :

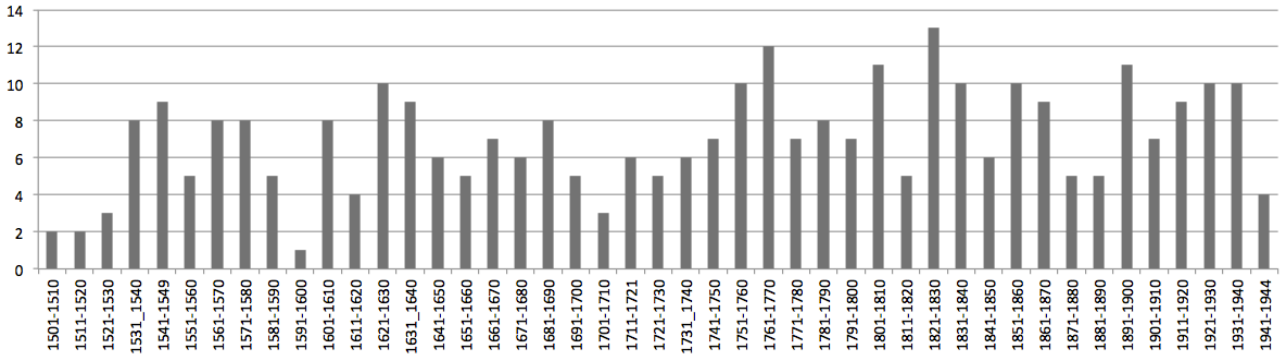


Diagramme 2
Répartition des textes par décennies dans le corpus Presto, niveau étendu (Presto–^{ETENDU})

Le diagramme suivant permet de comparer la répartition des 162 textes de Presto–^{CONTROLE} et celle des 315 textes de Presto–^{ETENDU}.

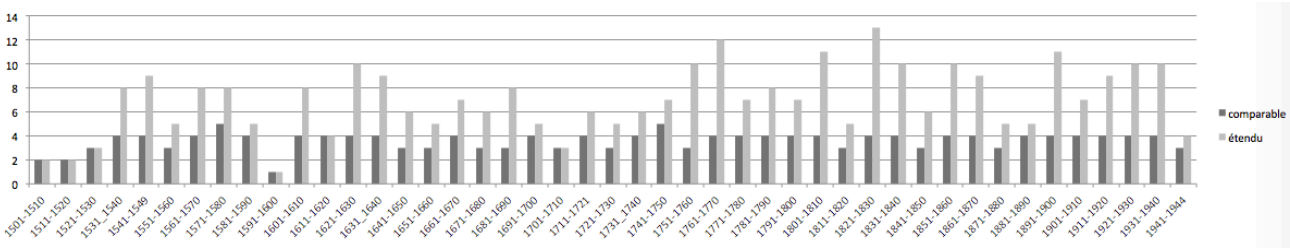


Diagramme 3
Répartition des textes par décennies dans les niveaux contrôlé (Presto–^{CONTROLE}) et étendu (Presto–^{ETENDU})

On observe à la fois le gain en nombre de textes (et de mots) que procure le niveau étendu du corpus, mais en même temps la perte en termes de comparabilité entre les tranches décennales notamment en taille de textes (et de mots).

Remarque :

Dans la suite de ce mémoire, nous désignerons par « corpus Presto » le niveau contrôlé dans sa version échantillonnée. La référence à d'autres niveaux ou versions spécifiques du corpus sera toujours stipulée (Presto–$^{ETENDU\_ECHANTILLONNE}$, Presto–$^{NOYAU\_INTEGRAL}$, etc.). Enfin, lorsque nous voudrons désigner de manière indifférenciée tous les niveaux corpus Presto (par ex. dans la section consacrée à la tokenisation et à l'annotation), nous parlerons du « corpus intégral Presto ».

1.4.1.4. **Corpus spécialisés** (= Presto–^{SPECIALISE})

Nous avons enfin constitué deux corpus spécialisés pour des études spécifiques conduites en vue notamment de la publication collective dans la revue *Langages* (206) et de l'ouvrage collectif à paraître chez P. Lang d'autre part.

Le premier corpus (Presto–^{FIGARO}) réunit 1277 numéros du quotidien *Le Figaro*, dont 964 sont datés du XIX^{e} siècle (années 1885, 1890, 1895, 1896) et 313 du XXI^{e} siècle (année 2002) :

Sous-corpus	**Années**	**Total numéros**	**Total mots**
Le Figaro XIX^{ème} siècle	1885, 1890, 1895, 1896	964	37.446.988
Le Figaro XXI^{e} siècle	2002	313	30.497.189
TOTAUX		1277	67.944.177

Tableau 1
Corpus spécialisé Presto–^{FIGARO}

Le second corpus (Presto–^{ENCYCLOPEDIE}) réunit les versions électroniques des trois encyclopédies suivantes (soit 121.956.209 mots):

- L'*Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers* dir. Diderot et d'Alembert (1751-1772 ; 1776-1780 pour les suppléments de C.-J. Panckoucke) à laquelle

nous avons accès grâce à notre collaboration avec l'ARTFL et son directeur R. Morrissey; [*in extenso*, soit 21.700.000 mots].
- L'*Encyclopædia Universalis* (édition numérique de 2005) ; [*in extenso*, soit 49.859.864 mots].
- *Wikipédia* (2001-), encyclopédie en ligne collaborative [version partielle constituée dans le cadre de Presto ; sélection aléatoire d'un article sur onze ; soit 50.396.345 mots].

Le corpus Presto–ENCYCLOPEDIE a aussi été intégré dans un projet de recherche déposé auprès du FNS par le Pr. C. Rossari de l'Université de Neuchâtel (« *Le positionnement énonciatif et ses variations dans le discours encyclopédique entre les XVIIIe et XXIe siècles* ») dans lequel plusieurs membres de Presto figurent comme collaborateurs scientifiques. Nous reviendrons sur ce projet dans la partie finale de ce mémoire intitulée « perspectives ».

Le tableau ci-dessous récapitule les différents niveaux et versions du corpus. Les cellules grisées permettent d'identifier le corpus que nous avons le plus utilisé et le plus travaillé, et que nous nommons par défaut « le corpus Presto ».

Niveaux / Versions	Noyau	Contrôlé	Étendu	spécialisés
textes échantillonnés	53 textes | 1.924.532 mots	162 textes | 5.358.382mots	315 textes | 11.002.199 mots	Presto–FIGARO : 1277 numéros, 67.944.177 mots Presto–ENCYCLOPEDIE : 3 œuvres, 121.956.209 mots
textes intégraux	53 textes | 6.820.161 mots	162 textes | 11.636.573 mots	315 textes | 28.309.240 mots	

Tableau 2
Récapitulatif des niveaux et des versions du corpus intégral Presto.

Rem : les œuvres du corpus « spécialisé » ne sont pas répertoriées dans l'ANNEXE 4.

1.4.2. Les « descripteurs » dans le corpus Presto

Il est impératif de documenter les corpus. Concernant Presto, outre les informations déjà délivrées, voici quelques informations relatives au jeu des métadonnées attaché aux textes du corpus et qu'on trouvera détaillées dans l'ANNEXE 1.

Ces métadonnées sont pour l'instant structurées autour de deux grands niveaux hiérarchiques : l'œuvre et l'exemplaire. *L'exemplaire* correspond à la réalisation matérielle d'une œuvre de l'esprit, *via* une édition ou un manuscrit, en l'occurrence celle utilisée dans le corpus.

Relèvent de *l'œuvre* les informations relatives à *l'auteur*, et de *l'exemplaire*, celles relatives à *l'éditeur scientifique*. Ce schéma minimal pourra être complexifié par la suite pour permettre la prise en compte du paratexte (préface, postface, etc.).

A ces quatre entités sont associées un ensemble de métadonnées que nous avons séparées entre métadonnées « minimales », qui doivent obligatoirement être recherchées et vérifiées pour tous les textes du corpus (et donc renseignées lorsqu'elles existent), et les métadonnées « maximales », qui ne seront pas obligatoirement renseignées et/ou vérifiées.

1.4.3. Population cible, stratification, échantillonnages et tailles du corpus

Dans cette section, on restreindra le propos au **seul corpus « Presto »** (= **niveau contrôlé, version échantillonnée**) dans la mesure où c'est sur lui qu'a porté l'essentiel de nos efforts en matière de construction du corpus.

1.4.3.1. Population

Population visée : au vu des contraintes que nous rappelons plus bas, nous avons choisi de définir la population du corpus suivant un critère de genres discursifs, plus précisément de « champ générique ». Ainsi avons-nous échantillonné pour la période 1501-1944 :

- **trois « champs génériques** » (au sens de F. Rastier 2011) relevant du *discours littéraire* (*ibid.*), à savoir *i)* les *genres narratifs* (romans, nouvelles, contes, …), *ii)* la poésie et *iii*) le théâtre.

- **un « champ générique** » (étiqueté comme tel par Frantext mais dont il reste à éprouver le bien-fondé), celui des « traités », qui s'avère « trans-dicours » puisqu'il peut s'agir de traités relevant des discours religieux, historiques, philosophiques, …

Le paramètre majeur qui a guidé le choix d'une telle population a été celui de la comparabilité entre tranches temporelles de dix ans, empan temporel que nous sommes fixé pour la structure temporelle interne de notre corpus.

Définition « opérationnelle » de la population (voir Biber 1993a & *supra*) : le projet soumis à l'ANR et la DFG déclarait sa volonté de coopérer avec des bases textuelles existantes. Ce furent Frantext (http://www.frantext.fr, V. Montémont, G. Souvay), les BVH (*Bibliothèques Virtuelles Humanistes*, http://www.bvh.univ-tours.fr - L. Bertrand, M.-L. Demonet), l'ARTFL (*American and French Research on the Treasury of the French Language*, http://artfl-project.uchicago.edu – R. Morrissey, M. Olsen) et plus marginalement le CEPM (*Corpus électronique de la première modernité*, http://www.cpem.paris-sorbonne.fr). La liste des textes sur laquelle nous avons opéré nos échantillonnages a été la liste agrégée des textes mis à disposition par ces différentes bases.

Cette liste a cependant été filtrée avant échantillonnage selon les critères suivants :

- utiliser des éditions de textes soit entièrement libres de droits (domaine public), soit sous licence libre (voir par ex. pour les BVH), soit pour lesquelles les éditeurs avaient disparu ;

- ne retenir que les premières éditions des oeuvres ou, à défaut, celles respectant le plus l'orthographe d'époque.

La liste des textes restants a constitué ce que Biber appelle le cadre d'échantillonnage (« *sampling frame* »).

1.4.3.2. Stratification

Nous avons adopté telles quelles les catégories génériques qui nous ont été communiquées par les bases citées ci-dessus. Autrement dit, aucun travail de définition des (sous-)genres n'a été conduit. Un tel travail - de première importance - demeure donc à faire, en bénéficiant notamment des réflexions menées par le groupe « typologie textuelle» dans le consortium CAHIER (https://cahier.hypotheses.org/groupe-typologie-textuelle).

1.4.3.3. Taille du corpus

Le nombre de mots pour chaque texte, le nombre de mots et de textes pour chaque champ générique, le nombre de mots total du corpus ont été fixés sans que nous ayons les moyens ni le temps de construire des procédures d'optimisation telle que celles conçues par exemple par D. Biber.

Pour nous donner un cap, nous sommes adossés à l'objectif de recherche linguistique déclaré dans Presto, à savoir l'étude des prépositions en diachronie. Il s'agit là d'une des catégories morphosyntaxiques les plus fréquentes dans tous les siècles de notre corpus (elle figure parmi les *common linguistic features*) et il paraît raisonnable de faire l'hypothèse que, comme en anglais contemporain (voir Biber *supra*), elles possèdent une *distribution linéaire stable* dans les textes et que, si leur distribution varie certainement suivant les genres, cette variation est moindre que pour des traits linguistiques rares.

Nous avons indiqué *supra* (§ 1.3.2.4.) les choix de taille en nombre de mots que nous avons faits pour l'échantillonnage des textes. Les deux diagrammes ci-dessous permettent de visualiser successivement i) le nombre de mots par tranche décennale dans le corpus, ii) le nombre de mots par siècles. Rappelons que le diagramme 1 *supra* présente le nombre de textes par tranche décennale.

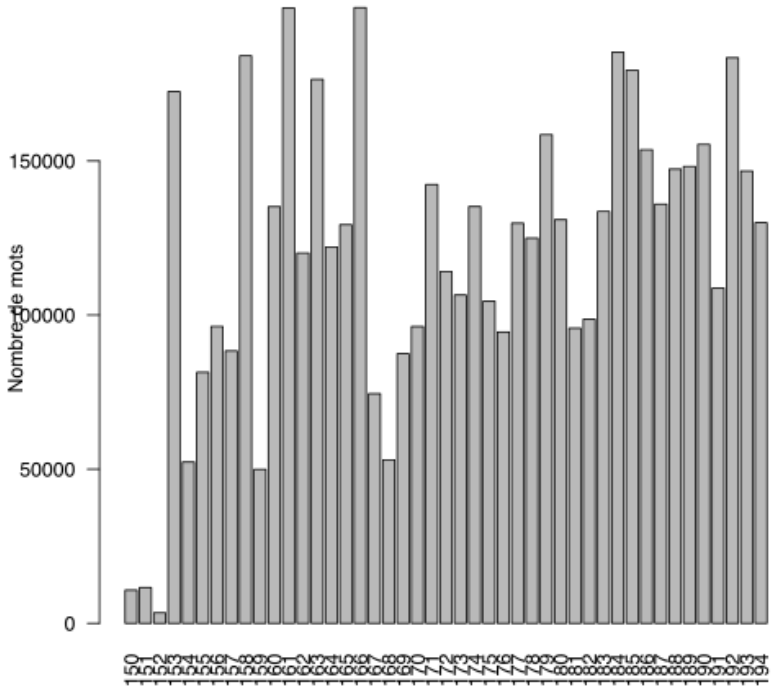


Diagramme 4.
Nombre de mots par tranche décennale dans le corpus Presto (1501-1944)

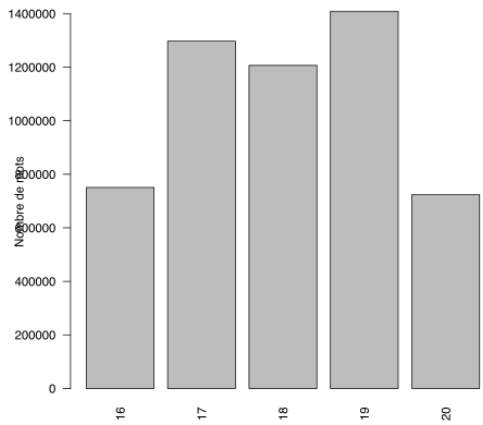


Diagramme 5.
Nombre de mots par siècle dans le corpus Presto

Il est aisé d'observer dans le diagramme 4 une faiblesse particulièrement sensible du nombre de mots pour les trois premières tranches décennales. Le diagramme 5 met quant à lui en lumière une disparité du XVI^e^ et du XX^e^ s. (toujours en termes de taille de mots) vis-à-vis des trois autres siècles. Il y a deux raisons distinctes à cette situation.

La taille du corpus XX^e^ s. est d'environ la moitié de celle du XIX^e^ s. pour une raison simple : il ne réunit que cinq décennies. Autrement dit, les principes d'échantillonnage ont été respectés pour ce corpus qui possède la même taille que les autres demi-siècles (excepté le XVI^e^ s.).

Concernant le XVI^e^ s. et notamment les trois premières décennies, le critère de comparabilité et les autres contraintes que nous nous étions fixées (droits juridiques, qualité philologique des textes) nous ont conduit à disposer d'une liste extrêmement réduite de textes candidats rendant impossible tout échantillonnage. La solution pour améliorer cette partie du corpus passera par trois voies qu'il conviendra de conjuguer : i) numériser de nouveaux textes, ii) éclaircir avec la base Frantext le statut juridique « flou » de certains textes dont on peut raisonnablement penser qu'ils ne sont plus sous droits et qu'ils pourraient être versés dans notre corpus, iii) rechercher dans Wikisource des textes numérisés et relus au moins par deux personnes et dont la qualité philologique s'approche au mieux de nos exigences.

Pour ce qui regarde la numérisation de textes, signalons que avons fait numériser par l'ATILF, dans le cadre de notre convention avec ce laboratoire, deux textes.

- Les *Œuvres complètes* d'Ambroise Paré. L'objectif consiste à disposer de la version du texte tel qu'il figurait dans l'édition originale de 1585.
- Le *Grand Dictionnaire Universel du XIX^e^ siècle* a lui aussi fait l'objet d'un travail de numérisation – cette fois, partielle. Il s'agissait d'étoffer le nombre d'articles prélevés dans cette œuvre monumentale comportant environ 75 millions de mots[141].

1.5. Tentative d'évaluation de la qualité actuelle du corpus

Pour évaluer la qualité actuelle du corpus Presto (niveau contrôlé, version échantillonnée), nous nous donnons les instruments de mesure suivants :

[141] Comme il n'existe pas encore de version numérique complète de cette œuvre, l'estimation de ce chiffre nous revient.

⇨ **Mesure de la CONTINUITE du corpus**

[C1] Continuité « temporelle » [142] : pour chaque décennie, dispose-t-on d'au moins
C1a : 1 œuvre (dont la date d'édition originale est incluse dans cette décennie) ?
C1b : 2 œuvres ?
C1c : 3 œuvres ?
C1d : 4 œuvres ?

[C2] Continuité « générique » : pour chaque décennie, dispose-t-on d'au moins
C2a : 1 champ générique commun avec toutes les autres décennies ?
C2b : 2 champs génériques communs avec … ?
C2c : 3 champs génériques communs avec … ?
C2d : 4 champs génériques communs avec… ?

⇨ **Mesure de la VARIÉTÉ du corpus**

[**C3] Variété des *genres discursifs*** : pour chaque décennie, dispose-t-on d'au moins
C3a : 1 champ générique ?
C3b : 2 champs génériques ?
C3c : 3 champs génériques ?
C3d : 4 champs génériques ?

[C4] Variété des *auteurs* : pour chaque décennie, dispose-t-on d'au moins
C4a : 2 auteurs distincts (entre eux) dans la décennie et distincts des auteurs présents dans la décennie précédente ?
C4b : 3 auteurs distincts … ?
C4c : 4 auteurs distincts … ?

Le diagramme suivant synthétise le degré (en %) de réussite[143] de chacun de ces critères :

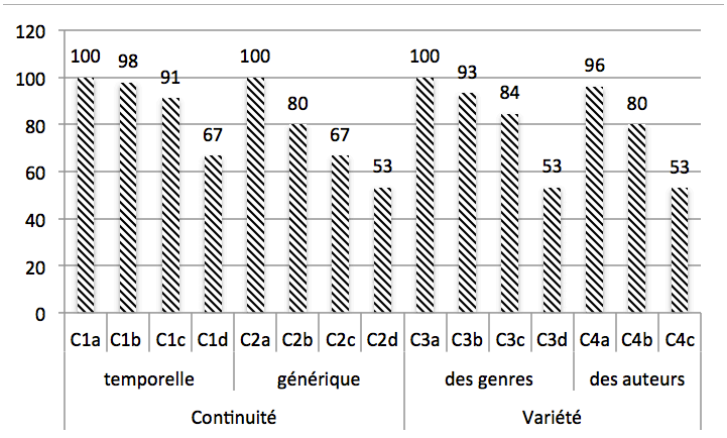


Diagramme 6
Pourcentages de validation des critères C1, C2, C3, C4 dans le corpus Presto

[142] Ce que nous appelons « continuité temporelle » correspond à ce que A. Condamines, J. Rebeyrolles & A. Soubeille (2004 : 548) nomme « diachronicité » d'un corpus historique : « Les textes qui le [= le corpus] composent devront nécessairement s'échelonner dans le temps afin de rendre possible l'observation de continuités, de ruptures et/ou d'évolutions des connaissances ».
[143] Au sens où il a été pleinement / partiellement / pas du tout atteint.

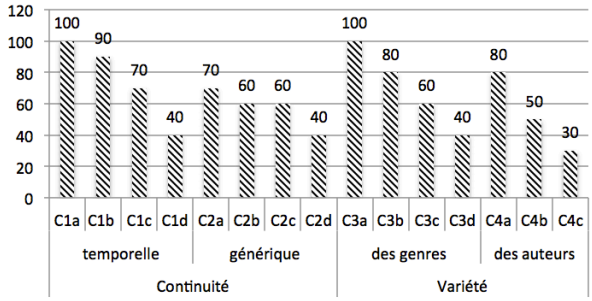
A la lecture de ce diagramme, on peut grosso-modo considérer que Presto réalise de manière satisfaisante (% ≥ 80) le « profil » suivant :

- ***Continuité*** :
  - temporelle : chaque décennie possède (presque toujours) au moins trois textes
  - générique : chaque décennie partage (presque toujours) au moins deux champs génériques avec les autres décennies
- ***Variété*** :
  - des genres : chaque décennie possède (presque toujours) au moins trois champs génériques
  - des auteurs : chaque décennie possède (presque toujours) au moins trois auteurs distincts au sein de la décade et distincts aussi par rapport à la décennie qui précède.

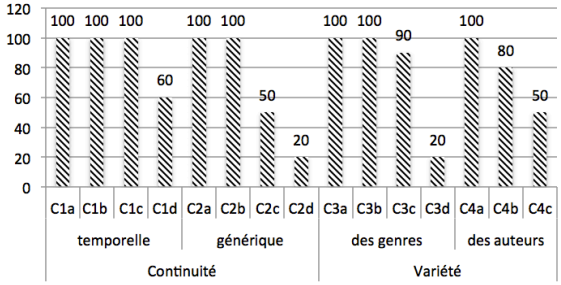
Quelles améliorations à apporter au corpus Presto dans un avenir proche ?

Il est utile, avant toute prescription, de mieux cerner pour chacun des siècles du corpus la « réussite » des critères énoncés grâce aux diagrammes suivants :

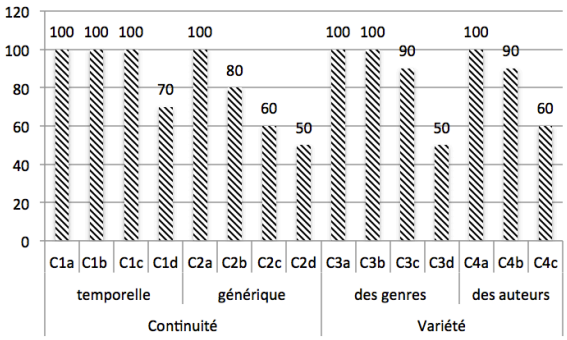
XVI^{e} s.



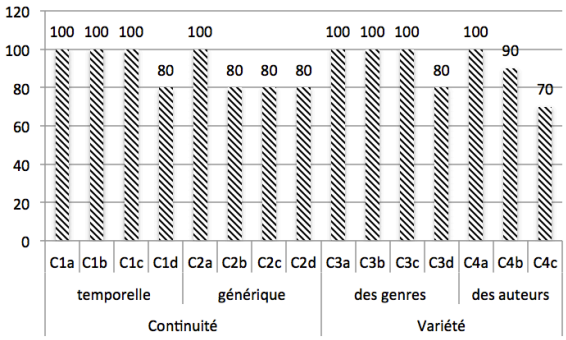
XVII^{e} s.



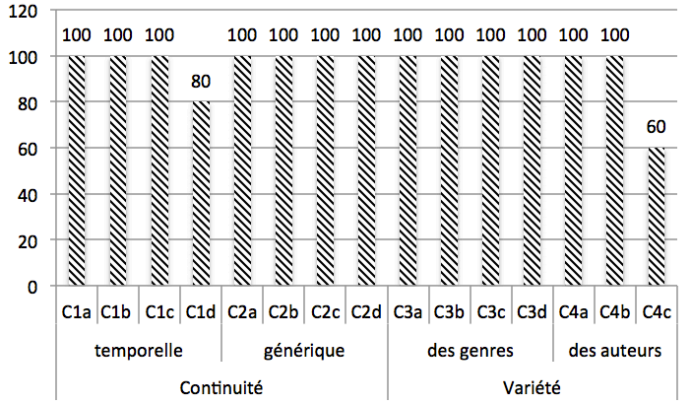
XVIII^{e} s.



XIX^{e} s.



XX^e s. (1901-1944)



Diagrammes 7 à 11
Pourcentages de validation des critères C1, C2, C3, C4 dans le corpus Presto respectivement pour les XVI^e s. (D7), XVII^e s. (D8), XVIII^e s. (D9), XIX^e s. (D10), XX^e s. (D11).

Il apparaît nettement (et sans grande surprise) que le sous-corpus du XVI^e s. vérifie le moins bien (< 80%) la plupart des critères (1/3 seulement ≥ 80 %). Plus on s'approche du XX^e s. plus les critères sont vérifiés.

Moins attendue peut-être est la mauvaise performance du sous-corpus XVII^e s. pour les critères C1d (quatre œuvres distinctes dans chaque tranche), C2c[144] (au moins 3 champs génériques communs avec toutes les autres tranches), C3d (au moins 4 champs génériques distincts) et C4c (4 auteurs distincts à la fois dans la décennie, et distincts des auteurs réunis dans la décennie qui précède).

L'amélioration de la qualité de notre corpus passera par les mêmes voies que celles décrites *supra* pour la tranche du XVI^e s : i) numérisation de nouveaux textes, ii) collaboration avec Frantext iii) recherches dans Wikisource.

1.6. Conclusion

« Il y a loin de la coupe aux lèvres » … Durant la constitution du corpus intégral Presto (i.e. envisagé dans ses différentes versions) qui nous a occupés environ deux années, nous avons souvent éprouvé la vérité de ce proverbe en mesurant la distance qui séparait le corpus « construit » du corpus « rêvé ». Et cela d'autant que nous nous étions engagés, lors du dépôt de projet auprès des organismes financeurs, à nous appuyer uniquement sur les bases textuelles existantes, ce qui nous ôtait à priori toute possibilité (sinon à la marge) de numériser des textes.

Pour autant, le travail accompli et le corpus auquel nous sommes parvenus nous paraissent dignes de considération. Ils nous ont permis d'aboutir à des résultats dont les publications récentes (en part. *Langages*, n° 206) illustrent, nous l'espérons, l'intérêt scientifique.

Nous allons maintenant nous tourner vers la chaîne des traitements appliqués à ce corpus. On examinera successivement sa tokenisation et son annotation (étiquetage morphosyntaxique, lemmatisation), avant de donner quelques informations sur son balisage TEI (*Text Encoding Initiative*).

[144] Et *a fortiori* C2d

2. Annoter et baliser le corpus intégral Presto[145]

G. Leech (1997 : 2) définit le processus d'annotation comme suit : « *adding such interpretative, linguistic information to an electronic corpus of spoken and/or written language data* ». Les informations linguistiques ajoutées par l'annotation relèvent de l'interprétation en ceci que « *annotation is, at least in some degree, the product of the human mind's understanding of the text* ». (*op. cit*.)

On est donc fondé, à la suite de T. McEnery, R. Xia & Y. Tono (2006 : 29), à distinguer entre *corpus annotation* (« *used in a narrow sense* ») et *corpus mark-up* – opérations que nous traduirons respectivement par *annotation de corpus* et *balisage de corpus*. Cette distinction est construite sur l'opposition entre le caractère *interprétatif versus objectif* de l'information délivrée. « *Corpus mark-up provides relatively objectively verifiable information regarding the components of a corpus and the textual structure of each text. In contrast, corpus annotation is concerned with interpretative linguistic information.* » (*op. cit*.). De fait, dans le cadre d'une annotation morphosyntaxique, affecter par ex. au mot *combien* le statut d'adverbe dans une interrogative partielle directe est une décision qui engage une interprétation d'ordre catégoriel elle-même adossée à un arrière-plan théorique (typologie des parties du discours) qui peut être contestable : ne vaudrait-il pas mieux ranger ce mot dans la catégorie des pronoms ? (voir par ex. M. Riegel, J. C. Pellat, R. Rioul, 2009 : 385, 649) ; ou dans la catégorie plus vaste des proformes ? A l'inverse, la plupart des informations relatives à l'auteur, au titre de l'œuvre, à sa forme (prose / vers) etc. apparaissent plus « objectives ». Ainsi les « descripteurs » évoqués *supra* de même que le contenu des en-têtes TEI relèvent-ils pour une grande part du *balisage*.

D'autres linguistes adoptent une définition plus « accueillante » de l'annotation, comme K. Fort par ex. (2012 :17) : « L'annotation recouvre à la fois le processus consistant à apposer (*ad*-) une note sur un support, l'ensemble des notes ou chaque note particulière qui en résulte et ce, sans préjuger a priori de la nature du support considéré (texte, vidéo, images, etc.), du contenu sémantique de la note (note chiffrée, valeur choisie dans un référentiel fermé ou texte libre), de son positionnement global ou local, ni de son objectif (visée évaluative ou caractérisante, simple commentaire discursif) ».

2.1. La tokenisation dans Presto

La tokenisation – qu'on peut considérer comme une segmentation réservée aux unités que sont en particulier les mots simples et les unités polylexicales[146] – consiste à isoler des unités – ou tokens – au sein du continuum textuel. Par token, on entend toute séquence de caractères (séquence éventuellement réduite à un singleton : *à, y, ...*) isolée par deux « séparateurs » (blanc, ponctuation, tiret, …) du reste du texte. Un token n'est pas nécessairement un « mot » tel qu'on l'entend habituellement. Ainsi, par ex., une ponctuation a dans Presto le statut de « token ».

Nous avons donc usé jusqu'ici d'une facilité de langage en parlant du nombre de *mots*. Désormais, nous parlerons de *tokens*[147].

---

[145] Je remercie vivement Achille Falaise, Ingénieur de Recherche sur le programma Presto à l'ENS de Lyon, pour sa relecture des sections 1 et 2 de cette deuxième partie et l'ensemble de ses explications, conseils et remarques relatifs à la chaîne de traitement mise en œuvre dans Presto (dont il a été l'un des grands architectes).

[146] Tel est le cas par ex. de B. Habert, A. Nazarenko & A. Salem (*op. cit.*: 166) : « La segmentation consiste à découper une suite de caractères en « unités » : mots simples ou unités polylexicales. » La segmentation du flux textuel en phrases est une opération différente que nous n'aborderons pas ici.

[147] Certes, un token peut être un « mot » au sens non technique. Mais pas nécessairement : les ponctuations, par ex., sont des tokens dans (tous les niveaux du corpus) Presto.

L'opération automatisée que constitue la tokenisation d'un corpus n'est en rien triviale. Comme le fait observer B. Habert (2005 : 13), « certains caractères sont tantôt séparateurs de mots ou de phrases tantôt composants de mots (l'apostrophe dans *l'ami vs. aujourd'hui* ; la virgule dans la phrase *vs.* dans les nombres ; le point entre les phrases *vs.* au sein d'abréviations ou de nombres notés à l'anglaise, etc.). » La question des unités polylexicales (*carte bleue, timbre poste, ...*) représente une autre difficulté : « jusqu'à un cinquième de la surface d'un texte peut relever des mots en plusieurs mots » (*op. cit.*). Il existe pour le français contemporain un dictionnaire électronique de « mots composés » réalisé au LADL (voir B. Habert, A. Nazarenko &A. Salem, *op. cit.* : 167). B. Habert (*op. cit.* : 13) signale en outre que les étiqueteurs INTEX et UNITEX intégrent ces inventaires. Mais dans la mesure où nous ne disposons pas actuellement de dictionnaires des « mots composés » pour des états anciens de la langue, décision a été prise dans Presto de ne pas chercher à repérer ces mots *a priori* dans le corpus – d'autant plus qu'un de nos objectifs consistait à déterminer à partir de quelle période une séquence linguistique donnée devenait figée.

Le processus de tokenisation dans Presto s'est déroulé en trois étapes successives, la sortie de l'étape n-1 constituant l'entrée de l'étape n.

Étape 1. Il a d'abord été procédé à une tokenisation « maximale » du corpus en ceci que les caractères séparateurs utilisés[148] ont été uniquement traités comme tels, sans considération pour leur rôle de « composant » possible. Autrement dit, *carte bleue* a été segmenté en deux tokens, *aujourd'hui* en trois (aujourd | ' | hui), etc. Une des difficultés rencontrée à la sortie de cette première étape est à mettre en rapport avec la grande variété des graphies au XVI^{e} s., certains mots pouvant contenir dans leur graphie un voire plusieurs caractères séparateurs. Le lemme <NAGUÈRE>[149] est un bon exemple, qui possède dans Presto les graphies suivantes: *naguère, naguére, naguiere, naguieres, nagyères, n'aguere, n'agueres, n'aguères, n'aguiere, n'aguieres, n'a gueres*[150]. À l'issue de l'étape 1 de la tokenisation, ces formes ont donc été segmentées en un (*naguére*, …), trois (*n'aguere*, …) voire quatre tokens (*n*|*'*|*a*|*gueres*) alors qu'elles doivent être rattachées à un seul lemme. Il nous a donc fallu effectuer un contrôle manuel de la tokénisation et en particulier de la fusion des tokens. Ce contrôle, illustré *infra* (§ 2.2.1), a été introduit en phase finale après l'annotation automatique par *TreeTagger* (H. Schmid, 1994).

Étape 2. Une nouvelle segmentation à partir de règles spécifiques a ensuite été appliquée. Celles-ci mettaient en jeu des listes de mots établies par l'ingénieur du projet (A. Falaise) à partir du principe de segmentation ainsi défini :

Une unité graphique est segmentée en plusieurs unités ssi cette unité se présente *uniquement* sous forme segmentée en français moderne

Exemple :

La forme *moymesme* courante au XVII^{e} s. a été segmentée. De même, les formes adjectivales et adverbiales, très courantes au XVI^{e} s., où l'adverbe *tres* est accolé à l'adjectif (*trecher, tresdoux, tresrude, ...*) ou à l'adverbe (*tresdoucement, ...*) qu'il modifie ont été segmentées.

[148] Dont voici la liste : ,;:.?!'"-«»()[]
[149] Conventionnellement, nous plaçons entre chevrons les métadonnées relatives à la *POS* (*part of speech*/ partie du discours) et au *lemme*, ce dernier étant par ailleurs graphié en petites capitales.
[150] « il m'escripvit n'a gueres qu'il estoit devenu patays » (1542, F. Rabelais, *Gargantua*)

(1) ***trescher*** *amy Antire, Pour te vouloir mon intention dire, J'ay desormais deliberé* (…) (M. Scève, *Saulsaye*, 1547)
(2) ***tresdouce*** *richesse ! Heureux repos eslongné de tristesse, Qui en Yver, Printemps, Automne (...)*(M. Scève, *Saulsaye*, 1547)
(3) *(...) dont le desir* ***tresdoucement*** *me mord.* (M. Scève, *Saulsaye*, 1547)

Rem : Les formes *ledit, notredit, votredit* etc. déterminant composé en ancien français ont été considérées comme appartenant au paradigme du déterminant composé défini « ledit » (voir jeu d'étiquettes Presto, *infra*).

Étape 3. Enfin, une étape de fusion automatique de tokens a été appliquée, s'appuyant sur l'algorithme dit "*longest matching*" ou "*longest match*" [151] (N. Y. Liang, 1986) et sur le lexique Presto (présenté *infra)*. Cet algorithme a ainsi permis par ex. de réunir les trois tokens obtenus à l'issue de l'étape 1 pour le mot « aujourd'hui » car le lexique Presto possédait la forme « aujourd'hui » (de même que la forme « hui » mais la règle qu'on peut traduire par « la plus longue chaîne d'abord » a conduit l'outil à sélectionner préférentiellement la forme « aujourd'hui »). En revanche, comme notre lexique ne contenait pas les formes *n'aguere, n'agueres, n'aguères, n'aguiere, n'aguieres, n'a gueres*, celles-ci n'ont pas été reconnues et ont donc dû subir un post-traitement (voir phase de contrôle manuel *infra*).

2.2. Annotation morphosyntaxique et lemmatisation dans Presto

On présente d'abord un tableau général des grandes étapes du processus d'annotation suivies dans Presto puis on évoque le jeu d'étiquettes utilisé.

2.2.1. Étapes du processus d'annotation

Campagne d'annotation manuelle par des experts portant sur la période du XVI^e^ s. au XVIII^e^ s. (période pour laquelle il n'existait pas avant le programme Presto de modèle de langage, du fait du manque de ressources (corpus d'apprentissage en particulier)). A l'issue du processus de tokenisation, une phase d'annotation manuelle d'une fraction du corpus Presto (sous-corpus dit « *gold standard* » : 5 textes échantillonnés, 62.000 tokens) a été accomplie en juillet-août 2014 par trois annotateurs experts distincts. Une partie des tokens avait préalablement été pré-annotée automatiquement par projection du lexique Presto sur le corpus, suivie d'une étape de désambiguïsation automatique, cela afin de faciliter le travail des annotateurs. L'environnement informatique choisi pour cette annotation manuelle a été le logiciel Analog conçu et développé par M.-H. Lay (Université de Poitiers) qui a participé à cette campagne. Cet outil permet d'éditer les corpus annotés (M.-H. Lay & B. Pincemin, 2010). Toutes les occurrences d'une erreur détectée sont localisées à l'aide d'un concordancier. La modification (correction) effectuée sur le résultat de la concordance porte en un seul temps sur toutes les occurrences similaires.

Entraînement de l'outil *TreeTagger*. A l'issue de cette campagne d'annotation, une fraction (80%) du sous-corpus « *gold standard* » ainsi que le lexique Presto construit parallèlement (voir *infra*) ont permis d'entraîner l'outil *TreeTagger* qui, à partir du « modèle de langage » qu'il a lui-même construit, a pu annoter automatiquement (étiquetage morphosyntaxique et lemmatisation) l'intégralité du corpus.

[151] Liang, N. Y. (1986) « On computer automatic word segmentation of written Chinese », *Journal of Chinese Information Processing* 1 (1).

Phase de contrôles manuels. A l'issue de cette étape, plusieurs post-traitements ont été appliqués : certains avaient trait à la **fusion** de tokens, d'autres à leur **segmentation.** D'autres enfin visaient à corriger des erreurs diverses (erreurs de segmentation, lemmes erronés, …).

Nous allons nous arrêter sur les étapes de fusion et de segmentation.

Contrôle manuel pour la fusion des tokens : nous nous sommes adossés aux trois principes de décision suivants :

- **Principe 1** : Une séquence de 2 unités graphiques (UG) est fusionnée ssi une au moins des unités qui la forme ne peut être traitée comme une unité linguistique car on ne dispose pas d'étiquette morphosyntaxique possible pour elle, quel que soit l'état de langue considéré. La recherche d'une telle étiquette doit être conduite manuellement en s'appuyant sur les dictionnaires de langue suivants : *Tobler-Lommatzsch*, DMF, Dict. Huguet, Dict. de l'Académie (1ère éd. 1694, 4ème éd. 1762)**,** TLFi. Cette fusion doit être réitérée jusqu'à ce qu'on aboutisse à une unité *linguistique* dotée d'une étiquette.

Pour appliquer ce principe, on recherche d'abord dans le corpus tokenisé les unités qui ne disposent pas, après projection du lexique *Presto* (voir *infra*), d'étiquette morphosyntaxique. On applique alors manuellement les principes de décision. Les tokens identifiés comme devant être fusionnés sont ensuite intégrés dans une liste pour un traitement automatique sur tout le corpus.

Voici deux exemples.

La segmentation de *parce que* en deux tokens aboutit à l'UG *parce* à laquelle on ne peut assigner aucune étiquette morphosyntaxique. Diachroniquement, cette unité résulte d'une fusion graphique de la préposition *par* et du pronom *ce*. Ici, deux positions seraient théoriquement possibles : on segmente *parce* en deux tokens [lemmes : <PAR> <CE><QUE> [152]] ou on fusionne *parce que* en un seul token. Il a été choisi de fusionner.

Autre exemple : il a été vu *supra* que les graphies normalement rattachables au lemme <NAGUÈRE> (*naguère, naguére, naguiere, naguieres, nagyères, n'aguere, n'agueres, n'aguères, n'aguiere, n'aguieres, n'a gueres*) ont été, à l'issue de l'étape 1 de la tokenisation, segmentées en un, trois ou quatre tokens.

Soient d'abord les graphies segmentées en trois tokens. Les unités graphiques « *aguere* | *agueres* | *aguères* | *aguiere* | *aguieres* » ne se voient attribuer aucune étiquette morphosyntaxique dans les dictionnaires cités. La fusion opérée avec l'apostrophe qui précède n'aboutit pas à une UG étiquetable. Elle est donc réitérée jusqu'à aboutir aux formes graphiques *n'aguere, n'agueres, n'aguères, n'aguiere, n'aguieres* dont la consultation de dictionnaires attestent l'existence[153].

Enfin, dans le cas de la séquence graphique *n'a gueres*, seule la fusion entre les tokens initiaux < *n* > et < *'* > est opérée comme partout ailleurs dans le corpus (variante graphique de *ne*). Dès lors, les tokens < *n'* >< a >< *gueres* > reçoivent tous une étiquette morphosyntaxique après projection du lexique Presto : <Rp><Vuc><Rg>. Aucune fusion n'a donc été opérée.

[152] Option adoptée par la BFM.

[153] Concrètement, on interroge la base du *Grand Corpus des dictionnaires* (*Classiques Garnier Numérique*) qui réunit les 24 dictionnaires les plus importants consacrés à la langue française, soit près de 200 000 pages. La liste des dictionnaires est disponible en ligne dans les « Principes d'édition » de ce grand corpus.

Doit-on s'alarmer du fait que la graphie *n'a gueres* n'ait ainsi pas été rattachée au lemme <NAGUÈRE> ? Nous nous en réjouissons au contraire. En effet, si un linguiste se saisit de l'idée d'étudier dans Presto le processus de figement (aboutissant à une agrégation des constituants initialement distincts) dont a été le siège cet adverbe, il retrouvera nécessairement sur sa route (c'est-à-dire dans sa concordance, pourvu qu'il examine – mais comment pourrait-il en être autrement ? – la séquence de lemmes <NE><AVOIR><GUÈRE>) la graphie en question (*n'a gueres*). Ce sera alors à lui de se prononcer sur le caractère figé de la séquence : nous ne l'aurons pas fait à sa place.

- **Principe 2** : On fusionne les séquences d'unités graphiques formant un nom propre dont l'orthographe moderne présente toujours une seule unité graphique.

Par exemple, *Ménil-montant* pour *Ménilmontant*

- **Principe 3** : Les séquences *de la, de l'*, dans les emplois où nous les avons analysés comme des variantes de l'article partitif DU actualisant des noms massif, ont été fusionnées.

Contrôle manuel pour la segmentation des tokens : à ce stade, le contrôle a porté, outre sur les erreurs de segmentation, sur les **amalgames.** Pour la période 1501-1944 couverte par le corpus, les seuls amalgames rencontrés concernaient les cas de préposition (*de, à, en*) suivi d'un article défini (*le, les*) ou d'un déterminant / pronom relatif (*duquel, auquel, esquels,* ...)[154]. La segmentation de ces amalgames était cruciale pour travailler sur les prépositions, en particulier pour explorer avec nos outils leurs contextes distributionnels. Dans un développement à venir (§ 2.2.4), on justifiera le choix opéré de ne pas intégrer la segmentation de ces amalgames dans le processus automatisé initial de tokenisation du corpus, préférant la réserver pour une phase plus tardive.

2.2.2. La construction du lexique Presto

La tokenisation et l'annotation du corpus ont nécessité à plusieurs reprises l'intervention du lexique Presto. Voici les grandes étapes qui ont présidé à sa construction.

Étape 1. Lexique de départ. Ce lexique, constitué de triplets lemme-forme-catégorie, a été élaboré pour sa plus grande part par G. Souvay. La ressource initiale sélectionnée a été le lexique *Lefff* (*Lexique des Formes Fléchies du Français*, B. Sagot : 2010). Puis ont été ajoutés à ce lexique, lorsqu'ils étaient manquants, des triplets lemme-forme-catégorie issus du logiciel *Morphalou* (http://www.cnrtl.fr/lexiques/morphalou/LMF-Morphalou.php, L. Romary, A.-S. Salman, G. Francopulo, 2004) dont les données sont tirées du TLF (*Trésor de la Langue Française*, http://atilf.atilf.fr), des triplets issus du DMF (*Dictionnaire du Moyen Français*, http://www.atilf.fr/dmf) et d'autres enfin tirés des lexiques morphologiques liés au lemmatiseur LGeRM (*lemmatisation de la variation graphique des états anciens du français et lexiques morphologiques*, Souvay & Pierrel, 2009). Des ajouts complémentaires ont été effectués par G. Souvay à partir d'autres lexiques plus spécialisés.

[154] L'annotation morphosyntaxique et la lemmatisation dans la BFM a dû en outre compter avec les phénomènes d'enclise absentes du corpus intégral Presto. Par ex. dans le cas du pronom relatif : enclise du pronom personnel précédé du pronom relatif : *kil, quil (« ki / qui + le »), ques (« que + les »), kis, quis (« ki/qui » + les »)* ; enclise du pronom adverbial précédé du pronom relatif : *quin (« qui + en »)*; etc. Voir http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf

Étape 2. Etant donné que nous avions réunis des lexiques d'origines très différentes, nous avons mis en œuvre une phase de **conversion des étiquettes morphosyntaxiques** utilisées dans ces lexiques en vue de les rendre compatibles avec le jeu d'étiquettes Presto défini pour le projet. De même, une étape de **normalisation des lemmes** a été réalisée.

Étape 3. Le lexique ainsi obtenu a ensuite été « **archaïsé** » en vue de couvrir au mieux l'extrême variété des formes graphiques qu'on trouve dans les textes notamment du XVI^{e} s. Pour ce faire, nous avons recouru à une sélection de règles élaborées par G. Souvay et issues du système *LGeRM*. Le tableau suivant, adapté de S. Diwersy, A. Falaise, M.-H. Lay & G. Souvay (2017 : 34) présente, en guise d'illustration, les règles appliquées pour produire dans le lexique des formes (tokens) « archaïques » pour des lemmes terminés par la séquence de caractères -UIT, comme *FRUIT*

Règles [R] utilisées		**Formes produites**
R1	si (finale (lemme) est UIT) alors UIT → UICT	*fruict*, *fruicts*
R2	si (finale (lemme) est UIT) alors UITS → UIS	*fruis*
R3	si (S en finale) alors S → Z	*fruitz*
R4	I → Y	*fruyt*, *fruyts*

Tableau 3
Règles utilisées pour l'archaïsation du lemme *FRUIT*.

Ces règles, itérées trois fois, ont été systématiquement appliquées aux formes issues de l'itération précédente. Ainsi la deuxième application de la règle 3 [R3] aux formes issues du tableau ci-dessus donne les (nouvelles) formes *fruictz, fruiz*, *fruytz* et de la règle 4 donne *fruyct, fruycts, fruys*. Une dernière itération produit les formes *fruyctz, fruyz.* Au final, au lemme *FRUIT* sont donc affectées dans le lexique Presto les formes *fruict*, *fruicts, fruis, fruitz, fruyt*, *fruyts, fruictz, fruiz*, *fruytz, fruyct, fruycts, fruys, fruyctz, fruyz.* Bien entendu, cette procédure d'archaïsation génère immanquablement des formes inexistantes dans le corpus. Mais il s'agit là d'un inconvénient mineur au regard de l'extension de la reconnaissance des formes qu'elle permet, comme l'illustrent S. Diwersy, A. Falaise, M.-H. Lay & G. Souvay (*op. cit*.: 35) en appliquant le lexique Presto sur le corpus Frantext (tranche 1451- 1751) :

> « Le lexique Presto comporte ainsi actuellement environ 3 200 000 entrées pour un lexique de départ comportant 450 000 entrées. Environ 2 000 règles d'archaïsation ont été utilisées. La figure [suivante] montre le gain en termes de couverture lexicale de l'archaïsation. L'analyse des lacunes du lexique montre une forte proportion de noms propres et de mots étrangers (essentiellement les mots latins). En termes de graphie on remarque une forte proportion d'hapax. Il s'agit souvent de variantes exotiques ou d'erreurs (numérisation, rupture de mots, impression). »

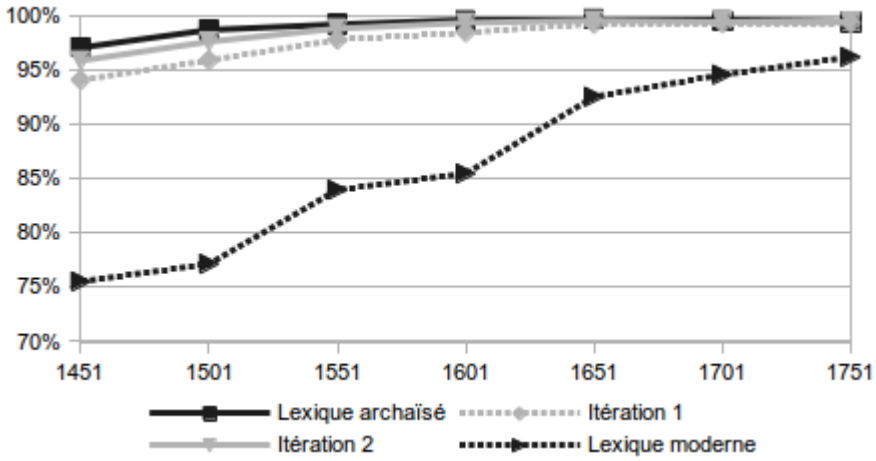


Figure 2

Couverture du lexique Presto sur le corpus Frantext *avant* archaïsation (lexique moderne), *après* archaïsation (lexique archaïsé), et étapes intermédiaires (itérations 1 et 2).

Étape 4. Vient ensuite une phase de **recherche des lemmes et des formes inconnus**. Le logiciel LGeRM analyse ces dernières en appliquant cette fois toutes les règles morphologiques dont il dispose et effectue des hypothèses sur leur lemme de rattachement, hypothèse qui sont ensuite évaluées manuellement par ordre de fréquence décroissante.

Étape 5. Phase de normalisation manuelle des lemmes: il s'agit d'appliquer à tout le lexique les mêmes conventions pour l'affectation d'un lemme à une forme donnée. Ces conventions ont fait l'objet d'une déclaration de règles qu'on trouvera dans l'ANNEXE 3. Lorsque le lemme est absent du lexique Presto, c'est le premier mot de l'entrée du TLFi qui est retenue, et si cette entrée n'existe pas, c'est le premier mot de celle du DMF et en dernier ressort, du *Tobler-Lommatzsch*.

Étape 6. Des contrôles manuels sont effectués sur le lexique, donnant lieu à des corrections et à des ajouts dans le lexique complémentaire.

Pour conclure, nous proposons de reproduire ci-dessous le schéma présenté dans S. Diwersy, A. Falaise, M.-H. Lay & G. Souvay (2015) qui récapitule la plupart des étapes évoquées ci-dessus (§ 2.1 et §2.2).

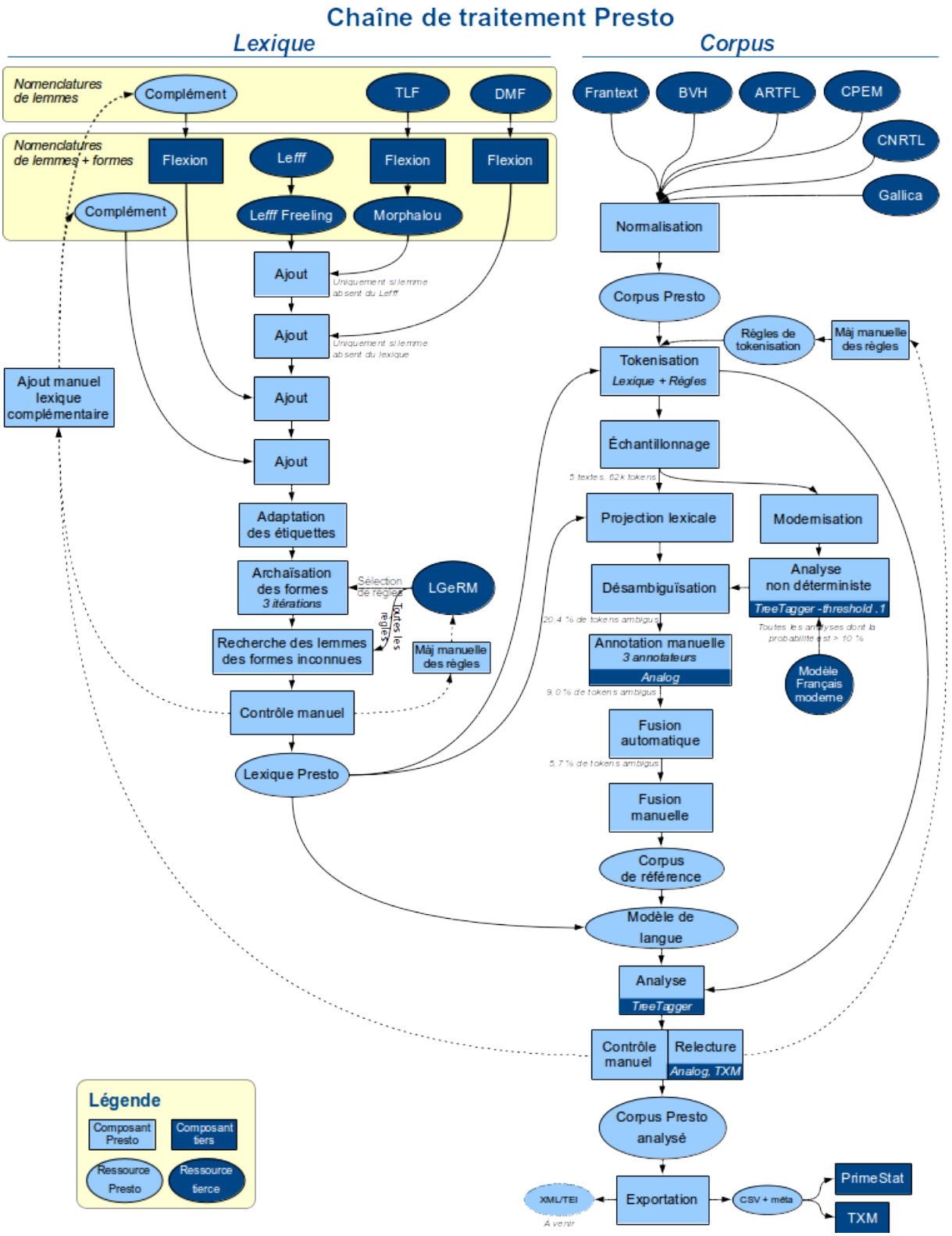


Figure 3
Chaîne de traitement utilisée dans le programme Presto.

2.2.3. Le jeu d'étiquettes Presto

Le choix d'un jeu d'étiquettes est toujours délicat pour l'annotation d'un corpus, car comme le soulignent S. Diwersy, A. Falaise, M.-H. Lay & G. Souvay (2017 : 31), « plus le jeu d'étiquettes est précis, plus la phase de désambiguïsation est complexe ; d'une part, parce que la plus grande précision du jeu d'étiquettes génère des ambiguïtés, multipliant ainsi les choix à faire et à valider ; d'autre part parce qu'il augmente la nécessité d'un annotateur expert capable de faire des choix complexes ». L'ambiguïté est d'autant plus importante dans un corpus historique que des mots peuvent changer de catégorie morphosyntaxique au cours du temps, ou que certaines distinctions catégorielles fondées sur des régularités morphosyntaxiques ne se stabilisent[155] que vers la fin du XVII^{e} s. Ainsi en va-t-il de la distinction entre participe présent, adjectif verbal et gérondif problématique en français classique : « la tripartition des formes en *–ant* n['y] va pas de soi (…) dans la mesure où la différence syntaxique et sémantique entre les trois catégories ne se marque pas formellement par une morphologie distinctive : le gérondif, invariable, se distingue mal du participe (au masculin singulier) du fait qu'il n'est pas régulièrement précédé de en ; le participe qui peut être variable en genre et en nombre, se distingue mal de l'adjectif verbal. (N. Fournier, 2002, § 421 : 291-292)

Nous avons donc décidé d'opter pour un jeu d'étiquettes « minimal » pour réduire le nombre d'ambiguïtés. Nous nous sommes appuyés sur les jeux d'étiquettes existants que sont MULTEXT *english* (2010) (http://nl.ijs.si/ME/V4/msd/html/msd-en.html) et GRACE (1997) (voir G. Adda, J. Mariani, P. Paroubek, M. Rajman, J. Lecomte, 1999). On trouvera dans l'ANNEXE 2 une présentation détaillée du jeu d'étiquettes mis au point pour le projet.

On formulera sur ce jeu deux commentaires : le premier est relatif à la catégorie « G », inexistante dans MULTEXT et GRACE; le second aux amalgames.

La catégorie « G » s'applique aux participes, adjectifs verbaux et gérondifs. Le choix de cette catégorie s'explique pour la raison évoquée *supra* : une distinction morphologique problématique jusqu'au début du XVIII^{e} s., source d'ambiguïtés et d'erreurs potentielles chez les annotateurs. Et à supposer que les annotateurs ne commettent aucune erreur, en l'état actuel des outils d'annotation automatiques, ces derniers ne parviendraient pas à lever ces ambiguïtés correctement. Pour éviter d'accumuler des erreurs d'étiquetage, nous avons donc préféré neutraliser la distinction en considérant que d'autres après nous - plus experts des états de langue considérés - pourront s'attaquer à ce problème spécifique. Comme le soulignent S. Diwersy, A. Falaise, M.-H. Lay & G. Souvay (*op. cit.* , 33), « on a (…) ici une zone de « diminution de la qualité globale » que nous avons préféré isoler afin de pouvoir lui consacrer des traitements spécifiques. »

Quant aux amalgames, ils ont reçu les étiquettes suivantes :

[155] Par ex. le couple *chaque / chacun* voir N. Fournier, 2002, § 200 : 140.

étiquettes	
S+Da	amalgame préposition + art. défini : *du, des, au, aux*
S+Pr	amalgame préposition + pronom relatif : *auquel, duquel, …*
S+Dr	amalgame préposition + déterminant relatif : *auquel, duquel, …*

Tableau 4

Lemmes provisoires affectés aux amalgames issus des prépositions *de, à, en* avant segmentation.

La segmentation de ces amalgames a constitué une étape incontournable pour nos travaux, portant sur l'étude de la distribution des prépositions. Le tableau ci-dessous indique la fréquence des amalgames dans Presto, rapportée (en pourcentages) à la fréquence d'emploi des articles définis d'une part, à la fréquence d'emploi de la classe des déterminants d'autre part.

Périodes	**% d'amalgames par / au nombre total d'*art. déf.*** (amalgamés ou non)	**% d'amalgames par / au nombre total des *déterminants*** (amalgamés ou non)
1501 -1600	23%	12%
1601-1700	21%	11%
1701-1800	21%	12%
1801-1900	22%	13%
1901-1944	21%	13%

Tableau 5

Proportions (exprimées en %) par siècles (1501-1944) du nombre de formes amalgamées issues de la combinaison (*de|à|en)+le|les* relativement au nombre d'occurrences i) des articles définis ii) des déterminants du nom – Corpus Presto.

La fréquence des amalgames *prép. + art. déf.* constitue une part significative dans les emplois des déterminants en français, avec une constance assez remarquable siècle après siècle.

On propose de s'arrêter dans les lignes qui suivent sur les difficultés rencontrées lors de la phase manuelle de segmentation de ces amalgames, étape qui s'est située tardivement dans le processus d'annotation, après la phase d'analyse par *TreeTagger*.

2.2.4. Segmentation des amalgames dans Presto

Comme on y reviendra dans la troisième partie de ce mémoire, les amalgames morphologiquement issus de *à* suivi de *le, les, lequel, lesquels, …* (soit : *au, aux, auquel, auxquels, …*) pouvaient être sémantiquement ambigus au XVI^{e} s car, outre leurs emplois attendus pour introduire les SP régis par *à*, ils se substituaient fréquemment aux amalgames issus de la préposition *en* suivie de *le, les, lequel, lesquels, …*(soit : *ou, on,… és, ès, es, … ouquel, onquel,… ésquels, èsquels, esquels, …*), souvent sentis comme archaïques. Voici quelques exemples où les formes *au, aux, auquel* (en gras) s'interprètent contextuellement comme équivalant à la préposition *en* suivie de *le|les|lequel*:

(4) *Face le ciel (quand il voudra) revivre / Lisippe, Apelle, Homere, qui le pris / Ont emporté sur tous humains espris / En la statue, **au*** tableau*, et **au** livre.* (1550, Du Bellay, *L'Olive*) (cité par S. Lardon & M.-C Thomine, 2009 : 393)

(5) ***Aux** braves exploits de sa vie, et en sa mort, on le* [=Caton] *sent toujours monté sur*

ses grands chevaux. (Montaigne, Essais, III) (cité par Gougenheim, [1951] 1974 : 183)

(6) *Le cerf **auquel** fut transformé Acthéon* (Rabelais, V, 33 ; cité par G. Gougenheim, *op. cit.*: 181)

A l'inverse, il arrive - notamment chez Rabelais - qu'un amalgame morphologiquement issu de *en* suivi de l'article défini *les* (= *es*) soit utilisé là où l'on attendrait (sur les plans syntaxique - constructionnel - et sémantique) normalement *aux* : par ex.

(7) *De tant loing que le vit Pantagruel, il dist **es** assistans : Voyez vous cest homme qui vient par le chemin du pont Charanton ?* (1542, F. Rabelais, *Pantagruel)*

Une question qui nous a occupés pendant plusieurs semaines a consisté à déterminer le meilleur parti à prendre pour la lemmatisation de tels amalgames. Pour les trois énoncés donnés ci-dessus par exemple ((4) à (6)), convient-il d'annoter les amalgames *au*, *aux* et *auxquels* au moyen (respectivement) des lemmes <EN + LE> et < EN + LEQUEL > ou <À + LE> et <À + LEQUEL> ? Selon le choix opéré, la préposition entrera après segmentation dans le paradigme soit de *en* soit de *à*. Or ce choix peut avoir ensuite une incidence significative (suivant le nombre d'occurrences concernées) sur l'exploration quantitative des contextes distributionnels préférés d'apparition de ces deux prépositions.

Les lignes qui suivent retracent les recherches qui nous ont permis d'aboutir à un choix éclairé. Nous nous concentrerons sur le seul cas des formes issues de l'amalgame entre les prépositions *de, à, en* et l'article défini *le, les*. En effet, la rareté fréquentielle des autres cas de figure (avec déterminant ou pronom relatifs) nous a conduit à considérer que les décisions prises pour le traitement des amalgames avec l'article défini s'appliquaient *a fortiori* aux autres, comme on y reviendra à la fin de ce développement.

Notre objectif, rappelons-le, consistait à segmenter de manière automatique tous les amalgames mettant en jeu les prépositions *de, à, en*. Une telle segmentation n'a pas posé de problème particulier pour les formes *du, des*. En effet, la désambiguïsation automatique des catégories morphosyntaxiques *article partitif* (catégorie : <Dp>[156], lemme : <DU>) *versus amalgame préposition de + article défini* (catégorie : <S+Da>, lemme <DE+LE>) était de bonne qualité à l'issue de l'étiquetage automatique opéré par *TreeTagger* que nous avions entraîné. La règle de segmentation suivie a été la suivante :

Avant segmentation : 1 mot			**Après segmentation : 2 mots**					
forme(s)	**pos**	**lemme**	**mot 1**			**mot 2**		
			forme	**pos**	**lemme**	**forme**	**pos**	**lemme**
du	<S+Da>	<DE+LE>	de	<S>	<LE>	le	<Da>	<LE>
des	<S+Da>	<DE+LE>	de	<S>	<LE>	les	<Da>	<LE>

Tableau 6
Règle de segmentation automatique appliquée aux formes amalgamées issues de la combinaison de *de+le* et *de+les* dans le corpus intégral Presto

[156] Voir jeu d'étiquettes, ANNEXE 3.

En revanche, nous ne disposions pas d'une qualité d'étiquetage morphosyntaxique acceptable pour les formes[157] *ou, on, és, és, es*. Divers dispositifs de correction ont donc dû être mis en place pour améliorer l'annotation de ces formes, doublés d'une réflexion de fond sur leur lemmatisation.

2.2.4.1. Désambiguïsation de la catégorie morphosyntaxique des formes *ou, on, és, ès, es* aux XVI^e^ s. et XVII^e^ s.

Une première désambiguïsation a dû être effectuée car:

- ***ou*** peut être
  - conjonction de coordination (étiquette MS <Cc>)
  - « adverbe » interrogatif (<Rt>) ou pronom relatif (<Pr>) dépourvus d'accent[158]
  - amalgame issu de préposition + article défini (<S+Da>)
- ***on*** peut être
  - pronom personnel (<Pp>)
  - amalgame de préposition + article défini (<S+Da>) ;
- ***es, és, ès*** peuvent être
  - forme verbale conjuguée du verbe *être* (<Vuc>)
  - amalgame de préposition + article défini (<S+Da>)
  - (pour *es*) forme verbale conjuguée du verbe *esse* en latin[159] (<Xe>)

Cette première difficulté a pu être levée facilement. *Treetagger* est apte à mettre en œuvre une procédure statistique automatique à même de désambiguïser la catégorie morphosyntaxique d'une forme en affectant à l'une et à l'autre catégorie envisageable un score de probabilité. Cette procédure, traditionnelle, requiert une phase préalable d'annotation manuelle d'un nombre relativement restreint (une vingtaine) d'occurrences de chacune des formes concernées au sein d'un sous-corpus tiré de Presto, les textes étant sélectionnés sur le critère du nombre le plus significatif d'occurrences des amalgames recherchés[160]. Voici la liste des textes retenus :

1537, *Les contes amoureux de madame Jeanne Flore*, J. Flore.
1542, *Pantagruel,* F. Rabelais.
1542[161], *Gargantua,* F. Rabelais.
1556, *Discours non plus mélancoliques que divers*, E. Vinet.
1572, *La Savoye,* J. Pelletier du Mans.
1603, *Le théâtre d'agriculture et mesnage des champs*, O. Serres.

---

[157] Il s'agit là de la liste exhaustive des formes amalgamées issues de *prép. en + art. déf.* que nous avons recensées dans le corpus intégral Presto.

[158] Cas très fréquent dans les textes entre les XVI^e^ et XVIII^e^ s. pour les éditions à orthographe non modernisée que nous utilisons.

[159] Le corpus contient de nombreuses citations latines.

[160] Cette caractéristique a pu être vérifiée au moyen du concordancier, en sélectionnant les formes recherchées au moyen de requêtes CQL adaptées puis en opérant des analyses linguistiques (amalgame ou non ?) à la volée. Il n'est évidemment pas indifférent que la liste des textes sélectionnés ne se prolonge pas au-delà de 1620: à partir de la fin de la première moitié du XVII^e^ s., la fréquence d'emploi de *ès, és, es* amalgames décroît de manière considérable, même si elle ne s'éteint que dans les débuts du XVIII^e^ s. Quant aux emplois de *ou, on* amalgames issus de *en+le*, ils disparaissent dès la seconde moitié du XVI^e^ s.

[161] Pour *Pantagruel* et *Gargantua*, nous faisons figurer la date de 1542, et non respectivement de 1532 (édition princeps de *Pantagruel*) et de 1535 (édition princeps de *Gargantua*). Le corpus Presto intègre en effet les deux versions de ces œuvres (BVH http://www.bvh.univ-tours.fr) éditées chez F. Juste à Lyon, toutes deux ayant été corrigées de la main de l'auteur (sur ce point, voir p 138-140 & passim, R. Lathuillère, 1981 : 129-145)

1619, *Introduction à la vie dévote*, St François de Sales.

L'annotation manuelle dans ces textes des amalgames visés a permis un nouvel entraînement de l'étiqueteur aboutissant à un nouvel étiquetage du corpus avec une qualité de performance tout à fait honorable.

2.2.4.2. Lemmatisation des formes amalgamées *au, aux, aus, és, ès, es*

Ces formes ayant été convenablement annotées en termes de catégorie grammaticale, on se trouve confronté au problème suivant : avant d'opérer une segmentation automatique des triplets lemme-forme-catégorie associées aux tokens *au, aux, aus, és, ès, es* identifiés comme des amalgames, se pose la question du lemme à leur affecter. Ces formes sont en effet sémantiquement ambiguës puisque, outre leurs emplois que nous nommerons « *standard* » (à savoir : *au < à+le ; aux, aus < à+les ; és, ès, es < en + les)*[162], elles peuvent avoir aussi des emplois « *non-standard* » : la forme *au* correspond alors à la combinaison *en* + *le*, les formes *aux, aus* à la combinaison *en* + *les*, les formes *es*, *és, ès* à la combinaison *à* + *les*.

La question est alors : aux formes amalgamées *au, aux, aus, és, ès, es* non-standard, faut-il affecter le lemme <À + LE> ou <EN + LE> ? Répondre à cette interrogation nous a conduit à examiner les deux points suivants :

- est-il légitime, sur le plan linguistique, d'introduire une telle différence d'annotation (au niveau des lemmes) entre emplois standard et non-standard de certains amalgames?
- Une telle annotation est-elle utile pour les calculs quantitatifs envisagés ensuite?

Sur le plan linguistique, d'abord, il nous est apparu légitime de coder les emplois non-standards des amalgames *au, aux, aus* au moyen du lemme <EN + LE> et ceux des amalgames *és, ès, es* au moyen du lemme <À + LE>. A l'appui de cette thèse, nous reprendrons l'argument de C. Molinier (1990) qui, travaillant sur la distribution de *en* et de *à* avec les noms de saisons en français contemporain (*en été, en automne|à l'automne, en hiver, au printemps*), fait observer : « la langue ne confond pas les deux *au*, comme le prouve le maintien de l'alternance *au* (*en le*) *N / en son N*, voir *au milieu de la pièce / en son milieu / *à son milieu, au sein du groupe / en son sein / *à son sein, au temps de mon grand-père / en son temps / *à son temps* et d'autre part *au sujet de Max / à son sujet, au sommet de la montagne / à son sommet* ». Or au XVI^{e} s. déjà, la langue distinguait tout aussi soigneusement les deux *au* qu'elle attachait à deux paradigmes distincts.

A cet argument, on peut en ajouter un second tiré d'un autre type d'alternance illustré par les deux exemples suivants déjà donnés *supra* :

(4) *Face le ciel (quand il voudra) revivre / Lisippe, Apelle, Homere, qui le pris / Ont emporté sur tous humains espris / **En la** statue, **au*** tableau*, et **au** livre.* (1550, Du Bellay, *L'Olive*)

(5) ***Aux** braves exploits de sa vie, et **en sa** mort, on le* [=Caton] *sent toujours monté sur ses grands chevaux.* (Montaigne, Essais, III)

On a ici affaire à des contextes syntaxiques à construction constante (ajouts[163] à sémantisme de domaine dans (4), à sémantisme temporel dans (5)) où *au* | *aux* alterne avec *en* + *Dét.*) Là

[162] Pour les amalgames *on, ou*, tous les emplois recensés dans le corpus se sont avérés « standard ». La question de leur lemmatisation ne se pose donc pas.

[163] Pour la notion syntaxique d'ajout, voir B. Lavieu (2006 : 133-136). Comme elle, nous considérons que l'*ajout*

non plus, la langue ne confond pas *au, aux* (<EN + LE>) et *au, aux* (<À + LE>), ce dont on peut se convaincre lorsqu'on tente de remplacer *en* par *à* :

(4') *Face le ciel (quand il voudra) revivre / Lisippe, Apelle, Homere, qui le pris / Ont emporté sur tous humains espris / ***À la** statue, **au*** tableau*, et **au** livre.*

(5') ***Aux** braves exploits de sa vie, et ***à sa** mort, on le* [=Caton] *sent toujours monté sur ses grands chevaux.*

La construction syntaxique s'opère bien au moyen de la préposition *en*, et l'amalgame *au* | *aux* entre donc dans le paradigme de cette dernière. De même, pour l'énoncé:

(7) *De tant loing que le vit Pantagruel, il dist **ès** assistans*

la substitution suivante serait exclue :

(7') *De tant loing que le vit Pantagruel, il dist ***en ses** assistans*

On a affaire à une construction dative et *ès* se rattache au paradigme de *à*.

Examinons maintenant la seconde question : est-il utile, sur le plan quantitatif, d'accomplir une telle lemmatisation visant à désambiguïser en contexte les formes examinées ici? Dit autrement, les emplois non-standard des amalgames *au, aux, aus, ès, és, es* sont-ils suffisamment nombreux dans le corpus pour qu'on mette au point une stratégie spécifique de désambiguïsation automatique de leurs lemmes, opération nécessairement coûteuse en temps? Pour répondre à cette interrogation, il convient de se faire une idée de la proportion qu'occupent ces emplois non-standard par rapport aux emplois standard. Dans ce but, nous avons annoté manuellement, dans un corpus de 6 textes (364.781 tokens) relevant du champ générique *genres narratifs* (discours *littéraire*), tous les amalgames de forme *ou, au, aux, aus, ès, és, es*.

Soulignons-le d'emblée : cette annotation n'est pas sans souffrir d'une faiblesse notable que nous aurons à cœur à corriger dans le courant du dernier trimestre civil 2017 : elle aurait dû être accomplie au moins par deux annotateurs experts de la langue préclassique, et travaillant en parallèle. Mais faute de temps et de ressources disponibles, nous avons opté pour une annotation par une seule personne : en l'occurrence, l'auteur de ces lignes…

Il convient par conséquent de prendre les résultats présentés ci-dessous avec de grandes précautions : ils devront être confirmés par une annotation par un voire deux experts du français préclassique.

Voici la liste des six textes qui constituent ce que nous avons appelé le « mini corpus » de test :

Mini-corpus de test

1530, *Ulenspiegel*, Anonyme, [27.215 tokens].
1542, *Gargantua,* F. Rabelais, [52.954 tokens].
1558, *Nouvelles récréations et joyeux devis*, B. Des Périers, [84.241 tokens].
1572, *Le Printemps*, J. Yver, [62.135 tokens].

(d'un SV ou de P, intraphrastique) est un constituant non régi par V (à l'inverse des *arguments* de ce dernier) et distinct de l'*incident* qui est un constituant extraphrastique placé *au-dessus de* P.

1624, *L'Endimion*, J. de Gombaud, [47.852 tokens].
1629, *Histoire indienne d'Alexandre et d'Orazie, où sont entremeslées les aventures d'Alcidaris, de Combaye et les amours de Pyroxène*, F. de Boisrobert, [90.384 tokens].

Le choix de ces textes a obéi aux critères suivants :

- *Critère temporel* : c'est durant la période 1500-1650 que l'on observe le plus grand nombre d'emplois non-standard des amalgames étudiés.[164] Par ailleurs, les œuvres choisies sont réparties de manière relativement régulière dans cet empan temporel : deux textes par demi-siècle.
- *Critère de champ générique* : recherche d'une certaine homogénéité générique (*genre narratif*).
- *Critère de taille* : c'est dans le champ générique « genres narratifs » que nous sommes parvenu à identifier des textes dont le nombre de tokens n'était pas trop disparate (de 27.307 à 90.405) lorsqu'on respectait les deux critères précédents.

Pour l'annotation manuelle de ce mini-corpus, voici les principes arrêtés :

2.2.4.3. Principes d'annotation des amalgames équivalant sémantiquement à *en/à + le/les*

Nous avons distingué deux cas :

A. Contextes d'occurrence où il est possible de désambiguïser le lemme

C'est le cas chaque fois que le sens contextuel et le contexte syntaxique, éventuellement associés à un contexte de coordination, de juxtaposition voire de comparaison, permettent d'identifier à quelle préposition on peut rattacher l'amalgame. S'il s'agit de *en*, le lemme sélectionné est <EN+LE>, s'il s'agit de *à*, c'est le lemme <À+LE>. Voici quelques exemples (outre ceux donnés *supra*) :

(8) (…) *comme tesmoigna le poisson pris, duquel on trouva* ***au*** [<EN+LE>] *ventre l'anneau tant precieux que ce fortuné Roy avoit jetté en la mer* (…) (1572, J. Yver, *Le Printemps*).
(9) *Voyre quand il tailloit un habillement pour soy, il luy estoit advis que son drap n'eust pas esté bien employé, s'il n'en eut eschantillonné quelque lopin : & caché* ***en la*** *liette, ou* ***au*** [<EN+LE>] *coffre des bannieres.* (1558, B. Des Perriers, *Nouvelles récréations et joyeux devis*).
(10) *Cuyde tu ces oultraiges estre recelles* ***es*** *esperitz eternelz, et* ***au*** [<EN+LE>] *Dieu souverain, qui est juste retributeur de noz entreprinses ?* (1542, F. Rabelais, *Gargantua*).
(11) *Pour moy je croyois bien surpasser tous les hommes,* ***aux*** [<EN+LE>] *vœux, et* ***en l'****affection de te servir;* (1624, J. de Gombauld, *L'Endimion*).

B. Contextes d'occurrence où il s'avère impossible de désambiguïser le lemme à affecter à la forme amalgamée rencontrée.

On rappellera en premier lieu l'exemple de l'énoncé cité par G. Gougenheim ([1951], 1974 : 183) destiné à illustrer, sous sa plume, un cas d'ambiguïté de *aux* :

[164] L'apparition puis la montée en puissance de *dans* – comme on y reviendra dans la troisième partie – en réduit considérablement le nombre après 1650

(12) *Amour ne change poinct le cueur, mais le monstre tel qu'il est, fol **aux** fols et saige **aux** saiges.* (1550, M. de Navarre, *l'Heptameron*)

Faut-il - s'interroge le linguiste - interpréter les SP *aux folles* et *aux saiges* comme des datifs régis par le verbe *monstre* (préposition *à*) ou bien comme équivalant à *en + les* au sens moderne de *chez* : *fou chez les fous, sage chez les sages*?

Il reste que, au regard du codage des occurrences que nous avons accompli, l'exemple avancé par G. Gougenheim ne nous apparaît pas du tout représentatif des configurations ambiguës les plus souvent rencontrées. Dans (12), chaque SP peut être candidat à deux constructions syntaxiques possibles. Autrement dit, à l'ambiguïté sémantique correspond une ambiguïté syntaxique. Or dans la plupart des énoncés que nous avons codés, l'ambiguïté sémantique n'avait aucune incidence sur la fonction syntaxique du SP qui restait identique. Voici un exemple :

(13) (…) *lesquelz pour faire l'honneur accoustumé à leur dame et maistresse, **veindrent** de bon matin **au chasteau**, chargés de rameaux, d'oiseaux en cage, de fueillade, de miel et de laictages de toutes façons* (…)(1572, J. Yver, *Le Printemps*)

Dans tous les cas, *au chasteau* est un ajout locatif qui exprime le lieu d'arrivée du procès de déplacement exprimé par *veindrent*. La question (uniquement sémantique) est : faut-il considérer que *au* équivaut ici à *en le* (entrée de la cible dans le site) ou *à le* (entrée possible mais non nécessaire)? Entre 1500 et 1650, on trouve en effet aussi bien <VENIR><À><DET><CHÂTEAU> que <VENIR> <EN><DET><CHÂTEAU> comme l'illustrent les énoncés suivants[165] :

(…) *la dame non seulement luy dit, qu'il y estoit bien venu, ains le voyant mal accommodé, qu'il **vint à son chasteau**, dequoy vaincu d'une gracieuse importunité, consentit en fin*. (1572, J. Yver, *Le Printemps*)

(…) *luy mandant qu' il vouloit communiquer avec luy d' affaires d' importance, et le priant **venir en un sien chasteau***. (1602, C. Fauchet, *Declin de la maison de Charlemagne*[166])

Il est donc impossible dans (13) de lever l'ambiguïté. Plutôt qu'ambiguïté, il vaudrait mieux d'ailleurs parler d'incapacité du système à mettre en œuvre dans ce contexte l'opposition qu'il permet d'exprimer ailleurs entre *venir **en Dét** NLieu* et *venir **à Dét** Nlieu*, du fait de la substitution des formes *au, aux* aux amalgames morphologiquement issus de *en le / *en les.

Nous avons donc fait le choix de recourir, pour la forme *au* dans (13), au lemme « structurellement » ambigu <À+LE|EN+LE>.

2.2.4.4. Observations quantitatives à l'issue de l'annotation manuelle du mini-corpus

[165] L'ambiguïté décrite ici dans un énoncé où figure le verbe *venir* peut être étendue à d'autres énoncés où le SP est un ajout locatif exprimant le lieu d'arrivée d'un procès de déplacement : *aller, conduire, mener, arriver*…

[166] Ce texte ne fait pas partie du mini-corpus mais de la version étendue du corpus.

L'annotation manuelle du mini-corpus accomplie et son import dans TXM opéré, nous avons dénombré les formes amalgamées qui s'y trouvaient et les lemmes qui leur avaient été affectés manuellement. Voici les valeurs obtenues :

Lemme / Forme	À+LE	EN+LE	À+LE|EN+LE	Total
au	1223	23	121	1367
aux	436	14	9	459
aus	1	0	0	1
ou	0	2[167]	0	2
és	0	1	0	1
es	42[168]	41	5	88
	1702	81	135	1918

Tableau 7
Répartition en fréquence brutes des lemmes affectés aux différentes formes amalgamées issues de *à|en+le|les* présentes dans le mini-corpus.

Les cellules grisées correspondent aux emplois non-standard et ambigus des amalgames.

La somme totale des occurrences relevées pour ces emplois (soit 214 occ.) constitue 11,2 % du total des formes amalgamées (soit 1918 occ.) issues de *en|à+le|les* et codées dans le mini-corpus.

Si l'on rapporte ce nombre de 214 occ. au total des occurrences des prépositions *en* et *à* (en situation d'amalgame ou non) présentes dans le mini-corpus (soit 10.976 occ.), on obtient une proportion d'environ 2 %. Autrement dit, si l'on avait étiqueté le mini-corpus au moyen d'un lemmatiseur ne traitant les amalgames qu'*en surface,* c'est-à-dire attribuant systématiquement aux formes *au, aux, aux* (standard, non-standard ou ambiguës) le lemme <À+LE> et aux formes *ou, es, ès, és* le lemme <EN+LE>, le mauvais étiquetage des formes dites non-standard et ambiguës n'aurait finalement affecté que 2 % du total des emplois de *en* et de *à*, ce qui apparaît quantitativement marginal[169]. En d'autres termes, et sous réserve qu'on puisse étendre à Presto–[1500-1670] les proportions observées dans le mini-corpus[170], nous avons conclu que la mise au point d'une lemmatisation plus sophistiquée que celle traitant *en surface* les formes amalgamées dans Presto[1500-1670] représenterait un coût exorbitant[171] au vu du caractère quantitativement marginal des emplois à désambiguïser.

[167] Uniquement chez F. Rabelais.
[168] Dont 41 occurrences chez F. Rabelais.
[169] Il se pourrait cependant que certains de ces emplois ne soient pas « qualitativement » si marginaux... Il faudrait en effet s'assurer que les contextes mis en jeu ne possèdent pas une spécificité sémantique qui leur conférerait, du coup, une singularité parmi les emplois de *en* ou de *à*. Nous avons à maintes reprises observé lors du codage que les verbes de déplacement, par ex., pour lesquels le lieu d'arrivée était exprimé au moyen de *au|aux*, engendraient une ambiguïté liée à la question : la « cible » en déplacement pénètre-t-elle ou non dans le site d'arrivée ?
[170] Une telle extension est justifiée pour ce qui concerne du moins le rapport entre le total des emplois amalgamés de *en* et *à*, et le total de leurs emplois amalgamés et non-amalgamés : il est de 22% dans Presto[1500-1650] et de 19% dans le mini-corpus.
[171] Nous n'avons rien dit, dans ce développement, des formes amalgamées issues de *en|à* + déterminant relatif ou pronom relatif de lemme <LEQUEL>. Le caractère fréquentiellement très marginal de ces amalgames nous a conduit à les écarter du champ de notre analyse et à conclure que leur codage en termes d'emploi standard/ non standard / ambigu serait encore plus inutile.

On peut ajouter une remarque à ce qui précède : les emplois non-standard de *es* - soit 42 occurrences - constituent dans le mini-corpus 46% du total des emplois des amalgames issus de *en+le/les*. Or 41 de ces 42 occurrences proviennent d'un seul texte : *Gargantua*. Une telle sur-représentation n'est pas une surprise : il est bien connu des seiziémistes que cet emploi non-standard de *es* se rencontre en particulier chez Rabelais (voir G. Gougenheim, [1951] :182 ; F. Brunot, 1967 : 278), auteur dont on sait en outre qu'il use davantage que la plupart de ses contemporains de cette forme amalgamée (voir F. Brunot, *ibid* : 277). Cette forte appétence de F. Rabelais pour les formes amalgamées *és*|*ès*|*es* – et de plus en emploi non-standard - invite à considérer que le pourcentage total d'amalgames non-standard issus de en|à+ le|les dans Presto$_{-1500\text{-}1670}$ doit être plus marginal encore que la projection des valeurs obtenues pour le mini-corpus ne le laisserait penser.

2.2.4.5. Conclusions sur la lemmatisation automatique des formes amalgamées *ou, au, aux, aus, és, ès, es*

Le recours à des calculs statistiques comme ceux présentés dans la troisième partie de ce mémoire nécessite de segmenter les amalgames issus de la combinaison *prép + art. déf.* Une telle opération ne pose pas de difficultés pour les formes *du, des* dans le corpus, dont la désambiguïsation morphosyntaxique d'avec les articles partitifs et indéfinis a été accomplie de façon satisfaisante par le modèle de langage utilisé pour entraîner le lemmatiseur. En revanche elle s'est avérée problématique pour les amalgames correspondant à *en*|*à + le*|*les*, et cela pour deux raisons.

En premier lieu, parce que la désambiguïsation morphosyntaxique entre les formes amalgamées recherchées et leurs homographes appartenant à d'autres classes grammaticales n'avait pas été correctement opérée. Une annotation manuelle d'un corpus de huit textes a permis d'accéder à un taux de désambiguïsation automatique réussie satisfaisant.

En second lieu, parce que l'affectation d'un lemme aux formes amalgamées *ou, au, aux, aus, ès, és, es* pose le problème de leurs emplois non-standard et ambigus. Nous avons vu que, si sur le plan linguistique il est justifié de vouloir disposer d'une annotation permettant de les identifier, ils s'avèrent sur le plan quantitatif suffisamment marginaux[172] pour autoriser le recours à une lemmatisation des amalgames *en surface*, c'est-à-dire se contentant d'affecter systématiquement aux formes amalgamées *au, aux, aus* le lemme <À+LE> et aux formes amalgamées *ou, ès, és, es* le lemme <EN+LE>.

2.3. Performance du modèle de langage construit par Presto

Pour finir cette seconde section consacrée à l'annotation dans Presto, on ne peut manquer de s'interroger sur la performance du modèle de langage obtenu. Pour répondre, nous cédons la parole à S. Diwersy, A. Falaise, M.-H. Lay & G. Souvay (*op. cit.* : 36)

« Le modèle obtenu, avec notre jeu de 38 étiquettes, atteint après plusieurs phases de correction et d'optimisation, un taux d'étiquetage correct moyen (*précision*) de 94,6 % (figure [ci-dessous]), mais qui s'avère sensiblement plus bas pour le XVI^{e} siècle (91,4 %). Pour ce siècle, le développement de ressources plus spécifiques permettrait probablement d'améliorer ce taux. »

[172] En nous fondant sur une projection accomplie à partir d'un mini-corpus de six textes où les formes amalgamées *ou, au, aux, aus, ès, és, es* ont été annotées manuellement.

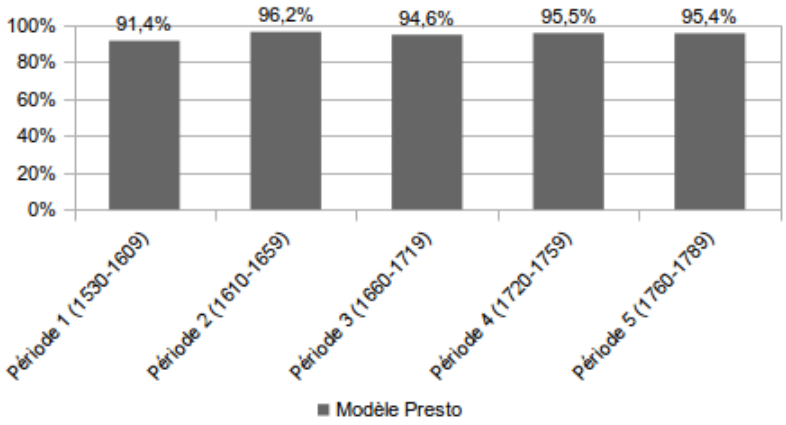


Figure 4

Taux d'étiquettes exactes dans le corpus Presto noyau, après l'annotation automatique.

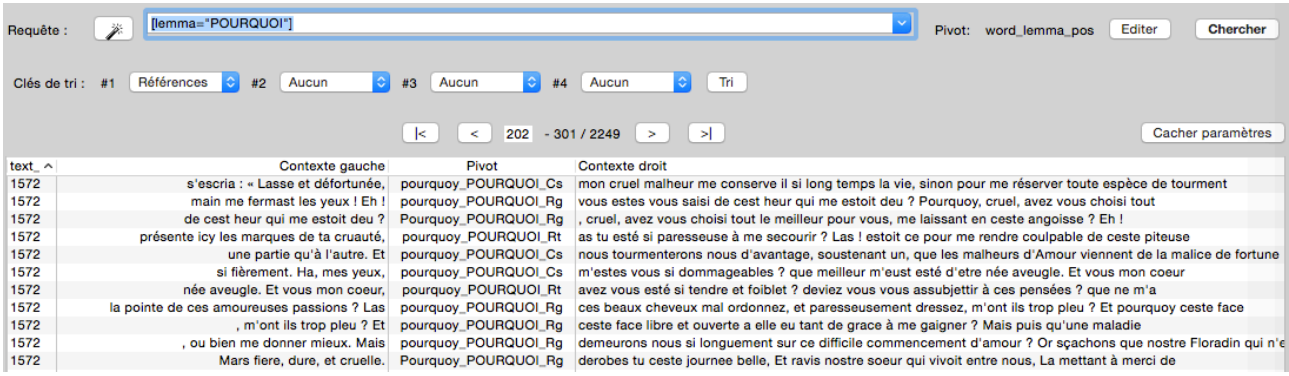
2.4. Conclusion

Grâce à ma participation comme linguiste aux différentes étapes de l'annotation - en particulier, la construction collective du jeu d'étiquettes et la mise au point des principes d'annotation – j'ai beaucoup appris sur certaines procédures de TAL en linguistique sur corpus. J'en ai aussi retiré la conviction qu'une participation étroite à la mise en place de la chaîne de traitement appliquée à un corpus – ou sinon une bonne connaissance de celle-ci - empêche d'en être ensuite un utilisateur « naïf ».

Mais il y a aussi un revers de la médaille : on devient du même coup terriblement conscient de la somme des erreurs, des imperfections voire des aberrations que le corpus qu'on manipule recèle. Que de choses à corriger et à améliorer!

Les outils d'exploration dont il va être question maintenant (§ 3) sont à cet égard des auxiliaires précieux. Il permettent en effet de traquer un bon nombre d'erreurs grâce à la concordance automatique. On trouvera ci-dessous deux captures d'écran qui illustrent ce type de recherche.

La première montre une concordance obtenue pour une requête dans TXM portant sur le lemme <POURQUOI> :



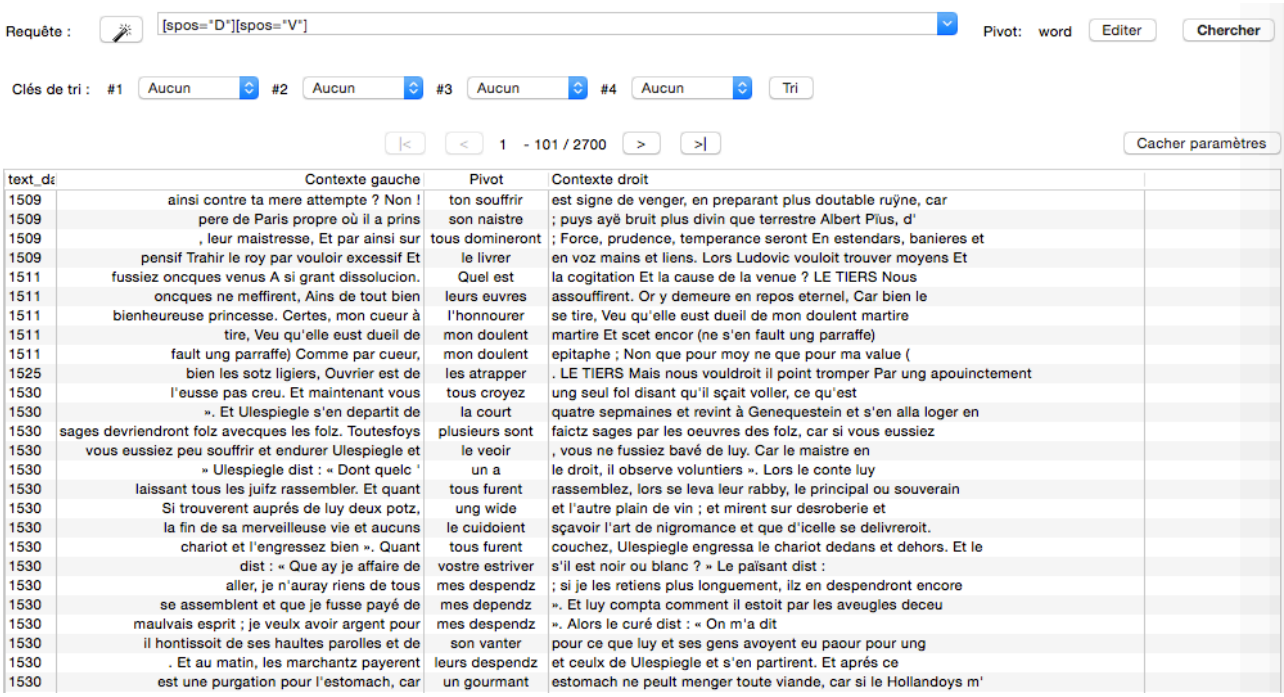
text_	Contexte gauche	Pivot	Contexte droit
1572	s'escria : « Lasse et défortunée,	pourquoy_POURQUOI_Cs	mon cruel malheur me conserve il si long temps la vie, sinon pour me réserver toute espèce de tourment
1572	main me fermast les yeux ! Eh !	pourquoy_POURQUOI_Rg	vous estes vous saisi de cest heur qui me estoit deu ? Pourquoy, cruel, avez vous choisi tout
1572	de cest heur qui me estoit deu ?	Pourquoy_POURQUOI_Rg	, cruel, avez vous choisi tout le meilleur pour vous, me laissant en ceste angoisse ? Eh !
1572	présente icy les marques de ta cruauté,	pourquoy_POURQUOI_Rt	as tu esté si paresseuse à me secourir ? Las ! estoit ce pour me rendre coulpable de ceste piteuse
1572	une partie qu'à l'autre. Et	pourquoy_POURQUOI_Cs	nous tourmenterons nous d'avantage, soustenant un, que les malheurs d'Amour viennent de la malice de fortune
1572	si fièrement. Ha, mes yeux,	pourquoy_POURQUOI_Cs	m'estes vous si dommageables ? que meilleur m'eust esté d'etre née aveugle. Et vous mon coeur
1572	née aveugle. Et vous mon coeur,	pourquoy_POURQUOI_Rt	avez vous esté si tendre et foiblet ? deviez vous vous assubjettir à ces pensées ? que ne m'a
1572	la pointe de ces amoureuses passions ? Las	pourquoy_POURQUOI_Rg	ces beaux cheveux mal ordonnez, et paresseusement dressez, m'ont ils trop pleu ? Et pourquoy ceste face
1572	, m'ont ils trop pleu ? Et	pourquoy_POURQUOI_Rg	ceste face libre et ouverte a elle eu tant de grace à me gaigner ? Mais puis qu'une maladie
1572	, ou bien me donner mieux. Mais	pourquoy_POURQUOI_Rg	demeurons nous si longuement sur ce difficile commencement d'amour ? Or sçachons que nostre Floradin qui n'e
1572	Mars fiere, dure, et cruelle.	Pourquoy_POURQUOI_Rg	derobes tu ceste journee belle, Et ravis nostre soeur qui vivoit entre nous, La mettant à merci de

Figure 5

Capture d'écran faisant apparaître une concordance dans TXM pour une requête opérée sur le corpus Presto visant à rechercher les occurrences du lemme <POURQUOI>.

On observe que la forme <*pourquoy*> employée dans des interrogatives directes a été codée sur le plan morphosyntaxique aussi bien <Rg> (adverbe « général »), <Rt> (adverbe « interrogatif ») et <Cs> (conjonction de subordination)… Or si l'on suit le manuel d'annotation, le codage catégoriel attendu était « adverbe interrogatif » <Rt> (soit 9 erreurs sur 11 pour la fenêtre capturée… !)

Le second exemple présente une concordance sur TXM pour une séquence catégorielle à priori impossible dans le corpus Presto: un déterminant suivi d'un verbe. En effet, chaque fois que l'infinitif est précédé d'un déterminant, il a dû être codé « Nom commun » (<Nc>)[173].



text_da	Contexte gauche	Pivot	Contexte droit
1509	ainsi contre ta mere attempte ? Non !	ton souffrir	est signe de venger, en preparant plus doutable ruÿne, car
1509	pere de Paris propre où il a prins	son naistre	; puys ayë bruit plus divin que terrestre Albert Pïus, d'
1509	, leur maistresse, Et par ainsi sur	tous domineront	; Force, prudence, temperance seront En estendars, banieres et
1509	pensif Trahir le roy par vouloir excessif Et	le livrer	en voz mains et liens. Lors Ludovic vouloit trouver moyens Et
1511	fussiez oncques venus A si grant dissolucion.	Quel est	la cogitation Et la cause de la venue ? LE TIERS Nous
1511	oncques ne meffirent, Ains de tout bien	leurs euvres	assouffirent. Or y demeure en repos eternel, Car bien le
1511	bienheureuse princesse. Certes, mon cueur à	l'honnourer	se tire, Veu qu'elle eust dueil de mon doulent martire
1511	tire, Veu qu'elle eust dueil de	mon doulent	martire Et scet encor (ne s'en fault ung parraffe)
1511	fault ung parraffe) Comme par cueur,	mon doulent	epitaphe ; Non que pour moy ne que pour ma value (
1525	bien les sotz ligiers, Ouvrier est de	les atrapper	. LE TIERS Mais nous vouldroit il point tromper Par ung apouinctement
1530	l'eusse pas creu. Et maintenant vous	tous croyez	ung seul fol disant qu'il sçait voller, ce qu'est
1530	». Et Ulespiegle s'en departit de	la court	quatre sepmaines et revint à Genequestein et s'en alla loger en
1530	sages devriendront folz avecques les folz. Toutesfoys	plusieurs sont	faictz sages par les oeuvres des folz, car si vous eussiez
1530	vous eussiez peu souffrir et endurer Ulespiegle et	le veoir	, vous ne fussiez bavé de luy. Car le maistre en
1530	» Ulespiegle dist : « Dont quelc '	un a	le droit, il observe voluntiers ». Lors le conte luy
1530	laissant tous les juifz rassembler. Et quant	tous furent	rassemblez, lors se leva leur rabby, le principal ou souverain
1530	Si trouverent auprés de luy deux potz,	ung wide	et l'autre plain de vin ; et mirent sur desroberie et
1530	la fin de sa merveilleuse vie et aucuns	le cuidoient	sçavoir l'art de nigromance et que d'icelle se delivreroit.
1530	chariot et l'engressez bien ». Quant	tous furent	couchez, Ulespiegle engressa le chariot dedans et dehors. Et le
1530	dist : « Que ay je affaire de	vostre estriver	s'il est noir ou blanc ? » Le païsant dist :
1530	aller, je n'auray riens de tous	mes despendz	; si je les retiens plus longuement, ilz en despendront encore
1530	se assemblent et que je fusse payé de	mes dependz	». Et luy compta comment il estoit par les aveugles deceu
1530	maulvais esprit ; je veulx avoir argent pour	mes despendz	». Alors le curé dist : « On m'a dit
1530	il hontissoit de ses haultes parolles et de	son vanter	pour ce que luy et ses gens avoyent eu paour pour ung
1530	. Et au matin, les marchantz payerent	leurs despendz	et ceulx de Ulespiegle et s'en partirent. Et aprés ce
1530	est une purgation pour l'estomach, car	un gourmant	estomach ne peult menger toute viande, car si le Hollandoys m'

Figure 6

Capture d'écran faisant apparaître une concordance dans TXM pour une requête opérée sur le corpus Presto visant à rechercher les occurrences du motif « déterminant suivi d'un verbe » (voir fenêtre de requête en langage de requête CQL : [spos="D"][spos="V]).

Dans le corpus Presto (niveau : contrôlé, version : échantillonnée), on trouve actuellement 2.700 occurrences de la séquence « déterminant suivi d'un verbe » qui correspondent toutes à des erreurs d'étiquetage. Si l'on se réfère à la vue sur la concordance de la figure précédente, on peut rapidement identifier que ces erreurs correspondent

- soit à un codage erroné du « déterminant » qui est en fait un pronom
  - indéfini : « sur *tous* domineront » (1509), « *plusieurs* sont faictz » (1530), …
  - personnel : « *le* livrer » (1509), « *l'*honnourer » (1511), …
- soit à un codage erroné du « verbe » qui est en fait
  - un nom commun (« ton *souffrir* » (1509), « leurs *œuvres* » (1511), …)
  - un adjectif verbal (code attendu : « G ») : « mon *doulent* martire » (1511), …

[173] On signalera *a contrario* que dans la BFM, un double codage catégoriel est proposé (http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf): l'un, « morphosyntaxique », étiquette l'infinitif derrière déterminant comme un nom commun (<NOMcom> Etiquette du jeu « Cattex09 » utilisé par la BFM), l'autre « morphologique » l'étiquette comme verbe (<VERinf>). Autrement dit, dans la BFM et sous réserve qu'on sélectionne l'étiquetage morphologique des tokens, il est possible de trouver des motifs vérifiant la séquence « déterminant suivi d'un verbe » sans que cela ne soit imputable à une erreur d'étiquetage.

- soit à une erreur de découpage : « *quelc'un a* » (1530, ligne 15) : en phase initiale de tokenisation (segmentation maximale), le mot « *quelc'un* » a été découpé en trois tokens (<quelc>< ' ><un>). L'application des règles de fusion (voir *supra*, étape 3 de la tokenisation) aurait dû faire de la suite « *quelc'un* » un unique token dans la mesure où, dans le lexique, la forme *quelc* n'est liée à aucun lemme. Après examen, il s'avère que la forme *quelqu'un* figure bien dans le corpus comme un seul token (la fusion a donc été opérée avec rattachement au lemme <QUELQU'UN>) mais que les deux occurrences de la forme « *quelc'un* » - absente du lexique - ont échappé aux règles de fusion automatique.

3. De deux plateformes d'exploration et de calcul en linguistique sur corpus outillée

Dans cette troisième section - modeste en taille - on présente en quelques paragraphes les deux plateformes d'exploration et de calcul utilisées dans le programme Presto : TXM et BTLC/Primestat. On situe ces deux plateformes dans l'ensemble plus vaste des outils automatiques mis à la disposition des chercheurs en linguistique sur corpus. On présente ensuite les fonctions d'exploration et de calculs statistiques auxquelles il a été le plus souvent recouru.

3.1. TXM et BTLC/Primestat dans le paysage plus vaste des outils automatiques d'exploration et de calcul sur corpus numérisés

Le consortium « CORLI » (*CORpus, Langues, Interactions*), issu des deux consortiums de linguistique « Corpus Ecrits » (http://corpusecrits.huma-num.fr) et « IRCOM » (http://ircom.huma-num.fr/site/accueil.php), héberge un groupe de travail « Exploration de corpus : pratiques et outils » (http://explorationdecorpus.corpusecrits.huma-num.fr) sur lequel on trouve recensés 19 outils pour l'écrit et 24 outils pour l'oral. Nous ne nous intéresserons ici qu'à la première catégorie (http://explorationdecorpus.corpusecrits.huma-num.fr/outils-logiciels-corpus-ecrits).

Si l'on convient de réunir l'essentiel des fonctionnalités mises à disposition par un outil logiciel dans les trois grandes familles suivantes, à savoir

i) l'**exploration** des corpus (visualisation, extractions, etc.),
ii) la mise en œuvre de **calculs statistiques** (calculs des spécificités, AFC, …),
iii) une **annotation** *en direct* par l'utilisateur du corpus au moyen d'étiquettes qu'il définit librement,

alors TXM et BTLC/Primestat permettent uniquement d'accéder aux deux premières familles de fonctionnalités. A cet égard (et à bien d'autres), ils s'approche beaucoup (mais avec des spécificités propres) des logiciels d'analyse textuelle *open-source*[174] que sont HYPERBASE (http://ancilla.unice.fr), LEXICO3 (http://www.lexi-co.com), IRaMuTeQ (http://www.iramuteq.org) et DtmVic (http://www.dtmvic.com). Ajoutons que certains outils proposent, outre les deux premières familles de fonctionnalités, une annotation en direct du corpus par l'utilisateur : tel est le cas de ANALEC par ex. (http://www.lattice.cnrs.fr/Telecharger-Analec).

Quelques mots maintenant à propos de chacun de ces deux outils.

[174] Nous ne citons pas ici ALCESTE par ex. à accès payant.

TXM. Créée fin 2008 à l'École Normale Supérieure de Lyon dans le cadre de l'ANR « textométrie » (2007-2010)[175], cette plateforme « open-source » développée par S. Heiden (http://textometrie.ens-lyon.fr/spip.php?article9) et son équipe d'abord à l'UMR ICAR (2007-2016) puis à l'UMR IRHIM (2016-…) a succédé au logiciel Weblex[176]. Disponible sous forme de portail web pour l'accès à la BFM, elle possède une version pour poste autonome utilisée pour la présente étude (version 07.7[177]). Comme les autres logiciels dédiés à l'analyse textuelle et librement accessibles, cette plateforme propose, outre des fonctionnalités d'exploration des corpus (concordances, retour plein texte, extraction de pivots…), des fonctionnalités d'analyse statistique (calcul des spécificités, calcul des cooccurrences, analyse factorielle des correspondances AFC, ….). Certaines de ces fonctionnalités mettent en jeu des calculs adossés à un modèle mathématique d'essence probabiliste et qui recourent à un indice statistique issu des spécificités de P. Lafon fondé sur la loi hypergéométrique. D'autres mettent en œuvre des méthodes classiques de la statistique descriptive multidimensionnelle : méthodes factorielles (AFC) et méthodes de classification (Classification ascendante hiérarchique, CAH). Dans une perspective élargie, il convient de rattacher le développement informatique de cette plateforme textométrique à l'histoire plus « longue » de la lexicométrie d'inspiration française, qui plonge ses racines dans les travaux fondateurs en statistique textuelle de P. Guiraud (1954, 1960) et de C. Muller (1973, 1977, 1992) puis des recherches menées au laboratoire « lexicométrie et textes politiques » de l'ENS Fontenay sous la direction de M. Tournier. A ce courant peuvent être rattachés à de nombreux égards les travaux de P. Lafon, L. Lebart, B. Pincemin, A. Salem, … Pour un historique partiel de la textométrie/statistique textuelle, voir J. Léon & S. Loiseau (éds)(2016). Comme ouvrage de référence sur la statistique textuelle, voir L. Lebart & A. Salem (1994).

BTLC/PrimeStat. Créée et développée d'abord à l'Université de Cologne par S. Diwersy (http://www.praxiling.fr/diwersy-sascha,405.html) dans le cadre des travaux conduits à l'Institut des langues romanes sous la direction de P. Blumenthal, cette plateforme est depuis 2015 développée à l'Université de Montpellier (UMR Praxiling). Elle propose la plupart des fonctionnalités d'exploration et d'analyse statistique offertes par TXM (avec des visualisations et des options de paramétrage différentes) mais aussi d'autres – par ex. l'échelonnement multidimensionnel (pour une présentation et utilisation, voir Blumenthal 2008 : 39 & sq.). Un grand avantage de cette plateforme est qu'elle permet de travailler sur de très gros corpus de plusieurs dizaines de millions de mots, ce que ne permet pas TXM. En termes enfin d'influence et d' « école », le développement de cette plate-forme s'inscrit davantage dans le courant des travaux du contextualisme britannique (pour une présentation générale, voir J. Léon : 2008, 2015), avec une influence particulière exercée par les recherches de M. Hoey (2005) sur l'amorçage lexical (lexical priming). Une dernière observation : TXM propose le langage de requête CQL (Corpus Query Language) tandis que BTLC/Primestat propose une interface avec des menus prédéfinis.

3.2. De quelques fonctionnalités particulièrement employées dans Presto

Fonctionnalités d'exploration du corpus. Nous avons déjà indiqué *supra* l'intérêt que revêt la **concordance** pour la recherche de pivots – qu'il s'agisse de tokens, de catégories morphosyntaxique, de motifs etc. C'est évidemment l'une des fonctionnalités majeures que

[175] Description du projet : http://textometrie.ens-lyon.fr/spip.php?rubrique115&lang=fr
[176] http://www.canal-u.tv/video/ecole_normale_superieure_de_lyon/instruments_et_resultats_presentation_de_weblex.5043
[177] Un manuel utilisateur de la version0.7. est disponible en ligne en français (http://textometrie.ens-lyon.fr/files/documentation/Manuel de TXM 0.7 FR.pdf).

nous avons utilisée, systématiquement doublée de la possibilité, précieuse, de faire correspondre au moyen d'un simple clic à une ligne de la concordance un extrait large du texte d'où elle est extraite, permettant ainsi une contextualisation.

Ainsi, en cliquant sur la première ligne de la concordance (TXM) dont on a proposé *supra* (figure 6) une capture d'écran, on obtient :

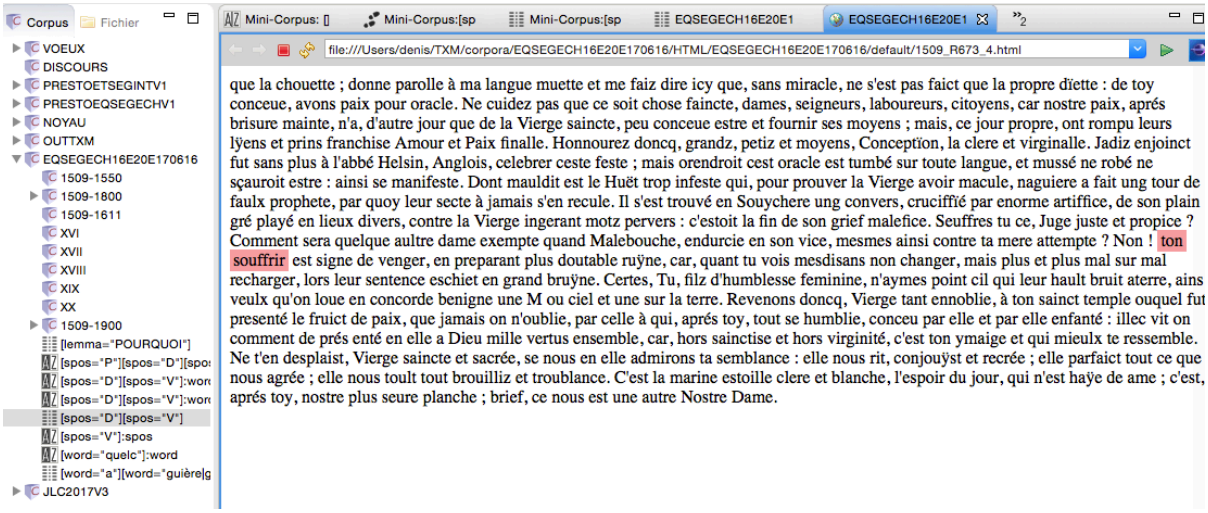


Figure 7
Capture d'écran sur TXM. Retour en plein texte à partir de la concordance présentée dans la figure 6.

Une autre fonctionnalité souvent utilisée en phase « préparatoire » d'études recourant à des fonctionnalités statistiques est la **progression** sur TXM. On en trouvera une illustration dans notre troisième partie (§ 1.1.2), dans le but d'observer l'évolution, décennie après décennie, de la fréquence absolue cumulée de la préposition *dans*.

Fonctionnalités statistiques. Dans le cadre de Presto, j'ai pour ma part utilisé uniquement le calcul des spécificités appliqué à une partition et le calcul des cooccurrences lexicales qui mettent tous deux en jeu les spécificités de P. Lafon. Je présente ces deux calculs dans la troisième partie à laquelle je me permets de renvoyer le lecteur. Je souhaite désormais étendre mon usage des fonctionnalités statistiques proposées par les deux plateformes à d'autres calculs, en particulier à l'Analyse Factorielle des Correspondances (AFC) et la Classification Ascendante Hiérarchique (CAH). Jusqu'à maintenant, je ne m'y suis jamais risqué seul, mais toujours « accompagné » par S. Diwersy. C'est ainsi que nous (S. Diwersy, A. Falaise, D. Vigier) avons présenté en juillet 2017 une communication aux *neuvièmes journées internationales de la linguistique de corpus* dans laquelle étaient analysées des classifications ascendantes hiérarchiques construites à partir d'une méthode décrite dans S. Gries & M. Hilpert (2008 ; 2012) sous le nom de *Variability-based Neighbor Clustering* (VNC). Ma préoccupation première vis-à-vis de ces fonctionnalités statistiques est d'en comprendre le plus en profondeur possible (compte-tenu de ma formation de linguiste et non de statisticien) l'esprit et le champ d'application.

3.3. Conclusion

En guise de conclusion à cette (courte) troisième section, je souhaiterais risquer quelque mots sur ce qui me paraît être, aujourd'hui, une collaboration fructueuse entre linguistes, informaticiens et statisticiens.

Dans la mesure où mon expérience dans ce domaine est encore fraîche, je me garderai de tout jugement définitif, préférant avancer quelques idées simples (peut-être trop ?) à propos de quelques points qui m'ont particulièrement frappé durant ces quatre années où mon activité de recherche a entièrement été dominée par Presto.

Le premier élément est qu'une telle collaboration, pour qu'elle puisse vivre et porter ses fruits, est avant tout subordonnée (mais n'est-ce pas toujours le cas ?) à des facteurs *humains* : les chercheurs qui y sont impliqués doivent entretenir entre eux, sinon de l'amitié, du moins une forme de sympathie, voire parfois d'empathie pour se comprendre et progresser ensemble.

Le second élément, plus polémique peut-être, est que le format « programme de recherche financé sur objectifs » (ANR, programmes européens, …) me paraît constituer aujourd'hui (mal gré que j'en ai à certains égards) un excellent moyen pour réunir, sur une période de temps relativement courte, des chercheurs appartenant à des équipes internationales qui se mobilisent sur un projet commun. Tirant de ce projet suffisamment de force et de motivation pour se colleter avec la masse immense des défis que constitue une recherche en linguistique empirique, ils parviennent en peu d'années à aboutir à des résultats publiés et tangibles pour la communauté.

Enfin, et c'est un point plus personnel, les linguistes impliqués dans de tels groupes doivent avoir la fibre mathématique… Pour ma part, je l'ai toujours eue, et comme j'y reviendrai à la fin de ce mémoire, j'ai la ferme intention au cours des années d'activité qui me restent, de développer le plus possible – en lien étroit avec certains projets de recherche à venir qui me tiennent à cœur - mes connaissances dans le domaine de la statistique textuelle.

Conclusion de la deuxième partie

La première section de cette deuxième partie, consacrée aux grandes étapes qui rythment la construction d'un corpus historique, a permis de s'arrêter sur la question centrale de la représentativité, traitée non seulement dans la perspective des corpus synchroniques destinés aux études des usages contemporains d'une langue mais aussi dans la perspective des corpus historiques. Une tentative d'évaluation de la qualité actuelle du corpus Presto, en termes de continuité temporelle et générique et de variété des genres discursifs et des auteurs, a été proposée.

De la deuxième section consacrée à l'annotation, on retiendra l'enjeu que constitue le recours à une tokenisation homogène pour toutes les périodes du corpus. Concernant la construction du lexique, la collaboration avec G. Souvay de l'ATILF a permis, grâce aux règles d'archaïsation tirées de LGeRM, d'accroître de façon spectaculaire la couverture du lexique Presto pour le français pré-classique et classique. La communauté disposera ainsi bientôt d'un logiciel d'étiquetage (morphosyntaxe et lemmes) panchronique du français (XVI^{e} s. - XX^{e} s.), *open source*, doté d'une performance tout à fait satisfaisante. Enfin, la réflexion et les travaux conduits sur la segmentation des amalgames *au, aux, és, ès, es* en français préclassique a mis en lumière certaines difficultés encore jamais traitée, nous semble-t-il, dans les travaux de linguistique quantitative en diachronie.

Le caractère plus modeste (en nombre de lignes) de la troisième section ne doit enfin pas atténuer l'importance majeure que les deux plateformes utilisées dans Presto ont revêtu pour nos travaux.

La troisième partie de ce mémoire se propose de rendre compte des travaux que nous avons menés sur le corpus Presto, en nous aidant des calculs de spécificité et de cooccurreence lexicale disponibles sur TXM et BTLC/Primestat.

TROISIÈME PARTIE

Les prépositions *en, dans, dedans* du XVI^{e} s. au XX^{e} s. Approche statistique en corpus

RAPPEL DE LA TABLE DES MATIÈRES DE LA TROISIÈME PARTIE

Introduction

Nous nous proposons d'étudier dans le corpus étiqueté et lemmatisé Presto[178] l'usage des prépositions *en, dans, dedans* entre le XVI^e^ et le XX^e^ s. à partir d'une étude probabiliste de leurs cotextes d'apparition. L'hypothèse directrice[179] qui sous-tend ce travail et sur laquelle nous reviendrons *infra* considère que l'étude statistique des environnements distributionnels (proches et distants) des prépositions livre des informations précieuses sur l'évolution de ces usages en diachronie.

Sur le plan linguistique, bien des raisons plaident pour que l'on traite dans un même ensemble les trois prépositions *en, dans, dedans*.

Leur origine étymologique commune d'abord[180] : toutes - au premier rang desquelles *en* - sont issues de la préposition latine *in*. *Dedans* vient de l'adverbe latin *intus*, composé de *in* et du suffixe *–tus* qui a donné en ancien français la forme *enz* le plus souvent employée comme adverbe (1), *enz en* jouant le rôle de préposition (2). Par allongement préfixal se sont formés les termes *deenz, denz, dens* qui pouvaient être adverbe[181] ou préposition.

(1) *Quant il furent venuz a lor nef si entrerent* ***enz*** (*La Queste del Saint Graal*, 1220, v. 1-2, p. 229)
(2) ***Enz en*** *l'oreille li conseilla souef: « Amis biaux frere, ou est Gombaus reméz?*» (*Ami et Amile*, 1200, v. 345-346, p. 12)

Par redoublement de la préfixation, on a abouti aux formes *dedenz, dedens, dedans* qu'on rencontre dans les textes dès le XII^e^ s. (3), éventuellement aussi associées à *enz* (4) :

(3) ***Dedens*** *son lit se rest assise* (Gautier d'Arras , *Eracle*, 1180, v. 183, p. 6)
(4) *Entrerent* ***enz dedenz*** *le mur Qui tuz ert faiz de cristal dur.* (Benedeit, *Voyage de saint Brendan*, 1100, v. 271-272, p. 37)

Deux hypothèses étymologiques sont en général avancées pour expliquer l'origine de *dans* : la première y voit un mot directement hérité de l'adverbe de l'ancien français *denz*. La seconde propose de dériver *dans* de la forme *dedans* par réduction, sur le modèle de *sous* en face de *dessous* etc.

Sur le plan sémantique, *en, dans* et *dedans* sont toutes trois aptes à opérer aux XVI^e^ et XVII^e^ siècles[182] une localisation spatiale analogue, à savoir localiser une cible dans les

[178] Pour une présentation détaillée du corpus Presto, voir la partie II. Dans cette étude, nous recourrons presque systématiquement au niveau « contrôlé » et à la version « échantillonnée » du corpus. Pour certaines études plus ciblées, nous recourrons au corpus étendu. En ce cas, nous le signalerons explicitement.

[179] Le présent travail s'inscrit – du point de vue de la méthode et des outils d'exploration et de calcul - dans le sillage des recherches menées par l'équipe de P. Blumenthal à l'Université de Cologne (voir par ex. P. Blumenthal 2007, 2011a, 2011b, 2013 ; P. Blumenthal, I. Novakova & D. Siepmann (éds) (2014)), et des travaux conduits dans plusieurs programmes dont EMOLEX (http://www.emolex.eu ANR & DFG 2009-2012) et Presto (ANR & DFG 2013-2017). Au cours de cette étude (voir § 2.), nous reviendrons sur cette « hypothèse directrice » pour en expliciter certains des fondements et la resituer dans le contexte des « linguistiques de corpus ».

[180] Pour plus de détails sur ce point, on se reportera à B. Fagard et L. Sarda, 2009.

[181] Nous maintenons dans ce travail (pour des raisons toutes pratiques : voir note suivante) la distinction traditionnelle entre emploi comme préposition (*il a voté contre ce candidat*) *versus* comme adverbe (*il a voté contre*).

[182] Période où *dedans* possède un fonctionnement prépositionnel suivi d'un régime nominal, qui perdurera (mais avec un usage de plus en plus rare) encore jusqu'au XVIII^e^ s. L'absence actuelle d'annotation syntaxique

frontières d'un site au terme (5-7) ou non (8-10) d'un franchissement de frontières[183].

(5) *Quant le mary veid qu'il en avoit bien faict son debvoir, entra **en la chambre** et le mercia de la peyne qu'il en avoit prinse* (1550, M. de Navarre, *l'Heptameron*)
(6) *Le mary **dedans la chambre** entre, prest de donner ventre sur ventre*. (1549, Anonyme, *Sottie pour le cry de la basoche*)
(7) *Le president entra **dans la chambre** et trouva sa femme et nicolas couchez ensemble*. (1550, M. de Navarre, *L'Heptameron*)
(8) *Mais tousjours demouroit **en la nef** entre les femmes* (1532, F. Rabelais*, Pantagruel)*
(9) *Or on ne peut cognoistre cela aux poissons, car on ne peut sçavoir leur aage, d'autant qu'ils vivent **dedans l'eau**.* (1556, B.-G. Gelli, *Les discours fantastiques de Julien Tonnelier*)
(10) *Ce pendant qu'on est en ce monde, On est **dans une mer profonde** (...)* (1587, Pierre de l'Estoile, *Registre-journal du regne de Henri III*)

Au-delà du XVII[e] s., *en* et *dans* (de même que *dedans* en emploi absolu, ou « adverbial ») continuent jusqu'en français contemporain à partager la capacité d'opérer ce type de localisation spatiale, mais de façon beaucoup plus marginale pour *en* dont la « valeur intrinsèque » (G. Gougenheim, (1950, [1970]) aurait migré vers des sens plus abstraits amplement étudiés depuis G. Guillaume (1919, [1975]).[184] D'après G. Gougenheim (*op. cit.*), ce déclin des emplois de *en* à valeur spatiale, étroitement lié au développement des usages de *dans* dans la seconde moitié du XVI[e] s., signalerait un changement du « *centre de gravité de la préposition* »[185] (56).

Outre le domaine spatial, ces trois prépositions peuvent partager aussi entre le XVI[e] s. et le XVIII[e] s. des valeurs sémantiques communes sur le plan temporel. Ainsi le futur peut-il par exemple être marqué par *dedans* ou *dans* :

(11) ***Dans demain** Nous deux mettrons icy la main Et ferons l'aoust sans ayde aucun.* (1542, G. Corrozet, trad. des *Fables d'Ésope*).
(12) (…) *ains vous convient sortir de la Thrace **dedans demain*** (…) (1557, A. Olvido, *Amadis de Gaula Traduit nouvellement d'Espagnol en Francoys*, livre 6).

Quant à la distinction moderne entre ***dans** trois jours / **en** trois jours,* on sait qu'elle a tardé à se fixer[186] puisque l'on trouve encore au XVIII[e] s. - chez Montesquieu par exemple – une alternance *en/dans* en tête de SP aspectuels exprimant l'intervalle de durée nécessaire à l'effectuation intégrale du procès dénoté par le verbe et ses compléments argumentaux (prérogative réservée uniquement à *en* en français standard contemporain) :

(13) *Il est plus facile à un Asiatique de s'instruire des mœurs des François **dans un an**, qu'il ne l'est à un François de s'instruire des mœurs des Asiatiques **dans***

---

disponible dans Presto nous a conduit à adopter, pour l'étiquetage de *dedans*, la terminologie traditionnelle de la grammaire (emploi sans régime = emploi adverbial) afin de bien séparer, lors de nos calculs, les usages de *dedans* suivis d'un régime (= préposition) de ceux où le régime de *dedans* est nul (= adverbe).

[183] Nous avons souligné dans la première partie de ce mémoire l'insuffisance d'une approche géométrique et topologique visant à décrire les valeurs sémantiques des prépositions. Nous l'adopterons provisoirement dans ce travail mais nous y reviendrons, dans notre partie finale « perspectives ».

[184] Voir notre article introductif au numéro 178 de *Langue Française* (D. Vigier, 2013).

[185] Thèse reprise et prolongée par W. de Mulder (2008) et W. de Mulder, D. Amiot (2013)

[186] Sur ce point, voir N. Fournier & D. Vigier (2017).

quatre. (1721, Montesquieu, *Lettres Persanes*).

Enfin, en français moderne, *en, dans, dedans* demeurent étroitement liés dans certaines de leurs conditions d'usage, même si leurs cotextes d'emploi ont considérablement changé par rapport au XVI^{e} s.

(14) *J'ai voyagé **en train**.*
(15) *J'ai voyagé **dans un train bondé**.*
(16) *Ce train bondé, j'ai voyagé **dedans**.*

Dans peut sous certaines conditions se substituer à *en* devant régime nominal actualisé par un déterminant comme dans (15); on croise là un principe définitoire adopté par E. Spang-Hansen (1963) pour distinguer les prépositions incolores (comme *en)* des colores (comme *dans*) :

> [...] nous proposons de définir les prépositions incolores comme les prépositions que la détermination plus précise d'un des termes reliés peut faire échanger contre d'autres prépositions (simples). (op. cit.: 21).

Dans (14), le SP *en train* permet de typer le procès exprimé par le verbe en le situant dans un paradigme de *manières-types* de voyager (*voyager en train* contraste avec *voyager en avion, en voiture, ...*). Ce typage est rendu possible par le blocage référentiel du régime nominal de *en* (absence de déterminant devant *train*) qui conduit à une saisie abstraite, qualitative, de ce dernier. Dans (15) en revanche, l'adjonction d'une expansion adjectivale (*bondé*) au nom *train* conduit à une particularisation du référent qui rend nécessaire sa saisie référentielle : d'où la substitution de *dans* à *en*[187] :

(14') ***J'ai voyagé **en train bondé***

Quant à *dedans*, il se présente (mais de manière non-systématique : voir A. Berthonneau, 1999) dans (16) du fait de la dislocation à gauche du SN, qui conduit la préposition à être suivie d'un régime nul - d'où l'emploi de *dedans* et non de *dans.*

Voilà à nos yeux de nombreuses raisons qui autorisent qu'on étudie ici ensemble les prépositions *en, dans, dedans*.

Dans la première section de cette troisième partie, nous traiterons de la « naissance » de la préposition *dans* en français préclassique. Nous brosserons d'abord un tableau d'ensemble de l'évolution quantitative des usages de *dans* et de *en* entre 1551 et 1940. Puis nous examinerons l'hypothèse avancée en 1885 par A. Darmesteter et qui a toujours cours dans les études diachroniques sur *en,* hypothèse visant à expliquer l'impressionnante montée en puissance fréquentielle de *dans* dans les textes littéraires du moins, à partir de 1550 environ. Nous montrerons que l'étude statistique en corpus des préférences combinatoires que cette préposition manifeste en français préclassique conduit à rejeter une partie de l'hypothèse de cet auteur au profit d'une autre voie d'explication possible.

La deuxième section, d'ordre méthodologique, s'arrêtera sur certains des principes qui fondent l'approche statistique et distributionnelle du sens développée dans cette troisième

[187] Toute expansion adjointe au nom régime de *en* ne déverrouille pas nécessairement le blocage référentiel du SN. Par exemple, l'adjonction du complément *de nuit* permet une typification procès avec *en* : *Je voyage en train de nuit* (voir P. Cadiot, 1997 : 217).

partie de notre mémoire. Nous nous intéresserons en particulier au principe suivant lequel le cotexte distributionnel d'une unité constitue une des voies d'accès privilégiée à son sens, ainsi qu'à l'épineuse question de l'interprétation (qualitative) de résultats quantitatifs fournis par certaines fonctionnalités statistiques disponibles sur la plateforme TXM (en particulier, le calcul de cooccurrence lexicale fondé sur le calcul des spécificités de P. Lafon (1980, 1984)).

Dans la dernière partie, nous mettrons au jour quelques grandes tendances dans l'évolution des usages des prépositions *en, dans, dedans* entre le XVI^{e} s. et le XX^{e} s. à partir d'une étude comparative de leurs cooccurrents nominaux les plus spécifiques dans notre corpus. Nous mettrons en lumière notamment certaines reconfigurations distributionnelles et sémantiques auxquelles la préposition *en* a été conduite, du fait de l'accroissement spectaculaire de l'usage de *dans* à partir de la première moitié du XVII^{e} s. Nous serons ainsi conduit à examiner la thèse de G. Gougenheim ([1950] 1970) selon qui *en* aurait vu son « centre de gravité sémantique » se déplacer entre le XVI^{e} s. et aujourd'hui.

1. Naissance de la préposition *dans*

1.1. Tableau général de l'évolution quantitative des usages de *en* et de *dans* entre 1501 et 1940

Avant de nous pencher sur la période temporellement limitée où l'on assiste au premier accroissement significatif de la fréquence d'usage de *dans* en français dans le discours littéraire, qu'on nous permette de brosser dans un tableau général l'évolution des usages de *en* et de *dans* sur cinq siècles environ. Pour ce faire, nous recourrons au calcul des spécificités[188] disponible sur la plateforme open source TXM[189] (calcul mis au point par P. Lafon (1980, 1984)) en l'appliquant à une partition[190] opérée sur[191] Presto$^{1551\text{-}1940}$. Précisons cursivement en quoi consiste ce calcul.

1.1.1. Le calcul des spécificités sur TXM

Le calcul des spécificités implémenté dans TXM permet de porter un jugement statistique sur la sous-fréquence f_{ij} d'une forme *i* dans une partie *j* d'un corpus, étant connus par ailleurs les paramètres suivants :

F_i : fréquence de la forme *i* dans le corpus ;
t_j : taille de la partie j ;
T : taille du corpus.

Les valeurs f_{ij}, F_i, t_j, T ayant été définies, est alors calculé un indice de spécificité dont la valeur numérique peut être de signe positif ou négatif. Plus la valeur absolue de cet indice est élevée, plus la chance que la sous-fréquence f_{ij} observée pour la forme *i* dans la partie *j* soit

[188] Pour une présentation détaillée du calcul des spécificités, voir L. Lebart et A. Salem (1994 : 172-185).
[189] http://textometrie.ens-lyon.fr. Je remercie B. Pincemin pour l'ensemble de ses conseils et de ses remarques relatifs aux fonctionnalités statistiques auxquelles cette plateforme donne accès.
[190] Par *partition* d'un corpus de textes, il faut entendre sa « division (…) en parties constituées par des fragments de texte consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus. » (L. Lebart & A. Salem, *op. cit.*: 316)
[191] Par convention, les dates figurant en exposants après le nom Presto stipulent les bornes temporelles du sous-corpus découpé dans le corpus intégral.

due au hasard est faible[192]. Un indice supérieur ou égal à 3 signale que la fréquence f_{ij} observée est significativement élevée par rapport à ce que le modèle hypergéométrique laissait prévoir. On dira donc que la forme *i* est *spécifique positive* pour la partie *j*. Inversement, un indice inférieur ou égal à -3 signale que la forme est sous-représentée par rapport à ce que le modèle laissait prévoir : on parlera de *spécificité négative*. Enfin, un indice situé entre -3 et +3 sera considéré comme inférieur au seuil de significativité statistique[193] et la forme *i* sera dite *banale* pour la partie *j*.

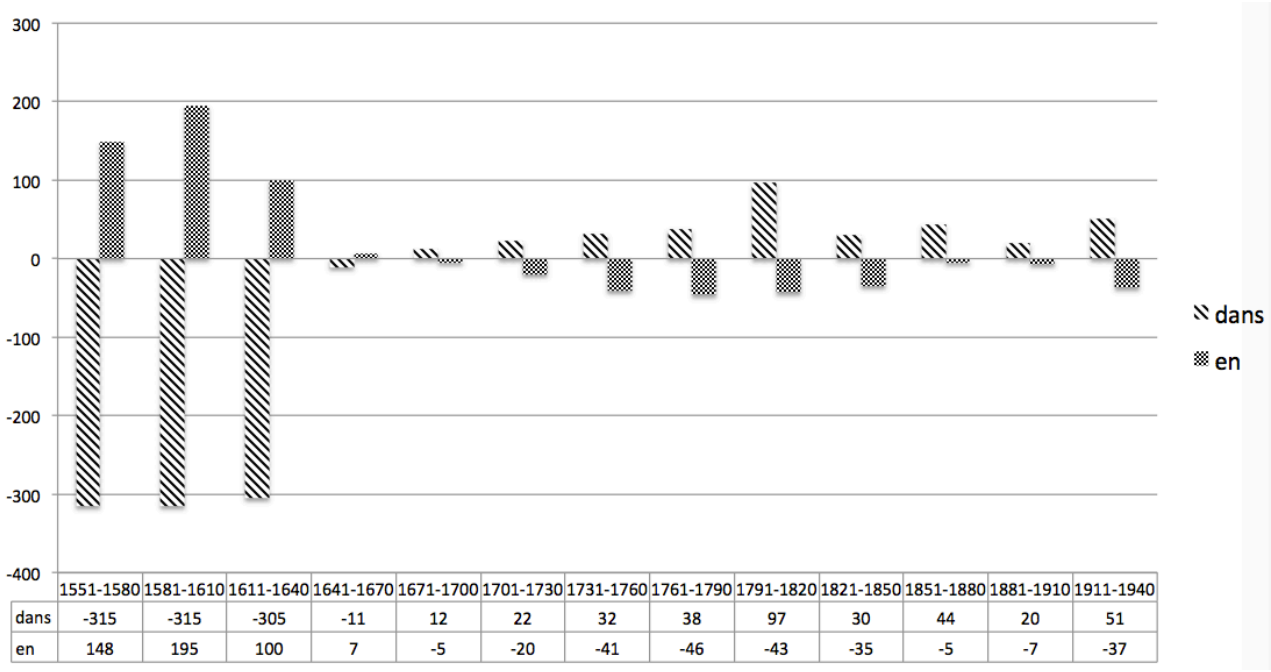
Pour le calcul visé, nous avons partitionné le corpus Presto^{1551_1940} en treize parties d'empan temporel identique, soit trente années. Notre objectif consistait à déterminer si les prépositions *en* et *dans* s'avéraient, dans l'une ou l'autre de ces parties, spécifiques positives, spécifiques négatives ou banales. Voici les paramètres définis pour ce calcul :

f_{ij} = sous-fréquence de la préposition sélectionnée pour le calcul (*en* ou *dans*) dans chacune des treize parties*;*
F_i = fréquence de cette même préposition dans le corpus total (Presto^{1551_1940}) ;
t_j = sous-fréquence de tous les mots[194] dans chacune des treize parties ;
T= fréquence tous les mots dans Presto^{1551_1940}.

1.1.2. Présentation et analyse des résultats

Voici les résultats obtenus.[195] Au diagramme est associée la table des valeurs (scores de spécificités) calculées.

[192] Le calcul des spécificités de Lafon est fondé sur la loi hypergéométrique, loi de probabilité correspondant à un tirage sans remise dans l'hypothèse d'indépendance (voir L. Lebart & A. Salem, 1994 : 174 ; C. Labbé & D. Labbé : 1994).
[193] Pour une discussion relative à la détermination de ce seuil, voir C. Reutenauer (2002 : 161).
[194] Par « mot », il faut entendre toute unité résultant de l'opération de segmentation automatique du flux textuel accomplie sur le corpus Presto.
[195] A. Salem (1988) a montré que le calcul des spécificités appliqué à des suites textuelles chronologiques pouvait gagner à être complété par le recours à des indicateurs (triangle des spécificités connexes, coefficient de Von Neumann, …) permettant de mieux appréhender la ventilation d'une forme dans les *n* parties du corpus. Ces indicateurs s'imposent lorsque le diagnostic de spécificité porté sur la sous-fréquence d'un terme dans une ou plusieurs parties d'une série chronologique aboutit à un jugement de banalité. Notre étude n'entre pas dans ce cas de figure puisque aucune sous-fréquence de *en* ou de *dans* n'apparaît comme statistiquement banale dans notre partition.



	1551-1580	1581-1610	1611-1640	1641-1670	1671-1700	1701-1730	1731-1760	1761-1790	1791-1820	1821-1850	1851-1880	1881-1910	1911-1940
dans	-315	-315	-305	-11	12	22	32	38	97	30	44	20	51
en	148	195	100	7	-5	-20	-41	-46	-43	-35	-5	-7	-37

Diagramme 1

Évolution des scores de spécificité affectés aux prépositions *en* et *dans* entre 1551 et 1940. Corpus Presto partitionné en 13 tranches de 30 ans

L'examen des spécificités positives et négatives dans ce tableau conduit aux interprétations suivantes : tandis que *dans* voit ses emplois gagner un terrain considérable en cinq siècles[196], ceux de *en* s'érodent tout aussi considérablement. *Dans* connaît en outre un accroissement fréquentiel très net à la fin de la première moitié du XVII^{e} s., comme le montre la chute de la valeur absolue de son score de sous-spécificité pour la période 1641-1670. Parallèlement, *en* connaît une baisse continue de sa fréquence d'emploi, avec une chute particulièrement sensible entre 1611 et 1670. Enfin, la période 1641-1700 apparaît comme charnière puisque *dans* et *en* voient leurs zones de spécificités s'inverser. À l'aube du XVIII^{e} s., le sort des deux prépositions est scellé : *dans* passe définitivement dans la zone de spécificité positive et *en* dans la zone négative.

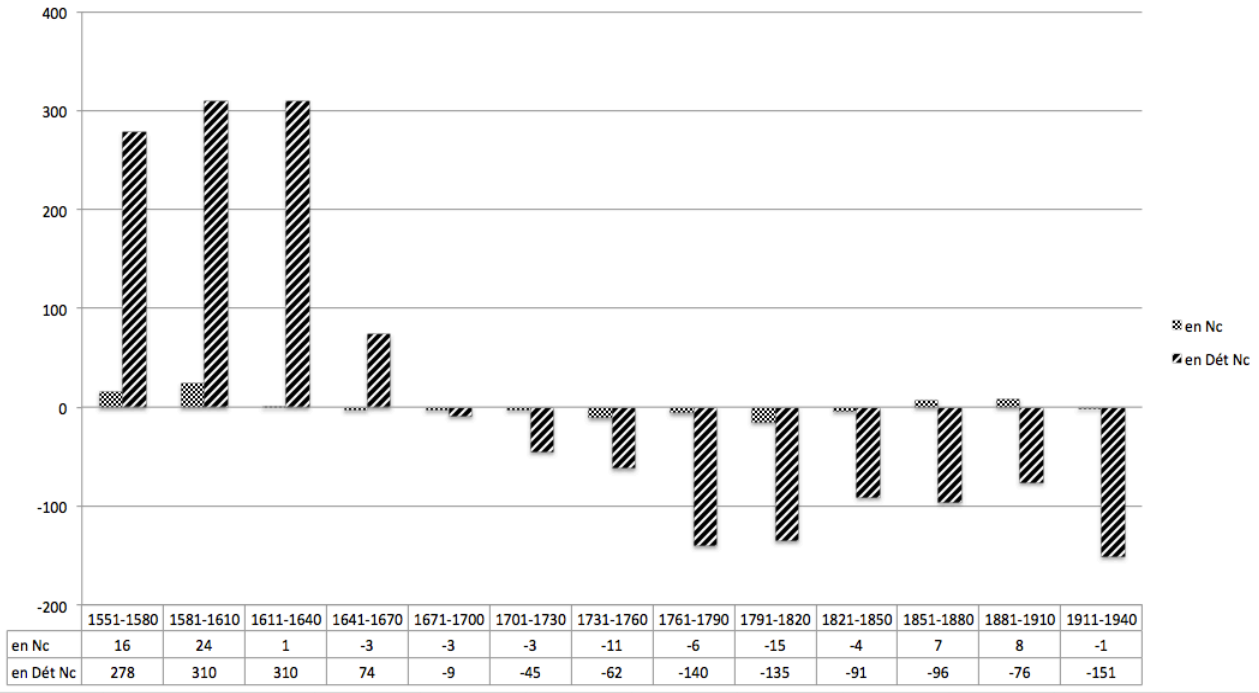
On peut examiner de plus près la trajectoire des emplois de *en* dans notre corpus en distinguant les cas où il est suivi d'un déterminant et d'un nom commun (*En Det Nc*) *versus* d'un nom commun nu (*En Nc*). Pour ce calcul, les paramètres sont :

f_{ij} = sous-fréquence observée dans chacune des treize parties de la séquence S_1 « en *suivi d'un déterminant et d'un nom commun* » ou de la séquence S_2 « en *suivi d'un nom nu* »;

[196] Cette tendance générale ne doit pas masquer des variations plus locales des valeurs du score calculé, pour lesquelles des études de détail seraient nécessaires et qui mettraient au jour certains effets de corpus. Par ex., le pic de +97 pour la période 1791-1820 s'explique en partie par la présence, dans cette tranche chronologique, d'un ouvrage de médecine (R. T. Laennec, *De l'auscultation médiate*) qui traite de l'invention du stéthoscope. Comme chaque bruit entendu *via* ce nouvel instrument doit être systématiquement localisé (concrètement – *dans* un organe – ou abstraitement – *dans* une maladie), le scripteur est conduit à sur-utiliser la préposition *dans* par rapport aux autres textes figurant dans cette même période, et même dans une période plus large (1750-1850). D'où une augmentation significative des emplois de *dans* pour cette tranche qui se traduit par un accroissement local du score de sur-spécificité.

F_i = fréquence de la séquence S_1 ou de la séquence S_2 dans le corpus total (Presto^{1551_1940}) ;
t_j = sous-fréquence de tous les mots dans chacune des treize parties ;
T= fréquence de tous les mots dans Presto^{1551_1940}.

Voici le diagramme des spécificités auquel on aboutit :



	1551-1580	1581-1610	1611-1640	1641-1670	1671-1700	1701-1730	1731-1760	1761-1790	1791-1820	1821-1850	1851-1880	1881-1910	1911-1940
en Nc	16	24	1	-3	-3	-3	-11	-6	-15	-4	7	8	-1
en Dét Nc	278	310	310	74	-9	-45	-62	-140	-135	-91	-96	-76	-151

Diagramme 2

Évolution des scores de spécificité affectés aux séquences *en Det Nc* et *en Nc* entre 1551 et 1940. Corpus Presto partitionné en 13 tranches de 30 ans

Les évolutions des séquences *En Det Nc* et *En Nc* sont clairement distinctes. La première se voit affecter des scores de spécificités positifs très élevés entre 1551 et 1670 (sur-représentation prononcée), puis des scores significativement négatifs à l'aube du XVIII^{e} s. Les scores de la seconde séquence oscillent quant à eux autour de l'axe des abscisses, leur valeur étant alternativement positive puis négative et demeurant à partir du début du XVII^{e} s. dans une frange de valeurs absolues qui excède rarement 10. En d'autres termes, tandis que la séquence S_1 « en suivi d'un déterminant et d'un nom commun » subit une chute très sévère dans son usage au cours de la seconde moitié du XVII^{e} s., l'emploi de *en* suivi d'un nom commun non déterminé, quoique soumis lui aussi à une certaine désaffection entre 1640 et 1850 environ, se maintient dans une amplitude de variation fréquentielle limitée et voit même son score de spécificité se redresser au cours du XIX^{e} siècle.

Récapitulons : l'étude quantitative du corpus Presto^{1551_1940} montre que *dans* y connaît un accroissement majeur de ses emplois vers la fin de la première moitié du XVII^{e} s. A l'inverse, la préposition *en* dont l'emploi s'érode globalement entre le début du XVII^{e} s. et nos jours, voit son usage avec un régime nominal déterminé subir un ralentissement spectaculaire au cours du XVII^{e} s. Suivie d'un nom commun *nu*, sa fréquence d'emploi connaît en revanche un ralentissement nettement moindre avec même une phase d'embellie au cours du XIX^{e} s. qu'il faudrait confirmer sur d'autres corpus.

On sait enfin grâce à de nombreuses études sur corpus concordantes, que l'usage de *dans* a commencé à s'accroître de manière remarquable dès 1550 [197], du moins dans le discours littéraire. La figure 1 ci-dessous, qui reproduit une graphique de progression accompli sur TXM, permet de suivre la fréquence absolue cumulée de *dans* au XVI^e^ s. On y observe que dans notre corpus, cette préposition commence à être timidement employée autour de 1530 puis de façon beaucoup plus significative à partir de 1570 environ.

C'est à cette première hausse fréquentielle de *dans*, située en plein cœur du XVI^e^ s., que nous allons maintenant nous intéresser.

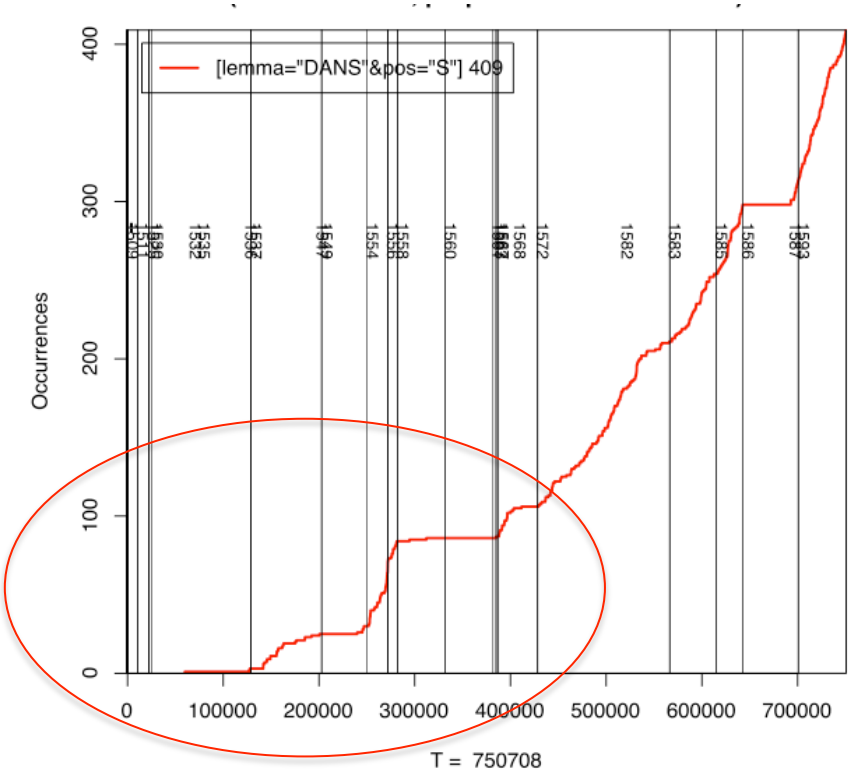


Figure 1
Graphe de progression obtenu sur TXM permettant d'observer l'évolution, décennie après décennie, de la fréquence absolue cumulée de la préposition *dans*. Corpus Presto–XVI

1.2. Fortune de *dans* à partir de 1550 : l'hypothèse de Darmesteter (1885)

On peut considérer pour acquis (voir les nombreuses études sur corpus qui toutes s'accordent sur ce point : A. Darmesteter 1885 ; G. Gougenheim [1945] 1970, [1951] 1974 ; F. Brunot 1967 ; B. Fagard & L. Sarda 2009 ; etc.) que c'est à partir du début de la seconde moitié du XVI^e^ s. que *dans* fait une entrée remarquable sur la scène des prépositions du français – du moins dans le discours littéraire. C'est en effet autour de 1550 que l'usage de cette préposition – qui auparavant « traîn[ait] une existence obscure » – s'accroît de manière spectaculaire, *dans* se voyant ainsi appeler « au plus brillant succès » (A. Darmesteter, 1885 : 185) pour figurer finalement, en français contemporain, au cinquième rang en termes de

[197] Ce premier « saut » demeurant insuffisant en termes de fréquences relatives pour être sensible dans les calculs de spécificités sur les partitions dont les résultats ont été présentés plus haut (diagramme 1).

fréquence d'emploi derrière *de, à, en* et *pour*[198]. Comment expliquer une fortune si étonnante et si rapide ?

1.2.1. Bref rappel de la situation des amalgames issus de la combinaison de *en* avec les formes de l'article défini *le* et *les* au XVI[e] siècle

Dès le Moyen Âge, *en* placé devant les formes de l'article défini *le* et *les* se contracte pour donner des formes amalgamées. En revanche, il ne se produit aucune contraction devant les formes *la* et *l'* de ce même article ni devant les autres déterminants. Une telle situation est parallèle à celle observée pour les prépositions *de* et *à* combinées avec les mêmes formes de l'article défini, qui donnent à partir du Moyen Âge respectivement les formes *del, deu, dau, du…* (< *de*+ *le)* et *dels, daus, des, …* (< *de* + *l*es), ainsi que les formes *al, au* (< *à*+ *le)* et *als, as, aus, aus, aux, …* (< *à*+ *les)*

Les formes amalgamées morphologiquement issues de *en* + *le/les* et les graphies correspondantes qu'on peut observer dans les textes entre les IX[e] s. et XVI[e] s. sont nombreuses. En partant des dictionnaires F. Godefroy (1891-1902) et E. Huguet (1925-1967), de la récente grammaire de S. Lardon & M.-C Thomine (2009) ainsi que de nos observations sur le corpus Presto, on peut proposer la liste (non-exhaustive) suivante :

en + *le* > *enl, el, eu, u, o, hou, hu, ou, on, om, un*
en + *les* > *ens, ans, eins, ons, eis, eus, aus, as, es, ès, és, ez*

A l'aube du XVI[e] s., certaines de ces formes furent senties comme archaïques[199] (par ex. *on* qu'on trouve chez Rabelais), d'autres s'éteignirent rapidement (*ou* par ex. n'est plus usité vers le milieu du siècle) voire avaient déjà disparu (la forme *enl* n'est attestée ni dans Frantext (1501-1600), ni dans la base *Epistemon* des BVH[200], ni dans notre corpus). On soulignera à l'inverse la relative résistance qu'opposèrent les amalgames *es, ès, és* à ce processus d'élimination : ils réussirent à traverser le siècle pour être presque définitivement abandonnés - sauf constructions archaïsantes ou figées – dans le courant du XVII[e] s.

Il se produisit en outre un fait majeur: la disparition progressive des amalgames morphologiquement issus de *en* entamée bien avant 1500 entraîna leur remplacement « soit par la préposition *dedans*, soit par *a* contracté avec l'article masculin : *au* (qui sonne presque comme *ou* = o) et *aux*. » (F. Brunot, *op. cit.*: 278). La situation devint alors la suivante au XVI[e] s. : « *Au* et *aux* se trouvaient donc correspondre à la fois à *à* + *le, à* + *les* d'une part et à *en* + *le, en* + *les* d'autre part. » (G. Gougenheim, 1951, [1974] : 182).

Voici quelques exemples[201] illustrant l'ambiguïté potentielle des morphèmes *au, aux* :

(17) *Face le ciel (quand il voudra) revivre / Lisippe, Apelle, Homere, qui le pris / Ont emporté sur tous humains espris / **En la** statue, **au*** tableau*, et **au** livre.* (1550, Du Bellay, *L'Olive*) (cité par S. Lardon & M.-C Thomine, *op. cit.*: 393)

(18) ***Aux** braves exploits de sa vie, et **en sa** mort, on le* [=Caton] *sent toujours monté sur ses grands chevaux.* (Montaigne, Essais, III) (cité par Gougenheim, *op. cit.*: 183)

(19) *Car ils profitent davantagte, et aident mieux la digestion qui se fait **en***

[198] Voir liste des 1063 mots les plus fréquents du français oral élaborée par le Centre du Français Elémentaire (G. Gougenheim *et al.* 1956) et C. Vaguer (2008 : 23).
[199] Sur ce point, voir entre autres F. Brunot (*op. cit.*: 277)
[200] http://www.bvh.univ-tours.fr/Epistemon/index.asp
[201] Dont certains ont déjà été présentés dans notre deuxième partie.

l'estomach, ***au*** *foye, et* ***és*** *veines*, (1589, J. le Paulmier, *Traité du vin et du sidre*, [Trad.])

(20) *Et puis à quel usage les deschiremens et desmembremens des Corybantes, des Menades, et, en noz temps, des Mahometans qui se balaffrent les visages, l'estomach, les membres, pour gratifier leur prophete, veu que l'offence consiste* ***en*** *la volonté, non* ***en*** *la poitrine,* ***aux*** *yeux,* ***aux*** *genitoires,* ***en*** *l'embonpoinct,* ***aux*** *espaules et* ***au*** *gosier.* (1582, M. de Montaigne, *Essais*)

La situation de confusion liée à l'ambiguïté sémantique des morphèmes *au, aux* était encore accrue du fait que la forme amalgamée *ès* pouvait (en particulier chez F. Rabelais) correspondre aussi à un amalgame équivalant sémantiquement à *à + les*. (Voir notre deuxième partie, § 2.2.4.2.& § 2.2.4.3)

1.2.2. Présentation de l'hypothèse d'A. Darmesteter

Cette thèse formulée en 1885 a été reprise par G . Gougenheim (*op. cit.*), F. Brunot (*op. cit.*) et L. Terreaux (1968) entre autres, et prévaut encore dans la littérature linguistique actuelle[202] sans être toujours cependant formulée dans ses détails[203]. On peut y distinguer deux volets. Le premier a trait à la cause de la fortune fréquentielle de *dans* à la fin de la première moitié du XVI^{e} s. : d'après l'auteur, la préposition *dans* aurait pris la place laissée vide par la disparition des amalgames morphologiquement issus de **en le, *en les* provisoirement remplacés par les morphèmes *au, aux*.

> « Là où nous employons *dans*, le Moyen Âge disait *ou, es* : *ou champ, es champs, es circonstances*. Ainsi *en* s'est maintenu dans les cas où il n'y avait pas lieu de le combiner avec l'article ; *dans* s'est substitué à *en* dans ceux où *en* se contractait avec *le* et *les*. (…) *Dans* est venu prendre la place laissée vide par la disparition de *ou* et de *es*, contractions de *en* et de *le* et *les*. (…) Il y a coïncidence entre la disparition de *ou* et *es* et le développement extraordinaire acquis par *dans*. L'une est la **cause**[204] de l'autre, il n'est pas difficile de le prouver *»* (A. Darmesteter, *op. cit.*: 184).

Cette preuve (c'est le second volet), l'auteur la tire de l'étude quantitative de la distribution des déterminants placés dans le régime de *dans* au sein de quelques textes de la seconde moitié du XVI^{e} s. En effet, déclare A. Darmesteter, « [Si] *dans* s'est substitué à *en* dans [les emplois] où *en* se contractait avec *le* et *les*, [alors] les premiers emplois ont dû être ceux où *dans* était suivi de l'article *le* et d'un mot commençant pas une consonne, ou de l'article *les* » (*op. cit.*). Ainsi accomplit-il un relevé des occurrences de cette préposition dans le premier volume de l'édition Blanchemain des *Amours* de P. de Ronsard.: « *Les exemples [tirés] de Ronsard confirment cette vue* [= notre hypothèse]*, puisque sur cinquante-quatre exemples*, dans *est suivi de* le *ou* les *dans* trente *cas et d'un autre mot quelconque dans* vingt-quatre *cas seulement*[205]. (185, 186)

---

[202] « La plupart des auteurs consultés établissent un rapport entre la création de la préposition dans et l'évolution des formes de *en* devant l'article défini. » W. de Mulder (*op. cit.*: 287). Voir aussi S. Lardon & M.-C. Thomine (*op. cit.*: 392).

[203] Séjour favori du diable, comme on le sait…

[204] C'est nous qui soulignons.

[205] On voit ici se profiler la faiblesse du raisonnement : l'auteur se fonde sur des fréquences absolues sans poser la question de la distribution des articles définis dans l'ensemble du corpus, quelle que soit leur position. Il va de soi qu'à une époque où les chercheurs ne disposaient pas de l'outil informatique, une telle question ne pouvait demeurer que sans réponse.

Soucieux d'étayer davantage cette thèse, les successeurs d'A. Darmesteter se sont employés à approfondir la piste quantitative en l'enrichissant de nouveaux relevés accomplis manuellement sur corpus. Ainsi G. Gougenheim (1950, [1970]) présente-t-il les résultats d'un dépouillement conduit sur les quatre premiers volumes des œuvres de Ronsard (collection de la *Société des Textes français modernes*) qui le conduisent à conclure que cet auteur serait le premier à opter préférentiellement pour *dans* plutôt que pour *en* devant *le* et *les*. L. Terreaux élargit le champ du dépouillement à toute l'œuvre de Ronsard et conclut dans le même sens (voir en part. *op. cit.*: 195-204).

Il reste que ces enquêtes, pour minutieuses qu'elles fussent, sont demeurées limitées dans leur ampleur : jusqu'à l'arrivée de l'outil informatique, le dépouillement manuel s'avérait une tâche extrêmement coûteuse en temps, donc limitée en termes de nombre de mots traités et non exempte d'erreur[206]… Avec l'avènement de l'informatique et la miniaturisation de ses composants, l'étude systématique du vocabulaire menée sur des corpus de plus en plus vastes peut s'effectuer aujourd'hui en quelques millisecondes et se trouve à la portée de tous les chercheurs. Il est ainsi possible de tester sur un corpus de grande ampleur l'hypothèse d'A. Darmesteter.

Une telle étude n'est en rien mineure ou anecdotique. Il s'agit de mieux comprendre par quelles voies une préposition du français contemporain a pu s'imposer récemment et de manière fulgurante dans le peloton de tête de ses mots les plus fréquents. Préposition faisant par ailleurs figure d'exception parmi les langues romanes si l'on en croit B. Fagard et B. Combettes (2013 : 93) : « Seul parmi les langues romanes, le français moderne délaisse en partie la préposition *en* (issue du latin *in*) au profit d'une autre, *dans.* »

1.2.3. Mise à l'épreuve de l'hypothèse de Darmesteter

On se propose d'éprouver sur le corpus Presto l'hypothèse d'A. Darmesteter en examinant les implications que l'on peut en tirer :

- la première **[I_1]** est formulée par l'auteur lui-même (cf. *supra*) : « Les premiers emplois [de *dans*] ont dû être ceux où *dans* était suivi de l'article *le* et d'un mot commençant par une consonne, ou de l'article *les* » ;

- Seconde implication **[I_2]** : on devrait observer une prévalence de la catégorie des articles définis au voisinage[207] de *dans* à l'inverse des autres sous-catégories de déterminants[208].

1.2.3.1. Examen de l'implication 1 [I_1]

Dans la mesure où notre étude se focalise désormais sur une période restreinte de notre corpus (correspondant en gros à la « naissance » de *dans*), nous avons travaillé sur le sous-corpus Presto$^{1501\text{-}1700}$, que nous avons lui-même scindé en six sous-corpus[209] : le premier

[206] Voir D. Vigier (2015).

[207] Par « voisinage de *dans* » nous entendons la position qui suit immédiatement la préposition (*dans le cas)* et la seconde position si la première est remplie par le (pré)déterminant « tout » (*dans tous les cas*) – sachant que dans notre corpus, la première occurrence de cette séquence (*dans tous les* N) apparaît en 1607.

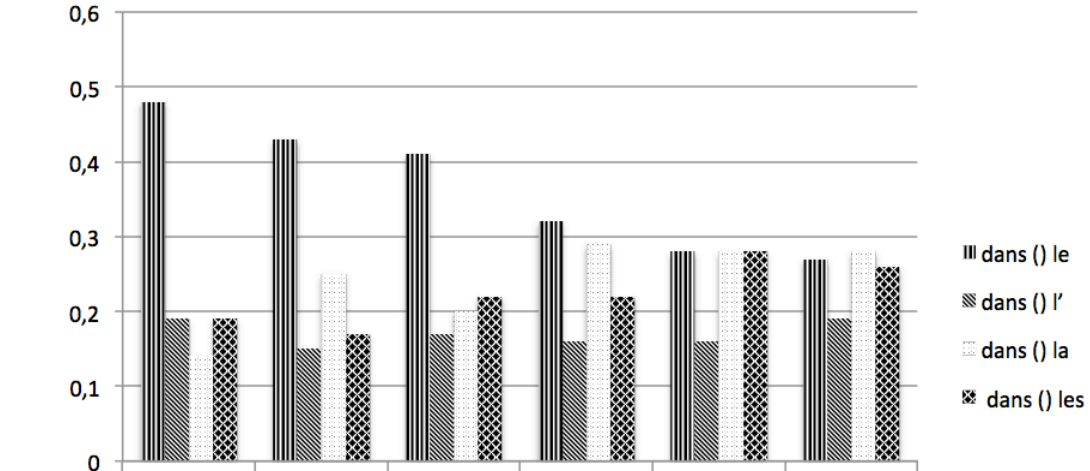
[208] Cette seconde implication est distincte de la première en ceci que [I_1] concerne le paradigme des formes de l'article défini (*le, les versus la, l'*) tandis que [I_2] nous place dans le paradigme des sous-catégories de déterminants susceptibles de suivre *dans*.

[209] Sans pouvoir développer ce point, on observera que nous travaillons désormais sur des « sous-corpus » et non sur des « parties » comme précédemment. Il importe de distinguer soigneusement « partie » et « sous-corpus » en textométrie. Une partie est issue d'une *partition* d'un corpus, de sorte que chaque « partie » implique les

couvre une période de cinquante ans[210] (1501-1550), les cinq autres une période de trente ans.

Nous avons ensuite calculé, dans chacun de ces sous-corpus, quelle était la fréquence relative[211] de chaque forme de l'article défini au voisinage de *dans*.

Voici sous forme de diagramme les résultats obtenus :



	1501-1550	1551-1580	1581-1610	1611-1640	1641-1670	1671-1700
dans () le	0,48	0,43	0,41	0,32	0,28	0,27
dans () l'	0,19	0,15	0,17	0,16	0,16	0,19
dans () la	0,14	0,25	0,2	0,29	0,28	0,28
dans () les	0,19	0,17	0,22	0,22	0,28	0,26

Diagramme 3

Fréquences relatives des formes de l'article défini au voisinage immédiat de *dans ;* corpus Presto–[1501-1700] divisé en 6 tranches temporelles.

Entre 1501 et 1610, la fréquence relative de la forme *le* domine nettement celle des trois autres formes. Cette prévalence va cependant décroissant et on arrive, pour la période 1641-1700, à une situation de quasi-égalité entre les fréquences affectées à *le, la, les*. A aucun moment en revanche, la fréquence de *les* ne se détache clairement de celle de *la*, l'une devançant faiblement l'autre tour à tour, et cela jusqu'en 1641-1670. Un tel constat pour la forme *les* déjoue clairement une partie des prévisions d'A. Darmesteter.

Est-ce à dire que l'on comptera parmi les arguments favorables à la thèse de ce dernier la seule, mais nette, prévalence de *le* – au vu des fréquences relatives – au voisinage de *dans* ? Notre réponse est négative et voici pourquoi. On peut formuler l'hypothèse que, dans chacun des six sous-corpus distingués *supra* (Presto[1501-1550], Presto[1551-1580], Presto[1581-1610], etc.), la prévalence de *le* observée au voisinage de *dans* n'est peut-être simplement que le reflet de la distribution de cette même forme dans tout le sous-corpus.

---

autres (la taille de toutes les parties étant égale à la taille du corpus). En revanche, un sous-corpus est le résultat d'un prélèvement opéré sur un corpus de départ sans impliquer d'autres sous-corpus. Si nous avons choisi de travailler ici sur des sous-corpus, c'est qu'il ne s'agit plus pour nous de comparer la distribution d'une forme *i* dans une *partie j* avec sa distribution dans le corpus entier, mais de comparer, au sein d'une tranche chronologique (= *sous-corpus*), la distribution d'une forme *i* (= forme de l'article défini) dans le voisinage d'un pivot (*dans*) par rapport à sa distribution dans le reste de cette même tranche quelle que soit sa position sur l'axe syntagmatique.

[210] Eu égard au nombre plus restreint de textes disponibles pour cette période.

[211] Cette fréquence correspond au quotient d'une fraction prenant pour numérateur la fréquence (absolue) de l'une des formes de l'article défini dans le voisinage de *dans* et pour dénominateur le nombre total d'apparitions de l'article défini (dans la tranche considérée) au voisinage de *dans*.

Pour vérifier une telle hypothèse (qui conteste à *dans* toute combinatoire singulière avec *le*, *contra* A. Darmesteter), nous avons recouru à un calcul des cooccurrences afin de déterminer si la distribution des formes de l'article défini présente ou non une déformation statistiquement significative au voisinage de *dans* par rapport à sa distribution dans l'ensemble du corpus.

Dans TXM, le calcul des cooccurrences peut être ramené à un calcul des spécificités sur un sous-corpus[212]. Nous nous proposons d'expliquer ici i) en quoi consiste ce calcul (des cooccurrences), ii) pourquoi il peut être ramené à un calcul des spécificités accompli sur un sous-corpus, iii) et son intérêt pour l'objectif que nous visons.

1.2.3.1.1. Le calcul des cooccurrences sur TXM

L'objectif de ce calcul consiste à déterminer les cooccurrents les plus spécifiques d'un « pivot » (mot, lemme, partie du discours, expression CQL, …) au sein d'un corpus donné. Ce calcul est précédé d'un paramétrage accompli par l'utilisateur afin de définir (notamment) i) le contenu du pivot (équation de requête CQL) ii) le contexte de cooccurrence à gauche et/ou à droite du pivot. D'autres éléments du calcul - dans le détail desquels nous n'entrerons pas ici (seuil de fréquence, seuil de spécificité, … : voir manuel TXM en ligne[213]) - sont définis par défaut dans la machine. Ajoutons qu'en l'absence actuelle de métadonnées relatives aux relations de dépendances syntaxiques entre les constituants dans le corpus Presto, le paramétrage du cotexte de cooccurrence du pivot s'opère sur des critères de nombre de mots, de leur rang et de leur position vis-à-vis du pivot.

Dans le cas qui nous occupe ici, pour chaque tranche temporelle sélectionnée, on accomplit un calcul des cooccurrences pour le pivot prépositionnel *dans*. Le paramétrage du cotexte de cooccurrence consiste à sélectionner une fenêtre de deux mots à droite du pivot : on peut ainsi faire porter le calcul non seulement sur les articles définis figurant dans la suite immédiate du pivot, mais aussi après le prédéterminant *tout.* Le calcul lancé permet de construire la liste hiérarchiquement ordonnée (classement des « indices de cooccurrences » par ordre de grandeur décroissant) des cooccurrents les plus spécifiques du pivot. Les indices de cooccurrence s'interprètent exactement comme les indices de spécificité.

Dans TXM, le calcul des cooccurrences est fondé sur le calcul des spécificités[214]. En effet, l'ensemble des cotextes de cooccurrences du pivot défini par l'utilisateur constitue une « partie[215] » de taille t_j du corpus de taille T, l'autre partie dans le calcul équivalant au corpus diminué de la partie et donc de taille T-t_j.

Illustrons (et précisons) ce point par un des calculs qui nous occupera dans cette

---

[212] Le calcul des cooccurrences implémenté dans TXM applique le calcul des spécificités de Lafon à une partie formée des contextes du pivot. Pour plus de précisions : https://groupes.renater.fr/wiki/txm-users/public/faq.

[213] http://textometrie.ens-lyon.fr/spip.php?rubrique64

[214] Comme dans la plupart des logiciels de textométrie actuels, le calcul des cooccurrences implémenté dans TXM applique le calcul des spécificités à une partie formée des contextes du pivot. « On calcule alors les spécificités de cette partie (constituée de tous les contextes), par rapport à l'ensemble du corpus : le calcul met en évidence les mots qui sont statistiquement sur-représentés dans la partie, c'est-à-dire -vue la manière dont a été constituée la partie- les mots qui sont sur-représenté au voisinage du pivot, qui semblent statistiquement attirés par ce pivot. » (https://groupes.renater.fr/wiki/txm-users/public/faq). Il reste que le calcul des préférences que manifeste dans un corpus donné un pivot pour ses cooccurrents peut s'accomplir sans recourir aux spécificités. Ainsi P. Blumenthal et S. Diwersy recourent-il au score du log-likelihood (voir T. Dunning 1993).

[215] « Pour le mot ou motif indiqué en requête (le "pivot" des cooccurrences), et pour la définition de contexte choisie via les paramètres (contexte en nombre de mots ou en structure, ex. phrase(s) ou paragraphe), on construit l'ensemble de tous les contextes des occurrences du pivot. Cet ensemble définit une partie du corpus. » (https://groupes.renater.fr/wiki/txm-users/public/faq)

section : celui visant à déterminer si, pour la période 1551-1580 du corpus Presto (soit, Presto$_{-1551_1580}$) la distribution de la forme *le* de l'article défini au voisinage de *dans* est ou non statistiquement banale par rapport à sa distribution dans le reste du corpus. Le paramétrage du calcul effectué, on peut considérer que la sous-fréquence des mots figurant dans l'ensemble des cotextes de cooccurrence du pivot constitue une partie de taille t_j au sein du corpus Presto$_{-1551_1580}$ de taille T (fréquence totale des mots qui y figurent). Dès lors, les paramètres du calcul de cooccurrence proposé par TXM sont :

f_{ij} = sous-fréquence de la forme *le* de l'article défini au voisinage de *dans* ;
F_i = fréquence de la forme *le* de l'article défini dans Presto$_{-1551_1580}$;
t_j = sous-fréquence de tous les mots au voisinage de *dans* ;
T = fréquence de tous les mots dans Presto$_{-1551_1580}$

Cependant, dans la mesure où nous nous intéressons aux variations possibles des distributions de quatre formes relevant d'une classe grammaticale fermée, nous avons choisi d'adopter une conception plus restrictive des tailles de la partie et du corpus mises en jeu dans le calcul en les ramenant, pour la première (t_j), à la sous-fréquence de la seule catégorie *article défini* au voisinage de *dans*, pour la seconde (T), à la fréquence de cette même catégorie de déterminants dans Presto$_{-1551_1580}$. En d'autres termes, nous avons modifié deux des paramètres présentés ci-dessus en leur substituant les suivants :

t_j = sous-fréquence de l'article défini au voisinage de *dans* ;
T = fréquence de l'article défini dans Presto$_{-1551_1580}$.

L'interprétation de l'indice de cooccurrence obtenu repose sur les mêmes principes que ceux stipulés *supra* § 1.1.1.

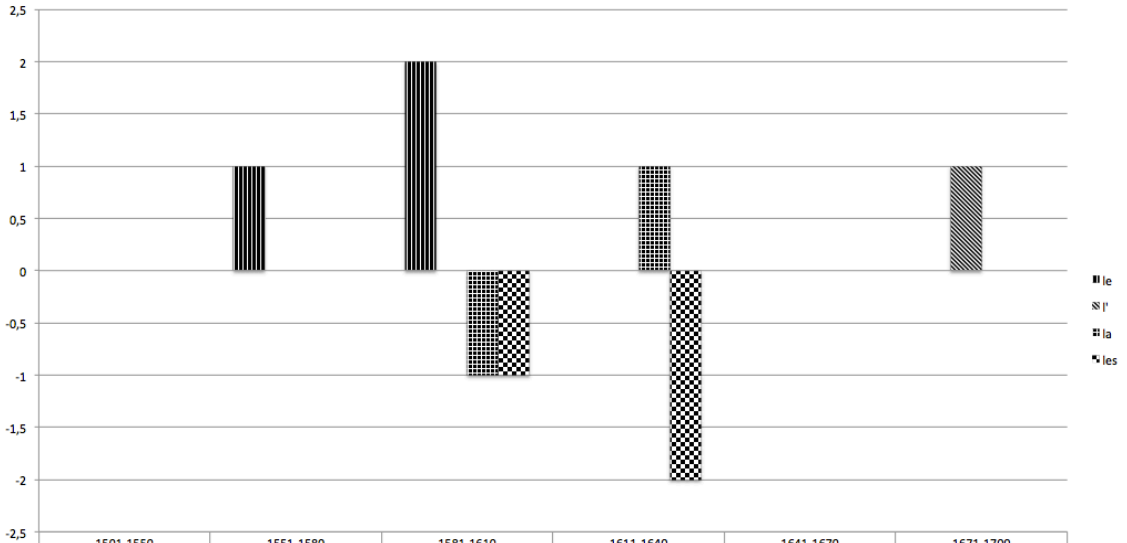
1.2.3.1.2. Présentation et analyse des résultats

Rappelons-le : nous souhaitons savoir si, entre 1551 et 1700, la distribution des formes de l'article défini présente ou non une déformation statistiquement significative au voisinage de *dans* par rapport à sa distribution dans l'ensemble du corpus.

L'ensemble des calculs de cooccurrences accomplis pour chacun des sous corpus distingués a pris pour paramètres les valeurs suivantes :

f_{ij} = sous-fréquence de la forme de l'article défini (*le*|*la*|*les*|*l'*) au voisinage de *dans* ;
F_i = fréquence de la forme (*le*|*la*|*les*|*l'*) de l'article défini dans le sous-corpus;
t_j = sous-fréquence de l'article défini au voisinage de *dans* ;
T = fréquence de l'article défini dans le sous-corpus.

Voici les résultats auxquels nous avons abouti :



	1501-1550	1551-1580	1581-1610	1611-1640	1641-1670	1671-1700
le	0	1	2	0	0	0
l'	0	0	0	0	0	1
la	0	0	-1	1	0	0
les	0	0	-1	-2	0	0

Diagramme 4

Calculs des cooccurrences appliqués au voisinage de la préposition *dans* au sein de cinq sous-corpus constitués dans Presto$^{-1551_1670}$. Indices de cooccurrence obtenus pour les diverses formes de l'article défini.

Tous les scores calculés évoluent dans une zone située en deçà du seuil de significativité statistique de valeur absolue 3 : signe que la distribution des formes de l'article défini au voisinage de *dans* ne présente en réalité aucune déformation statistiquement remarquable et donc qu'elle est à l'image de sa distribution dans le reste du corpus, quelle que soit la position syntaxique du SN dans la composition duquel il entre.

L'étude des données fréquentielles concernant la distribution des formes de l'article défini au voisinage de *dans* nous conduit donc à conclure que :

- contrairement à ce que prévoit l'hypothèse d'A. Darmesteter, la forme *les* n'est pas prévalente dans cette position au regard des trois autres formes relevant de cette catégorie ;
- les distributions des formes de l'article défini au voisinage de *dans* reflètent leurs distributions dans le reste du corpus; autrement dit la prévalence fréquentielle de *le* dans le voisinage de *dans* jusqu'en 1640 est à l'image de sa prévalence pour la même période dans le reste du corpus et ne constitue pas une caractéristique propre à *dans*.

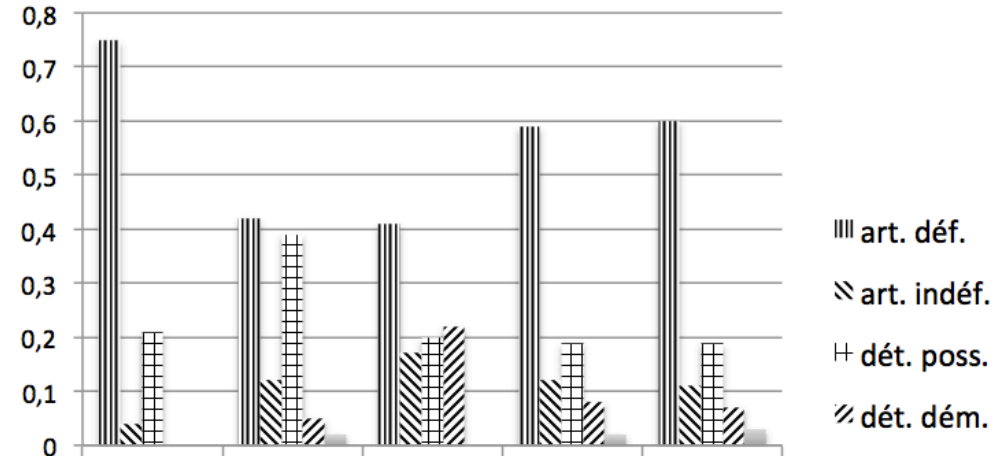
Tous ces éléments conduisent à reconsidérer l'hypothèse d'A. Darmesteter, ce que va confirmer l'examen de la seconde implication présentée *supra*.

1.2.3.2. Examen de l'implication 2 [I_2]

Rappelons pour mémoire le contenu de cette seconde implication : l'hypothèse de Darmesteter, si elle était vérifiée, devrait conduire à observer une prévalence des articles définis au voisinage de *dans* à l'inverse des autres sous-catégories de déterminants.

Un premier calcul des fréquences relatives pour les cinq sous-catégories de déterminants suivantes : *articles définis, indéfinis, déterminants possessifs, démonstratifs, indéfinis* amène à conclure que celles affectées aux définis l'emporte (plus ou moins

nettement) sur les quatre autres pour les tranches temporelles que nous avons distinguées entre 1501 et 1670 :



	1501-1550	1551-1580	1581-1610	1611-1640	1641-1670
art. déf.	0,75	0,42	0,41	0,59	0,6
art. indéf.	0,04	0,12	0,17	0,12	0,11
dét. poss.	0,21	0,39	0,2	0,19	0,19
dét. dém.	0	0,05	0,22	0,08	0,07
dét. indéf.	0	0,02	0	0,02	0,03

Diagramme 5

Fréquences relatives de l'article défini au voisinage immédiat de *dans ;* corpus Presto$^{1501\text{-}1670}$ divisé en 5 tranches temporelles.

Cependant, le recours au calcul des cooccurrences suivant change le point de vue qu'on peut porter sur la distribution des articles définis au voisinage de *dans*.

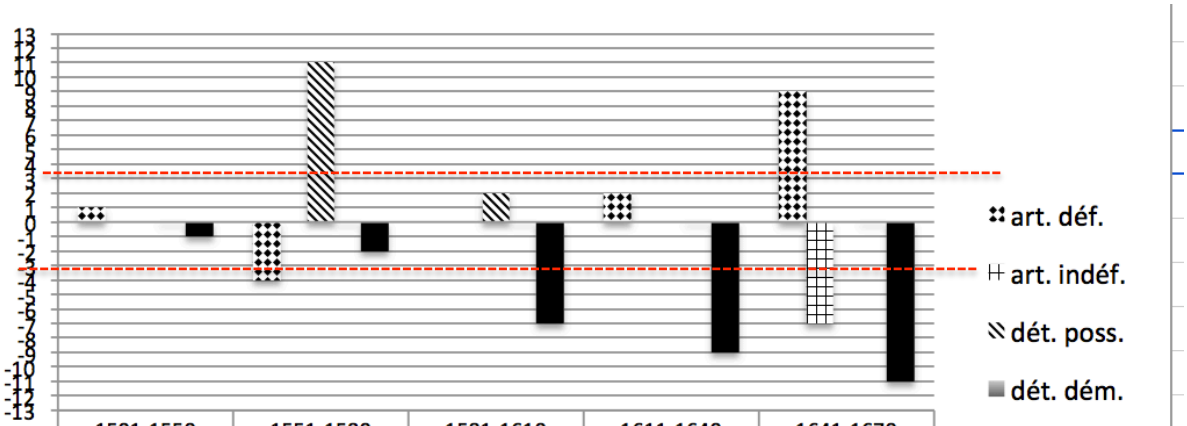
Les variables f_i, F, t_j, T prises en compte pour ce calcul ont été les suivantes :

f_{ij} = sous-fréquence de l'article défini | de l'article indéfini | du déterminant possessif | du déterminant possessif | du déterminant indéfini au voisinage de *dans ;*

F = fréquence de l'article défini | de l'article indéfini | du déterminant possessif | du déterminant possessif | du déterminant indéfini dans le corpus entier ;

t_i = fréquence cumulée des articles définis, indéfinis, déterminants possessifs, démonstratifs et indéfinis au voisinage de *dans ;*

T = fréquence cumulée des articles définis, indéfinis, déterminants possessifs, démonstratifs et indéfinis dans le corpus entier.



	1501-1550	1551-1580	1581-1610	1611-1640	1641-1670
art. déf.	1	-4	0	2	9
art. indéf.	0	0	0	0	-7
dét. poss.	0	11	2	0	0
dét. dém.	0	0	0	0	0
dét. indéf.	-1	-2	-7	-9	-11

Diagramme 6

Calculs des cooccurrences appliqués au voisinage de la préposition *dans* au sein de cinq sous-corpus constitués dans Presto^{1551_1670}. Indices de cooccurrence obtenus pour 5 sous-catégories de déterminants.

Pour la période 1501-1550, la distribution des cinq sous-catégories de déterminants considérées se situe au-dessous du seuil de significativité statistique. Il n'en va plus de même ensuite :

- l'article défini s'avère sous-représenté pour la période 1551-1580 et sur-représenté pour la période 1641-1670 ;
- le déterminant possessif connaît un brusque pic de sur-représentation (qu'il faudrait expliquer) pour la période 1551-1580, mais évolue sinon dans la zone de banalité statistique pour les autres périodes, tout comme les déterminants démonstratifs ;
- enfin, les déterminants indéfinis sont sous-représentés depuis la tranche 1581-1610 jusqu'à la tranche 1641-1670 : il semble qu'il y ait là une forme de récalcitrance croissante de *dans* à l'égard de cette sous-catégorie de déterminants.

Mais restons-en aux seuls articles définis qui nous intéressent au premier chef : si l'on se place dans le cadre de l'hypothèse d'A. Darmesteter, on peine à comprendre pourquoi ce n'est qu'à partir de la seconde moitié du XVII^{e} s. que l'article défini devient une classe de déterminants sur-utilisée au voisinage de *dans* par rapport à sa distribution dans le corpus entier. Le caractère très tardif du phénomène est incompatible avec le raisonnement tenu par l'auteur.

L'examen des deux implications tirées de l'hypothèse d'A. Darmesteter plaide donc pour le rejet de l'hypothèse d'A. Darmesteter. La ruine de la « preuve » quantitative entraîne-t-elle avec elle tout l'édifice? Notre point de vue n'est pas si radical et nous plaiderons plus loin pour un réaménagement de la thèse de l'auteur.

Avant de montrer quels éclairages nouveaux peuvent apporter nos outils sur l'environnement distributionnel de *dans* au XVI^{e} s., nous voudrions pour finir nous arrêter sur un point d'importance : quelle incidence le paramètre des genres discursifs (ou les « champs

génériques » de F. Rastier, 2011) a-t-il sur l'évolution des emplois de *dans* dans les œuvres de notre corpus ? Pour tenter de la mesurer, nous avons partitionné Presto$^{1501\text{-}1600}$ suivant le critère des « champs génériques » du discours littéraire disponibles comme métadonnées dans ce sous-corpus. Nous avons ainsi distingué trois parties : le théâtre [11 textes, 118.867 mots], la poésie [11 textes; 207.765 mots], le genre narratif [12 textes; 741.183 mots].

Voici sous forme de diagramme les résultats que nous avons obtenus :

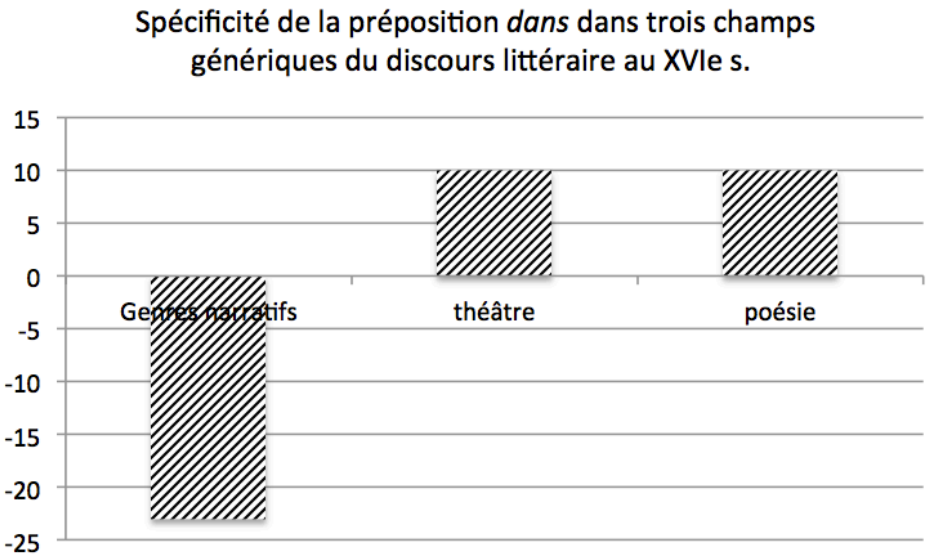


Diagramme 7
Calcul des spécificités accompli sur une partition du corpus Presto$^{1501\text{-}1600}$ réalisée sur le critère du *champ générique*

Tandis que *dans* apparaît nettement sous-représenté dans les genres narratifs par rapport à l'ensemble du corpus Presto$^{1501\text{-}1600}$, c'est l'inverse qui se produit pour la poésie et le théâtre. Tout porte à croire que la diffusion de *dans* au sein des œuvres littéraires du XVI^{e} s. suit un rythme distinct selon le champ générique considéré. Il faudrait vérifier ces tendances sur un corpus plus vaste, mais ces premières observations confirment l'importance du critère de généricité discursive dans la vitesse de diffusion d'un changement linguistique.

Venons-en maintenant à la présentation d'une hypothèse (partiellement) alternative à celle qu'avait argumentée A. Darmesteter et qui s'appuie sur une étude statistique des contextes d'apparition de *dans* au sein des œuvres de notre corpus.

1.3. Construction d'une hypothèse alternative

1.3.1. Point de départ : exploration statistique de la combinatoire *amont* de la préposition *dans* au XVI^{e} s.

La voie que nous nous proposons d'explorer a été initialement ouverte par B. Fagard et B. Combettes (2012) qui travaillaient sur des données extraites de FRANTEXT et ne disposaient donc pas pour leur étude de corpus étiqueté et lemmatisé.

Désireux d'étudier « l'impact du contexte sémantique plus large sur l'évolution de *en* et d'autres prépositions, entre XVI^{e} et XVII^{e} siècles » (102), les auteurs choisissent d'explorer la combinatoire aval des verbes *entrer* (« intransitif de mouvement »), *jeter, lancer*

(« mouvement causé »). Le choix de ces verbes est de leur propre aveu « en partie arbitraire, et doit être vu comme un point de départ pour une réflexion sur le fonctionnement de *en, dans* et *dedans* » (*op. cit.*). La conclusion à laquelle ils aboutissent est que dans ces contextes, on observerait pour l'essentiel un phénomène de remplacement de *en* par *dans*, accompagné d'un changement notable dans l'emploi du déterminant devant le régime nominal. « [On observe une] opposition frappante entre XVI^{e} et XVII^{e} siècles, sur plusieurs points. Le premier est que *dans* s'impose très rapidement comme préposition introduisant le complément des verbes *entrer* et *jeter* aux dépens de *en*. Le second est que *en* se spécialise dans l'introduction de N nus, i.e. sans déterminant. » (103) Et les auteurs de conclure (112) : « *dans* a progressivement remplacé *en* dans un grand nombre de constructions (au sens de la grammaire des constructions (…), avec une concurrence très passagère de *dedans* et, pour certaines constructions, d'autres prépositions. »

Dans le but à la fois d'affiner (sur le plan quantitatif) certains volets de cette étude et de consolider (voire d'élargir) la gamme de verbes recteurs dans le régime desquels on peut observer à partir de 1550 environ une préférence significative pour les SP à tête *dans*, nous avons accompli un calcul des spécificités visant à déterminer i) quelles catégories morphosyntaxiques ii) et quels lemmes s'avéraient sur-représentés dans le voisinage amont de *dans* au sein du corpus Presto étendu[216] pour la période 1550-1600. Le résultat est sans appel : la catégorie morphosyntaxique préférée pour ce cotexte de cooccurrence est celle des verbes ; les cinq premiers lemmes verbaux qui apparaissent par ordre de préférence décroissant sont : *entrer, mettre, jeter, enfermer, cacher*[217].

Cette liste de verbes préférés étant arrêtée, on peut renverser le calcul pour faire de ces verbes le pivot d'une analyse cooccurrentielle visant à étudier les préférences combinatoires qu'ils manifestent à l'égard de leur régime. Le diagramme ci-dessous présente les résultats des calculs opérés sur le cotexte droit (fenêtre de deux mots) de ces verbes saisis ensemble dans un seul pivot complexe, et cela pour la période temporelle 1501-1800. Le recours à ce type de diagramme permet de visualiser, pour chaque tranche temporelle, la part relative (exprimée en pourcentage) occupée par l'indice de cooccurrence affecté à chacune des trois prépositions (indice exprimé dans chaque segment de la barre) sur la somme totale de ces trois indices.

[216] Le recours à cette version du corpus s'explique pour des raisons évidentes de fréquence pour les séquences étudiées.

[217] Si l'on se réfère au classement des verbes notamment de *déplacement* vs *mouvement* proposé par M. Aurnague (par ex. 2012b), *entrer* ressortit aux verbes de déplacement, *jeter* aux verbes de déplacement causé, *mettre* aux verbes de changement d'état, *enfermer* et *cacher* aux verbes d'inclusion sans déplacement.

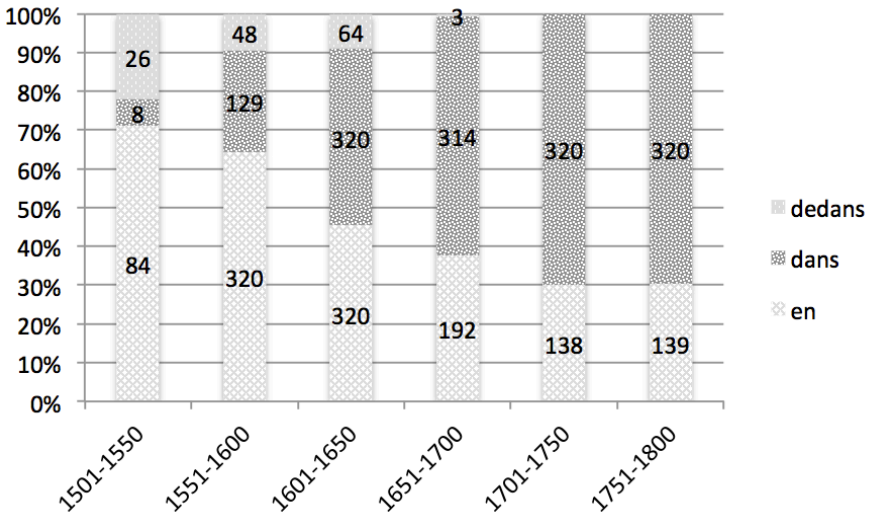


Diagramme 8

Evolution de la part occupée (en %) par chacun des indices de cooccurrence affectés aux prépositions *en, dans, dedans* à la suite du pivot complexe *[lemma=" ENTRER| METTRE|JETER|ENFERMER|CACHER"]*.
Contexte de cooccurrence de 2 mots à droite du pivot; Corpus Presto[1501-1800] étendu, période 1501-1800, tranches de 50 ans.

L'évolution des pourcentages occupés par les scores de spécificité de *dans* et de *en* est frappante : tandis que le premier (pourcentage) ne cesse de croître entre 1501[218] et 1750, le second ne cesse de baisser jusqu'à la même date. *Dans* occupe seul (et définitivement) le haut du podium des prépositions préférées du pivot à partir de la tranche temporelle 1651-1700. Quant à *dedans* il disparaît à l'aube du XVIII[e] s. de la liste des cooccurrents préférés.

Pour affiner notre compréhension de l'évolution des distributions de *dans* et de *en* au voisinage des verbes sélectionnés, nous avons fait porter notre enquête sur l'identification des cooccurrents nominaux les plus spécifiques situés dans le régime des verbes *entrer* et *mettre*[219].

Les deux tableaux suivants présentent les listes des cinq cooccurents nominaux les plus spécifiques (l'ordre des mots correspond à leur ordre hiérarchique de spécificité décroissante) obtenues pour chaque tranche temporelle considérée.

[218] La fait que *dans* puisse, dès la période 1501-1550, être doté d'un indice de spécificité déjà significativement élevé peut surprendre ; en réalité, l'essentiel de ses occurrences se trouve d'une part dans les *Quatre premiers livres des Odes* de P. Ronsard (1550) [cela confirme les observations d'A. Darmesteter et de G. Gougenheim] et – chose plus remarquable – dans *Les contes amoureux de Jeanne Flore* de J. Flore (1537).
[219] La fréquence moindre des trois autres verbes dans notre corpus n'a pas permis une application satisfaisante de nos outils.

ENTRER

	XVI	XVII	XVIII	XIX
Pivot	**Liste des 5 premiers cooccurrents nominaux (lemmes) les plus spécifiques[220] pour chaque pivot**			
Pivot 1 ***entrer en N***	*possession, propos, dispute, matière, colère*	*possession, matière, comparaison, religion, désespoir*	*lice, conférence, concurrence, négociation, matière*	*action, lice, partage, possession, ligne*
Pivot 2 ***entrer en dét. N***	*chambre, maison, église, entendement, ville*	*salle, chambre, pays, cabinet, possession*	∅	∅
Pivot 3 ***entrer dans dét. N***	*pays, chambre*	*chambre, ville, port, détail, sentiment*	*détail, chambre, ville, maison, composition*	*détail, chambre[221], composition, maison, voie*
Liste des cooccurrents nominaux spécifiques PARTAGÉS[222] par deux pivots dans une même tranche temporelle				
Pivots 1 & 3	∅	*composition défiance*	*composition*	*combinaison*
Pivots 2 & 3	*chambre*	*chambre, salle, cabinet, pays, esprit, maison, église, composition, place, royaume*	∅	∅

Tableau 1
Listes des lemmes nominaux les plus spécifiques situés dans le régime du lemme verbal *ENTRER*, corpus Presto-ETENDU_1501_1800.

[220] Répartis dans au moins *trois* textes distincts de façon à corriger les distorsions introduites par les cas de fréquences élevées, dans un seul texte, de mots par ailleurs peu usités dans le reste du corpus.
[221] La prévalence pendant cinq siècles du N *chambre* après *dans* a été remarquée par P. Blumenthal (2011 : 293) sur un autre corpus.
[222] Les lemmes partagés sont prélevés dans l'ensemble des cooccurrents les plus spécifiques des pivots concernés et non seulement parmi les cinq premiers. D'où la possibilité pour certains d'entre eux de ne pas figurer dans les lignes 3 à 5 du présent tableau.

METTRE

	XVI	XVII	XVIII	XIX
Pivot	**Liste des 5 premiers cooccurrents nominaux (lemmes) les plus spécifiques pour chaque pivot**			
Pivot 1 ***METTRE en N***	*oubli, pièce, route, campagne, chemin*	*peine, colère, état, œuvre, devoir*	*état, œuvre, prison, possession, usage*	*œuvre, jeu, rapport, route, mouvement*
Pivot 2 ***METTRE en dét. N***	*lieu, tête, main, place, état*	*place, main, bouche, état, rang*	*place, main*	∅
Pivot 3 ***METTRE dans dét. N***	*vaisseau*	*esprit, bouche, nécessité, disposition, lit*	*bouche, nécessité, main, tête, impuissance*	*bouche, poche, dépendance, cas, tête*
Liste des cooccurrents nominaux spécifiques PARTAGÉS par deux pivots dans une même tranche temporelle				
Pivots 1 & 3	∅	*danger, état*	*chemin, colère, prison, tête, état*	*action, tête*
Pivots 2 & 3	*vaisseau*	*bouche, cage, danger, fantaisie, état, perfection.*	*main*	∅

Tableau 2

Listes des coccurrents nominaux préférés situés dans le régime du verbe ***mettre***, corpus Presto-ETENDU_1551_1800.

Voici les éléments les plus saillants qui ressortent de l'examen comparé de ces deux tableaux :

- Au XVI^e^ s, lorsque *dans* entre dans la combinatoire de *entrer* et de *mettre*,
  - ces deux verbes présentent déjà une combinatoire nominale étoffée avec *en*;
  - la combinatoire nominale spécifique qu'ils sélectionnent avec *dans* réunit uniquement des noms dénotant des réalités dotées d'extension matérielle ou physique (*pays, chambre, vaisseau*[223]) qui jouent contextuellement le rôle de site de repérage spatial d'une cible;
  - il existe d'emblée une combinatoire nominale spécifique partagée pour les constructions V *en dét. N* et V *dans dét. N*.

- Le XVII^e^ s. constitue un siècle charnière dans l'évolution de la combinatoire de ces deux verbes avec *en* et *dans*.

  C'est en effet à cette époque qu'on observe :

[223] *Vaisseau*, au XVI^e^ s. peut signifier *vase, pot, récipient* (« Les prunes de damas seront mis dans des vaisseaux, et par dessus on jettera du vin nouveau » (1564, C. Estienne, *L'Agriculture et maison rustique*)) ou *navire* (« s'estans mis dans quelques vaisseaux à la merci de la mer auroyent esté jettez & seroyent abordez en ceste terre du Bresil » (1578, J. de Léry, *Histoire d'un voyage fait en la terre du Brésil*).

  - une combinatoire nominale étoffée pour *chacun* des trois pivots, ce qui ne sera plus le cas ensuite ; concernant *dans* plus particulièrement, apparaissent dans sa distribution spécifique, outre des noms d'entités spatiales, des noms abstraits dénotant des réalités sans extension matérielle ou physique : *détail, sentiment, esprit, nécessité, disposition*.
  - Les combinatoires nominales *partagées* s'avèrent remarquablement étoffées si on les compare aux autres siècles, en particulier pour les pivots 2 (V *en dét. N*) et 3 (V *dans dét. N*).

- Au delà, au XVIII[e] s. et au XIX[e] s., la construction (*entrer + mettre) en dét. N* s'éteint rapidement. Parallèlement, la combinatoire la plus spécifique de (*entrer + mettre) dans dét. N* continue à présenter une série hybride de noms dénotant des réalités contextuellement dotées – ou non - d'extension matérielle ou physique[224] (*chambre, ville, maison*) ou non (*détail, composition, nécessité, dépendance, ...)*. Enfin, la structure (*entrer + mettre) en N* sélectionne comme noms les plus spécifiques presque[225] uniquement des N (contextuellement) abstraits.

Ces observations mises en faisceaux permettent de décrire un processus global dans lequel on peut distinguer deux étapes :

Première étape : XVI[e]-XVII[e] s. : vers une ouverture maximale de la combinatoire nominale de (*entrer + mettre) (en+ dans)*

L'arrivée de *dans* au sein de la combinatoire de ces deux verbes conduit rapidement (XVII[e] s) à la cohabitation de trois combinatoires nominales qu'on peut répartir en termes de « polarité » concrète/ abstraite comme suit :

- deux combinatoires à polarité nominale hybride concrète-abstraite (noms dénotant contextuellement des réalités dotées ou non d'extension matérielle ou physique) : *(entrer + mettre) (en + dans) dét. N,*
- une combinatoire à polarité nominale abstraite : *entrer en N.*

Dès le XVI[e] s, la construction (*entrer + mettre) dans* conquiert une partie de sa combinatoire nominale spatiale sur celle de (*entrer + mettre) en + déterminant + N*. Au siècle suivant, *dans* s'approprie aussi une partie de la combinatoire abstraite de cette même construction. En ce sens, on peut dire qu'il y a eu un remplacement de *en* par *dans* devant certains noms.

Le résultat de ce premier processus conduit à une situation « exceptionnelle » au XVII[e] s. où les constructions (*entrer + mettre) en* et (*entrer + mettre) dans* partagent une combinatoire nominale spécifique particulièrement substantielle.

Seconde étape : XVIII[e] s. - XIX[e] s.. Processus de reconfiguration et de spécialisation de la combinatoire nominale de (*entrer + mettre) (dans + en)*.

A l'aube du XVIII[e] s., on observe une raréfaction puis une disparition de la combinatoire

[224] On observera au passage que trois des cooccurrents nominaux préférés de *entrer en dét N* au XVI[e] s. (*chambre, maison, ville*) sont passés sous la coupe de *entrer dans dét N*.

[225] Le cas de *(mettre en) prison* pour le XVIII[e] s. est à cet égard intéressant : même si le nom *prison* dénote une réalité étendue dans l'espace, la construction elle-même dénote autant la mise dans un lieu que l'affectation d'une qualité (celle de prisonnier).

nominale de la construction (*entrer* + *mettre) en* + *déterminant* + *N*. La combinatoire partagée entre les pivots 2 et 3 connaît le même destin.

Il s'opère alors une spécialisation des deux constructions restantes :

- (*entrer* + *mettre) dans* dotée d'une combinatoire à polarité hybride s'avère apte i) à opérer des localisations spatiales concrètes d'une cible au terme d'un déplacement ou d'un mouvement ; ii) à exprimer divers rapports plus abstraits (« *entrer dans la composition (de), (se) mettre dans le cas de,* etc.»)

- (*entrer* + *mettre) en N* doté d'une combinatoire nominale abstraite non actualisée, tend à se figer en des locutions verbales (« *mettre en usage, entrer en matière, … »*)

Comme nous le verrons dans la troisième section de cette dernière partie, ce processus en deux étapes que nous avons tenté de décrire ne s'observe pas seulement dans la construction des deux verbes *entrer* et *mettre*, mais concerne plus largement la combinatoire de *en* et de *dans* entre le XVI^{e} siècle et le XX^{e} siècle.

1.3.2. Vers la formulation d'une nouvelle hypothèse

Souvent, les grammairiens de la langue du XVI^{e} s. auxquels nous nous sommes référé dans ce travail soulignent la « *confusion* » engendrée par l'ambiguïté sémantique des morphèmes *au, aux* au XVI^{e} s. On peut insister en parallèle sur la forme d'*appauvrissement*, en termes de ressources linguistiques mises à disposition par le système de la langue, à laquelle furent confrontés les locuteurs de cette époque pour exprimer leur pensée. L'exemple des verbes de déplacement en fournit de bons exemples. Placés dans le régime de tels verbes, les SP locatifs ayant pour tête *à* peuvent exprimer la direction prise par la cible vers le site tandis que *en* y ajoute le trait de pénétration de la cible dans le site. Cette différence de valeur sémantique peut avoir une rentabilité forte dans certains contextes. En voici deux exemples :

(21) *Et en sortant de cheu le Magniffique, vit bien que la Gripa alloyt* ***a*** *l'Evesché et n'entra pas* ***en*** *l'Eveché.* (1542-1544, Collectivité, *Registres du Consistoire de Genève au temps de Calvin*.)

(22) *Henry* [le valet] *estant venu* ***a*** *la chambre de sa maistresse, luy demanda : Que vous plaist-il, Madame, me voici. (…) Le damoyselle lui respõdit : Allez vous coucher, car je ne vous demande pas. (…) Le povre Henry se retirant* ***en*** *sa chambre* (…). (1568, *Histoire Pitoyable Dv Prince Erastvs,* Trad.) [Le serviteur Henri ne pénètre pas dans la chambre de sa maîtresse, mais retourne en revanche *dans* la sienne.]

On voit tout le parti que le scripteur tire de l'opposition *à / en* pour l'expression de la localisation dans ces deux énoncés. Or dès que le nom régime de la préposition est masculin à initiale consonantique ou pluriel, et actualisé par un article défini, cette paire prépositionnelle n'est plus disponible pour produire de tels distinguos - même si *dedans* demeure un recours.

Autrement dit, il paraît incontestable que les locuteurs de l'époque ont été placés face à un déséquilibre du système engendré par la substitution des formes *au, aux* aux amalgames morphologiquement issus de **en le / *en les*. Le présent travail ne remet donc pas en cause

l'hypothèse selon laquelle la cause première (le *primum movens* en quelque sorte) du changement linguistique profond que marqua l'ascension fulgurante des emplois de *dans* en français préclassique fut cette « *surcharge* » due à l'ambiguïté de *au, aux* qu'évoque G. Gougenheim (1951, [1974] : 182). Nos conclusions divergent en revanche de celles d'A. Darmesteter et de ses successeurs concernant les premiers contextes d'occurrence de *dans* en français préclassique. Rien dans notre corpus ne permet en effet de considérer que cette préposition se serait d'abord préférentiellement substituée à *en* devant les formes *le* et *les* de l'article défini. Tout porte à croire que la diffusion de *dans* s'est d'emblée accomplie devant toutes les formes de déterminants[226], avec une préférence pour la position valencielle de certains verbes marquant notamment le mouvement, le déplacement, le changement d'état.

Ajoutons pour finir que très vite (dès la première moitié du XVII^{e} s.), la nouvelle venue *dans* s'affranchit des conditions de sa naissance et commence sa vie propre - tout comme *en* d'ailleurs, ainsi que l'avait observé F. Brunot : *« En* commence une autre histoire, du jour où cette préposition cesse de se contracter avec l'article en *ou* et *es*, et où *dans* entre en concurrence avec lui. *»* (*op. cit.*: 469). C'est là le chapitre d'une autre histoire que nous nous proposons d'ouvrir dans la troisième section de cette partie.

Auparavant, nous souhaiterions aborder quelques points d'ordre méthodologique en vue d'éclairer notre démarche à venir.

2. Cotexte d'une unité linguistique et accès à son sens

Dans cette troisième partie de notre mémoire, notre propos vise avant tout à montrer, par quelques études spécifiques, la fécondité de certaines méthodes d'analyse quantitative en linguistique de corpus que nous avons appliquées dans le cadre du programme Presto. Dans la troisième section (à venir) plus particulièrement, nous mettrons en œuvre une approche statistique des contextes cooccurrentiels d'emploi des prépositions *en, dans, dedans* au sein de notre corpus, dans le but d'étudier l'évolution sémantique de ces dernières. Or une telle approche s'adosse, sur le plan théorique, à certains principes de sémantique qui, bien avant les travaux de sémantique distributionnelle automatisée, ont fondé les études lexicologiques. Il n'en demeure pas moins que les travaux relatifs à la mathématisation du langage d'une part, les développements de l'informatique et de l'automatisation en sciences du langage d'autre part (voir J. Léon, 2008), ont profondément revivifié ces principes sur le plan théorique en leur ont offert des domaines d'application radicalement nouveaux, ouvrant du même coup des perspectives d'une grande fécondité pour la recherche en diachronie.

Ce sont quelques-uns de ces principes fondateurs que nous exposons dans cette deuxième section.

2.1. « Dis-moi qui tu fréquentes, … »

Les travaux de l'école « contextualiste » britannique issue de la *London school*[227] (J.-R. Firth (1890-1960), M.-A. K. Halliday (1925- …), J. Sinclair (1933-2007), G. Leech (1936- …), …), eux-mêmes influencés par les recherches philosophiques de Wittgenstein (1889-1951) sur le langage (en part. *Investigations philosophiques* publiées à titre posthume en 1953), ont largement contribué à développer et étayer l'idée suivant laquelle l'accès au sens

[226] Dans la thèse d'A. Darmesteter, l'ouverture de *dans* à tous les déterminants ne s'accomplissait que dans un deuxième temps : « La langue ne pouvait se résoudre à n'employer dans qu'avec l'article le ou les, exactement dans les cas où il représentait ou et es. C'aurait été imposer à ses habitudes un formalisme et une rigueur inconnue de l'esprit populaire. Celui-ci (…) étendit l'emploi de dans à toutes les expressions où le substantif est déterminé : *dans la maison, dans cet état, dans toute affaire, dans ces circonstances*, etc. » (*op. cit.*, 186)

[227] Pour une étude détaillée des « sources de la *corpus linguistics* », voir J. Léon (2008).

d'une unité linguistique passe par l'exploration de son contexte d'apparition.

On peut rappeler à ce propos la formule de J. Firth « *you shall know a word by the company it keeps* » (1957:11) qui n'est pas très éloignée de la phrase de L. Wittgenstein (*Investigations philosophiques*, 1953 : § 43) : « *La signification d'un mot est son usage dans le langage* »[228]. Une telle conception du sens conduit naturellement à établir une continuité entre lexique et grammaire : la 'grammaire' d'une unité du lexique - entendue comme l'ensemble de ses environnements distributionnels possibles et exclus - constitue la clef de sa signification.

On peut illustrer avec simplicité le lien qui unit lexique et grammaire au moyen d'un exemple emprunté à D. Leeman (1996 : 8-9) :

« Si je prends ainsi le premier mot de la lettre *a* du premier dictionnaire qui me tombe sous la main, je trouve *abaca*. A priori, ce mot ne me dit rien (il n'a pas de sens pour moi), mais si l'on m'explique que :

L'abaca pousse aux Philippines
Le fruit de l'abaca n'est pas comestible
On arrache (ou on plante) des abacas à Manille

ces contextes me permettent de déterminer que *abaca* est (sans doute) un nom d'arbre. Et si l'on me dit que :

L'abaca est extrait d'une sorte de bananier
Les paillassons peuvent être en abaca
Le tapis est tissé avec des fibres d'abaca

alors je conclurai que abaca peut aussi désigner une certaine matière. »

Si l'on souscrit à l'idée que l'accès à l'identité sémantique d'une unité linguistique dans une langue donnée passe par l'exploration la plus systématique de ses contextes d'usage, on conçoit aisément en quoi l'outil informatique qui a permis de constituer de grandes banques de données textuelles structurées et annotées, et d'automatiser des chaînes de calculs statistiques souvent longs et complexes, a constitué un pas décisif pour l'exploration en informatique linguistique de ces contextes dans les langues naturelles[229].

Certes, comme le soulignent J.-F. Haussmann et P. Blumenthal (2003 : 6), le recours au contexte – et notamment aux « collocations » – bénéficiait déjà dans le champ de la lexicologie générale d'une solide tradition (notamment en France) qui remonte bien avant l'invention de la science informatique. « *Dans la lexicographie générale, les collocations ont toujours eu une place de premier choix* ». Et les auteurs de citer un extrait de l'article *débat* tiré du premier *Dictionnaire de l'Académie française* de 1694. La tradition britannique en lexicographie historique n'était pas non plus en reste si l'on en croit D. Geeraerts (2010 : 169) : « *The great historical dictionary projects (...) were, in their own painstakingly manual way, corpus-based – a dictionary like the Owford English Dictionary (Murray 1884) rests on*

[228] On sait que J.R. Firth a été influencé par Wittgenstein qu'il cite plusieurs fois.

[229] Il reste bien entendu que l'exploration statistique du cotexte verbal d'une unité pivot au sein d'un corpus aussi vaste et varié soit-il, laisse ouverte cette question abyssale : « *is the corpus sufficient as a context ?* » (D. Geeraerts, 2010 : 178). Peut-être faut-il voir dans les recherches récentes de « sémantique distributionnelle multimodale », quoique embryonnaires, une tentative crédible de remédier à cette limite jusqu'ici infranchie. Par ex. E. Bruni, N. Tran et M. Baroni, (2014 : 38) : « *A multimodal distributional semantic model integrates a traditional text-based representation of meaning with information coming from vision. In this way, it tries to answer to the critique that distributional models lack grounding, since they base their representation of meaning entirely on the linguistic input, neglecting statistical information inherent in perceptual experience, that we humans instead exploit.* »

a huge collection of quotations extracted from historical texts – and the method used by the historical lexicographers for analysing and classifying those quotations was surely also based on the principle of interpretation in context, i.e. on the examination of the elements co-occuring with the target word ».

Mais seuls les développements de l'ordinateur et du traitement automatique du langage ont permis de traiter de grandes masses de données du langage naturel, offrant aux principes fondateurs firthiens (retravaillés et remodelés par ses successeurs, en particulier dans le cadre de l'école « *corpus-driven*[230] » - J. Sinclair, M. Hoey, …) les moyens de déployer toutes leurs potentialités. Comme l'écrivent encore J.-F. Haussmann et P. Blumenthal (*op. cit.*: 11)

« Corpus numérisés et outils lexicométriques ont profondément révolutionné la recherche linguistique et, plus particulièrement, lexicographique, en décuplant les possibilités d'observations. Vu l'impact de cette révolution, il ne paraît pas exagéré d'affirmer que l'avant et l'après s'opposent autant que l'œil nu s'oppose aux jumelles, télescope et microscope réunis ».

Le tournant qu'a représenté pour les sciences du langage, à partir des années 1990, l'accès à de très grands corpus n'a cependant pas entraîné de discontinuité dans le domaine des méthodes et des lois mathématiques mobilisées en linguistique de corpus, comme le montre J. Léon (2015) :

« Le tournant se situe plus au niveau de l'automatisation que de la mathématisation. Les méthodes sont de nature statistique et probabiliste, dans la continuité des méthodes ébauchées dans les années 1950-1960 à partir de la théorie de l'information. » (157)

En d'autres termes, la 'révolution' informatique des années 70 à 90 a introduit une transformation moins tant « qualitative » que « quantitative » dans l'approche contextuelle du sens. Et les sémantiques distributionnelles automatisées contemporaines - au-delà de leur arsenal mathématique et TAListe qui effraie souvent le linguiste - demeurent à bien des égards les héritières directes de la lexicographie et de la lexicologie à fondement collocationnel du XIX^{e} siècle. Comme l'écrit D. Geeraerts (*op. cit.*): « *The systematicity with which the data are collected and scrutinized may have improved, but the idea itself of using a large repository of real language data as the empirical basis for semantic descriptions is a continuation of the finest traditions of philological and lexicographical work rather than a radical break with the past.* » (169)

Depuis une quarantaine d'années se sont donc développés de plus en plus de travaux ressortissant à ce qu'il est convenu d'appeler les « *linguistiques de corpus* », mettant en jeu une approche statistique du langage naturel. Plus on s'avance vers les années 2000, plus ces approches se caractérisent par un « métissage[231] » grandissant et fécond des démarches, des

[230] Pour une présentation synthétique des deux « filiations » firthiennes que constituent les courants « corpus-driven » et « corpus-based » (dont le chef de file actuel est G. Leech), on se reportera à J. Léon (*op. cit.*: 161-165)

[231] Nous empruntons ce terme à B. Habert et P. Zweigenbaum (2003) qui l'utilisent plusieurs fois : « 1. Des croisements au métissage voulu des approches en TAL. Les croisements actuels des approches constituent une évidence empirique. Le TAL au sens large (traitement automatique des langues, ingénierie des langues, linguistique computationnelle) voit sinon converger du moins se rencontrer et se mêler des communautés naguère clairement séparées. (…) On observe (…) l'interpénétration des paradigmes, les données annotées s'adossant aux acquis de la description linguistique, et les approches statistiques ne pouvant dépasser leurs limites sans s'étayer de connaissances linguistiques. Un métissage conscient et actif.» (84)

outils, des méthodes, etc. Dans ce vaste et complexe paysage, on peut souligner la place prépondérante qu'ont occupé, en linguistique, les recherches sur les collocations[232] en France et en Europe entre 1990 et 2010 environ - que ce soit dans la mouvance de l'école contextualiste britannique, des travaux de Mel'čuk ou de ceux relevant d'une approche française parfois liée à la lexicométrie. Ces travaux ont largement démontré la fécondité du principe de détermination du sens des mots en contexte et par l'usage dans les langues naturelles. Depuis une dizaine d'années, on assiste à la montée en puissance des travaux de *sémantique distributionnelle automatique* en TAL (voir par ex. le numéro 56 : 2, 2015, de la revue TAL) qui conjoignent les héritages firthiens et harrissiens d'une part, le *behavioural profile*[233] d'autre part (K. Heylen et A. Bertels : 2016 : 51-53) et qui constituent une nouvelle étape dans l'exploration du sens contextuel des mots en corpus[234].

Dans ce mémoire, un de nos objectifs consiste à montrer quel parti on peut tirer de ce type d'approche pour l'étude des prépositions en diachronie.

Dans la première section de cette troisième partie, nous avons montré un usage possible de l'étude statistique du voisinage d'un pivot prépositionnel : déterminer si l'on observe ou non dans les premiers emplois de *dans* une déformation statistiquement significative, à son voisinage immédiat, de la fréquence de certains déterminants par rapport à leur distribution dans l'ensemble du corpus pris comme étalon. Il ne s'agissait pas, cependant, d'approcher la sémantique de *dans* ni de saisir à travers les variations de sa combinatoire lexicale de possibles changements intervenus dans son « identité » sémantique à travers le temps.

Dans la troisième section en revanche, nous chercherons à identifier des changements survenus dans l'identité sémantique de *en, dans, dedans* entre le XVI^{e} s. et le XX^{e} s., en explorant par des méthodes statistiques les contextes distributionnels de ces prépositions dans des tranches temporelles successives. L'hypothèse qui sous-tend cette démarche s'inscrit dans le sillage des positions théoriques cursivement exposées plus haut : l'apparition de changements notables dans la distribution statistiquement préférée d'une unité ouvre des pistes heuristiques vers la mise au jour d'évolutions ayant affecté son noyau sémantique « stable ».

2.2. Notre approche du contexte pour l'accès au sens

Dans les lignes qui suivent, nous proposons d'articuler le principe suivant lequel l'accès au sens d'une unité linguistique passe par l'exploration de son contexte d'apparition, avec certains des outils et des calculs statistiques utilisés dans Presto. Cette réflexion d'ordre méthodologique nous permettra d'éclairer les fondements de la démarche que nous mettrons en œuvre dans la troisième section pour l'étude des spécificités cooccurrentielles de *en, dans, dedans* et de leur évolution entre le XVI^{e} s. et le XX^{e} s.

2.2.1. Comment interpréter le calcul des spécificités de P. Lafon utilisé par la plateforme de calcul TXM ?

On rappellera préalablement que les principaux outils de lexicométrie (Hyperbase, Lexico,

[232] Notion déjà présente dans les travaux de H. E. Palmer dès 1933 (selon A. Tutin (2010 : 11)) mais qui acquiert une importance de premier plan dans les travaux de l'école de Londres grâce aux recherches de J. R. Firth. Selon les approches et les courants, la « collocation » a reçu des définitions variées (pour un point, voir notamment F. Grossmann et A. Tutin (éds)(2003), J.F. Hausman et P. Blumenthal (éds) (2006)).

[233] Voir par ex. S. Th. Gries (2010).

[234] « Ces méthodes et leurs résultats sont encore mal connus de la communauté des linguistes, alors qu'on pourrait souhaiter qu'ils permettent d'étendre la gamme des outils à leur disposition pour explorer le sens des mots à partir de leurs usages dans les corpus. » (C. Fabre, 2015 : 395)

TXM, BTLC, ….) aujourd'hui accessibles aux linguistes proposent des fonctionnalités statistiques permettant l'analyse quantitative des cooccurrences d'une unité pivot. Or les choix des formules, des lois de probabilité et des indices qui y sont mis en jeu pour mesurer le caractère plus ou moins significatif de la fréquence de rencontre entre un pivot et un cooccurrent dans un corpus donné sont variables.

Si nous restreignons notre propos aux seuls outils[235] que nous avons utilisés jusqu'ici dans le cadre du projet Presto – à savoir TXM et BTLC – on observe que le premier s'est jusqu'ici focalisé sur le seul indice des spécificités de P. Lafon adossé à la loi de probabilité hypergéométrique, tandis que le second propose à l'utilisateur une liste de plusieurs indices probabilistes possibles: t-score, z-score, rapport de log-vraisemblance (*log-likelihood ratio*), … Le choix de l'une ou de l'autre de ces mesures n'est pas indifférent et a à voir avec un certain nombre de paramètres[236] : [1] la taille totale du corpus traité (taille de la population), [2] la taille du sous-corpus (taille de l'échantillon), [3] le nombre total d'occurrences de l'unité observée dans le corpus et [4] son nombre d'occurrences dans le sous-corpus. Or si la taille du sous-corpus et le nombre d'occurrences de l'unité observée dans le sous-corpus sont particulièrement faibles, certains des indices probabilistes cités *supra* donnent des résultats moins concluants que pour les hautes fréquences, reproche dont se trouve exempt l'indice des spécificités de P. Lafon. Dans la mesure où, dans notre corpus, il nous arrive fréquemment de travailler sur des événements de faible fréquence au sein de sous-corpus réduits (par ex. *supra* les premiers emplois de chaque forme de l'article défini au voisinage de *dans* entre 1501 et 1550), nous avons choisi de privilégier le calcul des spécificités proposé par TXM. Cela dit, nous avons recouru aussi dans la troisième section (§ 3.) de cette troisième partie au « rapport de log-vraisemblance » (*log-likelihood ratio*) implémenté dans PrimeStat (BTLC) dans la mesure où nous travaillions sur des sous-corpus de tailles plus importantes et sur des fréquences plus élevées. Cela nous a permis de vérifier qu'en effet, les différences de résultats (calcul des *spécificités* de P. Lafon *versus ratio log-likelihood*) en termes de hiérarchisation des coocccurrents étaient globalement peu significatifs pour les coocccurrents les plus spécifiques [237].

Venons-en maintenant à l'esprit du calcul mis en jeu pour mesurer au sein d'un corpus donné « l'attraction » d'un coocccurrent pour un pivot. Si l'on s'en tient au seul calcul de P. Lafon (1980, 1984) implémenté dans TXM, l'indice des spécificités permet de hiérarchiser l'ensemble des cooccurrents du pivot sélectionné dans un cotexte de cooccurrence paramétré par avance. Chaque cooccurrent est classé parmi les autres sur le critère suivant : plus il est spécifique, plus la chance que sa sous-fréquence f_{ij} observée dans le contexte de cooccurrence défini soit due au hasard est faible. On conçoit dès lors l'intérêt que revêt une telle information pour le linguiste qui veut explorer le cotexte cooccurrentiel d'une unité : elle lui permet de trier et de hiérarchiser les cooccurrents d'un pivot sur le critère statistique de ses préférences combinatoires.

Il convient ici de faire deux remarques :

---

[235] Pour une revue des indices utilisés dans les principaux outils de textométrie actuels, ainsi que des formules et lois qui leur sont associées, on se reportera à C. Reutenauer (*op. cit.*: 155 et sq.).

[236] Voir C. Reutenauer (*ibid* : 146 et sq.).

[237] Un des horizons de la linguistique de corpus outillée demeure, si l'on en croit D. Geeraerts (*op. cit.*: 178), la mise au point d'un état de l'art couplant les principaux types d'objectifs poursuivis en matière de linguistique descriptive et les méthodologies les plus appropriées pour les atteindre. « *Distributional corpus analysis has not yet reached the stage where it can present a stable set of methodological procedures coupled to specific descriptive questions.* » Sur cette voie, il semble par exemple désormais acquis que: « *for the modelling of synonymy, (...) syntax-based word space models outperform all other approaches.* » (176)

<u>Remarque 1.</u> La haute spécificité (de signe positif) d'un cooccurrent n'implique pas que sa fréquence absolue soit élevée dans le corpus.

Un cooccurrent de fréquence absolue très élevée dans le corpus peut obtenir un score de spécificité faible voire nul au voisinage d'un pivot donné. Il n'y a là rien que de très attendu puisque les cooccurrents les plus pertinents pour l'analyse du pivot sont ceux dont la cooccurrence avec ce dernier est, dans le corpus considéré, statistiquement plus fréquente que ne l'aurait laissé attendre le hasard. « Les cooccurrents les plus fréquents ne donnent pas nécessairement le plus d'informations sur la signification du mot-cible. À l'instar des travaux sur les collocations, les méthodes d'analyse distributionnelle s'appuient sur des mesures d'association statistiques, ce qui permet de donner plus de poids aux cooccurrents qui apparaissent significativement plus souvent avec le mot-cible qu'attendu par le hasard. Ces cooccurrents avec un poids plus élevé fournissent plus d'informations sur le sens du mot-cible que les autres cooccurrents, quelle que soit leur fréquence absolue. » (K. Heylen et A. Bertels, *op. cit.*: 57).

Inversement, un cooccurrent de fréquence absolue faible dans le corpus peut se voir affecter un indice de spécificité positif très élevé au voisinage d'un pivot donné. C'est alors le signe que, quoique doté d'une fréquence peu élevée dans l'ensemble du corpus, ce cooccurrent est remarquable en ceci qu'il cooccurre presque uniquement avec le pivot considéré. Dans ce cas de figure, il est très important pour le linguiste de disposer systématiquement des fréquences du pivot et du cooccurrent dans l'ensemble du corpus ainsi que de la co-fréquence du cooccurrent dans le contexte paramétré. Nous avons ainsi été conduit à écarter quelques cooccurrents de *en* ou de *dans* qui, quoique dotés d'un fort indice (positif) de spécificité, s'avéraient employés dans un nombre trop restreint de textes dans le corpus. Les retenir aurait introduit un biais, du fait de leur répartition trop inégale dans le corpus entier. On aurait sur-valorisé une thématique ou un trait de style auctorial ou un champ générique alors que notre propos consiste à cerner des préférences combinatoires calculées sur la base de fréquences réparties sur un nombre minimal de textes (généralement ≥ 3) eux-mêmes idéalement répartis dans plusieurs genres, et produits par des auteurs différents.

<u>Remarque 2.</u> S'avèrent *remarquables* (et intéressantes au moins sur le plan heuristique pour le linguiste) non seulement les spécificités élevées de signe positif mais aussi celles de signe négatif.

Si l'on se place dans le cadre traditionnel des études lexicométriques qui visent à contraster des sous-parties d'un corpus global afin de mettre au jour des sur- ou des sous-emplois d'items donnés susceptibles de donner lieu à des interprétations éclairantes, le linguiste sera très intéressé par l'identification des sous-spécificités (spécificités élevées de signe négatif) significatives. Ainsi P. Lafon (2006) comparant les emplois des pronoms personnels déictiques *je, nous, vous* dans le corpus des « Vœux présidentiels sous la cinquième république de 1959 à 2001 » partitionné sur le critère du locuteur note-t-il : « La première surprise se lit sur la ligne du JE, qui s'avère être partout [de spécificité] positive, sauf chez De Gaulle. [indice= - 17] Le charisme naturel gaullien n'a manifestement pas besoin du renfort d'une énonciation répétitive en JE » (5). Dans la perspective qui est la nôtre – et qui vise le plus souvent à contraster la distribution d'une unité au voisinage d'un pivot avec sa distribution dans le reste du corpus (calcul des cooccurrences), les sous-spécificités significativement négatives peuvent aussi nous intéresser très directement. Ainsi a-t-on déjà observé *supra* que dès le XVIII^{e} s., la préposition *en* manifeste dans Presto une aversion particulière pour la catégorie « déterminant » (ce qui est parfaitement cohérent avec nos analyses présentées en première partie) et que cet indice de sous-spécificité croît en valeur absolue siècle après siècle entre le XVIII^{e} et le XX^{e} s. Il reste que dans cette deuxième partie,

nous n'étudierons pas les lexèmes nominaux sous-représentés, pour la raison qu'ils s'avèrent très peu nombreux et dotés d'un indice le plus souvent faiblement négatif.

2.2.2. De quel(s) phénomène(s) une sur-spécificité statistique calculée dans un corpus pour un collocatif au voisinage d'un pivot peut-elle être le signe ?

A quel phénomène linguistique peut correspondre un indice de spécificité élevé attribué à un cooccurrent d'un mot pivot au sein d'un corpus ? On touche ici au problème de l'interprétation des résultats. « *Distributional corpus analysis is primarly a method, not a model. It opens an impressive amount of empirical data, but how exactly those data may be interpreted is not always given by the technique itself*. » (D. Geeraerts, *op. cit.*: 177)

Les nombreux travaux qui ont fleuri depuis plusieurs décennies dans le champ des études sur la *phraséologie*[238] ont conduit les linguistes à élaborer des typologies plus ou moins fines et exhaustives pour rendre compte des phénomènes linguistiques qu'est susceptible de « capturer » l'étude statistique de l'environnement cooccurrentiel d'un pivot simple ou complexe[239]. Dans la mesure où, dans cette deuxième partie, nous nous cantonnerons essentiellement à l'étude des collocatifs nominaux des prépositions *en, dans, dedans* au sein de Presto, nous proposons de nous en tenir ci-dessous à une typologie réduite des « causes » possibles d'une sur-spécificité cooccurrentielle en corpus, telles que nous les avons identifiées dans notre étude.

Mais auparavant, voici deux points sur lesquels nous voudrions encore insister :

i) ces causes sont le plus souvent multiples, et leur effet cumulé peut parfois expliquer le caractère très élevé de l'indice. Les mesures statistiques « captent » en effet, par la mesure de la sur-utilisation statistique d'un mot au voisinage d'un autre dans un corpus, un ensemble de phénomènes qui peuvent relever de plusieurs domaines et niveaux d'analyse. Comme l'écrit S. Loiseau (2011 :73) « Les divergences systématiques identifiées par les tests statistiques mêlent toujours des choses qui doivent être distinguées dans l'interprétation des résultats. (…) Qualifier ces résultats, c'est également rapporter les faits observés à différents types d'explications : l'interprétation des faits doit distinguer ce qui relève de phénomènes phraséologiques, de normes textuelles, de faits historiques, de différences de variété, etc. Autrement dit, l'utilisation de ce type de données relève d'une herméneutique. »

ii) La recherche de ces causes conduit parfois à sortir des limites du champ de la linguistique *stricto sensu* pour gagner les terres plus inconnues (sinon pour le linguiste, du moins pour nous) des modes, des conceptions dominantes d'une époque et d'une catégorie sociale déterminée dont on peut penser qu'elles ont influé le cours de la langue *via* les usages, etc. (voir *infra*).

Venons-en à notre esquisse de typologie des « causes » possibles d'une sur-spécificité cooccurrentielle au niveau lexical (collocations) en corpus.

Dans le champ de la linguistique stricto-sensu, la mesure statistique d'une préférence

[238] « La phraséologie, que l'on peut définir en quelques mots comme le domaine qui traite les séquences lexicales perçues comme préconstruites » (D. Legallois et A. Tutin, 2013 :3).
[239] Voir par ex. la synthèse proposée par C. Bolly (2010 :16-18).

combinatoire d'un mot pour un autre peut signaler :

- Une contrainte sélectionnelle pesant sur la syntagmatique des classes de mots dans la langue considérée. On touche ici aux 'colligations' telles que l'entendent J.-R. Firth et J. Sinclair[240]. Si l'on prend l'exemple de *en*, on observe que si l'on accomplit un calcul des cooccurrences sur le cotexte aval (1 mot) de cette préposition pour la période 1900-1944, l'un des mots les plus spécifiques (rang 9) qui apparaît (pour Presto) dans la liste est le participe présent « *faisant* ». Inversement, les mots dotés des indices de spécificité les plus significativement négatifs sont les quatre formes de l'article défini *le, la, les, l'*. Une des informations que l'on peut extraire de ce classement, à savoir que *en* se combine préférentiellement avec des formes du participe présent (notamment) et qu'elle possède une aversion pour les déterminants, peut paraître triviale pour la connaissance du français contemporain. Cela devient nettement plus intéressant quand on adopte une perspective diachronique, et qu'on cherche par ex. à déterminer à partir de quelle époque *en* a cessé d'avoir pour accompagnateur préféré un déterminant (voir notre première partie).

- Un figement plus ou moins avancé, ce phénomène multifactoriel étant intrinsèquement scalaire.
    - Parmi les critères généralement retenus, trois jouissent d'un large consensus dans la littérature sur le sujet : i) le critère sémantique de *non compositionnalité* du sens ii) le critère lexical de *non substituabilité paradigmatique* iii) le critère de *non modifiabilité* des marques morphosyntaxiques. Le problème vient des séquences non prototypiques c'est-à-dire ne vérifiant pas les trois critères. Dans une perspective diachronique, le figement peut être le signe d'une grammaticalisation en cours voire aboutie.
    - Toute collocation (statistique) n'est cependant pas le symptôme d'un figement. « La haute spécificité d'une combinaison n'équivaut pas forcément à la présence de figement. » (P. Blumenthal, 2011 : 293). Autrement dit, il y a des combinaisons « libres » significativement récurrentes. Ces combinatoires préférentielles non figées peuvent donner des renseignements précieux sur la manière dont les locuteurs, à une époque donnée, « conceptualisent » les référents des lexèmes nominaux placés dans la dépendance d'une préposition (voir P. Blumenthal (2006) et le dernier item de la présente liste).

- Influence du global sur le local : une collocation peut aussi s'expliquer pour partie par les influences qu'exercent sur le niveau syntagmatique les discours, les champs génériques, les genres et sous-genres discursifs dont relèvent les textes du corpus. On peut citer ici pour illustration une étude récente, menée dans le cadre du projet ANR-DFG PhraseoRom, par L. Gonon, O. Kraif, I. Novakova, J. Piat & J. Sorba (2016) qui montre que le sous-genre du roman policier (période contemporaine) se caractérise par la présence d'une affinité du lexème *scène* pour le lexème *crime* et inversement, cette attraction mutuelle signant la récurrence particulièrement élevée du motif « *scène de crime* ».

- La prise en compte de la variation peut aussi s'avérer centrale pour comprendre un phénomène collocatif. Pour ce qui concerne la variation diachronique au centre de notre travail, l'apparition, le développement puis la disparition des collocations

[240] Voir respectivement D. Legallois (2012 : 37-38) et D. Geeraerts (2010 : 170).

> relevant ou non du figement linguistique nous intéressent au plus haut point. Là encore, diverses études ou observations peuvent justifier cet intérêt. On renverra par ex. à P. Blumenthal (2011) pour l'étude des liens associatifs étroits (relevant d'une combinatoire libre) au XVII^{e} s. entre *dans* et les noms *esprit* et *âme*, liens qu'on ne retrouve plus au XX^{e} s. Un autre exemple nous a été fourni, dans un travail que nous menons sur le discours encyclopédique, par la comparaison des coocccurrents préférés de *dans* de *L'Encyclopédie* de Diderot et d'Alembert (1751-1772) et de *L'Encyclopædia Universalis* (2005). On observe en effet que l'un des tout premiers collocatifs de cette préposition, dans *l'Encyclopædia Universalis*, est le nom *cadre* (rang 3)*,* alors qu'il est totalement inexistant de cette même liste des collocatifs pour *l'Encyclopédie*. Chose peu étonnante si l'on se réfère au *Dictionnaire Historique de la Langue Française* (DHLF) qui nous apprend (rubrique *cadre*) que le sens abstrait de ce nom - *« délimitation transposée sur un plan abstrait (...) à ce qui structure une pensée, sert de matière et de plan à une œuvre* » (320) n'est apparu en français qu'au XIX^{e} s. C'est donc récemment que la séquence « *dans dét. cadre* » a connu une étape de pragmaticalisation lui permettant d'exprimer non plus une valeur référentielle (par ex. *dans le cadre du tableau)* mais une valeur de cadrage « énonciatif », étape qui se signale par une attraction élevée du pivot pour son cooccurrent.[241]

Si l'on se place maintenant dans le cadre … d'une linguistique lato-sensu, il se peut que la collocation identifiée signale une trace dans l'usage (par répétitions significatives) de ce que nous nommerons des « conduites socio-culturelles historiquement situées » : phénomènes de mode, types de conceptualisation des référents dans une culture et/ou une époque données, …. Cette question des liens possibles à tisser entre changements linguistiques et arrière-plans socio-historiques a jalonné tout le déroulement du programme Presto. Nous ne développerons pas ce volet dans le présent mémoire mais nous nous permettons de renvoyer à l'introduction que nous avons rédigée avec P. Blumenthal dans un volume paru chez P. Lang[242] : *Etudes diachroniques du français et perspectives sociétales* (2017). Elle présente les enjeux et les difficultés d'une telle approche qui, à l'image de ce que fait la sociolinguistique historique, cherche à transcender la traditionnelle distinction entre « histoire interne » et « histoire externe ».

2.3. Programme de travail pour la troisième section

Nous comptons montrer d'abord qu'après avoir établi par un calcul statistique et pour des tranches temporelles successives, la liste des cooccurrents nominaux préférés de *en, dans, dedans* dans le corpus Presto, une approche préliminaire « à gros grain[243] » de ces cooccurrents permet de détecter des changements majeurs intervenus dans l'histoire de l'identité sémantique de *en* et de *dans*. Ce type d'approche, classique dans l'étude des collocations, peut être qualifiée plus précisément par les critères suivants : elle met en jeu i) un modèle à base de mots *(word-based) ii)* qui considère les cooccurrents graphiques au niveau des lexèmes (*type-level*). Aucun mode de représentation computationnelle sous forme

[241] Si l'on poursuit la lecture de l'entrée « *cadre* » du DHLF, on découvre une observation fort intéressante : « *un usage contemporain, la locution* « dans le cadre de »*, est devenu envahissant pour* « dans le domaine de », « en ». » (*op. cit.*) De fait, un rapide sondage sur *Google Books Ngram Viewer* montre une poussée spectaculaire de la séquence *dans dét. cadre* à partir de 1950 environ, confirmant que sa fortune dépasse le cadre strict du genre encyclopédique

[242] Cet ouvrage constitue les actes d'un colloque (P. Blumenthal, D. Vigier, à par. 2017) que nous avons organisé en mars 2016 à l'ENS de Lyon (https://clps2016.sciencesconf.org) et qui avait pour thème : *Changements linguistiques et phénomènes sociétaux.*

[243] En l'occurrence, une approche fondée sur l'opposition *noms abstraits / noms concrets*

de *graphe* ou de type *vectoriel* ne sera ici associé au traitement des résultats statistiques obtenus, qui sont analysés manuellement. Dans cette mesure, la démarche que nous adopterons relève de *l'analyse des collocations* telle que présentée dans le tableau suivant emprunté à K. Heylen et A. Bertels (*op. cit.*: 52) :

	Identification d'indices contextuels	Classification d'occurrences
Philologie classique	manuelle	manuelle
Analyse des collocations	statistique	manuelle
Analyse de profils comportementaux	manuelle	statistique
Méthodes d'analyse distributionnelle	statistique	statistique

Tableau 3
Aperçu des outils statistiques utilisés en sémantique lexicale
(tiré de K. Heylen et A. Bertels (*op. cit.*: 53))

Insistons pour terminer sur un point : les conclusions auxquelles nous parviendrons n'auront pas à nos yeux le statut de « vérités démontrées » aptes éventuellement à détrôner certaines des conclusions auxquelles sont parvenues d'autres recherches menées notamment dans d'autres cadres théoriques et méthodologiques. Nous pensons en particulier aux travaux des grammairiens et linguistes du siècle précédent (G. Gougenheim, F. Brunot, L. Huguet, …) qui certes ne pouvaient recourir à de grands corpus informatisés mais dont l'abyssale et fascinante familiarité avec les textes anciens leur permettait d'exprimer des jugements d'une sûreté extraordinaire sur les usages du temps et le système de la langue. Autrement dit, notre souci sera davantage d'entrer en dialogue avec ces auteurs, en vue de problématiser certaines de leurs conclusions en les confrontant à nos propres observations fondées pour partie sur des éléments d'ordre quantitatif.

3. Étude des spécificités cooccurrentielles propres à *en, dans* et *dedans* entre le XVI^{e} s. et le XX^{e} s.

Le tableau suivant présente pour chaque tranche temporelle constituées dans le corpus Presto$^{1509\text{-}1944}$ les dix[244] cooccurrents nominaux préférés (valeurs des indices de spécificité rangés par ordre de grandeur décroissante) identifiés pour chacune de ces trois prépositions. Ces listes ont été produites à partir d'un calcul des cooccurrences analogue à celui exposé dans la première partie. Pour chaque lemme nominal, on a calculé son score de spécificité en partant des valeurs suivantes :

- f_{ij} = sous-fréquence du lemme dans le cotexte de cooccurrence défini au voisinage de la préposition;
- F = fréquence du lemme dans le sous-corpus considéré (1509-1600, 1601-1700, …) ;
- t_j = sous-fréquence des noms communs dans le cotexte de cooccurrence défini au voisinage de la préposition;
- T = fréquence des noms communs dans le corpus Presto$^{1509\text{-}1944}$.

[244] Dix ou moins : voir le cas de *dedans*, dont le faible effectif dans le sous-corpus Presto$^{1509\text{-}1600}$ ne permet pas de faire émerger dix cooccurrents statistiquement spécifiques pour cette tranche temporelle.

dans				
1509-1600	**1601-1700**	**1701-1800**	**1801-1900**	**1901-1944**
cœur, sein, entrailles, âme, mer, château, feu, ville, sang, maison	*sein, cœur, chambre, monde, ciel, âme, cabinet, bois, écrit, ville*	*pays, sein, plaine, moment, maison, état, cas, bois, monde, suite*	*cas, état, ombre, chambre, maison, pays, salle, coin, salon, rue*	*chambre, cas, ombre, maison, bois, lit, mesure, domaine, pays, condition*
en				
***en* suivi d'un déterminant**				
lieu, endroit, chambre, maison, ville, place, sorte, présence, manière, instant	*lieu, endroit, état, façon, occasion, temps, âme, sorte, pays, place*	*mot, lieu, moment, faveur, endroit, cas, sorte, présence, pays, façon*	*sorte, moment, mot, sens, lieu, terme, faveur, cas, air, absence*	*cas, moment, sens, terme, air, sorte, lieu, manière, honneur, occurrence*
***en* suivi d'un nom sans déterminant**				
vain, bref, fin, dépit, pièce, général, terre, repos, haut, paix	*vain, fin, effet, suite, général, paix, colère, peine, repos, dépit*	*effet, vain, général, conséquence, faveur, vérité, état, paix, sûreté, œuvre*	*vain, dehors, face, général, présence, train, résumé, paix, définitive, vertu*	*effet, face, réalité, train, vérité, raison vain, général, arrière, temps*
dedans				
cœur, âme	*cœur, âme, yeux, chambre, sein, lit, fort, logis, air, eau*			

Tableau 4

Liste des dix premiers cooccurrents (lemmes) nominaux les plus spécifiques de *en, dans, dedans* pris comme pivots.

Calcul des spécificités accompli sur R à partir des valeurs f_{ij}, F, t_j, T calculées par TXM.

Corpus Presto[1509-1944].

On s'étonnera peut-être que nous ayons choisi de constituer des tranches temporelles dont les bornes initiales et finales coïncident peu ou prou[245] avec un découpage séculaire purement « arithmétique » de la période 1509-1944, choix qui apparaît (à juste titre) arbitraire relativement au phénomène linguistique que nous voulons étudier et à sa périodisation. Notre point de vue consiste à considérer qu'en première approche, la dimension du temps peut être appréhendée comme une simple « toile de fond (…) pour repérer une évolution ou au contraire une permanence (ou bien encore une évolution en « dents de scie », sans directionalité perceptible) » (S. Prévost, 2011 : 78). L'étude et la comparaison des cooccurrents nominaux les plus spécifiques de *en, dans, dedans* pour ces tranches temporelles devraient donc nous livrer de premières indications quant aux permanences et aux évolutions de leurs identités sémantiques au cours du temps. Il n'en demeure pas moins que cette

[245] Approximativement pour la première (1509-1600) puisque 1509 est la date de première édition de la première œuvre (chronologiquement parlant) qui figure dans notre corpus. Pour la dernière (1900-1944), il ne faut considérer que la borne initiale, le choix de la borne finale étant tout aussi arbitraire mais dicté par d'autres considérations (voir Partie II).

approche devra être complétée par une étude visant à une meilleure périodisation du phénomène.

Dans les commentaires du tableau 4 présentés ci-dessous, nous nous efforcerons de dégager quelques lignes de force. Chemin faisant, nous nous appuierons régulièrement sur l'opposition « ontologique[246] » entre « noms abstraits » et « noms concrets ». On sait combien est à la fois nécessaire et délicate cette opposition pour travailler sur l'organisation du lexique nominal. Pour un état récent de la question, on peut renvoyer à G. Kleiber & M. Vuillaume (2011) et R. Huyghe (2015). Dans le présent mémoire, nous adopterons comme ligne de démarcation entre noms concrets *versus* abstraits le critère d'accès aux sens[247] : un nom dénotant une réalité concrète devra donc vérifier le test : *Un N, (ça peut être vu / senti / touché / entendu) / ça possède une saveur*.

3.1. Les cooccurrents nominaux les plus spécifiques de *dans*

1501-1600	1601-1700	1701-1800	1801-1900	1901-1944
cœur, sein, entrailles, âme, mer, château, feu, ville, sang, maison	*sein, cœur, chambre, monde, ciel, âme, cabinet, bois, écrit, ville*	*pays, sein, plaine, moment, maison, état[248], cas, bois, monde, suite*	*cas, état, ombre, chambre, maison, pays, salle, coin, salon, rue*	*chambre, cas, ombre, maison, bois, lit, mesure, domaine, pays, condition*

Rappel du tableau 4

Du XVI^e^ au XX^e^ s., *dans* – à l'inverse de *en* – se caractérise par la présence continue et significative, parmi ses dix premiers cooccurrents nominaux les plus spécifiques, de noms dont le sens contextuel dénote majoritairement[249] une réalité concrète dotée d'une extension matérielle ou physique. A cet égard, la récurrence du nom *chambre* nous paraît symptomatique. La plupart de ces noms désignent, quel que soit le siècle, des lieux géographiques construits (*château, ville, maison, salon, ...*), plus rarement des lieux géographiques naturels (*mer, bois, plaine, ...*) voire des entités (*pays*) qui mêlent spatialité (=

[246] R. Huygues (2015 : 6-20) distingue les typologies « ontologiques » du spectre nominal et les typologies « fonctionnelle », les unes et les autres pouvant par ailleurs se combiner.

[247] Nous reconnaissons volontiers de son caractère rudimentaire et discutable. En tout premier lieu, il conviendrait de le raffiner. Pour un nom comme *cœur* par ex., on sait qu'il peut désigner au XVI^e^ s et ensuite, la *faculté* du courage c'est-à-dire une entité abstraite. Insistons sur le fait que nous avons systématiquement envisagé, pour cette étude, les emplois en contexte du N grâce aux concordances et aux retours possibles en plein texte. Ainsi, les emplois du SP *dans Dét cœur* dans nos sous-corpus des XVI^e^ s. et XVII^e^ s. dénotent-ils avant tout le siège (localisé à l'intérieur de l'enveloppe corporelle) des émotions et sentiments. En second lieu, on sait quels « brouillages » entre concret et abstrait peuvent introduire des entités comme les « idéalités concrètes » dénotées par des N comme *sonate, mot, phrase, récit, roman, ...* (voir N. Flaux 2002 par ex.). Il n'en demeure pas moins que, contrairement à ce que suggère cet auteur, nous considérons que le critère que nous utilisons possède une certaine pertinence. Il a en outre l'avantage d'être aisé d'application pour aboutir à un premier classement (perfectible) des cooccurrents nominaux les plus spécifiques.

[248] Dans les tableaux présent (4) et suivant (5), la polysémie du nom *état* n'a pas fait l'objet d'un filtrage ni a fortiori d'un codage. Le terme doit donc être entendu dans tous ses sens. Plus précisément, l'examen des concordances met en lumière que deux sens dominent: celui de « disposition » morale ou physique d'un individu d'une part, celui de « forme de gouvernement » d'autre part. On observe en outre que si au XVIII^e^ s. les deux sens sont présent de manière relativement équilibrée, celui de disposition morale ou physique l'emporte au XIX^e^ s. Probablement ce point serait-il à mettre en relation avec la représentation des discours de la science politique et de l'histoire dans notre corpus et plus largement peut-être avec les préoccupations des auteurs du XVIII^e^ s. *versus* du XIX^e^ s.

[249] Ainsi, le nom *sein* employé derrière *dans* désigne majoritairement au XVI^e^ s. la poitrine (partie du corps), et à partir du XVII^e^ s. majoritairement la partie intérieure d'une réalité abstraite (sein de l'église, etc.)

territoire) et caractéristiques abstraites. On observe plus particulièrement au XVI^e^ s. la présence de trois noms désignant des parties et des productions du corps humain (*entrailles, sang, sein)*, témoins de thématiques souvent liées à la guerre et à la maladie.

On soulignera aussi que dès le XVI^e^ s., *dans* (comme *dedans*) possède parmi ses cooccurrents les plus spécifiques un nom qui n'est pas sans rapport avec le corps[250] quoiqu'il désigne une entité immatérielle : *âme*. Il apparaît en outre clairement que plus on progresse dans le temps, plus cette préposition conquiert dans sa combinatoire hautement spécifique des noms abstraits (*suite, cas, mesure, ...*).

Concernant enfin la combinatoire nominale hautement spécifique que *dans* partage avec les deux autres prépositions, on remarque qu'elle est significative avec *dedans*[251] et avec *en+déterminant* [252] ; pour cette dernière, il est intéressant de contraster les lemmes nominaux du XVI^e^ s. (*ville, maison*) – concrets à extension matérielle ou physique – et celui des XIX^e^ et XX^e^ s.: *cas* (abstrait). Un seul cas enfin de nom hautement spécifique partagé avec *en* suivi d'un nom nu : *état* au XVIII^e^ s.

L'évolution générale de la combinatoire nominale spécifique de *dans* nous conduit à conclure que :

- lors de ses premiers emplois au XVI^e^ s., *dans* se trouve d'abord engagé dans une combinatoire nominale essentiellement concrète-spatiale. Il entre en compétition directe, sur le plan sémantique et distributionnel, avec *dedans* et avec *en + déterminant.* La même situation perdure au XVII^e^ s., à une exception près : sa combinatoire partagée avec *en* décroît.
- A partir du XVIII^e^ s., on observe un tournant : *dans* gagne dans sa combinatoire spécifique des noms abstraits (*état, cas, suite, ...*) et sa combinatoire spécifique partagée avec *en + déterminant*, si elle se prolonge (*cas, moment, pays* pour le XVIII^e^ s., *cas* pour les XIX^e^ et XX^e^ s.), concerne essentiellement des abstraits. Dans le même temps, *dans* conserve une combinatoire nominale hautement spécifique de nature spatiale-concrète (*chambre, cabinet, bois, maison, ...*) qui conduit à considérer que cette préposition possède désormais une combinatoire nominale spécifique hybride, concrète-abstraite.

Si l'on se tourne maintenant vers l'étude statistique du cotexte amont (immédiat) des SP formés de *dans* suivi d'un SN dont la tête est tel ou tel des cooccurrents nominaux réunis dans le tableau 4, on observe qu'à partir du XVIII^e^ s. et pour le seul SP *dans Det cas*, la catégorie qui apparaît comme la plus spécifique est la ponctuation. Au XIX^e^ s., cette situation perdure pour *dans Det cas* et s'étend au SP *dans Det état*[253]. Au XX^e^ s. enfin (1901-1944), on fait le même constat pour les SP *dans Det* (*cas / condition / domaine / mesure*). L'interprétation qu'on peut tirer de ces résultats après retour aux concordances, est que ces SP manifestent, dans chacun des siècles considérés, une disposition statistiquement remarquable au détachement qui les conduit souvent à fonctionner comme des introducteurs de cadres de discours (M. Charolles, 1997) : ils assurent alors au niveau du discours un rôle de marqueur de cohésion textuelle, plus précisément de marque d'indexation (voir M. Charolles, *op. cit.*; L. Sarda, D. Vigier, B. Combettes, 2016). Cela paraît évident pour le SP *dans Dét cas* (dès le XVIII^e^ s., que le N *cas* soit (23) ou non (24) suivi d'un complément, mais cela vaut aussi pour presque tous les autres SP envisagés, comme l'illustrent les exemples suivants où le SP voit

[250] Car localisée dans le corps.
[251] Pour 1501-1600 : *cœur, âme* ; 1601-1700 : *sein, cœur, chambre, âme*.
[252] Pour 1501-1600 : *ville, maison* ; 1601-1700 : *âme* ; 1701-1800 : *pays, moment, cas ;* 1801-1900 : *cas ;* 1901-1944 : *cas*.
[253] Alors qu'au siècle précédent, le SP apparaissait plus spécifiquement derrière des verbes en position liée.

sa portée sémantique (M. Charolles & D. Vigier, 2005) – matérialisée par deux crochets - s'étendre au-delà de sa phrase d'accueil.

(23) ***Dans le cas de l'ignorance***, [*nul doute, par exemple, qu'un conseil ne soit très-utile. Un avocat, un médecin, un philosophe, un politique, peuvent, chacun en leur genre, donner d'excellents avis.*] *Dans tout autre cas, le conseil est inutile. (...)* (1758, C.A. Helvetius, *De l'Esprit*)

(24) *Si, pendant le jour et au milieu du bruit, je réfléchis sur un objet, ce sera assez pour me donner une distraction, que la lumière ou le bruit cesse tout-à-coup.* ***Dans ce cas****, comme dans le premier,* [*les nouvelles perceptions que j'éprouve sont tout-à-fait contraires à l'état où j'étois auparavant. L'impression subite, qui se fait en moi, doit donc encore interrompre la suite de mes idées.*] (1746, Abbé de Condillac, *Essai sur l'origine des connaissances humaines*.)

(25) *On les nomme alors larves ou chenilles. Ils gardent cette forme plus ou moins longtemps après être sortis de l'œuf.* ***Dans cet état****,* [*les insectes sont recouverts d'une peau flasque et molle, divisée en segmens ou anneaux susceptibles de se mouvoir les uns sur les autres, à l'aide de bandelettes musculaires situées dans l'intérieur du corps. Souvent c'est sur ces anneaux seulement que l'insecte rampe, à la manière des reptiles, ou en appuyant alternativement chacun des segmens de son corps sur le plan qui le supporte*. (…)](1805, G. Cuvier, *Leçons d'anatomie comparée*).

(26) *Il faut observer toutefois que, lorsqu'on voit des groupes voisins rester à ce point distincts, c'est que le lien social est resté lâche et qu'il ne s'est point développé encore une force de civilisation capable de réunir et de fondre les contrastes.* ***Dans ces conditions****,* [*les particularités de tempérament sur lesquelles se greffent les habitudes prennent le dessus. Il peut arriver même que des causes artificielles de séparation telles que l'islam en a créées par rapport au christianisme tendent à perpétuer les divisions.*] (1927, J. Maritain, *Primauté du spirituel*).

(27) ***Dans ce domaine***[254] [de la sensation]*,* [*il est très visible que les subtilités, les traditions idéalistes de les philosophes, les ont amenés à de véritables contradictions, tandis que le réalisme naïf et irraisonné des savants les a mieux servis, ou plutôt les aurait beaucoup mieux servis s'ils n'avaient pas cru devoir faire les philosophes. Que l'on songe encore aux véritables absurdités où sont tombées les théories dites « génétiques » de l'espace (...)*] (1930, R. Ruyer, *Esquisse d'une philosophie de la structure*).

Le cas du nom *mesure* est à part : son fort indice de spécificité lorsqu'il suit *dans* signe le figement du syntagme *dans la mesure où* en une locution qui peut elle aussi entrer dans la formation d'un introducteur de cadre de discours :

(28) ***Dans la mesure*** *au surplus* ***où les théoriciens marxistes de l'URSS élaborent une métaphysique****,* [*c'est à une sorte d'hylozoïsme qu'ils reviennent, leur ligne générale philosophique demande qu'on attribue à la matière quelque chose*

[254] Pour un étude en diachronie de *dans dét domaine (de)* et d'autre marqueurs de topique de discours dans la presse nationale (*Le Figaro)* du XIX^{e} s. et au XX^{e} s., voir M. Charolles, S. Diwersy & D. Vigier (2017).

comme l'âme et la liberté ; ils n'osent pas dire encore, comme les vieux physiologues de l'Ionie : tout est plein d'âme et de divinité répandue, panta plêré théôn ; mais c'est bien dans ce sens qu'ils semblent se diriger.] (1930, J. Maritain, *Humanisme intégral).*

Ces observations croisent celles formulées par B. Fagard & L. Sarda (*op. cit.*), à savoir qu'entre le XVI^e^ s. et aujourd'hui, non seulement on assiste à une diversification croissante des sens de *dans* (spatial, temporel, abstrait, ….) mais que, détachés, les SP ayant *dans* pour tête sont rarement spatiaux et connaissent une très forte tendance au figement.

Nos analyses apportent cependant des éclairages supplémentaires qui nous semblent dignes d'intérêt.

En premier lieu, nos outils d'exploration et de calcul permettent de travailler au niveau des lexèmes nominaux et donc d'identifier, par l'étude des cooccurrents les plus spécifiques de *dans,* des champs lexicaux préférentiels, ce qu'une étude plus « traditionnelle » fondée sur le dépouillement des concordances par l'œil humain, ne permet pas dès qu'on atteint plusieurs milliers d'occurrences. Ainsi, le tableau 4 suggère-t-il la place significative que revêt le champ sémantique du corps humain au XVI^e^ s. dans le régime de la préposition, ou encore la permanence depuis le XVII^e^ s. du champ sémantique de l'habitat quotidien (*chambre, maison, lit*). Ou encore la place toute particulière qu'occupe *l'âme* dans les textes littéraires préclassiques et classiques, qui y apparaît - lorsqu'on fait retour aux concordances – avant tout comme le lieu où s'épanouissent et s'affrontent les passions, mais aussi comme la partie de l'homme reliée au divin :

(29) *Il n'y a rien de plus funeste que la guerre, c'est la Colere qui l'allume. Elle étouffe toutes les autres Passions, quand elle regne* ***dans une ame****.* (1640, J.-F. Senault, *De l'Usage des passions.*)

(30) *De grace qui a mis ceste loy à le fond de vostre cœur, à ce esté vostre pere, ou vostre mere ? Rien moins, car ils n'y pensoient seulement pas ; il faut donc que Dieu vous l'ait enpreinte* ***dans l'ame****, car vous ne la luy avez pas mise (…)* (1624, Le Père Marin Marsenne, *L'impiété des déistes, athées et libertins de ce temps*

« Ces observations valent pour votre corpus » nous objectera-t-on… Oui et non. Non, dans la mesure où elles convergent pour partie avec celles avancées par P. Blumenthal (2011) à partir de l'étude instrumentée d'un corpus pour partie distinct de Presto. Oui, parce que c'est le discours littéraire qui domine dans notre corpus, comme c'était le cas dans le corpus de P. Blumenthal. A cet égard, il est presque certain qu'un corpus de textes relevant d'autres discours – par ex. juridique, ou technique – révèlerait d'autres préférences statistiques.

En second lieu, nos outils permettent d'étudier finement des spécificités combinatoires propres à certaines positions syntaxiques occupées par le SP dans la phrase, et qu'un examen manuel des concordances peinerait à faire apparaître. Nous en avons donné un premier aperçu *supra* avec les SP « *dans Det. cas / état / condition* etc. *»*. On peut en donner une seconde illustration en examinant cette fois la liste des premiers[255] lemmes nominaux préférés qui

[255] Pour des raisons d'effectifs, nous avons travaillé pour cette construction sur le corpus Presto–^{ETENDU}. Pour les XVI^e^ s., XVII^e^ s. et XX^e^ s. (1901-1944) le nombre de collocatifs inférieur à 5 dans le tableau s'explique par : i) le faible effectif de la préposition (XVI^e^ s.), ii) par la réduction drastique des effectifs (f_{ij}, *F)* sur lequel porte le calcul, iii) le fait que nous avons écarté les constructions sur-représentées chez un auteur.

figurent après *dans* lorsque le SP occupe la tête de la phrase après ponctuation forte[256]. Voici les noms extraits :

1501-1600	1601-1700	1701-1800	1801-1900	1901-1944
antre	*temps*	*temps, cas, moment, pays, état*	*cas, pays, état, temps, circonstance*	*cas, condition*

Tableau 5

Liste des cinq premiers cooccurrents (lemmes) nominaux les plus spécifiques derrière *dans* lorsque le SP est placé en tête de phrase derrière ponctuation forte. Corpus Presto–ETENDU_1509-1944.

Si on compare cette liste à celles du tableau 4[257] (cellules affectées à *dans*), on peut insister sur les points suivants :

- globalement, l'évolution de la combinatoire nominale préférée de la préposition reste proche de ce qu'elle était dans le tableau 4 : on part d'un collocatif dénotant une réalité physique spatialisée (XVI^{e} s. : *antre)* pour aboutir à des entités nettement abstraites (XX^{e} s. : *cas, condition)* en passant par des N dénotant des entités temporelles (*temps, moment*).
- mais on observe aussi des différences notables
  - le N *temps*[258] apparaît préféré dans la tranche du XVII^{e} s., alors que dans le tableau 4, le champ sémantique du temps y était absent. La position initiale semble donner une prime à l'abstraction.
  - Cette dernière observation peut être reconduite pour les XVIII^{e}, XIX^{e} et XX^{e} s. où a disparu (*versus* dans le tableau 4) tout N dénotant une réalité spatiale naturelle ou construite, au profit de noms essentiellement temporels (*temps, moment)* ou abstraits (*cas, état, temps, circonstance, condition*) - hormis le N *pays* qui mêle spatialité et caractéristiques abstraites.

En d'autres termes, l'examen des lemmes nominaux préférés apparaissant dans la suite de *dans* lorsque la préposition figure en tête de phrase derrière ponctuation forte semble montrer que, contrairement aux emplois où le SP figure dans n'importe quelle position dans la phrase,

- le « tournant » vers l'abstrait de sa combinatoire préférée (en l'occurrence, l'abstraction temporelle) s'opère un siècle plus tôt (XVII^{e} s.) ;
- cette combinatoire s'avère dès le XVIII^{e} s. fondamentalement abstraite et ne présente pas le caractère hybride concret-abstrait que l'on observe lorsque le SP occupe n'importe quelle position dans la phrase.

---

[256] Cette enquête diffère de la précédente en ceci que nous partons cette fois du pivot : « ponctuation forte + *dans* » (requête : [pos="Fs"][lemma= "DANS"& pos="S"]) pour explorer ses préférences combinatoires nominales à droite dans une fenêtre de trois mots. Juste auparavant, nous avions exploré le contexte gauche des pivots constitués par la préposition *dans* suivi d'un SN dont la tête était l'un ou l'autre des cooccurrents préférés de cette préposition (quelle que soit la position syntaxique occupée par le SP). Autrement dit, l'objectif est distinct. Antérieurement, il s'agissait de déterminer à partir de quand les cooccurrents nominaux préférés de *dans* (toutes positions syntaxiques du SP confondues) forment avec lui des syntagmes qui marquent une disposition particulière à se placer en tête de phrase (en lien avec un figement). Dans le cas présent, il s'agit pour chaque siècle d'identifier quels sont les noms qui tendent préférentiellement à apparaître après *dans* (combinaisons libres ou figées) lorsque le SP est détaché en tête de phrase. Bien entendu, les résultats de ces deux enquêtes se croisent partiellement et l'on y retrouve, de fait, les noms *cas, état, condition.*

[257] *modulo* la différence entre les deux corpus considérés.

[258] La place éminente du N *temps* parmi les collocatifs de *dans* pour les XVII^{e} et XVIII^{e} s. est pour partie liée à l'usage dominant en position frontale de la locution figée *dans le temps que* qui disparaît au XIX^{e} s. (même si *temps* continue à être utilisé significativement dans cette position, mais avec des constructions plus variées).

Toutes choses parfaitement cohérentes avec ce que nous ont appris par ailleurs la grammaticalisation et l'étude des connecteurs, quant à la corrélation existant entre la position détachée des constituants en tête de phrase et leur tendance (sémantique) à l'abstraction (voir par ex. B. Lamiroy & M. Charolles, 2004 : 63).

3.2. Les cooccurrents nominaux les plus spécifiques de *en*

3.2.1. *En* suivi d'un nom actualisé par un déterminant

1509-1600	1601-1700	1701-1800	1801-1900	1901-1944
lieu, endroit, chambre, maison, ville, place, sorte, présence, manière, instant	*lieu, endroit, état, façon, occasion, temps, âme, sorte, pays, place*	*mot, lieu, moment, faveur, endroit, cas, sorte, présence, pays, façon*	*sorte, moment, mot, sens, lieu, terme, faveur, cas, air, absence*	*cas, moment, sens, terme, air, sorte, lieu, manière, honneur, occurrence*

Rappel du tableau 4

Au XVI[e] s., *en* suivi d'un nom actualisé par un déterminant présente parmi ses cooccurrents nominaux les plus spécifiques des noms concrets spatiaux (*maison, chambre, ville*). Comme observé à plusieurs reprises, ce siècle se caractérise (comme le XVII[e] s. : voir *infra*) par l'existence d'une zone sémantique et distributionnelle partagée par *en, dans, dedans* – ce dont témoignent aussi leurs cooccurrents nominaux les plus spécifiques (voir *maison, ville* partagés par *en* et *dans)* - , toutes trois pouvant exprimer la localisation spatiale d'une cible dans les bornes d'un site à l'issue ou non d'un déplacement. Par ex. :

(31) (…) *sans nul empeschement ni difficulté, nous* ***entrasmes*** *tous trois* ***dans ceste maison***. (1598 ; J. de Léry, *Histoire d'un voyage fait en la terre du Brésil)*

(32) *Quand elle* ***feut entree en sa maison*** (…) (1542, F. Rabelais, *Pantagruel)*

(33) (…) *Pour* ***entrer dedans la maison****, Luy a faict perdre la raison*. (1562, J. Grévin, *La trésorière*)

Comme nous l'avons indiqué ailleurs (L. Royer & D. Vigier, 2012 : 429), la distinction entre ces emplois quasi-synonymiques des trois prépositions est le plus souvent liée aux préférences observables en corpus quant à « la catégorie du déterminant qui actualise le nom régime de la préposition, autrement dit le mode de donation de la référence ». Ainsi observe-t-on dans Presto–[ETENDU] que pour la séquence *entrer dans*, le déterminant le plus spécifique pour actualiser le régime nominal au XVI[e] s. est l'article défini alors qu'il s'agit de l'article indéfini pour *entrer en*, la séquence *entrer dedans* ne manifestant pas de préférence statistiquement spécifique.

Toujours au XVI[e] s., la combinatoire nominale spécifique de *en* + *déterminant* se caractérise - par contraste avec celles de *dans* et *dedans* - par la présence des noms de lieu à valeur générique que sont *lieu, endroit*, *place*[259] (« noms généraux d'espace » pour R. Huygues, 2009), noms que l'on retrouve parmi les cooccurrents les plus spécifiques de *en* +

[259] Ce nom peut désigner aussi la place publique d'un village ou d'une ville. Mais l'examen des concordances montre que cette valeur ne domine pas.

déterminant dans les siècles suivants - jusqu'au XVIII[e] s. pour *endroit* et jusqu'au XX[e] s. pour *lieu*. Il y aurait là un point intéressant à creuser, en lien avec non seulement les travaux récents menés sur ces noms de localisation, mais aussi en relation avec la préposition *à* (voir les travaux de M. Aurnague, 2009, 2010), préposition dont on a vu *supra* combien le sort pouvait être lié, dans certains de ses emplois, à *en*[260].

Enfin, entre 1501 et 1600, *en* suivi d'un régime nominal actualisé manifeste une préférence très spécifique pour certains noms dénotant des réalités abstraites (*sorte, présence, manière, instant),* mots exclus de la combinatoire hautement préférée de *dans* ou *dedans* pour cette même période.

Pour conclure, si *en* + *déterminant* partage bien une zone distributionnelle et sémantique avec *dans* (et *dedans*) au XVI[e] s., *en* et *dans* se distinguent nettement l'une de l'autre en ceci que la première voit figurer aussi parmi ses cooccurrents les plus spécifiques des noms d'espace génériques auxquels *dans* n'a pas accès, ainsi que des noms dénotant des réalités abstraites. D'une certaine façon, *en* + *déterminant* manifeste au XVI[e] s. une polarité nominale spécifique hybride concret-abstrait qui deviendra, deux siècles plus tard, l'apanage de *dans*.

Le XVII[e] s. marque un tournant en ceci que la combinatoire nominale spécifique de *en* + *déterminant* tend à s'affranchir des noms dénotant des réalités non génériques pourvues d'une extension matérielle ou physique - tels *maison, chambre, ville* au XVI[e] s. Certes, le nom *pays* joue encore contextuellement le rôle de site spatial (jusqu'au XVIII[e] s.), mais les autres noms de lieu sont désormais des noms généraux d'espace (*endroit, place* et *lieu* qui perdure jusqu'au XX[e] s.). En outre, la part réservée aux noms abstraits (*état, façon, occasion, temps, âme, sorte*) s'accroît, et ce mouvement d'accroissement ne se démentira pas jusqu'au XX[e] s. (où l'on dénombre huit noms abstraits sur dix : *cas, moment, sens, terme, sorte, manière, honneur, occurrence*).

On peut être enfin frappé par la remarquable stabilité des noms *sorte* et *lieu* qui figurent parmi les dix premiers collocatifs préférés du pivot *en Det* durant cinq siècles. Un retour aux concordances prenant pour pivot la préposition *en* suivie d'un déterminant puis du nom *sorte/lieu* permet d'identifier, siècle après siècle, diverses expressions plus ou moins figées dont la haute fréquence confère une « prime » à l'indice de spécificité du cooccurrrent nominal de *en*. Le tableau suivant les récapitule (l'étendue de la flèche matérialise la période d'usage de l'expression).

[260] Dans D. Vigier (2015) nous avons exploré les diverses valeurs de la séquence *en cet endroit* au XVI[e] s. et souligné que *dans cet endroit* n'apparaissait qu'au XVII[e] s., pourvu d'une valeur essentiellement spatiale.

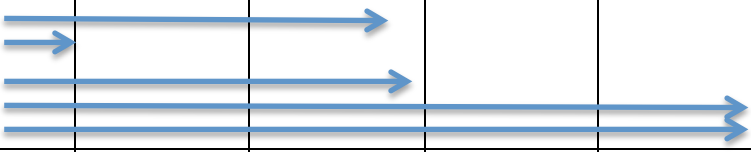
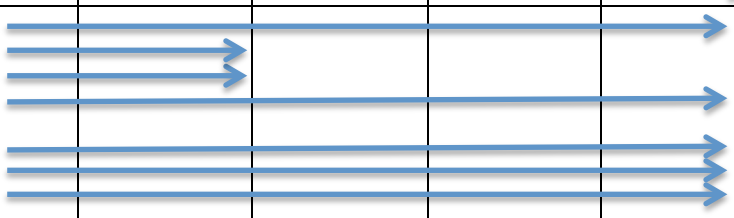
Syntagmes prépositionnels	XVI	XVII	XVIII	XIX	XX
en Det Poss lieu(x) *ès lieu(x)* *en quelque lieu* *en (tous/t, un) lieu(x)* *en (ce/ces) lieu(x)*					
en (cette, ces) sorte(s) *en la sorte* *en quelle sorte* *en quelque sorte* *en telle sorte* *en toute(s) sorte(s)* *en une sorte*					

Tableau 6
SP ayant pour tête *en* et pour régime nominal *lieu(x)/sorte* entre le XVI^e^ s. et le XX^e^ s.

Cet ensemble de syntagmes mériterait une étude détaillée, en vue d'examiner en particulier leurs positions dans la phrase, le figement dont certains d'entre eux sont le siège, ainsi que les variations de sens et de fréquence que ces figements engendrent.

Nous proposons d'illustrer ici une telle démarche par l'étude du SP *en quelque sorte* [261], fort intéressant en ceci qu'à son figement s'associe un phénomène de « pragmaticalisation[262] » (G. Dostie, 2004).

Insistons d'abord sur le fait que ce SP est construit avec un régime nominal actualisé par un déterminant indéfini *quelque*, issu du relatif *quel* qu'on trouve encore au XVI^e^ s. :

La seconde espece a grand efficace contre tout venin de bestes sauvages, principallement contre celuy des serpens et viperes, ***en quelle sorte que*** *vous la prenes, soit en boire, soit en menger, ou portée sur soy*. (1557, R. Dodoens, *Histoire des plantes*)

Ce terme *quel* est définitivement supplanté par *quelque* au XVII^e^ s.

Comme tout indéfini, *quelque* peut être regardé comme un « opérateur de parcours[263] » (P. Le Goffic, 1994 : 31).

Au XVI^e^ s., dans notre corpus, le SP *en quelque sorte* est d'emploi intraprédicatif et convoie un sens de *manière* : il participe, sur le plan sémantique, à la construction du sens référentiel (il est *endophrastique* au sens de C. Guimier 1996). De par la présence de l'indéfini *quelque*, le SP invite à opérer un parcours des « manières de faire » du procès verbal, comme le montre l'exemple suivant :

(34) *Toutesfois je n'impose point loy à ceux qui auront failly* ***en quelque sorte****, de faire tous un semblable vœu.* (1560, J. Calvin, *Institution de la religion chrestienne*)

[261] Le N *sorte* est issu du latin *sors, sortis* « sort ; rang, condition, catégorie ». Il prend à la fin du XV^e^ s. le sens de «manière de faire une chose, façon ». Le TLFi date la première occurrence de la locution adverbiale *en quelque sorte* au sens de « pour ainsi dire » en 1650 (P. Corneille, *Don Sanche d'Aragon*, I, 3, 251).

[262] Par « pragmaticalisation », il faut entendre que le SP passe d'une valeur sémantique en relation avec les dimensions du contexte référentiel à une valeur pragmatique, commentative, en relation avec l'énonciation. Voir sur ce point aussi B. Lamiroy & M. Charolles, *op. cit.*: 62.

[263] « Une opération de parcours consiste, comme son nom l'indique, à balayer toutes les valeurs possibles et imaginables susceptibles de vérifier (valider) une propriété. » (P. Le Goffic, 1994 : 32)

qu'on pourrait paraphraser comme suit :

> *Toutesfois je n'impose point loy à ceux qui auront failly* ***d'une manière ou d'une autre****, de faire tous un semblable vœu.*

Il est fréquent que *en quelque sorte* soit suivi d'une relative épithète :

(35) *Aucuns disent aussi, que* ***en quelque sorte que l'on prenne les fueilles ou la racine****, qu'il lache le ventre, et faict aller à chambre*. (1557, R. Dodoens, *Histoire des plantes*)

Il y a alors parcours de la manière du procès dénoté par le SV de la principale (*prenne les fueilles ou la racine*), parcours « ouvert » en ceci que le locuteur envisage toutes les façons possibles de s'administrer la plante en question (le « *piganum* ») sans en privilégier aucune et sans que ne soit jamais remise en cause sa vertu (*« il lache le ventre, et faict aller à chambre»*). On peut parler de relation de concession entre le SP et la principale dans la mesure où est identifiable en arrière-plan une règle d'inférence bloquée selon laquelle, dans le domaine de la médication, s'administrer une plante sous n'importe quelle forme met en péril sa vertu thérapeutique.

On observe en outre à cette époque la présence d'une expression renforcée ou le N *sorte* est coordonné avec le N *manière*, venant ensuite une relative épithète *:*

(36) *Car* ***en quelque sorte et maniere que le sort de cette guerre vienne à tomber****, quand tu auras ainsi departi ton or et ton argent entre les soldats, il n'est possible que le proffit ne t'en demeure* [= *de quelque manière que le sort de cette guerre vienne à tomber*] (1577, B. de Vigenère, *L'histoire de la décadence de l'empire grec, et établissement de celui des Turcs.*)

Il est intéressant de souligner enfin que cette expansion relative peut elle-même se figer en la relative *quoi que ce soit*, donnant alors la locution figée *en quelque sorte que ce (soit / fut / semi-aux. + être)* :

(37) *(...) et croyent que l'Homme ne peut avoir aucune chose de Felicité, sinon la delectation****, en quelque sorte que ce soit.*** (1577, L. L'Hébreu, *Philosophie d'amour*).

On sait que C. Muller entre autres (2006, 2007) voit dans cette relative une des structures possibles ouvrant en français à une interprétation dite « free choice » de l'indéfini, « *que ce soit* signifiant l'exhaustivité de la prise en compte du domaine référentiel du nom » (2007 : 94).

Cette expression figée disparaît presque entièrement au XVII^{e} s.

Pour en revenir à notre fil directeur de la « pragmaticalisation » de *en quelque sorte,* c'est au XVIII^{e} s. selon nous que se généralise pour cette expression le sens d'approximation (pragmatique) qu'on lui connaît dans son sens moderne, devenant ainsi *exophrastique.* Elle exprime désormais un commentaire sur l'énonciation et ne participe plus à la construction du sens référentiel. Ainsi dans :

(38) *Beaucoup de celles* [=les îles] *qui sont dans l'océan Indien, ont* ***pour ainsi dire*** *deux hémisphères, l'un oriental, l'autre occidental, divisés par des montagnes, qui vont du nord au sud (...) Nous venons de dire que la nature avoit donné* ***en quelque sorte*** *deux hémisphères aux premières* (…) (1784, B. de Saint Pierre, *Etudes de la nature*).

La succession reformulative qui lie « *pour ainsi dire* » à « *en quelque sorte* » met en lumière le rôle d'approximation - ou d' « enclosure[264] » - G. Kleiber & M. Riegel, 1978 ; D. Legallois, 2002 - joué par l'adverbial.

Le XVII^e^ s. apparaît quant à lui comme une période de transition (« bridging context »), en ceci que les contextes d'ambiguïté apparaissent très nombreux dans notre corpus. Ces ambiguïtés surgissent le plus souvent dans les emplois où *en quelque sorte* modifie un adjectif ou un participe, et où il devient difficile de faire le partage entre la manière et l'approximation. Par ex.

(39) *Je reconnoissois bien que la nature avoit* ***en quelque sorte*** *advantagé Celadon par dessus Lycidas, toutefois sans en pouvoir dire la raison, Lycidas m'estoit beaucoup plus agreable.* (1612, H. D'Urfé, *L'Astrée*)

(40) *Une deesse oublie tous les dieux, pour regarder un seul homme ; et trouve quelque chose en la terre, qui luy fait mespriser les cieux. Il n'est rien de plus bas, que ce qui retient ordinairement ses regards, et sa pensée ; et par son affection, elle devient* ***en quelque sorte*** *humaine, et mortelle.* (1624, J. de Gombauls, *L'Endimion*.)

Faut-il ici paraphraser *en quelque sorte* par « *d'une manière ou d'une autre, par un certain côté, ...* » (sens de *manière*) ou bien par « *pour ainsi dire* » ? Il semble parfois impossible de trancher.

Si nous ne faisons pas fausse route, nous conclurons provisoirement (sous réserve d'une étude plus approfondie) que c'est au XVII^e^ s. que le SP *en quelque sorte* a commencé à être le siège d'un processus de pragmaticalisation, processus en voie d'aboutissement au XVIII^e^ s. et achevé au XIX^e^ s.

[264] *« Dans le cadre de la logique floue, les enclosures deviennent des prédicats qui transforment la fonction d'appartenance à une classe »* (G. Kleiber & M. Riegel, 1978 : 93). Dans l'exemple présenté, les enclosures « *pour ainsi dire / en quelque sorte* » permettent de faire entrer dans la classe des « hémisphères » les deux zones géographiques que l'auteur distingue sur les îles dont il parle.

3.2.2. *En* suivi d'un nom nu

1501-1600	1601-1700	1701-1800	1801-1900	1901-1944
vain, bref, fin, dépit, pièce, général, terre, repos, haut, paix	*vain, fin, effet, suite, général, paix, colère, peine, repos, dépit*	*effet, vain, général, conséquence, faveur, vérité, état, paix, sûreté, œuvre*	*vain, dehors, face, général, présence, train, résumé, paix, définitive, vertu*	*effet, face, vain, réalité, arrière, train, haut, dehors, somme, définitive*

Rappel du tableau 4

La combinatoire nominale hautement spécifique de cet emploi de *en* apparaît très différente de celle examinée juste auparavant pour cette même préposition (même si au XVIII[e] s. le nom *état* apparaît partagé à la fois par *dans* et par *en* + *déterminant*)

Au XVI[e] s., on ne trouve pas - contrairement aux emplois de *en* suivi d'un déterminant – de noms de lieu dénotant des espaces construits (voir *chambre, maison, ville*). Pour autant, les noms concrets spatiaux ne sont pas absents : le nom *terre* en premier lieu. La préposition *en* possède, dans cet emploi, le plus souvent le sens de *sur* en français moderne, parfois de *en.* Par ex.

(41) *Allors descendit Gymnaste de son cheval, & montant au noyer souleva le moyne par les goussetz d'une main & de l'autre deffist sa visiere du croc de l'arbre, & ainsi le laissa tomber* ***en terre*** ([= sur le sol]*, & soy apres.* (F. Rabelais, 1542, *Gargantua*).

(42) (…) *tout ce qu'ils auront lié* ***en terre*** [= sur la terre] *sera lié au ciel.*(J. Calvin, 1560, *iNstitution de la Religion Chrestienne*).

(43) *S'un povre homme d'argent n'a poinct, Et qu'il advienne à la maleure Que sa bonne femme luy meure, Ja* ***en terre*** [= en terre] *on ne la mectra.* (Anonyme, 1530, *Six pièces polémiques du recueil La Vallière*)

Très fréquemment en outre, le SP *en terre* entre dans un réseau d'isotopies fortement polarisé lié au domaine religieux (terre/ciel, hommes/Dieu, mort/résurrection etc.). La « prime » donnée au nom *terre* inscrit dans un tel réseau n'est guère étonnante lorsqu'on sait que le premier lemme nominal en termes de fréquence pour la tranche Presto–[XVI] est *dieu.* Enfin, un calcul des cooccurrences prenant pour pivot l'expression *en terre* et pour contexte une fenêtre de dix mots en amont et en aval confirme ce que la lecture des extraits réunis dans le concordancier faisait soupçonner : le nom le plus spécifique qui apparaît dans le voisinage de *en terre* est le mot *ciel*.

Toujours parmi les coccurrents nominaux préférés dénotant un lieu, on soulignera la présence des *noms de localisation interne* (NLI) *haut, fin* (M. Aurnague, 1996 ; A. Borillo, 1998) et du nom de partie *pièce*[265]. Il semble qu'il existe dès le XVI[e] s. et au-delà (voir *face, haut* et *arrière* au XX[e] s.) un rapport privilégié entre la préposition *en* suivi d'un nom nu et la méronymie, plus particulièrement encore avec les noms de lieu (NLI).

Il n'en reste pas moins que dès le XVI[e] s. et plus encore ensuite, ce sont les noms abstraits qui dominent dans la liste des cooccurrents préférés de *en* suivi d'un nom nu.

Comme pour le pivot précédent *en Det*, on observe que certains termes figurent durant cinq siècles parmi les collocatifs préférés de *en* : tel est le cas de *vain* (*général* manquant de

[265] L'expression figée étant : *en pièces.*

peu la même trajectoire). Pourtant, la situation est fort différente puisque ce ne peut être, cette fois, qu'une seule et même expression figée (*en vain*) qui explique la « prime » donnée au terme nominal. Il conviendrait donc de se pencher sur les raisons qui font que, dans le discours littéraire essentiellement, cet adverbial ait eu un succès si constant.

3.3. Les cooccurrents nominaux les plus spécifiques de *dedans*

1501-1600	1601-1700
cœur, âme	*cœur, âme, yeux, chambre, sein, lit, fort, logis, air, eau*

Rappel du tableau 4

Nous n'aurons ici que peu de chose à dire, du fait d'une part que *dedans* préposition s'éteint dès la fin du XVII^e^ s[266]., d'autre part que nous avons déjà traité pour partie de sa combinatoire dans le paragraphe consacré à *dans*. Nous rappellerons simplement les deux points suivants :

- dès la seconde moitié du XVI^e^ s., *dedans* partage une partie remarquable de sa combinatoire nominale spécifique avec *dans ;*
- l'essentiel de cette combinatoire concerne des noms dotés contextuellement d'une extension matérielle ou physique, certaines désignant des parties du corps.

3.4. Conclusion de la troisième section

Au terme de ses deux études conduites sur *en* et *dans*, G. Gougenheim concluait :

« La naissance de *dans* à côté de *en* au XVI^e^ siècle a permis à la langue moderne de différencier ces deux outils grammaticaux. *Dans* a pris la valeur spatiale et matérielle de l'ancien *en.* » (1954, [1970] : 54)

« Le centre de gravité de la préposition [*en*] s'est en quelque sorte déplacé et l'emploi local et temporel de *en* n'est plus son emploi dominant. (…) *En* traduit [aujourd'hui] la tendance à l'identité, à la prise de possession par le dedans ». (1954, [1970] : 55, 65).

Notre propre étude permet de reconsidérer cette thèse sur certains points.

S'il est juste de considérer que *en* a pour partie abandonné à *dans* sa combinatoire spécifique avec les noms de lieu (par ex. pour les noms déterminés, *ville* et *maison* cessent de figurer parmi ses cooccurrents préférés dès le XVII^e^ s.), on peut affiner notre connaissance de cette évolution en observant que *en* a continué cependant d'entretenir jusqu'en français moderne un lien privilégié avec la localisation. Par le biais de la localisation « directionnelle » et « abstraite » d'une part. Nous avons constaté que derrière régime non actualisé, *en* manifeste une attraction spécifique pour les noms de localisation interne (NLI) avec lesquels le SP permet d'opérer des repérages directionnels, cela dès le XVI^e^ s. et jusqu'au XX^e^ s.. D'autre part, *en* conserve jusqu'au XX^e^ s. un lien hautement spécifique aussi avec les noms

[266] Le fait que l'on ne dénombre que peu de collocatifs nominaux pour cette préposition au XVI^e^ s. (*versus* au XVII^e^ s.) est avant tout dû au fait que i) elle est d'un emploi notablement plus rare que *en* et même *dans* ; ii) la taille du sous-corpus Presto–^XVI^ est nettement inférieure à celle du sous-corpus Presto–^XVII^ (respectivement 750.709 mots et 1.297.448 mots).

généraux d'espace *lieu, endroit, place* (en régime déterminé). Et cela même s'il faut tempérer aussitôt cette remarque en soulignant que ces noms prennent progressivement, au contact de *en*, une valeur abstraite non spatiale. Ainsi le nom *lieu* est-il entré dès le XVII^{e} s. dans un paradigme d'expressions figées jouant un rôle de marqueur d'intégration linéaire (i.e. de connecteur de cohésion textuelle). Par ex.

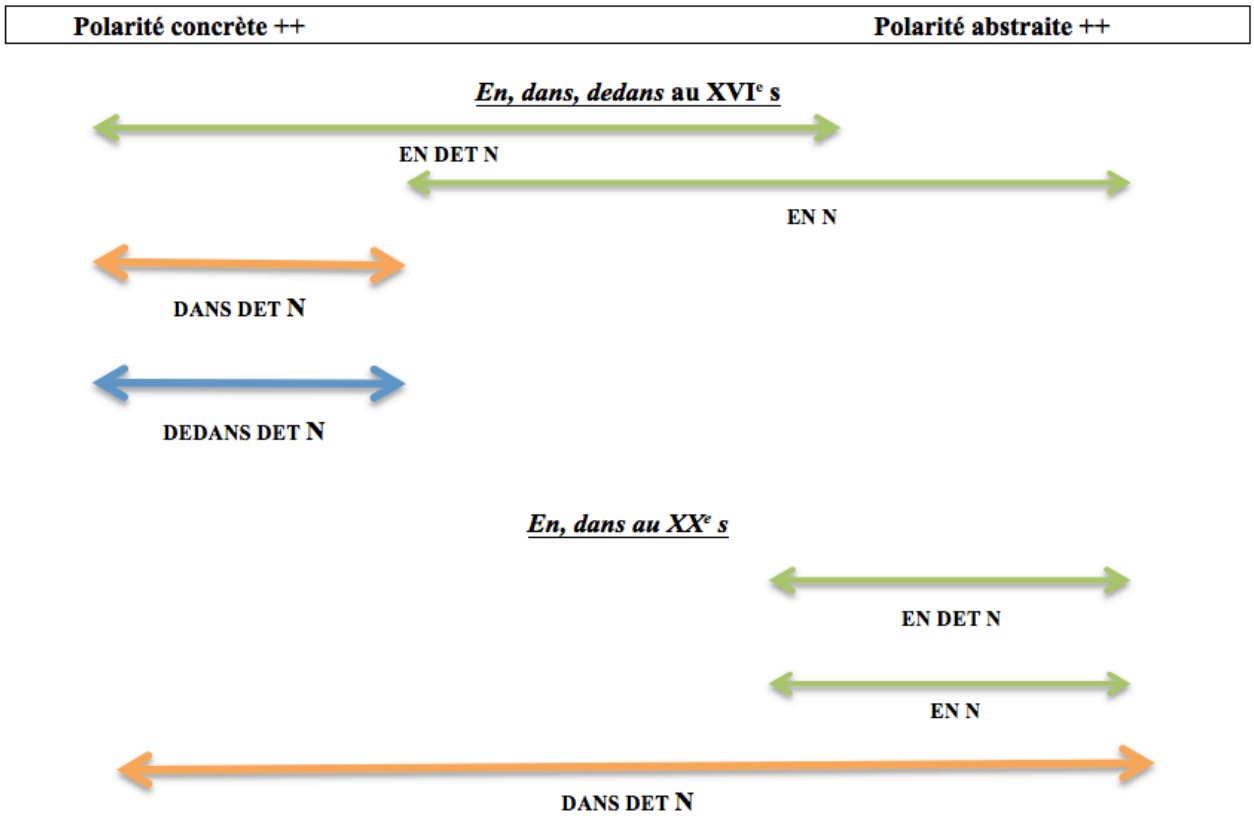
(44) *Pour l'entiere resolution d'un mariage, trois actions doivent entrevenir quant a la damoiselle que l'on veut marier : car* ***premierement****, on luy propose le parti ;* ***secondement****, elle aggree la proposition, et* ***en troisiesme lieu****, elle consent.* (1619, Saint François de Sales, *Introduction à la vie dévote*.)

Cet emploi « abstrait » de *lieu* était très marginal au XVII^{e} s., sa valeur spatiale dominant largement (*en ce(s) lieu(x), en tou(s)(t) lieu(x), en un lieu, ...)*. Ce n'est qu'au XIX^{e} s. que cet emploi de connecteur s'est imposé comme dominant derrière *en*.

Parallèlement, il est partiellement vrai seulement que *dans* aurait « *pris la valeur spatiale et matérielle de l'ancien* en *»*. Nous avons observé en effet que *dans* au XVIIIe s. conquiert une combinatoire nominale hautement spécifique à polarité abstraite qui le conduit à partager avec *en* (avec ou sans déterminant : voir *état*) plusieurs cooccurrents. En d'autres termes, si notre étude (focalisée sur les cooccurrents nominaux préférés de la triade *en, dans, dedans*) montre qu'il y a bien eu un déplacement du centre de gravité sémantique de *en* vers des valeurs abstraites, elle montre aussi que *dans,* après s'être en quelque sorte nourri des dépouilles de sa concurrente, a rapidement emprunté à son tour le chemin de l'abstraction.

Terminons avec *dedans* : il est frappant que G. Gougenheim ne l'inclue pas à son scénario. Or si *dans* a effectivement remplacé *en* dans certains de ses emplois spatiaux-temporels, sa proximité en termes de combinatoire nominale spécifique est encore plus forte avec *dedans*. De fait, si pour la période 1550-1700, on compare le pourcentage de cooccurrents nominaux spécifiques (indice de spécificité ≥ 3) que *dedans* partage avec *en* et *dans*, on trouve dans le premier cas 20% et dans le second 43%. Ces chiffres sont cohérents avec la comparaison des combinatoires partagées entre *en, dans* et *dedans* réduites aux dix premiers cooccurrents nominaux les plus spécifiques sur la période XVI^{e} s. et XVII^{e} s. (tableau 4). En d'autres termes, *dans* a aussi - et surtout… - pris « la valeur spatiale et matérielle de l'ancien » *dedans*…

Le tableau ci-dessous récapitule l'essentiel de ce que nous venons de dire en présentant les *zones d'emploi les plus spécifiques* de *en, dans, dedans* entre le XVI^{e} s. et le XX^{e} s, zones repérées ici selon un axe bipolaire *concret(spatial)-abstrait*.



Conclusion de la troisième partie

Cette troisième section avait pour principal objectif d'illustrer (quoique modestement) en quoi les analyses quantitatives sur corpus outillé peuvent apporter de nouveaux éclairages dans l'étude diachronique de la sémantique et de la pragmatique des prépositions. Nous avons notamment cherché à montrer qu'elles pouvaient permettre de trancher certains débats parfois anciens (à l'exemple de la naissance de « *dans* ») ou apporter des éclairages inédits (celui des cooccurrents spécifiques) à l'étude sémantique de telle ou telle préposition.

Il reste que les résultats et les analyses que nous venons de présenter sont entachés de nombreuses imperfections : nous voudrions en évoquer ici quelques-unes, sans prétention d'exhaustivité.

En premier lieu, l'analyse que nous avons proposée des cooccurrents nominaux de *en, dans, dedans* (§ 3.3.) ne peut se satisfaire d'une étude fondée seulement sur les dix collocatifs les plus spécifiques. Elle devra impérativement s'étendre à un nombre plus étoffé : *a minima*, une centaine probablement pour chaque siècle, pour ce qui regarde en tout cas *en* et *dans*. Nous envisageons ainsi d'opérer un codage sémantique de l'ensemble de ces cooccurrents nominaux en recourant d'une part au critère concret/abstrait vu plus haut, d'autre part à une grille des grandes catégories sémantiques mises en jeu dans les contextes d'emploi explorés[267] (espace, temps, notionnel - cause, conséquence, concession etc.) en vue de cerner ensuite quels sont les traits (et les complexes de traits) sémantiques dominants qui émergent de ces cooccurrents. Nous sommes certes conscient, là encore, du caractère « à gros grain » de ce

[267] Avec les difficultés de la polysémie que nous avons soulevée *supra* (ex. de *état* pour la séquence : *dans dét. état*) et à laquelle il nous faudra s'affronter dans ce codage.

type de codage[268] mais les résultats auxquels il permettrait d'aboutir constituerait déjà un degré supplémentaire de précision par rapport à l'étude présentée dans ce mémoire.

Il conviendrait en second lieu d'introduire de nouveaux calculs et de nouvelles méthodes statistiques dans nos études, jusqu'ici très cantonnées à ceux issus de la textométrie et mettant en jeu prioritairement les spécificités de P. Lafon. Cette restriction quelque peu drastique de notre spectre d'approche quantitative des phénomènes linguistiques dans ce mémoire est, précisions-le, parfaitement assumée et réfléchie. Nous avons avant tout voulu maîtriser au mieux le sens des calculs opérés de manière à les subordonner en toute connaissance de cause à nos objectifs de recherche linguistique. Mais nous sommes par ailleurs décidé à nous approprier d'autres calculs et d'autres méthodes dans un avenir proche. Par exemple, nous avons été particulièrement sensibilisé, du fait de nos collaborations avec S. Diwersy, à l'intérêt des travaux de S. F. Fries & M. Hilpert relatifs à la méthode de *classification hiérarchique ascendante par contiguïtés* (CHAC, de l'anglais *Variability based neighbour clustering* − voir S.-T. Gries, M. Hilpert 2008 ; 2012) qui vise à produire une périodisation automatique des phénomènes linguistiques étudiés. Il s'agit là, pensons-nous, d'un des moyens possibles d'affiner notre approche par trop rudimentaire des cooccurrents d'une préposition qui, pour le moment, prend pour « toile de fond » les siècles - c'est-à-dire un critère arithmétique arbitraire au regard des phénomènes linguistiques étudiés. C'est une telle piste que nous[269] comptons explorer lors des 9èmes Journées Internationales de la Linguistique de corpus organisées à Grenoble en juillet 2017 (https://jlc2017.univ-grenoble-alpes.fr).

Si l'on se tourne, enfin, vers les corpus, toute une série d'améliorations peuvent être apportées à ceux que nous avons explorés jusqu'ici. Sur le plan quantitatif : il nous paraît urgent d'étoffer notre actuel corpus Presto (point que nous avons abordé dans la partie II précédente) de manière à disposer d'occurrences plus nombreuses sur lesquelles appliquer nos méthodes d'analyse. On peut, pour illustration, évoquer le cas de *dedans* dans le discours littéraire, ou encore celui des expressions figées dont l'étude en diachronie nécessite qu'on dispose d'un nombre suffisant d'occurrences (et donc d'œuvres) pour chaque tranche temporelle que l'on compare. Sur un plan qualitatif, il nous apparaît en outre nécessaire d'introduire dans Presto de nouveaux *discours* qu'il conviendra d'étudier pour eux-mêmes. Tout porte à croire en effet que l'évolution de la sémantique et de la pragmatique des prépositions diffère selon qu'on a affaire au *discours* littéraire, juridique, scientifique, … ou encore selon les champs génériques (voir § 1.2.3.2.). Il convient donc de se tourner aujourd'hui, notamment en diachronie, moins vers la constitution de corpus « représentatifs » (dont nous avons dit *supra* le caractère problématique) que de corpus « spécialisés » réunissant un ensemble d'œuvres relevant d'un discours, voire d'un champ générique spécifiques. Nous serions à cet égard extrêmement désireux d'explorer à brève échéance des corpus historiques de discours scientifique, en particulier dans le domaine médical[270]. Signalons que d'ores et déjà, dans le cadre d'un programme de recherche élaboré avec C. Rossari à l'Université de Neuchâtel, nous allons travailler sur un corpus d'encyclopédies françaises couvrant une période d'environ trois siècles (1751 à 2001)[271].

---

[268] Même si nous comptons nous appuyer sur des grilles existantes : nous songeons en particulier à la grille sémantique mise au point dans le cadre de l'ANR (2006-2009) « SFA » (Spatial Framing Adverbials) à laquelle nous avions participé. (Coordonnateurs : Michel Charolles et Laure Sarda (UMR LaTTICe).

[269] Sascha Diwersy, Achille Falaise et moi-même.

[270] Ce qui nous semble faisable à brève échéance dans le cadre de notre collaboration avec l'ATILF. Cela permettrait de prolonger des travaux conduits avec ce laboratoire dans le cadre de Presto (en part. avec V. Montémont et G. Souvay) ayant abouti non seulement à la constitution d'une grande part de notre corpus, mais aussi, dans le cadre d'une convention, à la numérisation dans une édition libre de droits d'une partie du premier volume (750 pages) des *Œuvres complètes* (1585) d'Ambroise Paré.

[271] Nous allons y revenir dans la partie 4 consacrée aux « perspectives ».

Comme on le voit, cette troisième partie ne constitue à nos yeux qu'une première étape dans un projet plus vaste qui s'ouvre devant nous et qui devrait nous occuper pour plusieurs années. C'est par l'évocation de quelques-unes des perspectives (provisoires) qui s'associent à ce projet que nous voudrions achever ce mémoire.

PERSPECTIVES DE RECHERCHE

Dans cette dernière (et courte) partie, j'évoquerai quatre axes structurants pour mon activité de recherche dans les années à venir.

Approche cognitive et fonctionnelle de la sémantique des prépositions *en* et *dans*

Je souhaite m'approprier dans un délai rapide le cadre théorique et les outils méthodologiques mis au point par C. Vandeloise afin de réinterroger la définition de l'identité sémantique de *en* en synchronie du français contemporain, et d'enrichir ainsi le travail que j'ai mené par d'autres voies dans ma première partie. Mes échanges et discussions avec M. Aurnague au cours de la rédaction de ce mémoire m'ont convaincu de l'intérêt d'une telle étude, encore à ma connaissance jamais conduite.

L'intérêt d'une telle étude cognitive et fonctionnelle de la préposition *dans* peut paraître en revanche moins immédiatement évident dans la mesure où ce morphème a constitué un objet d'analyse récurrent dans les travaux de C. Vandeloise (voir la bibliographie complète de l'auteur présentée dans M. Aurnague, 2010b). En réalité, la perspective que j'ai en tête est cette fois d'ordre diachronique. Il s'agirait, en partant justement des travaux de C. Vandeloise (mais aussi de ceux d' A.-M. Berthonneau, de D. Leeman, de L. Sarda, …), d'observer en corpus l'apparition, la répartition et l'évolution entre le XVI^{e} s. et le XX^{e} s. des principaux traits de la relation contenance/contenu qui décrit, selon C. Vandeloise, la préposition *dans*. Cette étude, telle que je l'envisage actuellement, prendrait essentiellement appui sur l'examen de la distribution nominale « préférée » du pivot prépositionnel *dans* dans le corpus Presto, distribution dont le § 3.1. de notre troisième partie a pu donner une première idée. Le détail des méthodes d'analyse adoptées reste encore à définir, en lien avec Sascha Diwersy notamment.

Une telle approche, développée et expérimentée pour l'étude de la sémantique de *dans* en diachronie, et sous réserve qu'elle s'avère féconde, pourrait ensuite être appliquée à *en* dans le même but d'analyser l'apparition, la répartition et l'évolution des principaux traits que l'étude évoquée *supra* (pénultième paragraphe) aura permis de mettre au jour. L'ensemble que constitue mes études sur *en* – auquel je joins par avance ces deux études à venir – devrait me conduire, dans un avenir raisonnablement proche, à la publication d'un ouvrage dont une grande partie de ce mémoire fournirait l'essentiel de la trame et du contenu.

Vers une maîtrise plus vaste et plus approfondie des méthodes d'analyse statistique des données textuelles

M'aguerrir à l'analyse statistique importe à mes yeux pour assurer d'une part mon autonomie dans l'utilisation des fonctionnalités de calcul mises à disposition par les plateformes d'exploration automatique des corpus, d'autre part ma capacité à pratiquer un langage commun pointu et efficace avec les informaticiens et les statisticiens impliqués dans les projets de linguistique quantitative auxquels je souhaite encore participer dans l'avenir. J'ai donc le projet de continuer à me former dans le domaine de la statistique, en particulier en statistique descriptive multidimensionnelle. Parallèlement, je compte très vite mettre sur pied avec d'autres linguistes impliqués dans des objets de recherche communs (notamment C. Rossari, de l'Université de Neuchâtel et S. Diwersy de l'Université de Montpellier) un groupe de travail visant à partager et à interroger la pratique que nous avons de la statistique appliquée à la recherche linguistique. Ce groupe réunirait des enseignants-chercheurs, des doctorants et des post-doctorants issus des universités de Montpellier, de Grenoble, de Lyon et de Neuchâtel.

Presto : et maintenant ?

Le projet a administrativement pris fin le 1er avril 2017. Le rapport final, visé par l'ANR, devrait être bientôt disponible en ligne. Pour autant, ce programme demeure d'actualité. Grâce au recrutement de notre ingénieur de recherche (A. Falaise) sur un nouveau contrat financé par le LaBEx ASLAN (http://aslan.universite-lyon.fr) jusqu'en décembre 2018, le suivi informatique du programme est assuré jusqu'à cette date.

Dès l'écriture de ce mémoire achevée, je compte m'atteler à quatre objectifs majeurs :

i) La mise en ligne sous licence libre avec téléchargement possible :
   - ✓ du corpus « noyau » accessible en versions (a) intégrale *versus* échantillonnée, (b) nue *versus* annotée ;
   - ✓ des logiciels d'annotation élaborés au cours de Presto,
   - ✓ de son lexique.

ii) Une nouvelle version du corpus Presto (= contrôlé et échantillonné) sera produite, qui devra satisfaire 100% des critères énoncés dans le § 1.5. de notre deuxième partie.

iii) L'amélioration qualité actuelle de l'annotation (étiquetage MS et lemmatisation) du corpus. Le lexique sera corrigé en parallèle. Les ambiguïtés engendrées par l'application des règles d'archaïsation sur des périodes non pertinentes du corpus (XVIIIe s. au XXe s.) seront levées.

iv) L'annotation syntaxique du corpus devra être énergiquement mise en œuvre. L'objectif d'une publication en 2019 dans une revue anglo-saxonne, exposant de nouveaux résultats obtenus à partir d'un corpus étiqueté en morphosyntaxe *et* en syntaxe[272], devra être atteint.

Au-delà de ces objectifs à court terme, je compte continuer à utiliser les précieuses ressources de ce corpus pour travailler sur le changement linguistique en lien notamment avec la théorie de la grammaticalisation.

Vers un projet européen consacré au discours encyclopédique

En 2013, une collaboration entre le programme Presto et l'équipe de linguistique française de l'Université de Neuchâtel, dirigée par C. Rossari, a été mise en place, donnant lieu à des interventions dans la formation doctorale en Sciences du langage de la « Conférence universitaire de Suisse occidentale » (CUSO https://langage.cuso.ch) et à plusieurs journées d'études. La dernière[273], organisée en novembre 2016 à Neuchâtel, réunissait – outre des membres de Presto – des chercheurs issus de diverses disciplines (littérature, philosophie, linguistique) pour communiquer sur un objet commun : le discours encyclopédique. Cette journée, initiée et organisée par C. Rossari et dont plusieurs contributions devraient être réunies dans un numéro à venir de la revue *Langue Française*, était aussi destinée à alimenter la réflexion conduite conjointement par l'équipe de linguistique française de Neuchâtel et par Presto autour du positionnement énonciatif dans le discours encyclopédique. Ce thème a donné lieu à un projet de recherche intitulé « Le positionnement énonciatif et ses variations dans le discours encyclopédique entre les XVIIIe et XXIe siècles », déposé auprès du Fonds National Suisse par C. Rossari en mars 2017 et

[272] Objectif annoncé dans notre demande de prolongement administratif d'une année du programme Presto auprès de l'ANR.

[273] https://www.unine.ch/files/live/sites/islc/files/shared/linguistique%20francaise/Phenomenes_Enonciatifs_detail.pdf

dans lequel plusieurs membres de Presto sont impliqués comme collaborateurs scientifiques. Ce projet a pour but « d'analyser les formes linguistiques qui donnent des indications sur la façon dont un locuteur se positionne face à un certain contenu, dits « indices du positionnement énonciatif », dans le discours encyclopédique entre les XVIII[e] et XXI[e] siècles[274]. » Il cherchera à « cerner le type de rapport qu'un locuteur donné manifeste avec le savoir qu'il transmet et ce, indépendamment du positionnement idéologique propre aux auteurs. » Le corpus mobilisé dans ce projet coïncide en grande partie avec le corpus spécialisé Presto[-ENCYCLOPEDIE] puisqu'il réunit (partiellement ou intégralement) les œuvres suivantes :

- L'*Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers* dir. Diderot et D'Alembert (1751-1772) ;
- Le *Grand Dictionnaire Universel du XIX[e] siècle* – plus communément appelé *Grand Larousse du XIX[e] siècle* (1866-1876) ;
- L'*Encyclopædia Universalis* (1968-2012 pour l'édition papier) ;
- *Wikipédia*, encyclopédie en ligne collaborative (2001-) ;

Les marques modales et énonciatives étudiées seront pour l'essentiel : les verbes et adverbes modaux, les introducteurs de point de vue, les connecteurs, les indications évidentielles, les pronoms personnels renvoyant à l'instance de prise en charge et les marques dialogiques d'interlocution (formes pronominales ou nominales qui renvoient à l'instance du lecteur). Ces marques seront extraites et analysées au moyen de la plateforme BTLC/Primestat dans le but de construire (pour ensuite les comparer) leur « profil combinatoire » - non seulement pour chaque œuvre mais aussi dans le cadre de partitions constituées sur des critères autres : critères des *discours*[275] (F. Rastier, 2011), critères thématiques etc. L'acquis des méthodes construites dans Presto sera donc mis à contribution dans ce projet. Je compte m'investir pleinement dans ce projet qui, s'il est financé, s'étendra sur la période 2018-2021.

Par ailleurs et enfin, en m'appuyant sur le réseau international constitué grâce à Presto et grâce à notre collaboration avec l'Université de Neuchâtel, je souhaiterais concourir à l'élaboration d'un nouveau projet de recherche transdisciplinaire (linguistique, littérature, philosophie) de dimension européenne, consacré à l'émergence et à la maturation du projet encyclopédique en Europe, depuis le *Dictionnaire historique et critique* de P. Bayle en 1696 jusqu'à la Wikipédia d'aujourd'hui en passant (notamment) par la *Cyclopædia* d'E. Chambers (1728) en Angleterre, L'*Encyclopédie* d'*Yverdon* publiée entre 1770 et 1780. Ce projet, actuellement à l'état pré-embryonnaire oserais-je dire, nécessitera une phase préparatoire lourde. Une demande de montage d'un réseau scientifique international (MRSEI) devrait être déposé auprès de l'ANR en 2018 pour permettre à cette phase de se dérouler au mieux.

[274] Les citations sont extraites du résumé initial figurant dans le projet déposé par C. Rossari au FNS.
[275] Le « discours encyclopédique » hébergeant lui-même une pluralité d'autres discours qu'il surplombe : discours religieux, politique, historique, etc.

BIBLIOGRAPHIE

Adda, G., Mariani, J., Paroubek, P., Rajman, M. & Lecomte, J. (1999), « L'action GRACE d'évaluation de l'assignation des Parties du Discours pour le Français », *Langue,* 2 (2), 119-129.

Amiot D., de Mulder W. (2011), « L'insoutenable légèreté de la préposition *en* », *Studii de linguistica* 1, 9-27. [http://studiidelingvistica.uoradea.ro/arhiva-fr-1-2011.html]

Anscombre, J.-C. (2001), « L'analyse de la construction 'En tout N' par D. Leeman: quelques remarques », *Travaux de linguistique*, 42-43, 183-197.

Auer A., Peersman C., Pickl S., Rutten G. & Vosters R. (2015), « Historical sociolinguistics: the field and its future », *Journal of Historical Sociolinguistics*, 1(1), 1-12.

Aurnague M. (1996), « Les Noms de Localisation Interne : tentative de caractérisation sémantique à partir de données du basque et du français », *Cahiers de Lexicologie*, 69, 2, 159-192.

Aurnague, M. (2004), *Les structures de l'espace linguistisque : regards croisés sur quelques constructions spatiales du basque et du français,* Leuven/Paris : Peeters.

Aurnague M. (2009), « A cet endroit vs. dans un tel endroit : ce que à nous dit d'endroit et vice-versa », *Langages*, 173, 34-53.

Aurnague, M. (2010a), « Places-repère, localisation et routines : lorsque l'analyse du nom *place* rejoint celle de la préposition *à* », *Corela*, n° spécial (Espace, préposition, cognition, Hommage à Claude Vandeloise, Col G. & C. Collin (éds)), https://corela.revues.org/919

Aurnague, M. (2010b), « Claude Vandeloise : bibliographie des travaux / bibliography of his works », *Corela*, n° spécial (Espace, préposition, cognition, Hommage à Claude Vandeloise, Col G. & C. Collin (éds)), http://corela.revues.org/1755

Aurnague, M. (2012a), « Quand la routine s'installe : remarques sur les emplois de « à » de type « routine sociale » », *Revue Romane*, 47 (2), 189-218.

Aurnague M. (2012b), « De l'espace à l'aspect : les bases ontologiques des procès de déplacement », *Corela* [En ligne], HS-12 | 2012, mis en ligne le 04 avril 2013, consulté le 30 septembre 2016. URL : http://corela.revues.org/2846 ; DOI : 10.4000/corela.2846

Aurnague, M. & Vieu, L. (2013), « Retour aux arguments : pour un traitement « relationnel » des prépositions spatiales », *Faits de langues*, 42, 17-38.

Azzopardi, S. (2010), « Présentation : La linguistique « de » corpus au-delà des champs disciplinaires : questions et enjeux transversaux », *Cahiers de praxématique,* 54-55, 11-24.

Bat-Zeev Shyldkrot, H. (2008), « Complétives introduites par *Prep que P* vs Complétives introduites par *Prép ce que P* », *Langue française*, 157, 1, 106-122.

Berthonneau, A.-M. (1999), « A propos de dedans et de ses relations avec dans », *Revue de Sémantique et de Pragmatique*, 6, 13-41.

Biber, D. (1989), « A typology of English texts », *Linguistics*, (27), 3-43.

Biber, D. (1993a), « Representativeness in Corpus Design », *Literary and Linguistic Computing,* 8 (4), 243-257.

Biber, D. (1993b), « Using Register-Diversified Corpora for General Language Studies », *Journal Computational Linguistics*, Special issue on *using large corpora*, Vol. 19 (2), 219-241.

Biber, D. , Joahnsson, S. , Leech, G. , Conrad, S. & Finegan, E. (1999), *Longman Grammar of Spoken and Written English*, London : Longman.

Biber**,** D., & Conrad**,** S**. (**2009**),** *Register**, genre, and style*, Cambridge: Cambridge University

Press.
Blumenthal, P. (2006), « De la logique des mots à l'analyse de la synonymie », *Langue française*, 150, 14-31.
Blumenthal P. (2007), « Sciences de l'Homme vs sciences exactes : combinatoire des mots dans la vulgarisation scientifique », *Revue de linguistique appliquée*, XII, 2, 15-28.
Blumenthal, P. (2008), « Combinatoire des prépositions. Approche quantitative », *Langue Française,* 157, 37-51.
Blumenthal P. (2011a), « Le figement. Du XVII^{e} s. au français contemporain », *in* J.-C. Anscombre & S. Mejri (éds), *La parole entravée*, Paris : H. Champion, 283-302.
Blumenthal P. (2011b), « Odeur – évolution des profils combinatoires », *Langages*, 181, 53-71.
Blumenthal P. (2013), « La préposition *en* dans la francophonie africaine », *Langue française*, 178, 117-131.
Blumenthal, P., Novakova I. & Siepmann D. (éds.) (2014), *Les émotions dans le discours*, Paris/Berne : P. Lang.
Blumenthal, P. & Vigier, D. (2017), « Présentation », *Langages*, 206, 5-20.
Blumenthal, P. & Vigier, D. (à par. 2017), *Etudes diachroniques du français et perspectives sociétales*, Paris/Berne : P. Lang.
Bolly, C. (2010), « Flou phraséologique, quasi-grammaticalisation et pseudo marqueurs de discours : un no man's land entre syntaxe et discours ? », *Linx*, 62-63, 11-38. URL : http://linx.revues.org/1356 ; DOI : 10.4000/linx.1356
Borillo, A. (1991), « De la nature compositionnelle de l'aspect », in *Les typologies de procès*, C. Fuchs (éd.), Paris : Klincksieck, 97-102.
Borillo, A. (1996), « La relation partie-tout et la structure [NI à N2] en français », *Faits de langues*, 7, 111-120.
Borillo, A. (1998), *L'espace et son expression en français*, Paris : Ophrys.
E. Bruni, N.-K. Tran, & M. Baroni, (2014), « Multimodal distributional semantics », *Journal of Artificial Intelligence Research*, 49, 1-47, 2014.
Brunot F. (1967), *Histoire de la langue française. Des origines jusqu'en 1900*. Tome II: *le seizième siècle*, Paris : A. Colin.
Buridant, C. (2000), Grammaire *nouvelle* de l'ancien français, Paris : SEDES.
Cadiot, P. (1991), *De la grammaire à la cognition : la préposition* pour, Paris : Éditions du CNRS.
Cadiot, P. (1997a), *Les prépositions abstraites en français*, Paris : A. Colin.
Cadiot, P. (1997b), « Les paramètres de la notion de préposition incolore », *Faits de langues*, 9, 127-134
Cadiot, P. (2002), « Schémas et motifs en sémantique prépositionnelle : vers une description renouvelée des prépositions dites « spatiales » », *Travaux de Linguistique,* 44, 1, 9-24.
Cadiot, P. & Visetti, Y.-M. (2001), *Pour une théorie des formes sémantiques – motifs, profils, thèmes*, Paris : P.U.F.
Cervoni, J. (1991), *La préposition. Etude sémantique et pragmatique*, Paris : De Boeck-Duculot.
Charolles, M. (1997), « L'encadrement du discours : univers, champs, domaines et espaces », *Cahier de recherche linguistique*, 6, 1-73.
Charolles M. & Vigier D. (2005), Les adverbiaux en position préverbale : portée cadrative et organisation des discours. *Langue Française*, 148, 9-30. (http://halshs.archives-ouvertes.fr/halshs-00373342/en/)
Charolles M., Diwersy S. & Vigier D. (2017), « Evolution des emplois des marqueurs de topiques de discours dans *Le Figaro* de la fin du XIX^{ème} et du début du XXI^{ème} siècle »,

Langages, 206, 85-104.
Claridge, C. (2008), « Historical corpora », in *Corpus linguistics. An international handbook*, A. Lüdeling & al. (eds.), Berlin/New-York : Mouton de Gruyter, 242-259.
Combettes, B. (2012), « De quelques problèmes spécifiques à l'élaboration d'une grammaire historique », *Langue française*, 176, 69-83.
Combettes, B. & Marchello-Nizia, C. (2010), « La périodisation en linguistique historique : le cas du français préclassique », in B. Combettes et al. (éds), *Le changement en français*, Paris/Berne : P. Lang, 129-141.
Condamines, A. (2000), « Les bases théoriques du groupe toulousain « Sémantique et Corpus » : ancrages et perspectives », *Cahiers de Grammaire*, 25, 5-28.
Condamines, A., Rebeyrolle J. & Soubeille, A. (2004), « Variation de la terminologie dans le temps : une méthode linguistique pour mesurer l'évolution de la connaissance en corpus », *Actes d'Euralex International Congress*, Lorient, 6-10 juillet 2004, 547-557.
Corblin, F. (2011), « Des définis para-intensionnels : *être à l'hôpital, aller à l'école* », *Langue Française*, 171, 55-75.
Cori, M. (2008), « Des méthodes de traitement automatique aux linguistiques fondées sur les corpus », *Langages*, 171, 95-110.
Cori, M. & David, S. (2008), « Les corpus fondent-ils une nouvelle linguistique ? », *Langages* 171, 111-129.
Cori, M., David, S. & Léon, J. (2008), « Présentation : éléments de réflexion sur la place des corpus en linguistique », *Langages*, 171, 5-11.
Creissels, D. (1999), « Parfait et statif en tswana », *Cahiers Chronos* 4, 185-202.
Creissels, D. (2006), *Syntaxe générale*, *une introduction typologique*, t 1, 2. Coll. *Langues et syntaxe*.
Creissels, D. (2014), « Approche typologique de la notion de sujet », Colloque international *Du Sujet et de son absence dans les langues*, Université du Maine, 27-28 mars 2014. http://www.deniscreissels.fr/public/Creissels-appr.typ.suj.pdf
Croft, W. & Cruse, A. (2004), *Cognitive Linguistics*, Cambridge: Cambridge University Press.
Culioli, A. (1990), *Pour une linguistique de l'énonciation*, Paris : Ophrys.
Dagnac, A. (2009), « Elle a teint ses rideaux en rouge : entre manière et résultativité », *Langages,* 175*,* 67-83.
Dardel, R. de, Banniard, M. & Combettes, B. (éds.)(2011), *Périodisation(s)*, *Diachroniques*, 1, Paris : Presses Universitaires de la Sorbonne.
Darmesteter A. (1885), *Notes sur l'histoire des prépositions françaises en, enz, dedans, dans*, Paris: Le Cerf.
De Mulder, W. (2001), « La linguistique diachronique, les études sur la grammaticalisation et la sémantique du prototype : présentation », *Langue française*, 130, 8-32.
De Mulder, W. (2008), « *En* et *dans:* une question de « déplacement » ? ». Dans: O. Bonami, S. Prévost, M. Charolles, J. François et C. Schnedecker (éds.), *Discours, diachronie, stylistique du français: études en hommage à Bernard Combettes*. Bern: P. Lang, 277-291.
De Mulder, W. & Stosic, D. (2009) (éds), *Approches récentes de la préposition*, *Langages,* 173.
De Mulder, W. & Stosic, D. (2009), « Présentation », *Langages,* 173, 3-13.
De Mulder, W. & Amiot, D. (2013), « En : de la préposition à la construction », *Langue française*, 178, 21-39.
Desclés, J.-P. (1991), « Archétypes cognitifs et types de procès », in *Les typologies de procès*, C. Fuchs (éd.), Paris : Klincksieck, 171-196.
Desrosières, A. (1988), « La partie pour le tout : comment généraliser ? La préhistoire de la

contrainte de représentativité », *Journal de la société statistique de Paris*, 129 (1-2), 96-115.

Di Meola, C. (2000), *Die Grammatikalisierung deutscher Präpositionen*, Tübingen : Stauffenburg.

Diwersy, S., Falaise, A. & Vigier, D. (2017), « Étude de l'évolution sémantique des prépositions *à, en, dans, dedans* du français. Quel(s) apport(s) d'une périodisation automatique ? » in *9èmes Journées Internationales de la Linguistique de corpus*, juillet 2017, Grenoble.

Diwersy, S., Falaise, A., Lay, M.-H. & Souvay, G. (2015), Traitements pour l'analyse du français préclassique, in Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles, Caen : ATALA, 565-571.

Diwersy, S., Falaise, A., Lay, M.-H. & Souvay, G. (2017), « Ressources et méthodes pour l'analyse diachronique », *Langages*, 206, 21-44.

Dodge, Y. (2007), *Statistique. Dictionnaire encyclopédique*, Paris : Springer.

Dostie, G. (2004), *Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique*, Bruxelles, Duculot / De Boeck.

Ducos, J. & Soutet, O. (2012), *L'ancien et le moyen français*, Paris : PUF.

Dunning T. (1993), « Accurate methods for the statistics of surprise and coincidence », *Computational Linguistics*, 19-1, 61-74.

Eskénazi, A. (1987), « Député de Saône-et-Loire – Préfet du Rhône – En Vendée », *LINX* 16-I : 28-69.

Fagard, B (2006), « La grammaticalisation en question : du latin aux langues romanes modernes», *Modèles linguistiques*, XXVII/1, 53, 91-110.

Fagard B. & Sarda L. (2009), « Etude diachronique de la préposition *dans* », in J. François, E. Gilbert, C. Guimier & M. Krause (éds..), *Autour de la préposition: position, valeurs, statut et catégories apparentées à travers les langues, Bibliothèque de Syntaxe et Sémantique,* Caen : PUC, 225-236.

Fagard B. & Combettes B. (2013), « De *en* à *dans*, un simple remplacement ? Une étude diachronique », *Langue Française* 178, 95-119.

Fabre, C. (2015), « Sémantique distributionnelle automatique : la proximité distributionnelle comme mode d'accès au sens », *Études de linguistique appliquée*, 4, 180, 395-405.

Fabre, C. & Lenci, A. (éds.) (2015), *Numéro spécial sur la sémantique distributionnelle*, 56-2.

Falaise, A. & Leeman, D. (2017), « Prépositions et noms de régions anciennes : évolution des emplois et représentations socio-culturelles », in *Etudes diachroniques du français et perspectives sociétales*, Blumenthal, P. & Vigier, D. (éds.), Paris/Berne : P. Lang.

Firth J. R. (1957), « A synopsis of linguistic theory 1930-1955 », *Studies in linguistic analysis*, Oxford: Blackwell, 1-32.

Flaux, N. (2002), « Les noms d'idéalités concrètes et le temps », in *Temps et aspect : de la grammaire au lexique*, Lagae V., Carlier A. & Benninger C. (éds.), *Cahiers Chronos*, 10, 65-78.

Fong, V. (2003), « Resultatives and depictives in Finnish », in Nelson, D. & S. Manninen (eds.), *Generative approaches to Finnic and Saami linguistics*, CSLI, Stanford, 201-233.

Fort, K. (2012), *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus*. Traitement du texte et du document. Université Paris- Nord - Paris XIII < tel-00797760v2>

Fournier N. & Vigier D. (à par. 2017), « Contribution à l'étude de la quantification de la durée entre le XVI^{e} s. et le XX^{e} s. », *in* P. Blumenthal & D.Vigier (éds) (2017), *Études diachroniques du français et perspectives sociétales*, Actes du colloque « Changements

linguistiques et phénomènes sociétaux » (CLPS), 7-9 mars 2016, ENS Lyon. Paris/Berne : P. Lang.

Franckel J.-J. & Lebaud D. (1991), « Diversité des valeurs et invariance du fonctionnement de en, préposition et préverbe », *Langue française*, 91, 56-79.

Franckel, J.-J. & Paillard, D. (2007), *Grammaire des prépositions,* t. 1, Paris : Ophrys.

Fuchs, C. (1997), « L'interprétation des polysèmes grammaticaux en contexte », in Kleiber G. & *al*., *Les formes du sens*, Bruxelles : De Boeck Supérieur « Champs linguistiques », 127-133.

Fuchs, Catherine. 1999. Les tours qualifiants en *comme N* : *Jean travaille comme maçon*. in *Les opérations de détermination : quantification / qualification*, A. Deschamps & J. Guillemin-Flescher (éds.), 63-82, Paris : Ophrys.

Furukawa, M. (2010), « L'article défini et le problème dit de l'unicité : quantité ou qualité? », *Bulletin d'Études de Linguistique Française*, 44, 65-82.

Gardelle, L. & Vigier, D. (2014), (2014a), *La Prédication*, *Verbum*, XXXVI, 2, Presses Universitaires de Nancy.

Geeraerts D. **(***2010)***,** *Theories of Lexical Semantics*, Oxford: Oxford University Press.

Geyken, A. (2008), « Quelques problèmes observés dans l'élaboration de dictionnaires à partir de corpus », *Langages*, 171, 111-129.

Godefroy F. (1891-1902), *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IXe au XVe siècle*, Paris.

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

Goldberg, A. (2006), *Constructions at Work. The Nature of Generalization in Language*, Oxford: Oxford University Press.

Gonon, L., Kraif, O., Novakova, I., Piat, J. & Sorba, J.. (2016), « Sur la scène de crime… Enquête sur les enjeux linguistiques et stylistiques de motifs récurrents dans le *thriller* contemporain », *in Actes du 5e Congrès mondial de linguistique française – CMLF 2016*. URL : http://www.shs-conferences.org/articles/shsconf/pdf/2016/05/shsconf_cmlf2016_06006.pdf

Gougenheim G. (1945, [1970]), « Les prépositions « en » et « dans » dans les premières œuvres de Ronsard ». *in Etudes de grammaire et de vocabulaire français, réunies sur l'initiative de ses collègues et amis pour son soixante-dixième anniversaire*, Paris : Picard, 66-76.

Gougenheim, G. (1950, [1970]), « Valeur fonctionnelle et valeur intrinsèque de la préposition « en » en français moderne », in *Etudes de grammaire et de vocabulaire français*, Paris : Picard, 55-65.

Gougenheim G. (1954, [1970]), « Tant de royaumes réunis 'dans' une vaste monarchie », *in Etudes de grammaire et de vocabulaire français, réunies sur l'initiative de ses collègues et amis pour son soixante-dixième anniversaire*, Paris : Picard, 54.

Gougenheim G. (1951, [1974]), *Grammaire de la langue française du seizième siècle*, Paris : Picard.

Gougenheim G., Michéa, R., Sauvageot, A. & Rivenc, P. (1956, [1964]), *L'élaboration du français fondame*ntal, Paris : Didier.

Gries, S. Th. (2010), « Behavioral Profiles: a fine-grained and quantitative approach in corpus-based lexical semantics ». *The Mental Lexicon*, 5(3), 323-346.

Gries, S. T. & Hilpert, M. (2008), « The identification of stages in diachronic data: Variability-based neighbour clustering », *Corpora* 3 (1), 59-81.

Gries, S. T. & Hilpert, M. (2012), « Variability-based neighbor clustering: A bottom-up approach to periodization in historical linguistics », in T. Nevalainen & E. Traugott

(eds), *The Oxford Handbook of the History of English*, Oxford: Oxford University Press, 134-144.
Gross, M. (1986), *Grammaire transformationnelle du français. 3. Syntaxe de l'adverbe*, Paris : Asstril.
Grossmann, F. & Tutin, A. (éds.), (2003), *Les collocations : analyse et traitement, Travaux et recherches en linguistique appliquée*, Amsterdam : de Werelt.
Guillaume, G. (1919, [1975]), *Le problème de l'article et sa solution dans la langue française*, Paris : Nizet.
Guillot, C., Heiden, S., Lavrentiev, A. & Marchello-Nizia, C. (2008), « Constitution et exploitation des corpus d'ancien et de moyen français », *Corpus* [Online], 7 | 2008, Online since 13 November 2009, connection on 15 August 2017. URL : http://corpus.revues.org/1495.
Guimier, C. (1978), « *En* et *dans* en français moderne », *Revue des langues romanes*, 83 (2) 277-306.
Guimier, C. (1996), *Les adverbes du français : le cas des adverbes en –ment*, Paris : Ophrys.
Guimier, C. (2007), « La préposition et la tradition grammaticale anglaise », *Langages*, 167, 85-99.
Guiraud, P. (1954), *Les caractères statistiques du vocabulaire*, Paris : Presses Universitaires de France.
Guiraud, P. (1960), *Problèmes et méthodes de la statistique linguistique*, Paris : Presses Universitaires de France.
Habert, B. (2000), « Des corpus représentatifs : de quoi, pour quoi, comment ? », *in Linguistique sur corpus. Etudes et réflexions*, M. Bilger (éd..), Perpignan : Presses Universitaires de Perpignan, Collection Cahiers de l'université de Perpignan, 31, 11–58.
Habert, B. (2005), *Instruments et ressources électroniques pour le français*, Paris : Ophrys.
Habert, B., Nazarenko, A. & Salem A. (1997), *Les linguistiques de corpus*, Paris : A. Colin.
Habert, B. & Zweigenbaum, P. (2003), « Classer les mots : sémantique à gros grain et méthodologie harrissienne », *Revue de Sémantique et Pragmatique*, 2, 25–45.
Hagège, C. (1982, [2013]), la *Structure* des langues, Paris : P.U.F, coll. Que sais-je ?
Hagège, C. (1997), « Les relateurs comme catégorie accessoire et la grammaire comme composante nécessaire », *Faits de langues*, 9, 19-28
Halliday, M.A.K. & Hasan, R. (1976), *Cohesion in english*, London : Longman.
Haussmann, J. & Blumenthal, P. (2003), *Collocation, corpus, dictionnaire*, *Langue Française*, 150.
Haussmann, J. & Blumenthal, P. (2003), « Présentation : collocation, corpus, dictionnaire », *Langue Française*, 150, 3-13.
Heylen, K.& Bertels, A. (2016), « Sémantique distributionnelle en linguistique de corpus », *Langages*, 201, 51-64
Hoey, M. (2005), *Lexical Priming*, London : Routledge.
Huguet E. (1925-1967), *Le dictionnaire de la langue française du seizième siècle*, 7 volumes, Paris : Champion / Didier.
Hunston, S. (2008), « Corpus compilation and corpus types. Collection strategies and design decisions », in *Corpus linguistics. An international handbook*, A. Lüdeling & al. (eds.), Berlin/New-York : Mouton de Gruyter, 154-168.
Huyghe, R. (2009), *Les noms généraux d'espace en français. Enquête linguistique sur la notion de lieu*, Bruxelles : De Boeck Duculot.
Huyghe, R. (2015), « Les typologies nominales : présentation », *Langue française*, 185,1, 5-27.
Jacques, M.-P. (2005), « Pourquoi une linguistique de corpus ? », in G. Williams (éd.), *La linguistique de corpus*, Rennes : Presses universitaires de Rennes, 21-30.

Kayne, R. (1977), *Syntaxe du français. Le cycle transformationnel*, Paris : Seuil.
Khammari, I. (2008), « Les compléments de verbes régis par *en* », *Langue Française,* 157 : 52-73.
Kleiber, G. (1981), *Problèmes de référence. Descriptions définies et noms propres*, Paris : Klincksieck.
Kleiber, G. (1990), *La sémantique du prototype. Catégories et sens lexical*, PUF : Paris
Kleiber, G. (1994), *Anaphores et pronoms*, Bruxelles: Duculot.
Kleiber, G. (1997), « Sens, référence et existence : que faire de l'extra-linguistique ? », *Langages*, 127, 9-37.
Kleiber, G. (2007), « En passant par le gérondif, avec mes (gros) sabots », *Cahiers Chronos*, 19, p 93-125.
Kleiber, G. (2008), « Petit essai pour montrer que la polysémie n'est pas un sens interdit », in J. Durand, B. Habert & B. Laks (éds.), *Congrès Mondial de Linguistique Française* – CMLF 2008, Paris : Institut de Linguistique Française, 87-101.
Kleiber, G. & Riegel, M. (1978), « *Les « grammaires floues* » », in R. Martin (éd.), *La notion de recevabilité en linguistique*, Paris : Klincksieck, 67-123.
Kleiber, G. & Vuillaume, M. (2011), « Sémantique des odeurs », *Langages*, 181, 17-36.
Koch, P. & W. Oesterreicher (2001), « Gesprochene Sprache und geschriebene Sprache / Langage parlé et langage écrit », in Holtus G., Metzeltin M. & Schmitt Ch. (éds), *Lexikon der Romanistischen Linguistik*, Bd. I/2, Tübingen, Niemeyer, p. 584-627.
Kupferman, L. (1991), « Structure événementielle de l'alternance un / Ø devant les noms humains attributs », *Langages* 102, 52-75.
Labbé C. & Labbé D. (1994), « Que mesure la spécificité du vocabulaire ? », Grenoble, CERAT. Repris dans : *Lexicometrica* 3, 2001.
Labov, W. (1994), *Principles of linguistic change*. *Internal factors*. Oxford: Blackwell.
Lafon P. (1980), « Sur la variabilité de la fréquence des formes dans un corpus », *Mots*, 1, 127-165
Lafon P. (1984), *Dépouillements et statistiques en lexicométrie*, Genève/Paris : Slatkine Champion.
Lafon, P. (2006), « Statistique et lexicométrie : position des problèmes », *Projet ATHIS : Atelier n°2, L'historien, le texte et l'ordinateur*, École normale de Lyon, 27-28 novembre 2006. En ligne : http://www.menestrel.fr/IMG/pdf/LAFON.pdf
Lagarde J.-P. (1988), « Les parties du discours dans la linguistique moderne et contemporaine », *Langages*, 92, 93-108.
Lamiroy B. & Charolles M. (2004), « Des adverbes aux connecteurs: *simplement, seulement, malheureusement, heureusement* », *Travaux de linguistique*, 49, 57-79.
Lardon, S. & Thomine, M.-C. (2009), *Grammaire du français de la Renaissance*, Paris : Garnier.
Lathuillère, R. (1981), « Les problèmes de l'édition de Rabelais », *Cahiers de l'Association internationale des études francaises*, 33, 129-145.
Lavieu, B. (2006), « De la difficulté à distinguer entre groupes prépositionnels régis et non régis », *Modèles linguistiques*, XXVII/1, 53, 130-143.
Lay, M.-H. & Pincemin, B. (2010), Pour une exploration humaniste des textes. In *Statistical Analysis of Textual Data -Proceedings of 10th International Conference JADT 2010***.** (http://lexicometrica.univ-paris3.fr/jadt/jadt2010/allegati/JADT-2010-1045-1056_106-Lay.pdf)
Lebart, L. & Salem, A. (1994), *Statistique textuelle*, Paris : Dunod.
Leeman, D. (1994), *Les circonstants en question(s)*, Paris : Kimé.
Leeman, D. (1995), « Pourquoi peut-on dire *Max est en colère* mais non **Max est en peur* ?

Hypothèses sur la construction *être en* », *Langue française*, 105, 55- 69.
Leeman, D. (1996), *Vêture, coiffure, chaussures et autres ... coquetteries* », in *Sémiotique, phénoménologie, discours. Du corps présent au sujet énonçant. Hommages à Jean-Claude Coquet*, Paris : L'Harmattan, 77-89.
Leeman, D. (1996), « Le « sens » et l' « information » chez Harris », *Linx*, 8, 209-220.
Leeman, D. (1997), « Sur la préposition *en* », *Faits de langues*, 9, 135-145.
Leeman, D. (2006, 2007) (éd.), *la préposition en français,* I, II, III, *Modèles linguistiques* 53, 54, 55, tomes XXVII-I & II et XXVIII-I.
Leeman, D. (2006), « La préposition française : caractérisation syntaxique de la catégorie », *Modèles linguistiques*, XXVII/1, 53, 7-18.
Leeman, D. (2008) (éd.), *Enigmatiques prépositions, Langue Française*, 157.
Leeman, D. (2008), « Prépositions du français, état des lieux », *Langue Française*, 157, 5-19.
Leeman, D. (2015), « La préposition *en* et les noms de pays », in *Phraséologie et profils combinatoires. Lexique, syntaxe et sémantique. Hommages à P. Blumenthal*. Paris :H. Champion.
Leeman, D. (2016), « Hypothèse de résolution du problème posé par l'emploi des prépositions devant les noms de pays », *Lingvisticæ Investigationes Supplementa*, 32, 107–124
Leeman, D. & Falaise, A. (2017), « « Les prépositions devant les noms de région et de département français », *Langages,* 206.
Leeman, D. & Vaguer, C. (2014), « La préposition peut-elle être prédicative ? Le cas de la préposition *en* », *Verbum, XXXVI, 2,* Nancy : PUN.
Leech, G. (1991), « The state of art in corpus linguistics », in K. Aijmer & B. Altenberg (eds), *English Corpus Linguistics*, 8-29, London : Longman.
Legallois, D. (2002), « Incidence énonciative des adjectifs vrai et véritable en antéposition nominale », *Langue française*, 136, 1, 46-59.
Legallois, D. (2012), « La colligation : autre nom de la collocation grammaticale ou autre logique de la relation mutuelle entre syntaxe et sémantique ? », *Corpus* [En ligne], 11 | 2012, mis en ligne le 21 juin 2013, consulté le 03 juin 2017. URL : http://corpus.revues.org/2202
Legallois, D. & Tutin, A. (2013), « Présentation : Vers une extension du domaine de la phraséologie », *Langages*, 189, 1, 3-25.
Le Goffic, P. (1993), *Grammaire de la phrase française*, Paris : Hachette.
Le Goffic, P. (1994), Indéfinis, interrogatifs, relatifs (termes en Qu-) : parcours avec ou sans issue, *Faits de langues*, 4, 31-40.
Léon, J. (2008), « Aux sources de la « Corpus Linguistics » : Firth et la London School », *Langages*, 171, 12-33.
Léon, J. (2015), *Histoire de l'automatisation des sciences du langage*, Paris : ENS Éditions.
Léon, J. & Loiseau, S. (éds) (2016), *Studies in Quantitative Linguistics* (24): *History of Quantitative Linguistics in France*, Lüdenscheid: RAM-Verlag.
Liang, N. Y. (1986), « On computer automatic word segmentation of written Chinese », *Journal of Chinese Information Processing* 1 (1).
Lodge, R.-A. (2009), « La sociolinguistique historique et le problème des données», in *Sociolinguistique historique du domaine gallo-roman: Enjeux et méthodologies*, Aquino-Weber D. & al. (éds), Paris/Berne : Lang, 199-219.
Loiseau, S. (2011), « Les faits statistiques comme objectivation ou comme interprétation : statistiques et modèles basés sur l'usage », *Travaux de linguistique*, 62, 1, 59-78.
McEnery, T., Xia, R. & Tono, Y. (2006), *Corpus-Based Language Studies. An advanced resource book,* London/New-York : Routledge.
Marchello-Nizia, Ch. (1997), *La Langue française aux XIV e et XV e siècles*, Paris :Nathan

Marchello-Nizia, C. (2002), « Prépositions françaises en diachronie : une catégorie en question », *Linguisticae Investigationes* XXV, 2, 205-221.
Marchello-Nizia, C. (2004), « *Linguistique historique, linguistique outillée »* : les fruits d'une tradition, *Le français moderne*, 72(1),58–70.
Marchello-Nizia, C. (2011), « *Écrire une nouvelle grammaire historique du français à* **la** *lumière de l'histoire des descriptions de la langue* », in *Vers une histoire générale de la grammaire française. Matériaux et perspectives,* B. Colombat, J.-M. Fournier, V. Raby (éds), Paris : H. Champion, 45-60. https://bcl.cnrs.fr/rubrique137?lang=de
Martinie, B. & Vigier, D. (2013), « Le régime nominal de la préposition *en* dans la construction *être en + N abstrait* : une étude aspectuelle », *Langue Française,* 178*,* 59-79.
Mayaffre, D. (2002), « Les corpus *réflexifs* : entre architextualité et hypertextualité », *Corpus* [En ligne], 1 | 2002, mis en ligne le 15 décembre 2003, consulté le 15 août 2017. URL : http://corpus.revues.org/11.
Mazouer C. (2010), « La tragédie religieuse de la Renaissance et le mystère médiéval : l'attirance d'un contre-modèle », *Seizième Siècle*, 6, 95-105.
Melis, L. (1986), *Les circonstants et la phrase*, Louvain : PU de Louvain.
Melis, L. (2001), « La préposition est-elle toujours la tête d'un groupe prépositionnel ? », *Travaux de linguistique*, 42-43, 11-22.
Mellet, S. (2002), « Corpus et recherches linguistiques », *Corpus* [En ligne], 1 | 2002, mis en ligne le 15 décembre 2003, consulté le 15 août 2017. URL : http://corpus.revues.org/7.
Merle, J.-M. (2008), « Prépositions et aspect », *L'Information grammaticale*, 117, 52-56.
Melis, L. (2003), *La préposition en français*, Paris : Ophrys.
Molinier, C. (1990), « Les quatre saisons : à propos d'une classe d'adverbes temporels », *Langue française*, 86, 46-50.
Molinier, C. (2006), « Les termes de couleur en français. Essai de classification sémantico-syntaxique », *Cahiers* de *grammaire* 30, pp. 259-275.
Molinier, C, Levrier, F. (2000), *Grammaire des adverbes en –ment. Description des formes en –ment*, Genève : Droz.
Muller, C. (1973, [1992]), Initiation aux méthodes de la statistique linguistique, Paris : Champion.
Muller, C. (1977), *Principes et méthodes de statistique lexicale*, vol. 1, Paris : Hachette.
Muller, C. (1992), *Principes et méthodes de statistique lexicale*, vol. 2, Paris : Champion.
Muller, C. (2006), « Polarité négative et *free choice* dans les indéfinis de type *que ce soit* et *n'importe* », *Langages*, 162, 7-31.
Muller, C. (2007), « Les indéfinis *free choice* confrontés aux explications scalaires », *Travaux de linguistique*, 54, 83-96.
Paillard D. (2001), « À propos des verbes « polysémiques » : identité sémantique et principes de variation », *Syntaxe et sémantique*, 2, 99-120.
Paris, G. (1900), Préface à l'ouvrage de J. Bédier *Le roman de Tristan et Iseut*, Paris : L'Edition d'Art, H. Piazza, 5-17.
Pearson, J. (1998), *Terms in Context*, Amsterdam: John Benjamins.
Péry-Woodley, M.-P. (1995), « Quel corpus pour quels traitements automatiques ? », *Traitement Automatique des Langues*, 36 (1-2), 213-232.
Péry-Woodley, M.-P. (2000), *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle*. Linguistique. Université Toulouse le Mirail - Toulouse II.<tel-00410572>
Péry-Woodley, M.-P., Afantenos, S. D. Ho-Dac, L.-M. & Asher N. (2011), « La ressource ANNODIS, un corpus enrichi d'annotations discursives », *Traitement Automatique des Langues*, 52, 3, 71-101. <halshs-00935201>

Pincemin, B. (1999), *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat en Linguistique, Université Paris IV Sorbonne.

Popper, K.R. (1956, [1985]), « Trois conceptions de la connaissance », *Conjectures et réfutations*, Paris : Payot.

Pottier, B. (1974), *Linguistique générale, théorie et description*, Paris : Klincksieck.

Pottier, B. (1997), « Le cognitif et le linguistique dans l'expression des relations », *Faits de langues*, 9, 29-38.

Prévost, S. (2008), « Corpus informatisés de français médiéval : contraintes sur leur constitution et spécificités de leurs apports », Corpus [Online], 7 | 2008, Online since 13 November 2009, connection on 15 August 2017. URL : http://corpus.revues.org/1500.

Prévost, S. (2011), *Expression et position du sujet pronominal du 12ème au 14ème siècle : une approche quantitative (recherche inédite)*. Linguistique. Ecole normale supérieure de Lyon - ENS LYON, 2011. <tel-00667183>

Prévost, S. (2015), « Diachronie du français et linguistique de corpus : une approche quantitative renouvelée », *Langages*, 197, 23-45.

Quirk R. & al. (1985), *A Comprehensive Grammar of the English Language*, London/New York : Longman.

Rastier, F. (2011), *La mesure et le grain. Sémantique de corpus*, Paris : H. Champion.

Reutenauer C. (2002), *Vers un traitement automatique de la néosémie*, Thèse nouveau régime, Université de Lorraine.

Riegel, M. (1985), *L'adjectif attribut,* Paris : PUF.

Riegel, M., Pellat, J.-C. & Rioul, R. (2009), *Grammaire Méthodique du Français*, Paris : PUF, coll. Quadrige.

Rissanen, M. (2008), « Corpus linguistics and historical linguistics », in *Corpus linguistics. An international handbook*, A. Lüdeling & al. (eds.), Berlin/New-York : Mouton de Gruyter, 53-68.

Romary L., Salman, A.-S, Francopulo G. (2004), « Standards going concrete: from LMF to Morphalou », *Workshop on Electronic Dictionaries* , Coling 2004, Geneva.

Rossari, C. (1997), *Les opérations de reformulation. Analyse du processus et des marques dans une perspective contrastive français-italien*, 2e éd., Berne : P. Lang.

Royer, L. & Vigier D. (2014), « Les collocatifs nominaux des prépositions *en, dans, dedans* au XVIe s », in *Nouvelles perspectives en sémantique lexicale et en organisation du discours*, I. Novakova & P. Blumenthal (eds), Grenoble : PUG, 423-434.

Sagot, B. (2010), « The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French », in *Proceedings of the 7th international conference on Language Resources and Evaluation* (LREC 2010), Istanbul, Turkey.

Salem A. (1988), « Approches du temps lexical. Statistique textuelle et séries chronologiques », *Mots* 17, 105-143.

Sarda, L. (2010), « Les adverbiaux prépositionnels en *dans* : exploration en corpus de la notion de contenance », *Corela* [En ligne], HS-7 | 2010, mis en ligne le 31 mai 2010, consulté le 07 mai 2017. URL : http://corela.revues.org/911 ; DOI : 10.4000/corela.911

Sarda, L., Vigier, D. & Combettes, B. (2016), *Connexion et indexation. Ces liens qui tissent le texte*, collection *Langages*, Lyon : ENS éditions.

Sartre, J.-P. (1943), *L'Être et le Néant*, Paris : Gallimard.

Schmid, H. (1994), « Probabilistic part-of-speech tagging using decision trees », *Proceedings of International Conference on New Methods in Language Processing* – NeMLaP (Manchester, England), 44-49.

Sinclair, J. (1991), *Corpus Concordance Collocation*, Oxford: Oxford University Press.

Sinclair, J. (1996), *Preliminary Recommendations on Corpus Typology. Rapport technique,*

EAGLES (Expert Advisory Group on Language Engineering Standards). Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale. Pise. URL : http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Sinclair, J. (2004), « Corpus and Text - Basic Principles », in *Developing Linguistic Corpora: a Guide to Good Practice*. http://ota.ox.ac.uk/documents/creating/dlc/index.htm

Spang-Hanssen, E. (1963), *Les prépositions incolores du français moderne*, Copenhague : G.E.C GAD Forlag.

Spang-Hanssen, E. (1993), « De la structure des syntagmes à celle de l'espace. Essai sur les progrès réalisés dans l'étude des prépositions depuis une trentaine d'années », *Langages* 110, 12-26.

Stosic, D. (2009a), « Comparaison du sens spatial des prépositions « à travers » en français et « kroz » en serbe », *Langages*, 173, 15-33.

Stosic, D. (2009b), « La notion de « manière » dans la sémantique de l'espace », *Langages* 175, 103-121.

Stosic, D. & Fagard, B. (2012), « Formes et sens : de l'unicité à la variabilité », *Langages*, 188, 3-24.

Talmy, L. (1985), « Lexicalization patterns : Semantic structure in lexical forms », in Shopen T. (éd.), *Language Typology and Syntactic Description*. Vol. 3 : *Grammatical Categories and the Lexicon*, New York, Cambridge University Press : 57-149.

Talmy, L. (2000), *Toward a Cognitive Semantics*, Cambridge, MA, MIT-Press.

Terreaux L. (1968), *Ronsard correcteur de son œuvre. Les variantes des Odes et des deux premiers livres des Amours*, Genève : Droz.

Touratier, C. (2010), *La sémantique*, A. Colin.

Tutin, A. (2010), *Sens et combinatoire lexicale : de la langue au discours*, dossier en vue de l'habilitation à diriger des recherches, vol. 1, Université Stendhal Grenoble 3.

Valin, R., Hirtle, H. & Lowe, R. (1997), *Leçons de linguistique de Gustave Guillaume*, 1946-1947 et 1947-1948, Québec : Les Presses Universitaires de Laval.

Vandeloise, C. (1986), *L'espace en français : sémantique des prépositions spatiales*, Paris: Seuil.

Vandeloise C. (1988), « Les usages statiques de la préposition *à* », *Cahiers de Lexicologie* 53,119-148.

Vandeloise, C. (1993), « Les analyses de la préposition *dans* : faits linguistiques et effets méthodologiques », *Lexique*, 11, 15-40.

Vandeoise, C. (1999), « Quand *dans* quitte l'espace pour le temps », *Revue de Sémantique et Pragmatique*, 6, 145-162.

Vandeloise, C. (2000), « Verbes de changement, de transformation et de génération », *Cahiers de Lexicologie*, 77: 117-136.

Vandeloise, C. (2001), *Aristote et le lexique de l'espace : rencontres entre la physique grecque et la linguistique cognitive*, Stanford, CA: CSLI (Langage et Esprit).

Van Peteghem, M.. (2006), « Le datif en français: un cas structurel », *Journal of French Language Studies*, 16, 93–110.

Vendler, Z. (1957), « Verbs and Times », *Philosophical Rewiew*, 66, 143-160.

Victorri, B. (1999), « Le sens grammatical », *Langages,* 136, 85-105.

Victorri, B. (2003), « Langage et géométrie: l'expression langagière des relations spatiales », *Revue de Synthèse*, Paris : Springer Verlag/Lavoisier, 119-138.

Victorri, B. & Fuchs, C. (2006), *La polysémie, Construction dynamique du sens*, Paris : Hermès

Vigier, D. (2003), « Les syntagmes prépositionnels en « en N » détachés en tête de phrase référant à des domaines d'activité », *Lingvisticae Investigationes*, 26 (1), 97 - 122.

Vigier, D. (2004), *Les groupes prépositionnels en « en N » : de la phrase au discours*, Thèse

de troisième cycle, Université Paris 3 - Sorbonne Nouvelle.

Vigier, D. (2008), « Contribution à une étude des constructions antéposées du type : « En homme intelligent et humain, il partagea tout de suite l'inquiétude de Marcel » (J. Verne) », *Discours* [En ligne], 2 | 2008, mis en ligne le 03 juillet 2008. URL : http://discours.revues.org/863 ; DOI : 10.4000/discours.863.

Vigier, D. (2013a), « Comportement, déguisements, rôles, … De quelques emplois intraprédicatifs de *en* », *Lingvisticae Investigationes*, 36 (1), 1-19.

Vigier D. (2013b), « Sémantique de la préposition *en* : quelques repères », in *Langue Française*, 178*,* 3-19.

Vigier D. (2015), « Les prépositions *en*, *dans* et *dedans* au XVIe s. Approche statistique et combinatoire », *Le Français moderne*, 2, 230-247.

Vigier, D. (2017), « La préposition *dans* au XVI^{e} siècle. Apports d'une linguistique instrumentée », *Langages*, 206, 105-122.

Vigier, D. (à par. 2017), « Autour des SP adverbiaux aspectuels *en DétQuant Ntps* », in *Quand les formes prennent sens: la préposition envers et contre tout*, ouvrage d'hommage à D. Leeman, C. Vaguer (éd.), Limoges : Lambert-Lucas.

Wittgenstein L. (1953-1961), *Tractatus logico-philosophicus* suivi de *Investigations philosophiques*, Paris : Gallimard (TEL).

Waugh, L. (1976), « Lexical meaning : the prepositions *en* et *dans* in french », *Lingua*, 59, 69-118.

ANNEXES

ANNEXE 1

Métadonnées documentaires pour Presto [276]

1. Métadonnées associées à l'auteur

1.1. Métadonnées minimales

- **Identifiant de l'auteur** : identifiant IDREF de l'auteur lorsqu'il existe (ex : Scève, Maurice (1511?-1564?)), ou identifiant formé sur le même modèle s'il n'existe pas dans cette base.
- **Lien notice IDREF** : lien permanent vers la notice de l'auteur dans la base IDREF, qui contient son identifiant numérique (ex, pour Maurice Scève : http://www.idref.fr/027125343)
- **Nom de l'auteur :** contient le nom de famille de l'auteur si la forme normalisée du nom retenu (nom de famille IDREF) le permet, ou bien l'intégralité de l'identité, pour les auteurs médiévaux par exemple (Chrestien de Troyes), pour lesquels on ne distingue pas le nom et le prénom. Autre cas : les pseudonymes, du type Molière, seront entrés comme nom de famille, le prénom restant vide. Dans ce dernier cas, le véritable nom de l'auteur (Jean-Baptiste Poquelin) ne sera pas consigné dans nos métadonnées, mais le lien vers la notice IDREF permettra cependant d'y remonter.
- **Prénom de l'auteur :** Contient le prénom de l'auteur, lorsqu'il en a un (cf. auteurs médiévaux et pseudonymes). Dans le cas contraire ce champ est laissé vide.
- **Genre social de l'auteur** : homme, femme, multi-auteur (pseudonyme cachant plusieurs individus de sexe potentiellement différent), anonyme.
- **Date de naissance de l'auteur** : l'année suffit. Un point signifiera qu'un nombre est inconnu (ex : 198.), un point d'interrogation que la date est incertaine (1511 ?). Un tiret permet de donner un intervalle (1201-1205).
- **Date de mort de l'auteur** : l'année suffit. Un point signifiera qu'un chiffre est inconnu (ex : 168.), un point d'interrogation que la date est incertaine (1564 ?). Un tiret permet de donner un intervalle (1201-1205).

1.2. Métadonnées maximales

- **Lieu de naissance de l'auteur :** pays, ou ville et pays si connu. *(déjà présent dans la base)*
- **Lieu de mort de l'auteur :** pays, ou ville et pays si connu. *(déjà présent dans la base)*
- **Langue d'expression de l'auteur :** telle qu'indiquée dans l'IDREF. *(déjà présent dans la base)*
- **Langue maternelle de l'auteur**
- **Liens :** *(déjà présent dans la base)*
    - **VIAF**
    - **ISNI**
    - **BNF**
    - **LCCN**
    - **GND**
    - **NLA**

[276] La rédaction de la plus grande part de ce document a été assurée par V. Goossens (http://vannina.goossens.free.fr), Post-doctorante dans le programme Presto du 01/11/2013 eu 01/11/2014.

- **WorldCat**

2. Métadonnées associées à l'Œuvre

2.1. Métadonnées minimales

- **Titre de l'œuvre**
- **Date de la première édition de l'œuvre :** date de référence à partir du 16e siècle (cf. réflexion du consortium CAHIER). Un point signifiera qu'un nombre est inconnu (ex : 198.), un point d'interrogation que la date est incertaine (1511 ?). Un tiret permet de donner un intervalle (1201-1205).
- **Domaine de l'œuvre** : domaine dont relève l'œuvre. Si l'œuvre possède des domaines multiples non hiérarchisés, définir un domaine dominant puis les autres.
- **Genre de l'œuvre** : genre textuel dont relève l'œuvre. Si l'œuvre possède des genres multiples non hiérarchisés, définir un domaine dominant puis les autres.
- **Forme de l'œuvre** : forme textuelle dans laquelle est rédigée l'œuvre, soit en vers soit en prose. Si l'œuvre possède des formes multiples, on ne consigne que la forme dominante. On peut aussi proposer trois catégories (vers régulier, forme versifiée et prose), du fait de l'importance de la métrique.
- **Langue dominante de l'œuvre** : langue dans laquelle est rédigée l'essentiel de l'œuvre. Utiliser les codes ISO 639-2 (http://fr.wikipedia.org/wiki/Liste_des_codes_ISO_639-2): par exemple *fra* pour le français, *frm* pour le moyen français (1400-1600) et *fro* pour l'ancien français (842-1400).
- **Langue d'origine de l'œuvre** : si l'œuvre est une traduction. Utiliser les codes ISO 639-2 (http://fr.wikipedia.org/wiki/Liste_des_codes_ISO_639-2): par exemple fra pour le français, frm pour le moyen français (1400-1600) et fro pour l'ancien français (842-1400).
- **Nom du traducteur**

2.2. Métadonnées maximales

- **Date de composition de l'œuvre libre** : date à laquelle a été rédigée l'œuvre. souvent hypothétique, peut se composer d'une approximation ou d'une tranche temporelle, notés librement (vers 1803, ca 1803, 1803-1890, etc.).
- **Date de composition de l'œuvre** : date à laquelle a été rédigée l'œuvre, notée de manière formalisée pour permettre d'éventuels calculs, notamment la moyenne sur un intervalle de date. C'est la date de composition qui fait office de date de référence pour les textes médiévaux (cf. réflexion du consortium CAHIER). L'année suffit. Un point signifiera qu'un nombre est inconnu (ex : 198.), un point d'interrogation que la date est incertaine (1511 ?). Un tiret permet de donner un intervalle (1201-1205).
- **Genre libre** : permet d'entrer d'autres mots-clés, pour compléter le genre normalisé. Peut permettre de garder en mémoire les genres affectés par les différentes bases sources. Utile à la recherche documentaire.
- **Nom de la série dans laquelle s'inscrit l'œuvre** : par exemple, *Les Rougon-Macquart.*
- **Identifiant BNF de l'œuvre :** s'il existe.

3. Métadonnées associées à l'exemplaire

3.1. Métadonnées minimales

- **Titre de l'exemplaire :** souvent identique au titre de l'œuvre.
- **Date d'édition de l'exemplaire** : date de l'édition utilisée.
- **Nom de la maison d'édition/de l'imprimeur/du libraire de l'exemplaire :** par défaut c'est l'éditeur ; sinon, préciser le rôle entre parenthèse après le nom ; on peut préciser plusieurs rôles, séparés par des |. Exemple : « Éditions Toto | Presses Tutu (imprimeur) ».
- **Source de l'exemplaire** : base textuelle, collègue ou institution ayant fourni l'exemplaire numérisé utilisé, ou bibliothèque possédant l'exemplaire numérisé par PRESTO.
- **Licence de l'exemplaire** : licence précise régissant les droits pesants sur la version numérisée de l'exemplaire utilisé (ex : CC BY-NC-SA), pour les textes qui seront redifusés par PRESTO uniquement.
- **Type de licence** : « libre » ou « non libre ».
- **Cote Frantext**
- **Cote BVH**
- **Nom du fichier**

3.2. Métadonnées maximales

- **Nature de l'exemplaire** : manuscrit, revue, ouvrage imprimé… (liste à stabiliser)
- **Date du manuscrit** : date importante pour les textes médiévaux, mais aussi pour les manuscrits non édités (ex : Manuscrits de Stendhal).
- **Exemplaire révisé/remanié par l'auteur** : oui, non, non renseigné.
- **Source 2 (source de la source) :** source auprès de laquelle la source (cf. ci-dessus) s'est procuré l'exemplaire.
- **Droits pesant sur l'exemplaire** : librement utilisable ou non (permet de faire des requêtes même si la licence n'est pas remplie). Se référer à la licence pour plus de précisions sur les possibilités juridiques permises.
- **Titre long de l'exemplaire** : uniquement pour les textes possédant un titre abrégé et un titre complet (ex : textes des BVH).
- **Titre du volume contenant l'exemplaire** : lorsque le texte a été édité au sein d'un recueil, de type œuvres complètes.
- **Numéro du volume contenant l'exemplaire** : dans le cas où le texte fait partie d'un recueil, noter l'éventuel numéro du volume le contenant.
- **Nombre de volumes** : nombre de volume total du recueil.
- **Caractères modernisés** : indication concernant les interventions ayant pu intervenir sur les caractères typographiques de l'exemplaire, que ce soit par l'éditeur de la version imprimée ou bien celui de la version numérisée (ex : transformation des *v* en *u*).
- **Orthographe modernisée** : indication concernant les interventions ayant pu intervenir sur l'orthographe de l'exemplaire, que ce soit par l'éditeur de la version imprimée ou bien celui de la version numérisée (ex :).
- **Ville maison d'édition/imprimeur/libraire de l'exemplaire**
- **ISBN de l'exemplaire** : pour les documents imprimés uniquement, hors revues.

4. Métadonnées associées à l'éditeur scientifique

Aucune métadonnée ne figure actuellement dans le corpus PRESTO

Métadonnées maximales

- **Nom de l'éditeur scientifique**
- **Prénom de l'éditeur scientifique**
- **Date de mort de l'éditeur scientifique** : à inscrire si des recherches sont effectuées pour la détermination des droits d'auteur.
- **Numériseur** : personne ayant numérisé le texte (ex : Marie-Lucie Demonet)

5. Remplissage et vérification des métadonnées

Pour les auteurs, la ressource qui fera autorité pour la complétion et la vérification des métadonnées associées aux texte est la base IDREF : http://www.idref.fr/autorites/autorites.html. Dans le cas ou un auteur n'apparaitrait pas dans cette base, d'autres ressources peuvent être envisagées en complément (indiquer en commentaire la provenance des informations) :

- La **BNF** : http://data.bnf.fr
- Le **DEAF**, pour les auteurs médiévaux : http://www.deaf-page.de/fr/index.php
- Des **manuels**, comme le *Dictionnaire universel des littératures* (sous la direction de Béatrice Didier, PUF, 1994), disponible à la bibliothèque Diderot. Compléter par d'autres références éventuelles.

ANNEXE 2

Jeu d'étiquettes Presto_min

Nous avons choisi de distinguer deux niveaux dans le jeu d'étiquettes PRESTO : Presto_min (jeu minimal) et Presto_max (jeu maximal).

Pour Presto_min, les champs utilisés sont <catégorie> <type> <mode>.

Pour chaque catégorie, nous spécifions les modifications apportées par rapport à MULTEXT *english* (2010) (désormais MULTEXT (*en*) : http://nl.ijs.si/ME/V4/msd/html/msd-en.html) et GRACE (1997).

CATEGORIE	**Valeur**	**Code**
CATEGORIE	Nom	N
CATEGORIE	Verbe	V
CATEGORIE	Adjectif	A
CATEGORIE	Pronom	P
CATEGORIE	Déterminant	D
CATEGORIE	Participe-Adjectif-Gérondif	G
CATEGORIE	Adverbe	R
CATEGORIE	Adposition	S
CATEGORIE	Conjonction	C
CATEGORIE	Numéral	M
CATEGORIE	Interjection	I
CATEGORIE	Résidu	X
CATEGORIE	Ponctuation	F

1. Noms (Nouns)

P	Attribut	Valeur	Code	Exemple
0	CATEGORIE	Nom	N	
1	Type	commun	c	*livre*
		propre	p	*Jean*

Comparaison par / à MULTEXT *(en)* et GRACE

ATTRIBUT : TYPE

- MULTEXT *(en)* propose deux valeurs : « common » (c), « proper » (p).
- GRACE , Outre les valeurs « common » (c), « proper » (p), propose la valeur « cardinal» (k). En effet, GRACE prend le parti de supprimer la catégorie « Numéral » proposée par MULTEXT *(en)* au profit d'une valeur additionnelle « cardinal » (notée : k) à l'attribut type des différentes catégories syntaxiques pouvant intégrer des emplois de numéraux cardinaux.
- PRESTO-MIN propose deux valeurs : « common » (c), « proper » (p) et opte comme MULTEXT *(en)* pour une catégorie « Numéral » (M).

2. Verbes (Verbs)

P	Attribut	Valeur	Code	Exemple
0	CATEGORIE	Verbe	V	
1	Type	être & avoir	u	*ai, suis*
		autre	v	*pars*
2	VForme	V conjugué à un mode personnel	c	*avons, étions, partirai*
		infinitif	n	*être, avoir, partir*

Comparaison par / à MULTEXT (en) et GRACE

ATTRIBUT : TYPE

- MULTEXT *(en)* propose 4 valeurs : « main» (m), « auxiliary» (a), « modal » (o), base (b).
- GRACE propose 2 valeurs : « main» (m), « auxiliary» (a).
    - PRESTO-MIN propose 2 valeurs : « être / avoir » (u), « Autre verbe » (v). On ne tranche pas entre emplois d'auxiliaires pour *être/avoir* et emplois de verbes pleins. Cette décision est directement liée au traitement appliqué dans Presto_min aux participes passés, pour lesquels nous avons décidé de ne pas trancher entre participes et adjectif (cf. *infra*). Or cette décision implique de ne pas trancher entre les structures du type *NO être Participe / NO être Adj.*, c'est-à-dire entre *être* auxiliaire et *être* verbe copule.

ATTRIBUT : FORME VERBALE (VFORM)

- MULTEXT *(en)* propose 4 valeurs : « indicative » (i), conditional (c), infinitive (n), participle (p)
- GRACE propose 6 valeurs : « indicative » (i), conditional (c), « subjonctive » (s), « imperative » (m), infinitive (n), participle (p).
- PRESTO-MIN propose deux valeurs pour l'attribut VForme
    - « Verbe conjugué à un mode personnel» (c) : cette valeur a été empruntée à Cattex09min (http://bfm.ens-lyon.fr/article.php3?id_article=176) qui lui a affecté le code (cjg) ;
    - « infinitif » (n).

Rem : Le mode non personnel « participe » ne donne pas lieu à une valeur car il est traité dans la catégorie G qui ne distingue pas entre participes (présent ou passé), adjectifs verbaux et gérondifs.

3. Adjectifs (Adjectives)

P	Attribut	Valeur	Code	Exemple
0	CATEGORIE	Adjectif	A	
1	Type	général	g	*aimable, municipal, futur, tel, …*
		possessif	s	*(un) mien (cousin)*

Comparaison par / à MULTEXT *(en)* et GRACE

ATTRIBUT : TYPE

- MULTEXT (*en*) propose 1 valeur : « qualificative » (q).
- GRACE distingue 5 valeurs: « qualificative » (q), « ordinal » (o), « cardinal » (k), « indefinite » (i), « possessive » (s).
- PRESTO-MIN propose 2 valeurs.
    - la valeur « général » (g) se substitue à « qualificatif » (q). Cette valeur de type rassemble, outre les traditionnels adjectifs qualificatifs, d'autres sous-catégories qui ne présentent pas les mêmes caractéristiques syntaxiques et distributionnelles que les traditionnels qualificatifs : les adjectifs « relationnels » (*municipal,* ...), les adjectifs du « troisième type » (Schnedeker (éd.) (2002) ; Riegel & al. 2009 : 634).
    - la valeur « possessif » (s) est conservée telle quelle.

Rem 1 : Sont éliminées les valeurs « ordinal » et « cardinal », les « adjectifs » correspondants étant placés dans la catégorie « numéral » (M) ; quant aux traditionnels « adjectifs indéfinis », la plupart sont versés dans la catégorie « Déterminants » (D).

Rem 2: Est conservée l'étiquette « adjectif possessif » pour les occurrences de *mien, tien, sien , ...* dans des contextes comme « *un mien cousin* », pour des raisons d'ordre diachronique et distributionnelle. Il s'avère en effet que ces formes sont combinables avec un adjectif qualificatif épithète dans un GN aux XVIe et au XVII e s. du moins

> « *Je propose les fantasies humaines et miennes, simplement comme humaines fantasies* », M. de Montaigne, *Essais : t. 1 (livres 1 et 2)*, 1592
>
> « *tirée de ceste cordiale et mienne bénéficence* » (lettres missives de Henri IV, t VII, p. 623. 23 octobre 1608. Citée dans *Henri IV et sa politique*, Charles Mercier de Lacombe, 1860, p. 814.

4. Pronoms (Pronouns)

P	Attribut	Valeur	Code	Exemple
0	CATEGORIE	Pronom	P	
1	Type	personnel	p	*je, le, en*
		démonstratif	d	*ce, celui*
		indéfini	i	*certains, plusieurs*
		possessif	s	*(le) mien*
		interrogatif	t	*qui, que*
		relatif	r	*qui, lequel*

Comparaison par / à MULTEXT *(en)* et GRACE

ATTRIBUT : TYPE

- MULTEXT (*en*) Propose 7 valeurs : « personal » (p), « possessive » (s), « interrogative » (q), « relative » (r), « reflexive » (x), « general » (g), « ex-there » (t).
- GRACE propose 8 valeurs: « personal » (p), « demonstrative » (d), « indefinite » (i), « possessive » (s), « interrogative » (t), « relative » (r), « reflexive » (x), « cardinal »

(k).

- PRESTO_MIN propose 6 valeurs : « personal » (p), « demonstrative » (d), « indefinite » (i), « possessive » (s), « interrogative » (t), « relative » (r).

  Rem 1 : Sont éliminées les valeurs :
  - « reflexive » (x) : valeur fondue dans la valeur « personnel » (p)
  - « general » (g) qui correspond à un choix propre à MULTEXT que nous ne suivons pas[277]
  - « ex-there» (t) : non pertinent pour le français
  - « cardinal » (k) : les pronoms cardinaux sont versés dans la catégorie englobante « numéral » (M)

  Rem 2 : pour la valeur « interrogatif », le code (t) adopté est repris de GRACE.

5. Déterminants (Determiners)

P	Attribut	Valeur	Code	Exemple
0	CATEGORIE	Déterminant	D	
1	Type	article défini	a	*le, la, l', les*
		démonstratif	d	*ce, cet, cette, ...*
		possessif	s	*mon, ta, leur, ...*
		article indéfini	n	*un, une, des, de, d'*
		article partitif	p	*du, de la , de l', des*
		indéfini	i	*quelque(s) N, tout N, chaque N, ...*
		relatif	r	*lequel, laquelle, ...*
		interrogatif/ exclamatif	t	*quel, quelle, ...*

Comparaison par / à MULTEXT *(en)* et GRACE

ATTRIBUT : TYPE

- MULTEXT (*en*) Propose 4 valeurs : « demonstrative » (d), « indefinite» (i), « possessive» (s), « général » (g).
- GRACE propose 7 valeurs : « article » (a), « demonstrative » (d), « possessive» (s), « indefinite» (i), « interr./excl.» (t), « relative» (r), cardinal (k).
- PRESTO_MIN propose 10 valeurs :
  - Sont retenues les 5 valeurs « demonstrative » (d), « possessive» (s), « indefinite» (i), « interr./excl.» (t), « relative» (r)
  - Est modifiée la valeur (« article défini ») associée au code (a)
  - Sont ajoutées les valeurs :

[277] *"General" pronouns are those which are not personal, possessive, demonstrative or reflexive. The choice of these four categories is based on distributional facts, though at a rather high level of abstraction. They enter into anaphoric dependencies which are signalled morphosyntactically and are therefore (in principle) more amenable to automatic detection. Most general pronouns do not, although they too sometimes encode number information.*

- « article indéfini » (n)
- « article partitif » (p)
- « négation » (n) : réunit les emplois de *de* sous la portée de la négation : *Je n'ai pas de voiture* / *Pas de nuages à l'horizon.*
- déterminants « complémentaires » (c) : réunit l'ensemble des prédéterminants, postdéterminants et identificateurs qui entrent dans la composition des groupes déterminants définis et indéfinis sans en constituer la tête (Riegel & al. 2009 : 304-305).

6. Participes, adjectifs verbaux, gérondifs

Cette catégorie est inexistante dans MULTEXT (*en*) ET dans GRACE.
Ce choix s'explique par le constat que la distinction entre les trois classes de mots : *participes, adjectifs verbaux, gérondifs* pose des problèmes nombreux en synchronie et en diachronie.

- *En synchronie*, la mise au point de procédures de décisions pour les emplois ambigus (nombreux) nécessitent plusieurs tests (Riegel & al. 2009 : 737-738) qui augmentent les chances de divergences entre annotateurs. En outre, ces tests ne garantissent pas la mise à l'écart de toute appréciation subjective : *Selon les cas (le type de verbe, le contexte), ils [les participes] sont sentis comme plus ou moins « verbaux » ou « adjectivaux » (avec une marge appréciable de liberté d'interprétation)* (P. le Goffic 1993, § 134 : 201)
- *La dimension diachronique* ajoute une difficulté supplémentaire car la distinction morphologique entre participe présent, adjectif verbal et gérondif est problématique. *En français classique, la tripartition des formes en –ant ne va pas de soi (…) dans la mesure où la différence syntaxique et sémantique entre les trois catégories ne se marque pas formellement par une morphologie distinctive : le gérondif, invariable, se distingue mal du participe (au masculin singulier) du fait qu'il n'est pas régulièrement précédé de* en *; le participe qui peut être variable en genre et en nombre, se distingue mal de l'adjectif verbal.* (N. Fournier, 2002, § 421 : 291-292)

P	Attribut	Valeur	Code	Exemple
0	CATEGORIE	Participe, adjectif verbal, gérondif	G	
1	Type	participe présent - adjectif verbal - gérondif	a	*chantant, (en) chantant*
		participe passé - adjectif verbal	e	*instruit*

7. Adverbes

P	Attribut	Valeur	Code	Exemple
0	CATEGORIE	Adverbe	R	
1	Type	général	g	*fortement, hier, ici*
		particule	p	*ne, n'*
		interro-exclam	t	*où, quand, comment, pourquoi, …*

Comparaison par / à MULTEXT *(en)* et GRACE

ATTRIBUT : TYPE

- MULTEXT (*en*) Propose 2 valeurs : « modifier » (m), « specifier » (s),
- GRACE propose 3 valeurs : général (g), particle (p), interro-exclam (x)
- PRESTO_MIN reprend les 3 valeurs proposées par GRACE mais affecte le code (t) aux adverbes interro-exclam. pour conserver une cohérence avec le code utilisés pour les pronoms et les déterminants interro-exclam.

8. Prépositions (Adpositions)

P	Attribut	Valeur	Code	Exemple
0	CATEGORIE	Préposition	S	

Comparaison par / à MULTEXT *(en)* et GRACE

ATTRIBUT : TYPE

- MULTEXT (*en*) Propose 2 valeurs : « preposition » (p), « postposition » (t),
- GRACE propose 2 valeurs : « préposition » (p), « déictique » (d)
- PRESTO_MIN propose la seule valeur « préposition ».

9. Conjonctions (conjunction)

P	Attribut	Valeur	Code	Exemple
0	CATEGORY	Conjonction	C	
1	Type	coordination	c	*mais, et, &, ...*
		subordination	s	*que*

Pas de modifications pour les attributs du « type » par rapport à MULTEXT (*en*) et GRACE.
Rem : *donc* est traité comme un adverbe

10. Numéral (Numeral)

P	Attribut	Valeur	Code	Exemple
0	CATEGORY	Numeral	M	
1	Type	cardinal	c	*deux*
		ordinal	o	*deuxième*

Pas de modifications par rapport à MULTEXT (*en*).
Pour GRACE, voir 1.1.

Rem : le mot *dernier* est codé Ag (n'est pas un numéral).

11. Interjections (Interjections)

P	Attribut	Valeur	Code	Exemple
0	CATEGORY	Interjection	I	Hep !

Pas de modifications par rapport à MULTEXT (*en*) et GRACE

12. Ponctuations (Punctuations)

P	Attribut	Valeur	Code	Exemple
0	CATEGORY	Numeral	F	
1	Type	forte	s	*. ! ?*
		faible	w	*, : ;*
		Autre (trait d'union, tiret, ponctuation parenthétique, …)	o	*- () []*

Catégorie absente de MULTEXT (*en*)
Présente dans GRACE, qui ne propose pas de champ <type>.

13. Résidu (Residual)

P	Attribut	Valeur	Code	Exemple
0	CATEGORY	Résidu	X	
1	Type	abréviation	a	*Dir.*
		mot étranger	e	*linguistics*
		symbole	s	*@*
		préfixe	p	*hyper-, ex-*
		consonne intercalée	i	*a-**t**-on, **l'**on*

Cette catégorie, absente de GRACE [catégorie « unknown » (?)], est empruntée à MULTEXT (*en*), qui ne propose pas de champ <type>.

Rem : Les abréviations que nous faisons figurer comme « valeur » pour l'attribut « type » de la catégorie « résidu » font l'objet d'une catégorie spécifique dans MULTEXT (*en*):
« Abbreviation » (Y).

INDEX

Index des attributs

Attribut	**Catégorie**	**Position**
Type	Adjectif	1
Type	Adposition	1
Type	Adverbe	1
Type	Conjonction	1
Type	Déterminant	1
Type	Interjection	1
Type	Nom	1
Type	Numéral	1
Type	Participe-Adjectif-Gérondif	1

Type	Ponctuation	1
Type	Pronom	1
Type	Résidu	1
Type	Verbe	1
Vforme	Verbe	2

Index des valeurs

Valeur	Code	Attribut	Catégorie
abréviation	a	type	résidu
article défini	a	type	déterminant
article indéfini	n	type	déterminant
article partitif	p	type	déterminant
autre	o	type	ponctuation
autre	v	type	verbe
cardinal	c	type	numéral
commun	c	type	nom
consonne intercalée	i	type	résidu
coordination	c	type	conjonction
démonstratif	d	type	déterminant
démonstratif	d	type	pronom
être & avoir	u	type	verbe
faible	w	type	ponctuation
forte	s	type	ponctuation
général	g	type	adjectif
général	g	type	adverbe
indéfini	i	type	déterminant
indéfini	i	type	pronom
infinitif	n	Vforme	verbe
interrogatif/ exclamatif	t	type	adverbe
interrogatif/ exclamatif	t	type	déterminant
interrogatif	t	type	pronom
mot étranger	e	type	résidu
ordinal	o	type	numéral
participe passé - adjectif verbal	e	type	Participe, adjectif verbal, gérondif
participe présent - adjectif verbal - gérondif	a	type	Participe, adjectif verbal, gérondif
particule	p	type	adverbe
personnel	p	type	pronom
possessif	s	type	adjectif
possessif	s	type	déterminant
possessif	s	type	pronom
préfixe	p	type	résidu
propre	p	type	nom
relatif	r	type	déterminant
relatif	r	type	pronom
subordination	s	type	conjonction
symbole	s	type	résidu

V conjugué à un mode personnel	c	Vforme	verbe

ANNEXE 3

Règles d'affectation des lemmes dans Presto

1. Noms (Nouns)

Pour le *nom commun*, forme masc. sing.
Pour le *Np*, forme rigide. Rem : variation graphique en diachronie des toponymes prise en compte dans les critères de fusion (contrôle manuel).

2. Verbes (Verbs)

Forme à l'infinitif simple

3. Adjectifs (Adjectives)

Forme masc. sing.
Rem : les formes comparatives et superlatives synthétiques sont conservées. *Meilleur* aura pour lemme MEILLEUR, *mieux* pour lemme MIEUX, etc.

4. Pronoms (Pronouns)

- Pro. personnel :
  *on conserve la **forme atone sujet** pour chaque **personne grammaticale** sans considération pour la fonction syntaxique. Par ex. lemme JE pour les formes *je, j', me, m', moi* - etc.
  * **on, en, y** ont un lemme spécifique (ON, EN, Y)
  * les pro. réfléchis n'ont pas de lemme spécifique

CCL : total des lemmes possibles: JE TU IL NOUS VOUS ILS ON EN Y

- Pro. démonstratif : on distingue la forme fléchie CELUI {*celui, celle, ceux, celles*} de la forme neutre CE {*ce, c},* CECI, CELA.
- Pro. indéfini : forme masc. sing. quand une flexion est possible
- Pro. possessif : on conserve un lemme (forme masc. sing.) par personne grammaticale => lemmes possibles : MIEN, TIEN, SIEN, NÔTRE, VÔTRE, LEUR
- Pro. interrogatif : LEQUEL (flexion possible), QUI, QUE (forme tonique *quoi* ramenée à la forme atone). Rem : les pronoms interrogatifs renforcés sont segmentés.
- Pro. relatif : idem interrogatifs

5. Déterminants (Determiners)

Forme masc. sing. pour chaque type quand la flexion est possible (*versus chaque*, par ex.)

			LEMMES
1	Type	article défini	LE
		démonstratif	CE
		possessif	MON, TON, ….

		article indéfini	UN
		article partitif	DU
		indéfini	CHAQUE, TOUT, CERTAIN, …
		relatif	LEQUEL
		interrogatif/ exclamatif	QUEL

6. Participes, adjectifs verbaux, gérondifs

Forme masc. sing. pour chaque type

7. Adverbes

RAS

8. Prépositions (Adpositions)

Rem : les formes *dedans, dessus, ...* ne sont pas traitées comme des allomorphes de *dans, sur, ...* mais se voient affecter un lemme spécifique (DEDANS, DESSUS, …).

9. Conjonctions (conjunction)

RAS

10. Numéral (Numeral)

Rem : le mot *dernier* est codé Ag (n'est pas un numéral).
Un lemme spécifique, en chiffres arabes, est affecté aux formes. Par ex. *trois, III, 3, ...* a pour lemme « 3 ».

11. Interjections (Interjections)

RAS

12. Ponctuations (Punctuations)

RAS

13. Résidu (Residual)

RAS

ANNEXE 4

CORPUS ÉTENDU PRESTO