Une grammaire fondée sur un corpus numérique

Sophie Prévost

Chapitre 3
Une grammaire fondée sur un corpus numérique

3.1 Une histoire de la langue française

Une langue, même appréhendée dans une perspective syntchronique, n'est pas un objet simple : elle revêt des formes diverses en fonction de différents paramètres, tels que le registre, la situation de communication, mais aussi la forme des textes, leur domaine, et, le cas échéant, le dialecte. Parler de LA langue française est donc une commodité de langage : sous cette appellation se dissimulent des variétés, des usages, tous un peu différents, mais néanmoins suffisamment similaires pour que l'on puisse les considérer comme autant de mises en œuvre d'une même langue, le français.

Une grammaire peut vouloir dégager les points communs entre les différents usages d'une langue, mettre au jour ce qui les réunit en une sorte de *koinè*, et laisser à la marge ce qui les distingue les uns des autres ; elle peut même, dans une perspective plus prescriptive que descriptive, et comme cela fut le cas pendant des siècles, ne s'intéresser qu'à certains usages, représentatifs de la norme, et tout simplement ignorer les autres ou ne les mentionner que comme étant déviants. Elle peut aussi, à l'inverse, prendre en compte la diversité des usages.

L'émergence et le développement de la linguistique, en particulier de la sociolinguistique, ont largement modifié la conception et l'approche de la langue. Beaucoup en sont désormais convaincus : faire la grammaire d'une langue, c'est tout autant dire expliciter ce qui devrait être que constater ce qui est réellement, dans toute sa variété, et tenter de l'expliquer. Il reste cependant difficile de déterminer – c'est-à-dire identifier de manière contrastive – les classes grammaticales à recenser et à représenter : sur ce point, notre connaissance, même si elle progresse sans cesse, reste encore incomplète, aussi bien pour les états de langue modernes que pour les états plus anciens (et ce pour des raisons différentes : exploration inachevée d'un ensemble fini de documents, dont certains sont encore à découvrir, et disparition de certains autres pour la langue ancienne, prolifération continue de documents pour les états de langue modernes).

Une grammaire qui a pour objectif de restituer l'évolution de la langue française doit, plus encore qu'une grammaire synchronique, nécessairement prendre en compte la variété de la langue. Il est en effet désormais admis que la plupart des changements résultent d'un fait de variation préalable (➔ 2.4): une forme ou une construction nouvelle, un paradigme nouveau, correspondent le plus souvent à une variante qui s'est imposée. On ne saurait donc considérer les seules normes (qui n'ont en outre pas été édictées avant le 16^{e} s.), il convient au contraire d'envisager aussi les usages déviants. Cela signifie que, à chaque étape, se trouve répétée la complexe identification de l'ensemble des usages, afin de redéfinir l'objet d'étude.

Une grammaire qui adopte une perspective diachronique couvrant plusieurs siècles est par ailleurs confrontée à la complexité du rapport à la langue. Celui-ci est en effet fort différent selon que l'on considère les états de langue anciens, pour lesquels il n'existe plus de locuteurs, ou au contraire le français moderne, pour lequel nous bénéficions de notre compétence de locuteur. L'absence de locuteur n'a certes pas le même impact selon la période considérée. En effet, alors qu'un locuteur moderne ne peut comprendre un texte du 13^{e} s. s'il n'a pas *appris* l'ancien français, il peut lire et comprendre un texte du 18^{e} s. même s'il ne

possède pas à proprement parler la compétence de la langue de cette époque, et n'en maîtrise donc pas les nuances (d'où de probables contresens, souvent légers). Se pose ici la question de déterminer la frontière temporelle à partir de laquelle un locuteur moderne ne peut plus comprendre un état de langue passé, et a besoin de l'apprendre comme une langue étrangère. La perte de compréhension est progressive : alors qu'on lit parfaitement Zola, très bien Musset, un peu moins bien La Bruyère et Marguerite de Navarre, la compréhension des *Cent nouvelles nouvelles* (15e s.) devient difficile, et celle de la *Chanson de Roland* (début 12e s.) impossible pour les non-initiés. Ainsi notre sentiment de familiarité avec la langue décline au fur et à mesure que l'on recule dans le temps, mais, quelle que soit l'époque considérée, l'absence de compétence pour l'état de langue qu'on se propose de décrire oblige à se fonder sur des données attestées : le recours aux textes, au sens large de ce terme, s'impose tout autant pour décrire la langue du 18e s. que celle du 13e s.

Il est possible de rédiger une grammaire du français moderne en se fondant principalement sur sa compétence de locuteur enrichie de l'expérience que l'on a des différents usages et de leur confrontation, et en se passant donc de l'appui sur des données attestées (même si l'on perd ainsi de précieux éléments de diversité). Une telle démarche est exclue pour les états de langue anciens puisque seuls les textes nous fournissent les données langagières à décrire ; en effet, les *Manières de langage*, qui apparaissent vers le 14e s., et les grammaires, qui se développent à partir du 16e s., contiennent des éléments d'information mais il est difficile d'évaluer dans quelle mesure ces ouvrages reflètent vraiment l'usage, la probabilité étant grande que, comme encore souvent en français moderne, leur valeur soit plus prescriptive que descriptive. Le corpus utilisé peut ne pas apparaître de manière explicite, ou ne pas présenter un caractère raisonné (en particulier lorsque la description et / ou l'analyse des faits langagiers s'est appuyée sur des études antérieures et que les sources textuelles de celles-ci ne sont pas mentionnées, ou bien quand les textes fonctionnent comme un simple réservoir à exemples) : les données attestées, et donc les textes, ont néanmoins toujours joué un rôle essentiel dans la découverte et l'appréhension des états de langue passés, ainsi que dans la compréhension de leur *grammaire*.

En revanche, ce qui est relativement récent, c'est l'importance accordée à la quantification des données, et à leur caractérisation selon des critères fixes, la fréquence plus ou moins élevée d'une construction ne pouvant en effet s'apprécier de manière intuitive, même d'ailleurs pour sa propre langue. C'est une démarche décisive pour pleinement rendre compte de l'émergence et de la propagation, ou au contraire du recul, d'une construction, ainsi que des faits de concurrence entre constructions, c'est-à-dire, d'une manière générale, de la variation et des changements qui en résultent.

3.2 Les grammaires et les corpus

Dans les ouvrages de grammaire sur la langue ancienne édités depuis un siècle, les textes qui sont utilisés sont toujours cités, ce qui permet d'observer des écarts assez importants quant à leur nombre. Cela est en partie dû, à l'époque moderne, à des choix personnels de la part des auteurs de ces grammaires, mais aussi et surtout aux possibilités nouvelles offertes par la numérisation des textes et à l'automatisation de certaines tâches de recherche et d'extraction.

L'évolution est nette lorsque l'on considère quelques ouvrages de référence, en s'en tenant par exemple à la période médiévale. Ainsi, pour l'ancien français, la *Petite syntaxe de*

l'ancien français de L. Foulet s'appuie de manière explicite sur 13 textes (essentiellement du 13e s., et quelques-uns de la seconde moitié du 12e s.) pour sa première édition en 1919 (19 textes pour la 3e édition en 1930), qui ont permis dès cette époque – et c'était alors tout à fait inédit – de calculer la fréquence de certains faits linguistiques. La *Grammaire de l'ancien français* de G. Moignet (^{2}1984 [1973]) se fonde quant à elle sur 102 textes, tandis que la *Grammaire nouvelle de l'ancien français* de C. Buridant (2000a) mentionne plus de 350 textes auxquels s'ajoutent des index et des concordances électroniques, ainsi que des fréquences issues de nombreuses études portant sur des thèmes précis. Pour le moyen français, la *Syntaxe du moyen français* (1980) de R. Martin et M. Wilmet, qui couvre une brève synchronie (1455-1465), repose sur 24 textes (dont 3 numérisés et analysés exhaustivement) ; *La langue française aux XIVe et XVe siècles* de C. Marchello-Nizia (^{2}1997a [1979]), qui se place dans une perspective partiellement diachronique (14e-15e s.), mentionne de son côté 69 textes, qui constituent deux corpus, dont l'un a fait l'objet de dénombrements systématiques. Pour ce qui est des grammaires portant sur les périodes ultérieures, la *Grammaire de la langue française du seizième siècle*, de G. Gougenheim (^{2}1974 [1951]) mentionne 33 textes, tandis que la *Grammaire du Français classique*, de N. Fournier, écrite quelques décennies plus tard (1998), en cite plus de 300.

Si le nombre de textes retenus est variable (on aimerait d'ailleurs aussi connaître leur nombre de mots), l'utilisation qui en est faite l'est tout autant. En effet les textes peuvent avoir subi un dépouillement exhaustif, ou avoir été l'objet de sondages réguliers. Ils peuvent au contraire n'être que des réservoirs à exemples (cela semble être les cas dans les grammaires de G. Gougenheim et de N. Fournier, qui dressent la liste des textes « cités »). De plus, les textes utilisés n'ont pas nécessairement été tous soumis au même traitement. Ces points ne sont guère explicités (C. Marchello-Nizia (^{2}1997a) distingue néanmoins les « textes dépouillés ou consultés » et ceux « auxquels il est fait occasionnellement référence »). On peut le regretter, car ces choix ne sont pas sans conséquence sur les descriptions proposées.

Il est un autre aspect qui varie assez fortement d'une grammaire à l'autre : il s'agit de la diversité des textes retenus. Ainsi, chez Foulet, les textes sont littéraires, alors que la grammaire de Marchello-Nizia s'appuie sur des genres plus variés, prenant par exemple en compte des textes historiques. Cette caractéristique qualitative est elle aussi susceptible de peser assez lourdement sur les analyses proposées (et cela d'autant plus que les textes ne sont pas de simples réservoirs à exemples mais étayent véritablement l'analyse).

A ce jour, aucune grammaire sur le français moderne ne s'appuie sur un corpus explicite et raisonné, qui serait exploité de manière systématique (ou non) pour l'ensemble des faits décrits et analysés : les données attestées qui sont mentionnées ne viennent qu'exemplifier la description et l'analyse. De ce point de vue, la grammaire publiée il y a vingt ans pour l'anglais contemporain, *The Longman Grammar of Spoken and Written English* de D. Biber et *al.* (^{4}2004 [1999]), constitue un ouvrage remarquable. Trois aspects en particulier méritent d'être signalés : la prise en compte systématique de différents registres, écrits et oraux, l'établissement de données quantifiées, et enfin l'exploitation à géométrie variable des corpus (40 millions de mots), en fonction des points à étudier. La démarche à l'œuvre traduit un traitement raisonné et explicité des textes, qui a été permis par une vaste entreprise de numérisation et de catégorisation des textes.

Il n'était pas possible, dans le cadre du présent projet, d'envisager une entreprise de constitution de corpus analogue à celle réalisée pour la grammaire de Biber et *al.*, et ce d'autant que notre projet ne s'en tient pas à trois décennies, mais couvre une période de 12 siècles. C'est donc, pour une large part, sur les ressources existantes que nous nous

sommes appuyés, à savoir des textes déjà numérisés, et enrichis morpho-syntaxiquement (parfois syntaxiquement) pour certains d'entre eux.

3.3 La représentativité du corpus

La difficulté de délimiter l'objet qu'une grammaire se donne à décrire trouve un écho direct dans la constitution du corpus, qui, idéalement, devrait regrouper toutes les variétés, tous les usages recensés dans la langue. Trouver des textes, en nombre suffisant, et les représentant tous, pose un problème particulièrement aigu pour les états de langue anciens. En effet, certains usages ne nous sont parvenus qu'à travers de rares textes, d'autres ne nous sont connus qu'indirectement, par des témoignages. Il est probable aussi que nous ignorons tout simplement certains usages, dont nulle trace n'est arrivée jusque nous (cela tient parfois aussi aux éditeurs, dont certains ont tendance à « corriger » comme déviants des termes ou des constructions non déjà recensés). C'est le cas en particulier des usages oraux, dont nous n'avons, pour la plupart des périodes, aucune trace directe : nous ne pouvons nous fonder que sur d'éventuels témoignages ou sur des représentations écrites. Or il est possible que ces dernières tirent les usages réels vers la caricature ou au contraire vers une normalisation sur le modèle de l'écrit. Ainsi, en adoptant une démarche sur corpus, on explicite une difficulté qui serait sinon éludée : la représentativité du corpus, et du même coup l'aptitude de la *grammaire* à être généralisable au-delà des seuls textes sur lesquels elle se fonde.

La représentativité du corpus est à envisager d'un double point de vue, quantitatif et qualitatif. Sur le plan quantitatif, il faut décider si l'on travaille sur des textes intégraux ou sur des échantillons, ou bien en combinant les deux, selon la taille des textes. On peut ainsi décider d'échantillonner les textes lorsqu'ils excèdent un certain nombre de mots. Il convient dans ce cas de fixer un seuil ainsi que les modalités de l'échantillonnage (un seul échantillon ou au contraire plusieurs échantillons répartis dans le texte). Sur le plan qualitatif, les critères qui s'avèrent décisifs dans un projet de grammaire historique, sont les suivants : la date, le domaine ou le genre textuel, le dialecte, la forme du texte. Tous ne sont pas pareillement importants tout au long de la période envisagée, comme on le verra plus bas en présentant les choix qui ont été faits. D'autres sont importants aussi, tels que le registre et la classe sociale, l'âge de l'auteur, la distinction entre région de l'auteur et région du copiste… Mais l'impossibilité d'accéder pour les périodes les plus reculées à, par exemple, des écrits familiers ou émanant des classes sociales peu lettrées ne permet pas d'en faire des critères de sélection systématiques. Par ailleurs, l'ampleur du projet, qui porte sur 12 siècles, nous a obligés à restreindre l'ensemble des critères qui auraient pu être envisagés, pour des raisons pratiques de faisabilité. En effet, à la difficulté de déterminer (en termes de dates, genres, dialectes…) les textes qu'il faudrait idéalement verser au corpus, s'ajoute celle d'y accéder, et en nombre suffisant.

L'objectif était d'obtenir un corpus aussi représentatif et équilibré que possible de l'objet *langue française*, dans toute la diversité qu'on lui présuppose. Plus un corpus est jugé représentatif, plus il est légitime de généraliser les résultats obtenus au-delà des seuls textes qui le constituent. Mais il convenait par ailleurs de constituer un corpus qui reste maniable, non seulement du point de vue de l'exploration des textes (certains faits sont plus faciles à repérer automatiquement que d'autres), mais aussi du traitement des données extraites. Selon les phénomènes étudiés, on a ainsi affaire à quelques occurrences (faits peu fréquents, hapax), ou à quelques centaines, ou bien encore à des dizaines de milliers :

étudier *quant à*, *désormais*, ou l'ordre des mots, les déterminants, n'a évidemment pas les mêmes implications de ce point de vue.

Il a donc fallu trouver un compromis acceptable entre le corpus idéal (que l'on sait inaccessible, mais dont il faut se rapprocher autant que possible), le corpus souhaité, et le corpus possible et raisonnable. La constitution d'un corpus à géométrie variable a permis de résoudre en partie les difficultés liées à la variation des modalités d'exploration des corpus et de traitement des résultats.

3.4 Nos choix, notre démarche

3.4.1 Un corpus à géométrie variable

La *Grande Grammaire Historique du Français* (GGHF) couvre plus de 12 siècles : le rapport du locuteur moderne aux états langagiers successifs n'est pas le même, de même que varient le rôle des textes dans notre accès à la langue, la disponibilité des données textuelles, le rapport des genres entre eux, etc. L'un des défis a consisté, pour élaborer le corpus, à dépasser cette hétérogénéité, et à adopter une démarche aussi homogène que possible à travers les siècles.

La GGHF s'est donné un corpus à géométrie variable, tant du point de vue de sa constitution que de son utilisation. En effet, pour chaque période (voir ci-dessous 3.4.2.2 a. pour la délimitation des périodes), un double corpus a été élaboré : un corpus « noyau » et un corpus « complémentaire ». Le premier répond à des critères de composition stricts quant à la taille des textes et quant à leur diversité.

Pour ce qui est de la taille, nous avons fait le choix de retenir les textes dans leur intégralité lorsqu'ils n'excèdent pas 45 000 « occurences » (mots et ponctuation, soit un peu plus de 40 000 mots). Pour les textes dépassant ce seuil, nous avons sélectionné trois échantillons d'environ 15 000 occurrences en début, milieu, et fin de texte. Toutefois, pour certains textes, jugés répétitifs du point de vue de leurs structures morphosyntaxiques, la taille de l'échantillon a été réduite à 20 000 mots. C'est le cas, par exemple, du *Registre criminel du Chatelet*, au 14^{e} s. Pour chaque période le corpus noyau comprend entre 200 000 et 245 000 mots, hormis pour la période la plus ancienne, avant 1100, pour laquelle la quasi-totalité des textes disponibles a été retenue, l'ensemble ne dépassant pas 10 000 mots. C'est dans le corpus noyau, dont certains textes bénéficient d'un étiquetage morpho-syntaxique, qu'ont prioritairement été effectués les calculs de fréquence. Le corpus noyau échantillonné contient 2055891 mots (et le corpus noyau non échantillonné 9 millions de mots).

Le corpus complémentaire a été conçu plus particulièrement pour l'étude des faits peu fréquents, susceptibles donc d'être peu représentés dans le corpus noyau, pour confirmer, ou non, une hypothèse développée à partir du corpus noyau, et il a de plus fourni un vaste réservoir d'exemples, permettant ainsi de diversifier les sources citées. A l'image des objectifs qui lui ont été fixés, sa constitution n'a pas été soumise aux mêmes contraintes que celle du corpus noyau : la taille des textes n'a pas été limitée, et les autres critères ont été appliqués avec une rigueur moindre. Le corpus complémentaire contient 4571477 mots. Le corpus intégral (corpus noyau non échantillonné et corpus complémentaire) comprend donc 13,5 millions de mots.

Par ailleurs pour certains chapitres et / ou pour certaines périodes, les contributeurs ont parfois fait usage, en plus du corpus noyau, ou même du corpus complémentaire, de corpus

spécifiques. C'est en particulier le cas pour la partie 4, qui porte sur les codes de l'écrit, et pour laquelle il s'est avéré nécessaire de recourir à des textes plus variés ou à des éditions plus sûres, voire aux manuscrits eux-mêmes. C'est aussi le cas, dans une certaine mesure, pour l'analyse des changements à partir du 17e s., des écrits relevant de registres plus familiers étant accessibles à partir de cette période.

Le corpus joue un rôle décisif dans la GGHF, et c'est là l'une des innovations de cet ouvrage. Il ne saurait cependant être question de faire table rase des études qui ont précédé : de nombreux phénomènes linguistiques ont déjà été bien décrits, et il ne s'agit pas de tout réécrire. Nous avons donc exploité plusieurs études, ainsi que les données quantifiées qui les accompagnent, le cas échéant. Ces dernières ont parfois été complétées par de nouveaux relevés, opérés dans notre corpus. Les études inédites, ou partiellement inédites, se sont beaucoup plus largement appuyées sur l'exploitation du corpus. Les modalités de ces relevés ont pu varier, de même que le traitement qui a été fait des données collectées. En effet, comme cela a déjà été souligné, la complexité à collecter des constructions, de quelque nature qu'elles soient, varie fortement selon leur caractère plus ou moins abstrait et selon le degré d'enrichissement morphosyntaxique des textes : il est plus aisé d'établir la fréquence des adverbes en *-ment* que celle des sujets nominaux. Les phénomènes qui relèvent de l'énonciation ou de la textualité, en particulier, se prêtent bien plus difficilement à une quantification des faits concernés. Par ailleurs, les modalités de traitement des données ont pu varier, selon leur nombre, leur analyse exhaustive n'étant pas possible au-delà de certains seuils.

3.4.2 Les critères de sélection des textes

Différents critères ont été retenus pour la sélection des textes. Certains, les *descripteurs*, ont pour but de caractériser le contenu des textes, sous différents aspects. D'autres, d'ordre en quelque sorte « paratextuels », relèvent davantage du point de vue que le locuteur moderne porte sur ces textes.

3.4.2.1 Les critères paratextuels

Il nous a ainsi paru important que le corpus de la GGHF comprenne, pour chaque période, quelques textes de référence, à côté de textes moins connus (et souvent aussi – car les deux sont de fait liés – moins littéraires). La notion de texte de référence peut certes varier, et elle est en partie subjective, mais l'on peut cependant identifier comme tel quelques oeuvres, en particulier pour la période médiévale. Il n'était ainsi pas concevable que le corpus du 12e s., par exemple, ne contienne pas la *Chanson de Roland* et un roman de Chrétien de Troyes ; pour le 13e s., la *Queste du Graal* et le *Roman de la Rose* se sont d'emblée imposés, bien qu'il ne s'agisse pas des seuls textes de référence pour les périodes concernées.

La sélection s'est révélée plus difficile au fur et à mesure que l'on avance dans le temps et que se multiplie la production écrite. Le choix a nécessairement été partial, mais néanmoins influencé par la prise en compte d'un autre paramètre : la qualité des éditions. Pour les textes les plus anciens, cette exigence nous a conduits à privilégier les éditions les moins interventionnistes, et pour la période suivante (du 16e au 19e s.), des textes non (ou très peu) modernisés (l'examen des graphies est un bon indice).

Il est enfin un critère pratique qui est intervenu dans nos choix, conjointement à ceux précédemment mentionnés et aux descripteurs qui vont être évoqués ci-après. Il s'agit de

l'existence d'une version numérisée (disponible) des textes, au moins pour ceux qui appartiennent au corpus noyau et qui ont fait l'objet de quantifications. Pour la période médiévale, nous nous sommes très largement appuyés sur les textes de la *Base de Français Médiéval* (BFM), dont certains sont enrichis linguistiquement (étiquette morpho-syntaxique, et syntaxique pour certains) et deux textes proviennent du corpus *Modéliser le Changement : les Voies du Français* (http://www.voies.uottawa.ca/index.html). Pour la période suivante, nous avons majoritairement sélectionné les textes dans la base Frantext (http://www.frantext.fr/), mais aussi dans la base Epistemon (http://www.bvh.univ-tours.fr/Epistemon/index.asp).

Les textes retenus l'ont été aussi, et prioritairement, parce qu'ils contribuaient à construire le corpus diversifié et représentatif que nous souhaitions, au regard des critères qui nous semblaient les plus pertinents, et qui sont présentés ci-dessous.

3.4.2.2 Les descripteurs

a. *La date des textes*

La GGHF se distingue d'autres ouvrages diachroniques, en particulier l'*Histoire de la Langue française* de F. Brunot, en ce qu'elle est organisée, en premier lieu, non par siècle ou par grande période mais par grands domaines de la langue (phonétique, morphologie, sémantique,…) : c'est au sein de chacune des questions abordées qu'intervient la perspective chronologique. Chaque phénomène a sa propre temporalité : il n'est donc pas possible d'établir un découpage chronologique adapté à l'évolution de l'ensemble des phénomènes. Nous avons déterminé, pour le corpus, un cadre chronologique très général, en délimitant des périodes de manière arbitraire, suivant pour cela un simple découpage par siècles, cette division ne correspondant en aucun cas à une quelconque présupposition quant à la périodisation des évolutions individuelles.

Le corpus a donc été organisé par siècles, en sélectionnant des textes qui s'échelonnent du début à la fin de chaque siècle. La période qui précède le 12^{e} s. fait exception : en raison du petit nombre de documents qui nous sont parvenus, et de leur brièveté, les quelques témoins dont nous disposons ont été regroupés ensemble. Pour eux, la mise en oeuvre des autres critères n'est donc pas pertinente : la *Séquence de Sainte Eulalie* a été retenue non pas parce que c'est un texte en vers qui relève du domaine religieux, mais simplement parce que c'est, avec les *Serments de Strasbourg*, le seul texte en français du 9^{e} s.

Pour chaque siècle envisagé, les textes ont été choisis en fonction de trois critères – forme, domaine et genre, dialecte – en faisant en sorte que l'ensemble composé soit diversifié. On ne peut cependant éviter, pour les périodes reculées, un certain *parasitage* entre les critères, en raison du nombre trop peu élevé de documents, ou simplement de leur absence. Ainsi, jusqu'à la fin du 12^{e} s., les textes qui nous sont parvenus sont très majoritairement en vers, et le dialecte anglo-normand ou normand est particulièrement représenté jusqu'au milieu du siècle ; de même les textes qui relèvent du domaine historique ne se rencontrent guère avant le 13^{e} s.

b. *La forme des textes : vers / prose*

La distinction entre textes en vers et textes en prose recouvre des réalités différentes selon les périodes considérées. Jusqu'au 12^{e} s., la grande majorité des textes s'est écrite en vers (décasyllabes, puis octosyllabes), qu'il s'agisse de récits épiques, de « romans », de récits hagiographiques… Au fil des siècles l'écriture versifiée va reculer, conjointement au développement de la prose à partir du 13^{e} s., pour finalement se voir réservée aux textes de poésie et de théâtre,

ainsi qu'aux chansons (types de textes qui peuvent aussi, surtout depuis le 20e s., être écrits en prose). La place respective faite à la prose et au vers n'est donc pas la même dans le corpus selon les siècles considérés : les textes en vers sont très largement majoritaires jusqu'au 12e s., puis ils cèdent une place croissante aux textes en prose, pour n'être plus associés, à partir du 18e s., qu'à certains genres : le théâtre et la poésie (chansons incluses).

c. *Les dialectes*

Le critère dialectal occupe une position à part parmi les critères retenus. Tout d'abord, il n'est véritablement pertinent que jusqu'au 15e s. environ, et déjà bien moins discriminant à cette époque qu'au 12e s. Par ailleurs il n'est pas toujours facile de définir le dialecte d'un texte, et il est fréquent d'opter pour un dialecte « non défini ». Enfin, et cela résulte en grande partie de la remarque précédente, ce n'est pas sur la base de leur dialecte que nous avons prioritairement sélectionné les textes. Il se trouve néanmoins que les textes retenus présentent, pour les périodes où cette distinction est pertinente, une relative diversité. Sont en particulier bien représentés l'Anglo-normand et le Picard, dont on sait qu'ils présentent plusieurs traits linguistiques spécifiques.

d. *Les domaines et les genres*

Nous nous sommes appuyés, pour déterminer de grands domaines, sur la classification qui a été proposée pour les textes d'ancien et de moyen français par l'équipe de la *Base de Français Médiéval*. Le *domaine* est défini comme un trait fonctionnel qui correspond à la destination principale du texte et au domaine d'activité auquel il se rattache. Dans cette perspective, les domaines retenus sont les suivants :

– littéraire : divertir
– didactico-scientifique : enseigner, instruire
– religieux : édifier (concerne le rituel et la diffusion du message chrétien)
– historique : consigner / relater les événements du passé
– juridique : réguler la vie sociale

Deux autres domaines ont été ajoutés pour les textes à partir du moyen français : il s'agit des domaines épistolaire et argumentatif, qui ne trouvent pas d'instanciation, dans les textes qui nous sont accessibles, avant le 14e s.

Contrairement à la liste des domaines, celle des *genres* est ouverte : roman, nouvelle, mémoire, chronique, lapidaire, traité, hagiographie, miracle, lyrique ... Cela tient principalement au fait que les genres ne sont pas nécessairement les mêmes d'une période à l'autre : certains apparaissent (« mémoires »), d'autres disparaissent (chanson de geste). De plus, les genres n'ont pas tous le même statut : certains sont en effet emblématiques d'une époque (les mémoires et les chroniques au 15e s., les nouvelles au 16e…), tandis que d'autres traversent les siècles, mais en connaissant des transformations radicales. Ainsi, la dénomination « roman » recouvre des réalités bien différentes au 13e s. et au 20e s.

3.4.3 La représentation de l'oral

Nous n'avons pas accès à la réalité matérielle, prosodique, de la langue orale avant le début du 20e s., mais nous pouvons faire l'hypothèse que, hier comme aujourd'hui, ses différents registres et genres (car, comme la langue écrite, la langue orale n'est pas homogène) présentent des spécificités qui les distinguent de ceux de la langue écrite (➔ chap. 37).

Nous n'avons pas non plus tenté de constituer, pour la période moderne, un corpus de données orales. L'oral n'en est pas moins présent dans la GGHF, indirectement. Il l'est tout d'abord à travers l'exploitation que nous avons faite des travaux portant sur l'oral contemporain. Par ailleurs, certains phénomènes et leur évolution ont été étudiés en observant de manière spécifique leur actualisation dans le discours direct (et en comparant la langue de ces épisodes en discours direct avec celle du récit dans lequel ils s'insèrent), et pour certains plus largement dans ce que C. Marchello Nizia a proposé d'appeler l'*oral représenté*. Il s'agit d'une partie des discours directs dans un récit (roman, chanson de geste, chronique, etc.), linguistiquement balisés (annonce, incise..), ce qui permet une comparaison entre la langue du récit enchâssant et celle du discours direct enchâssé.

Sans prétendre rendre compte de ce qu'a pu être réellement le français oral dans les siècles passés, nous avons fait l'hypothèse que, comme aujourd'hui, les réalisations écrites et orales de la langue avaient dû différer, et que les secondes avaient pu être pionnières de certains changements. De fait, cette procédure de comparaison permet de voir que de nombreuses innovations sont apparues d'abord dans les épisodes en discours direct enchâssé, c'est-à-dire en « oral représenté ».

Le corpus sur lequel s'est appuyé la GGHF n'est évidemment pas parfait. Il constitue cependant un compromis raisonnable entre l'exigence de représentativité des données et les contraintes liées à la fois à leur accessibilité et à leur traitement.

On le sait : nous n'accèderons jamais à la représentation des états passés du français dans toute leur diversité (si tant est qu'on puisse prétendre y accéder pour le français moderne). Certains aspects nous en resteront probablement inconnus à jamais, c'est une nécessité inhérente au fait d'étudier des états de langue révolus. Notre démarche, tant par la constitution du corpus que par le traitement qui en a été fait, a cependant tenté de réduire au mieux la part des zones d'ombre.

La constitution de ce corpus n'aurait pas été possible sans l'immense travail réalisé par les responsables de la Base de Français Médiéval (ENS Lyon, anciennement ICAR UMR 5191, désormais IHRIM UMR 5317), plus spécifiquement Céline Guillot-Barbance pour son expertise, le choix des textes de la période médiévale, et Alexei Lavrentiev pour l'échantillonnage puis l'intégration des textes à la plateforme TXM (http://textometrie.ens-lyon.fr/), créée par Serge Heiden, qui a permis leur exploitation. Le projet de la Grande Grammaire Historique du Français leur doit beaucoup, et nous les en remercions très sincèrement.

3.5 Liste des textes du corpus de la GGHF

Le tableau 1 ci-dessous présente la liste des textes du corpus et leurs caractéristiques. Les références complètes des textes sont données dans la bibliographie générale.

Les abréviations suivantes sont utilisées :

- *Forme* : P (prose) ; V (vers) ; M (mixte).
- *Domaine* : A (argumentatif) ; D (didactique) ; E (épistolaire) ; H (historique) ; J (juridique) ; L (littéraire) ; R (religieux).
- *Genre* : corresp. (correspondance) ; dramat. (dramatique) ; chroniq. (chroniques) ; hagiog. (hagiographie)
- *Dialecte* : champ. (champenois) ; angl.norm. (anglo-normand) ; ND (non défini) ;

Les textes dont le nombre de mots est suivi d'un astérisque ont été échantillonnés.

Sigle	Titre	Auteur	Date	For-me	Do-maine	Genre	Dialecte	Nombre de mots
Avant 1100								
Corpus noyau avant 1100 :								9588
Strasbourg	*Serments*		842	P	J	serments	ND	115
Eulalie	*Eulalie*		881	V	R	hagiog.	ND	188
Passion	*Passion de Clermont*		*ca* 1000	V	R	dramat.	ND	2904
StLegier	*Vie saint Léger*		*ca* 1000	V	R	hagiog.	ND	1406
StAlexis	*Vie saint Alexis*		*ca* 1050	V	R	hagiog.	normand	4975
Corpus complémentaire avant 1100 :								
Jonas	*Sermon sur Jonas*		entre 938 et 952	P	R	sermon	Flandre	815
12e siècle								
Corpus noyau 12e s. :								204980
Roland	*Chanson de Roland*		*ca* 1100	V	L	épique	normand	30039
Eneas1	*Eneas* (1)		*ca* 1155	V	L	roman	normand	35152
Beroul Tristan	*Tristan*	Beroul	entre 1165 et 1200	V	L	roman	franco-picard	27708
Pont-StMaxence Becket	*Vie de saint Thomas Becket*	Guernes de Pont Sainte Maxence	1172-1174	V	R	hagiogr.	ouest	39145*
TroyesYvain	*Yvain*	Chrétien de Troyes	1177-1181	V	L	roman	champ.	42331
Lapidaire	*Lapidaire en prose*		mi.-12e	P	D	lapidaire	angl. norm.	4781
AmiAmil	*Ami et Amile*		*ca* 1200	V	L	épique	ND	25824
Corpus complémentaire 12e s. :								160918
Benedeit Brendan	*Voyage de St Brendan*	Benedeit	déb. 12e	V	R	hagiogr.	angl. norm.	10955
Thaon Comput	*Comput*	Philippe de Thaon	1113 ou 1119	V	D	comput	angl. norm.	14678
Descri Engleterre	*Descripion d'Engleterre*		peu ap. 1139	V	H	histoire	angl. norm.	1303
Psaut Cambridge	*Psautier de Cambridge*		entre 1155 et 1160	P	R	psautier	angl. norm.	4312
Eneas2	*Eneas* (2)		*ca* 1155	V	L	roman	angl. norm.	24965
WaceBrut2	*Brut*	Wace	achevé en 1155	V	H	chroniq.	angl. norm.	15675
Adgar Miracles	*Collection de Miracles*	Adgar	3e tiers du 12e	V	R	miracles	angl. norm.	49330
SteMaure Chron Normandie	*Chronique des ducs de Normandie*	Benoît de Sainte Maure	1174	V	H	chroniq.	poitevin	25285
Béthune Chansons	*Chansons*	Conon de Béthune	*ca* 1180-1190	V	L	lyrique	picard	2687
Charte Chièvres	*Charte de Chièvres*		1194	P	J	charte	traits picards	1282
Bodel Nicolas	*Jeu de Saint Nicolas*	Jehan Bodel	entre 1191 et 1202	V	L	dramat.	artois	10446

Sigle	Titre	Auteur	Date	For-me	Do-maine	Genre	Dialecte	Nombre de mots
13e siècle								
Corpus noyau 13e s. :								223 298
ClariConstan-tinople	*Conquête de Constantinople*	Robert de Clari	ap. 1205	P	H	chroniq.	picard	34 292
RenartDole	*Roman de la rose ou de Guillaume de Dole*	Jean Renart	1210 ou 1228	V	L	roman	picard	35 050
Aucassin	*Aucassin et Nicolete*		dernier ¼ 12e ou 1ère moit. 13e	M	L	récits brefs	traits picards	10 078
CoinciMira-cles1/2 3/ 4	*Miracles de Nostre Dame*	Gautier de Coinci	1218-1227	V	R	miracles	picard	37 789*
Graal	*Queste del Saint Graal*		*ca* 1225	P	L	roman	ND	41 462*
LorrisRose	*Roman de la Rose*	Guillaume de Lorris	entre 1225 et 1230	V	D	roman	orléanais	24 325
MeunRose 1/2/3	*Roman de la Rose*	Jean de Meun	entre 1269 et 1278	V	D	roman	ND	19 563*
Beaumanoir Beauvaisis	*Coutusme de Beauvaisis*	Philippe de Beaumanoir	1283	P	J	traité	traits picards	20 739*
Corpus complémentaire 13e s. :								166 046
Renart10/11	*Roman de Renart* branches X-XI		déb.13e	V	L	récits brefs	ND	22 300
TristanProse	*Tristan en prose*		ap. 1240	P	L	roman	picard	75 186
MenestReims	*Récit d'un Ménestrel de Reims*		*ca* 1260	P	H	chroniq.	ND	50 046
CharteParis	*Chartes de la région parisienne*		1250 (1200-1299)	P	J	charte	Ile de France	18 514
14e siècle								
Corpus noyau 14e s.								217 486
Joinville Mémoires	*Mémoires ou Vie de saint Louis*	Jean de Joinville	entre 1305 et 1309	P	H	mémoires	champ.	40 707*
Machaut Fortune	*Remede de Fortune*	Guillaume de Machaut	1341	V	L	lyrique	champ.	25 265
Froissart Chroniques	*Chroniques*	Jean Froissart	entre 1369 et 1400	P	H	chroniq.	picard	40 512*
Registre Chatelet1	*Registre criminel du Chatelet*		1389	P	J	procès	Ile de France	19 623*
Mesnagier	*Mesnagier de Paris*		1393	P	D	manuel	ND	19 993*
Griseldis	*Estoire de Griseldis en rimes et par personnages*		1395	V	L	dramatiq.	traits picards	16 249
Manières 1396/1399	*Manières de langage*		1396, 1399	M	D	manuel	angl. norm.	20 315

Sigle	Titre	Auteur	Date	For-me	Do-maine	Genre	Dialecte	Nombre de mots
QuinzeJoies	*Quinze Joyes de Mariage*		*ca* 1400	P	L	nouvelles	ouest	34 822
Corpus complémentaire 14e s. :								555 328
Bersuire Décades1/9	*Les Décades de Titus Livius*, I,1 et I, 9	Pierre Bersuire	1354	P	H	histoire	ND	71 667
Berinus1/2	*Berinus*		*ca* 1370	P	L	roman	Ile de France	150 628
Oreste Aristote Commentaire	*le Livre de Ethiques d'Aristote, Commentaire*	Nicole Oresme	1370	P	A	traité	ND	124 719
Phoebus Chasse	*Le Livre de chasse*	Gaston Phebus	1387	P	D	traité	picard	77 798
Arras Mélusine	*Melusine*	Jean d'Arras	1392	P	L	roman	ND	124 929
Deschamps ArtDictier	*L'art de dictier*	Eustache Deschamp	1392	M	D	traité	champ.	5587
15e siècle								
Corpus noyau 15e s. :								201 214
Gerson Sermon	*Sermon pour le Fete de la Sainte trinité*	Jean Gerson	1402	P	R	sermon	ND	6915
PizanCité	*Le livre de la Cité des dames*	Christine de Pizan	entre 1404 et 1405	P	D	exemples	Ile de France	40 734*
Orléans Ballades	*Ballades*	Charles d'Orléans	1415	V	L	lyrique	orléanais	22 251
Manières1415	*Manières de langage*		1415	M	D	manuel	angl. norm.	3156
Pathelin	*Farce de Maitre Pathelin*		1456-1469	V	L	dramatiq.	Ile de France	10 752
Cent Nouvelles	*Cent nouvelles nouvelles*		1456-1467	P	L	nouvelles	picard	39 449*
LouisXI Lettre223	*Lettres de Louis XI*	Louis XI	1461-1465	P	E	corresp.	Ile de France	2362
LouisXI Lettre234	*Lettres de Louis XI*	Louis XI	1465-1469	P	E	corresp.	Ile de France	3143
Archier Baignollet	*Le franc Archier de Baignollet*		1468	V	L	comique	Ile de France	2500
LouisXI Lettre248	*Lettres de Louis XI*	Louis XI	1469-1472	P	E	corresp.	Ile de France	3419
Commynes Mémoires	*Mémoires*	Philippe de Commynes	*ca* 1490-1505	P	H	mémoires	ouest	40 435*
JehanParis	*Roman de Jehan de Paris*		1494	P	L	roman	Ile de France	26 098
Corpus complémentaire 15e s. :								364 247
SaleSaintré	*Jehan de Saintré*	Antoine de la Sale	1456	P	L	roman	Ile de France	92 056
Bueil Jouvencel1/2	*Le Jouvencel*	Jean de Bueil	1461	P	D	roman	ND	123 452

Sigle	Titre	Auteur	Date	For-me	Do-maine	Genre	Dialecte	Nombre de mots
Villon Testament	*Testament*	François Villon	1461	V	L	lyrique	Ile de France	13 378
Commynes Lettres	*Lettres*	Philippe de Commynes	1478-1511	P	E	corresp.	ouest	20 019
Phares Astrologues	*Recueil des plus celebres astrologues et quelques hommes doctes*	Simon de Phares	1494-1498	P	D	traité	ND	115 342
16e siècle								
Corpus noyau 16e s. :								245 202
Vigneulles Nouvelles	*Cent Nouvelles Nouvelles*	Philippe de Vigneulles	1515	P	L	nouvelles	standard	40 321*
Calvin Lettres	*Lettres à monsieur et madame de Falais*	Jean Calvin	1549 (1543-1554)	P	E	corresp.	standard	33 735
DuBellay Défense	*Defense et Illustration de la langue française*	Joachim du Bellay	1549	P	A	traité	standard	19 691
DuBellay Olive	*L'Olive*	Joachim du Bellay	1550	V	L	lyrique	standard	21 214
Ronsard Misères	*Discours des misères de ce temps*	Pierre de Ronsard	1563	V	L	lyrique	standard	5934
LaTailleSaül	*Saül le furieux*	Jean de la Taille	1572	V	L	dramatiq.	standard	15 416
LEstoile Registre1/2/3/4/5	*Registre-journal du regne de Henri III* (t.1-5)	Pierre de l'Estoile	1574-75 1576-78 1585-87	P	H	registre	standard	28 842*
LéryBrésil	*Histoire d'un voyage fait en la terre du Brésil*	Jean de Léry	1578	P	L	récit de voyage	standard	40 278
Montaigne Essais	*Essais*	Michel de Montaigne	1592	P	A	traité	standard	39 771*
Corpus complémentaire 16e s. :								578 220
Consistoire Genève	*Minutes du consistoire de Genève*		1542	P	J	procès	standard	136 182
DesPériers Récréations	*Nouvelles récréations et joyeux devis*	Bonaventure des Périers	1561	P	L	nouvelles	standard	75 543
Palissy Recepte	*Recepte veritable*	Bernard Palissy	1563	P	D	traité	standard	58 418
Etienne Agriculture	*L'Agriculture et maison rustique*	Charles Estienne	1564	P	D	traité	standard	120 199
Vigenère Décadence	*L'Histoire de la decadence de l'Empire grec*	Blaise de Vigenère	1577	P	H	traité	standard	187 878
17e siècle								
Corpus noyau 17e s. :								233 009
UrféAstrée	*L'Astrée*, 2nde partie	Honoré d'Urfé	1610	P	L	roman	standard	39 939*
Beroalde Parvenir	*Le moyen de parvenir*	François Béroalde de Verville	1616	P	L	roman	standard	39 326*

Sigle	Titre	Auteur	Date	Forme	Domaine	Genre	Dialecte	Nombre de mots
SorelBerger	*Le berger extravagant*	Charles Sorel	1627	P	L	roman	standard	20308*
CorneilleCid	*Le Cid*	Pierre Corneille	1637	V	L	dramatiq.	standard	18160
Descartes Discours	*Discours de la méthode*	René Descartes	1637	P	A	traité	standard	23142
Coeffeteau Histoire	*Histoire romaine*	Nicolas Coëffeteau	1646	P	H	traité	standard	20302*
Assoucy Poësies	*Poësies et lettres … contenant diverses pièces héroïques, satiriques et burlesque*	Charles Coypeau d'Assoucy	1653	V	L	lyrique	standard	28855
Molière Précieuses	*Les Précieuses ridicules*	Molière	1660	V	L	dramatiq.	standard	7127
Rabutin Lettres2/3/1/4	*Les Lettres de messire Roger de Rabutin, comte de Bussy* (t. 1-4)	Roger de Bussy-Rabutin	1672 1681 1686 1692	P	E	corresp.	standard	20019*
Racine Athalie	*Athalie*	Jean Racine	1691	V	L	dramatiq.	standard	15831
Corpus complémentaire 17e s. :								928801
Serres Agriculture1 Serres Agriculture2	*Le Théâtre d'agriculture et mesnage des champs*	Olivier de Serres	1603	P	D	traité	standard	518641
Gerhard Heroard	*Journal d'Heroard*	Jean Héroard	1601-1610	P	H	journal	standard	195930
Lafayette Clèves	*La Princesse de Clèves*	Madame de La Fayette	1678	P	L	roman	standard	65255
Bossuet Discours	*Discours sur l'histoire universelle*	Jacques-Bénigne Bossuet	1681	P	H	traité	standard	140524
Fléchier Oraison	*Oraison fuèbre de Marie-Thérèse d'Autriche reine de France*	Valentin-Esprit Fléchier	1691	P	R	oraison	standard	8451
18e siècle								
Corpus noyau 18e s. :								244966
Regnard Légataire	*Le légataire universel*	Jean-François Regnard	1708	V	L	dramatiq.	standard	17874
Montesquieu Lois	*L'Esprit des lois*	Montesquieu	1755	P	A	traité	standard	19922*
Voltaire Essay	*Essay sur l'histoire générale et sur les moeurs et sur l'esprit des nations*	Voltaire	1756	P	A	traité	standard	39938*
Prévost Mémoires	*Le Monde moral ou Mémoires pour servir à l'histoire du coeur humain*	L'Abbé Prévost	1760	P	L	roman	standard	26455*

Sigle	Titre	Auteur	Date	For-me	Do-maine	Genre	Dialecte	Nombre de mots
RetifBretonne Paysan	*Le Paysan perverti ou les Dangers de la ville*	Nicolas Rétif de La Bretonne	1776	P	L	roman	standard	19 183*
Mirabeau Lettres	*Lettres originales écrites du donjon de Vincennes pendant les années 1777, 1778, 1779, 1780*	comte Honoré de Mirabeau	1780	P	E	corresp.	standard	19 546*
Beaumarchais Figaro	*La folle journée ou le Mariage de Figaro*	Beaumarchais	1785	V	L	dramatiq.	standard	34 722
Robespierre Discours	*Discours*	Maximilien de Robespierre	1793	P	A	discours	standard	39 832*
Delille Géorgiques	*L'homme des champs ou les Géorgiques françaises*	Jacques Delille	1800	V	L	lyrique	standard	27 494
Corpus complémentaire 18e s. :								333 748
VaubanDixme	*Projet d'une dixme royale qui, supprimant la taille, les aydes, les doüanes d'une province à l'autre, les décimes du Clergé, les affaires extraordinaires...produiroit au Roy un revenu certain et suffisant*	Sébastien de Vauban	1707	P	D	traité	standard	39 737
Ramsay Cyrus	*Les voyages de Cyrus*	André-Michel de Ramsay	1727	P	L	roman	standard	66 578
Condillac Essai	*Essai sur l'origine des connaissances humaines*	Étienne Bonnot de Condillac	1746	P	D	traité	standard	83 684
Rousseau Discours	*Discours sur les sciences et les arts*	Jean-Jacques Rousseau	1750	P	A	traité	standard	9015
Diderot Essais	*Essais sur la peinture / Salon de 1759 / Salon de 1761 / Salon de 1763*	Denis Diderot	1759-1766	P	A	traité	standard	69 310
Buffon Epoques	*Des époques de la nature*	Georges-Louis de Buffon	1778	P	D	traité	standard	65 424
19e siècle								
Corpus noyau 19e s. :								235 688
CodeCivil	*Le Code civil des Français*		1804	P	D	traité	standard	19 692*
Château-briandGénie	*Le Génie du christianisme*	François-René de Châteaubriand	1803	P	H	traité	standard	19 596*
Musset Articles	Articles publiés dans la *Revue des deux mondes*	Alfred de Musset	1832	P	L	presse	standard	43 347

Sigle	Titre	Auteur	Date	For-me	Do-maine	Genre	Dialecte	Nombre de mots
FlaubertCor-respondance	*Correspondance (1830-1839)*	Gustave Flaubert	1839	P	E	corresp.	standard	16 888
Hugo Hernani	*Hernani*	Victor Hugo	1841	V	L	dramatiq.	standard	27 048
DuCampNil	*Le Nil, Egypte et Nubie*	Maxime Du Camp	1854	P	L	récit de voyage	standard	37 135
Baudelaire Fleurs	*Les Fleurs du Mal,*	Charles Baudelaire	1861	V	L	lyrique	standard	26 491
Goncourt Journal2/3/4	*Journal : Mémoires de la vie littéraire* (t. 2,3,4)	Edmond et Jules de Goncourt	1870 / 1890 / 1896	P	L	mémoires	standard	19 286*
Clemenceau Iniquité Clemenceau Réparation	*L'iniquité* et *Vers la réparation*	Georges Clémenceau	1899	P	L	mémoires	standard	26 205*
Corpus complémentaire 19e s. :								706 196
Procès TribunalSeine	*Procès instruit par le tribunal criminel du département de la Seine*		1801	P	J	procès	standard	22 873
StaelCorinne	*Corinne ou l'Italie*	Germaine de Staël	1807	P	L	roman	standard	192 672
Scribe Mariage	*Le mariage de raison*	Eugène Scribe	1826	P	L	dramatiq.	standard	16 900
Béranger Chansons	*Chansons*	Pierre-Jean de Béranger	1829	V	L	chanson	standard	22 898
Barbey Memorandum 1/2/3/4	*Memorandum* 1, 2, 3, 4,	Jules Barbey d'Aurevilly	1838-1858	P	L	mémoires	standard	138 865
DuCamp Hollande	*En Hollande, Lettres à un ami*	Maxime Du Camp	1859	P	E	corresp.	standard	53 750
Guyot Rapport	*Rapport de l'état de l'agriculture en Lorraine : 1789-1889*	Charles Guyot	1889	P	D	traité	standard	11 499
Claudel TêteOr	*Tête d'or*	Paul Claudel	1890	P	L	dramatiq.	standard	40 933
ZolaDébâcle	*La débâcle*	Emile Zola	1892	P	L	roman	standard	205 806
20e siècle								
Corpus noyau 20e s. :								239 897
Rolland JChristophe	*Jean-Christophe I : L'Aube*	Romain Rolland	1904	P	L	roman	standard	30 008*
VidalBlache Tableau	*Tableau de la géographie de la France*	Paul Vidal de la Blache	1908	P	D	traité	standard	26 679*
Apollinaire Alcools	*Alcools*	Guillaume Apollinaire	1913	V	L	lyrique	standard	17 121
Feydeau Maxim	*La dame de chez Maxim's*	Georges Feydeau	1914	P	L	dramatiq.	standard	32 136
Alain BeauxArts	*Système des Beaux-Arts*	Alain	1920	P	D	traité	standard	26 483*

Sigle	Titre	Auteur	Date	For-me	Do-maine	Genre	Dialecte	Nombre de mots
MaletVie	*La vie est dégueu-lasse*	Léo Malet	1948	P	L	roman	standard	29 451*
Green Journal 1/2/4/5	*Journal* (t. 1, 2, 4, 5)	Julien Green	1934 1939 1946 1950	P	L	journal	standard	29 258*
Sartre Lettres1 Sartre Lettres2	*Lettres au castor et à quelques autres* (vol.1, 2)	Jean-Paul Sartre	1932 (1926-39) 1951 (1940-63)	P	E	corresp.	standard	19 777*
Perec ModeEmploi	*La vie mode d'emploi*	Georges Perec	1978	P	L	roman	standard	28 984*
Corpus complémentaire 20^{e} s. :								777 158
Poincaré Electrtricité PoincaréVa-leurScience Poincaré Mécanique Poincaré Leçons	*5 traités*	Henri Poincaré	1901-1911	P	D	traité	standard	202 654
Fournier Correspon-dance	*Correspondance avec Jacques Rivière*	Alain-Fournier	1905-1914	P	E	corresp.	standard	367 117
GideCongo	*Voyage au Congo*	André Gide	1927	P	L	récit de voyage	standard	60 870
DabitHôtel	*L'Hôtel du Nord*	Eugène Dabit	1929	P	L	roman	standard	39 758
Césaire Discours	*Discours sur le colonialisme*	Aimé Césaire	1955	P	A	traité	standard	12 873
Zitrone Courses	*Léon Zitrone vous emmène aux courses*	Léon Zitrone	1962	P	D	traité	standard	7094
Jaccottet Chants	*Chants d'en-bas*	Philippe Jaccottet	1977	V	L	lyrique	standard	2263
Koltès Solitude	*Dans la solitude des champs de coton*	Bernard-Marie Koltès	1986	P	L	dramatiq.	standard	10 820
IzzoKhéops	*Total Khéops*	Jean-Claude Izzo	1995	P	L	roman	standard	73 709

Tableau 1 : Liste des textes des corpus noyau (échantillonné) et complémentaire de le GGHF

Sophie Prévost