La recherche linguistique à partir de l'exploitation des banques de données électroniques : élaboration des dictionnaires et concordanciers généraux ou spécialisés

Javier Sanchez, Christian Boely, Christine Bonneville, Philippe Galiana, Slimane Zamoun

La recherche linguistique à partir de l'exploitation des banques de données électroniques : élaboration des dictionnaires et concordanciers généraux ou spécialisés

Javier Sanchez et Collaborateurs [1]

1. Introduction

En tant que linguistes, quand il s'agit d'analyser le texte de grande dimension, nous nous confrontons inévitablement aux moyens d'analyse que notre discipline centrait traditionnellement autour de l'étude de l'énoncé. C'est ainsi que, en voulant accéder aux grands corpus, nous avons intégré depuis une quinzaine d'années [2] les nouvelles technologies dans nos recherches avec l'élaboration d'un protocole de recherche permettant la description et l'analyse des formes et des typologies des discours. C'est ainsi qu'il est actuellement possible d'étudier et de comparer des corpus de nature et d'objectif différents tels que : la langue standard, la langue littéraire ou encore très particulièrement la langue scientifique qui constituent une orientation multilingue des nos recherches (mémoires de traduction et de terminologie trilingue). Afin de donner un aperçu de notre méthodologie d'exploration contextuelle, nous l'illustrerons à partir de l'élaboration d'une banque de données électronique en langue espagnole sur les douze « Nouvelles Exemplaires » de Miguel Cervantes. Ce corpus, mettant en évidence la complexité syntaxique de l'espagnol, constitue notre première référence mais d'autres recherches sont en cours sur l'étude comparée de la langue orale/écrite de l'espagnol, l'élaboration des mémoires de traduction et terminologies spécialisées quadrilingues (français/espagnol/anglais/allemand) sur corpus scientifiques et économiques.

2. Constitution des Banques de données électroniques

La première application sur la constitution d'une banque de données électroniques a été réalisée sur la langue Cervantine étant donné l'intérêt que celle-ci apporte à la recherche sur la syntaxe et la typologie des discours. Ce travail nous a permis de mettre en place une méthodologie d'analyse textuelle [3] applicable actuellement sur corpus spécialisés dont nous axons surtout les applications dans le domaine scientifique étant donné le site privilégié dans ce domaine à l'Université d'Évry. Ce type de recherche s'adressant en particulier aux chercheurs souhaitant explorer le lexique des grands corpus ou élaborer des terminologies et des dictionnaires spécialisés, est de plus en plus répandu

grâce à la facilité de stockage et d'accès aux données particulièrement grâce à Internet. Ainsi, la création d'un fichier électronique permet l'interrogation et la création de listes de formes, globales ou sélectives, permettant de parcourir rapidement le vocabulaire sous ses différents critères de classification. L'exemple que nous citons ici sur la création d'une Banque de données sur les oeuvres de Miguel de Cervantes n'est qu'un exemple parmi d'autres dont nous allons également proposer des résultats. Nous parlerons ici des « Nouvelles Exemplaires » de Cervantès qui ont bénéficié du codage lexicométrique permettant l'exploitation structurelle et syntaxique des formes. Quant aux textes scientifiques que nous étudions, ils sont constitués de documents trilingues qui sont étudiés en parallèle afin de proposer des traductions et des terminologies multilingues attestées par les spécialistes du domaine. Nous ne proposerons ici qu'une illustration de l'étude structurelle et syntaxiques des corpus à partir de la Banque électronique Cervantine. Les recherches dans le domaine du discour scientifique donneront lieu à d'autres présentations pointues.

2.1. Protocole de recherche

La méthodologie utilisée pour la création des banques de données électroniques s'appuie donc sur les évolutions technologiques dans différents domaines d'application de la micro-informatique :

1. L'accès aux corpus à partir d'Internet ou la numérisation et la reconnaissance optique des caractères [4] : Internet constitue pour le chercheur une énorme source d'informations spécialisée en particulier sur les sites proposés dans le domaine scientifique. La saisie automatique des documents papiers est réservée exclusivement aux documents n'existant que sur version papier. Ainsi par exemple, les banques de données cervantines ou encore des documents issus d'archives spécialisées seront traités avec les logiciels de reconnaissance automatique des caractères. Le but est donc de construire un corpus électronique en optimisant la vitesse d'enregistrement de ce type de documents. Ainsi par exemple, nous collaborons avec le Laboratoire de recherche d'Histoire de l'Université d'Évry qui s'intéresse aux textes techniques et financiers étrangers (allemands, anglais et espagnols). Nous procédons à la traduction et à l'élaboration de terminologies permettant à l'équipe d'historiens de pouvoir étudier les corpus car ils ont été préalablement traduits en français.

2. L'utilisation des logiciels de traitement lexicométrique des corpus [5] et d'élaboration des mémoires de traduction et de terminologie : les logiciels lexicométriques permettent, à partir des Codifications Éditoriale et Structurelle de réaliser les différents traitements lexicométriques : génération automatique des listes de mots (formes) existant dans les corpus. Il est possible de construire plusieurs types de listes :

alphabétiques (dans l'ordre de l'alphabet) ou hiérarchiques (par ordre de fréquence), en contexte (concordances) ou hors contexte (index). Les index et les concordances peuvent faire référence à la globalité du corpus (globaux) ou à une partie seulement (sélectifs). Les logiciels de traduction et terminologie (Trados par exemple) sont le complément indispensable pour la gestion des traductions et des termes constitutifs du corpus spécialisé. Ils permettent également de créer automatiquement une sortie des données analysées afin de les présenter aux chercheurs linguistes ou du domaine.

3. Actuellement, le traitement de texte Word est l'outil de base utilisé pour l'importation des données issues d'Internet ou la correction des erreurs générées par la lecture automatique, l'insertion de la codification éditoriale ou structurelle ou le travail de traduction et de terminologie. En effet, le logiciel Trados s'intègre lors de l'installation dans un menu spécifique de Word permettant ainsi le travail de recherche. La base de données Multiterm intégrée à Trados permet le travail terminologique et l'insertion de toutes les informations dont le chercheur a besoin pour élaborer les glossaires de spécialité.

2.2. Codification &Eeacute;dito-Structurelle des corpus

Nous avons vu que grâce aux progrès de la micro-informatique, nous disposons d'une série d'outils capables de nous aider à étudier les textes de très grande dimension. Mais l'exploitation linguistique des textes suppose au préalable l'insertion dans le corpus électronique d'une codification représentant les informations extralinguistiques et intra-linguistiques liées à chaque texte, soit respectivement les marques éditoriales et la différenciation des types de discours. Cela nous permet d'atteindre les différentes contextualités des formes que nous appellerons ici : le contexte éditorial, le contexte structurel et le contexte syntaxique.

Le contexte éditorial [6] tient compte de l'édition source des corpus à partir de ses différentes marques : auteur(s), oeuvre(s), volumes, parties, chapitres, paragraphes, pages et lignes. Elles ont été indispensables pour la construction de la banque de données Cervantine que nous utilisons ici comme exemple. Ainsi chaque forme est accompagnée des références qui identifient à quelle oeuvre, page et ligne elle appartient.

En ce qui concerne le contexte structurel il met en évidence non seulement la syntaxe des formes ou contexte syntaxique, mais également la syntaxe des différents blocs textuels, en permettant d'examiner les formes du point de vue général et du point de vue sélectif. Les blocs textuels les plus rencontrés dans le corpus cervantin sont, pour l'interlocution : les rôles individuels et collectifs ; pour la narration : le récit syntaxiquement dominant (Rdo) et le récit syntaxiquement dépendant (Rdp). La dénomination dominant//dépendant fait

référence à la relation syntaxique qu'a le bloc narratif par rapport au bloc interlocutif. Ainsi, en effet, quand le récit a comme objet direct du verbe élocutif introductif le discours direct des personnages, il s'agit d'un récit syntaxiquement dominant. Dans l'exemple suivant :

> *« (...) comenzó a templar su guitarra, y sintió que el negro estaba ya atento, y llegándose al quicio de la puerta, con voz baja dijo :*
> *- ¿ Será posible, Luis, darme un poco de agua, que padezco de sed y no puedo cantar ? »* ;

la séquence « ¿ Será posible, Luis, darme un poco de agua, que padezco de sed y no puedo cantar ? », même si elle est structurellement autonome, est syntaxiquement dépendante du récit qui l'introduit (« dijo : ... »). Le récit syntaxiquement dépendant présente une configuration différente car le verbe qui introduit le discours direct est inséré et parenthésé entre deux ponctuations de la façon suivante :

> « -No -*dijo el negro*-, porque no tengo la llave desta puerta, ni hay agujero por donde pueda dárosla. »

Dans ce exemple, le discours direct du personnage est dominant, par rapport au récit qui lui est syntaxiquement dépendant : « -dijo el negro- ». Nous voyons comment le discours direct peut être dominant ou dépendant selon le type de syntaxe qui le lie à la narration, ce qui montre l'utilité du contexte structurel.

Enfin, sans nous attarder, nous rappellerons que nous avons également défini le contexte situationnel [7] qui constitue des sous-jalons du contexte structurel et qui vise à améliorer l'étude de l'interlocution avec la définition du personnage source et du personnage destinataire, ainsi que celle de la reprise du discours, permettant d'aller au-delà des possibilités offertes par le contexte structurel. À titre d'exemple, le contexte situationnel offre la possibilité d'étudier le lexique d'un personnage X qui s'adresse à un personnage Y, alors que le contexte structurel ne tient compte que du vocabulaire du personnage X tout interlocuteur confondu.

Le système de clés se combine avec des références textuelles et des symboles délimiteurs de codification selon le principe suivant :

1. un crochet d'ouverture < qui marque le début d'une codification,
2. une clé représentée par une lettre qui permet de hiérarchiser la codification,
3. un blanc séparateur pour indiquer la fin de la clé,
4. une chaîne de caractères qui représente la référence textuelle associée à chaque forme dans les Index ou Concordances (généralement six caractères au maximum),

5. un crochet de fermeture > qui marque la fin de la codification (retour au texte).

Par exemple dans le codage suivant : <C Rdo>, *C* est la clé qui permet de hiérarchiser le codage et *Rdo* est la référence textuelle qui définit le Récit Syntaxiquement Dominant.

Nous illustrons ce principe par un échantillon de codification issu de la nouvelle « El Celoso Extremeño » :

Récit dominant <C Rdo>, Page 111 <P Pg111> et ligne 1 <L 1>	ï	<C Rdo><P Pg111><L 1> Llegándose una noche, como solía, a la puerta, comenzó a templar su guitarra, y sintió que el negro estaba ya atento, y llegándose al quicio de la puerta, con voz baja, dijo :
Personage Loaysa <C Loay>	ï	<C Loay> ¿ Será posible, Luis, darme un poco de agua que padezco de sed y no puedo cantar ?
Personage Luis <C Luis>	ï	<C Luis>
Récit dépendant <C Rdp> et tengo la retour à Luis <C Luis>	ï	- No <C Rdp> -dijo el negro-,<C Luis> porque no llave desta puerta, ni hay agujero por donde pueda dárosla. (...)

Toutes ces informations sont intégrées à la demande du chercheur dans les concordances qui réaliseront l'exploration syntaxico-structurelle car cet index contextuel comporte les références textuelles (page, ligne, type de discours, etc.), la forme en étude (ou forme pôle) et son environnement contextuel.

3. Étude typologique des discours et des formes

À partir de l'interrogation de la banque, nous allons présenter sommairement quelques exemples d'application sur la typologie du discours et des formes cervantines. Nous commencerons par une illustration des possibilités offertes par cette banque en matière de statistiques textuelles globales ou sélectives qui constituent les premières observations objectives obtenues par la méthodologie et qui permettent d'orienter en amont la recherche linguistique.

Ensuite nous présenterons les grandes lignes dégagées par l'exploration structurelle et syntaxique du corpus en ce qui concerne la typologie des formes démonstratives dans les différents types de discours cervantins.

3.1. Statistiques globales et sélectives

L'interrogation de la Banque nous révèle que cet ensemble d'oeuvres comporte 15 504 formes et 179 944 occurrences (autrement dit qu'il existe 15 504 mots différents qui se répètent 179 944 fois). Les index sélectifs quant à eux ont permis de définir la dimension de chacune des douze Nouvelles afin de calculer leur temps de parole. Nous avons obtenu le tableau suivant :

Nom de la Nouvelle	Formes	Occurrences
La Gitanilla	4 416	23 472
El Amante Liberal	3 361	18 519
Rinconete y Cortadillo	3 032	13 682
La Española Inglesa	3 044	16 036
El Licenciado Vidriera	2 556	9 494
La Fuerza de la Sangre	1 916	7 687
El Celoso Extremeño	2 830	13 684
La Ilustre Fregona	3 813	20 026
Las dos Doncellas	2 908	14 943
La Señora Cornelia	2 645	14 340
El Casamiento Engañoso	1 441	5 027
Coloquio de los Perros	4 656	23 034
TOTAL GÉNÉRAL :	15 504	179 944
MOYENNE	1 294	14 995

Ce tableau nous montre que les « Nouvelles Exemplaires » sont de dimension très variable. Par exemple, « La Gitanilla » (23 472 occ.) est quatre fois et demie supérieure à « El Casamiento Engañoso » (5 027 occ.), alors que « El Celoso Extremeño » (13 684 occ.) et « Rinconete y Cortadillo » (13 682 occ.) ont, à deux occurrences près, la même dimension.

Pour mieux comparer les dimensions des Nouvelles, nous avons créé une représentation graphique des données brutes accompagnées des pourcentages correspondants. Nous obtenons ainsi le graphique suivant :

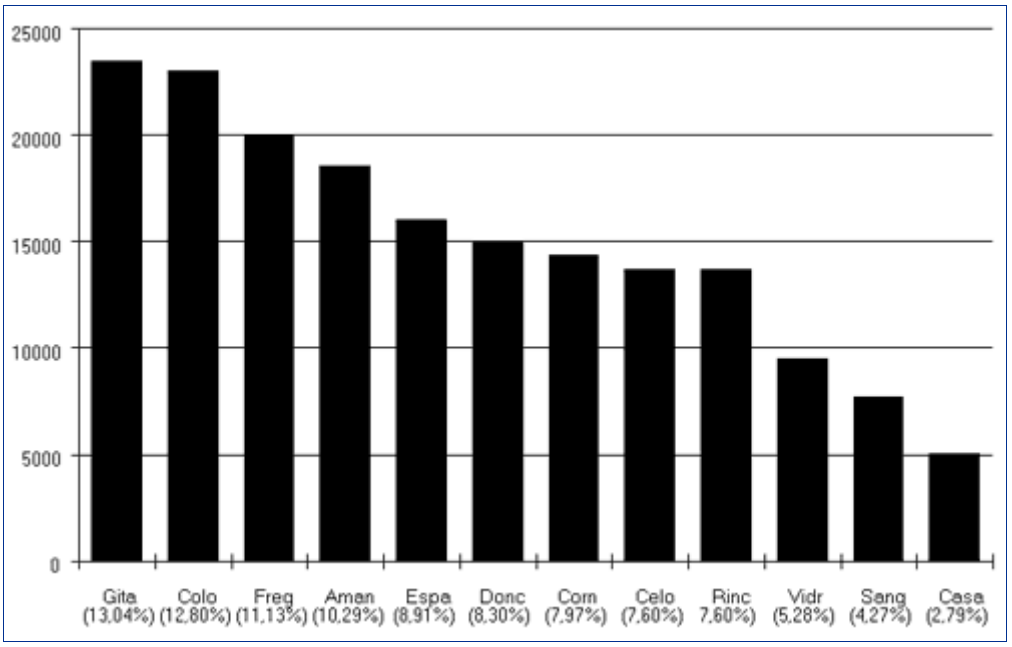

Figure 1 : Classement hiérarchique de la dimension des Nouvelles Exemplaires.

Comme l'indiquait le tableau précédent, le nombre moyen d'occurrences du corpus est de 14 995, ce qui correspond à un temps de parole moyen de 8,33 %. En fonction de cette moyenne, et au vu du graphique ci-dessus, nous pouvons répartir les douze « Nouvelles Exemplaires » en trois catégories :

1. un premier groupe, au-dessus de la moyenne des occurrences, est composé par « La Gitanilla », « El Coloquio de los Perros », « La Ilustre Fregona » et « El Amante Liberal » (entre 10 % et 13 %) et représente les nouvelles ayant les dimensions les plus importantes. Ces quatre nouvelles totalisent en effet presque la moitié de l'ensemble du corpus (85051 occurrences, soit 47,26 % des 179 944 occurrences recensées par l'index global).

2. un deuxième groupe, autour de 8,33 %, est constitué par « La Española Inglesa », « Las dos Doncellas », « La Señora Cornelia », « El Celoso Extremeño » et « Rinconete y Cortadillo », situées entre 7,60 % et 8,91 %. Ce sont des nouvelles de dimension moyenne.

3. enfin, nous pouvons distinguer un troisième ensemble qui regroupe « El Licenciado Vidriera », « La Fuerza de la Sangre » et « El Casamiento Engañoso » correspondant aux nouvelles de petite dimension (entre 2,79 % et 5,28 %).

Le nombre des formes et des occurrences de chacune des nouvelles nous permet de connaître la répétition moyenne d'une forme dans une nouvelle, ce qui nous fournit un indice quant à la richesse lexicale. En d'autres termes, nous dirons qu'un texte est d'autant plus riche que son vocabulaire se répète peu, ce qui correspond, sur le plan statistique, à un coefficient faible. Le classement hiérarchique des douze nouvelles selon leur coefficient de richesse lexicale est le suivant :

Nom de la Nouvelle	Coefficient
El Amante Liberal (Aman)	5,51
La Señora Cornelia (Corn)	5,42
La Gitanilla (Gita)	5,32
La Española Inglesa (Espa)	5,27
La Ilustre Fregona (Freg)	5,25
Las dos Doncellas (Donc)	5,14
El Coloquio de los Perros (Colo)	4,95
El Celoso Extremeño (Celo)	4,84
Rinconete y Cortadillo (Rinc)	4,51
La Fuerza de la Sangre (Sang)	4,01
El Licenciado Vidriera (Vidr)	3,71
El Casamiento Engañoso (Casa)	3,49
Coeff. moyen du Corpus	4,78

La lecture du tableau nous indique que « El Casamiento Engañoso » est la nouvelle la plus riche lexicalement et que « El Amante Liberal » est la nouvelle la moins riche. Cette information ne comporte pas d'intérêt en soi si l'on se limite à ce simple constat. En revanche, la connaissance de la richesse lexicale devient intéressante dès lors qu'elle peut être mise en relation avec d'autres variables, de type explicatif. D'une part, la taille de la nouvelle est en relation avec la richesse lexicale, dans le sens où les nouvelles les plus petites ont tendance à être les plus riches (c'est le cas de « El Casamiento Engañoso » ou « El Licenciado Vidriera ») et les nouvelles les plus grandes sont les moins riches (comme « La Gitanilla », « La Ilustre Fregona » ou « El Amante Liberal »). D'autre part, d'autres paramètres interviennent dans l'explication de cette variabilité, en particulier la répartition des types de discours, à savoir l'équilibre entre le Récit et l'Interlocution dans les douze nouvelles.

Dans cette optique l'examen des index sélectifs et globaux nous a permis de mettre en évidence pour l'ensemble du corpus que, sur les 179 944 occurrences existantes, 86 683 correspondent au Récit et 93 261 à l'Interlocution, soit respectivement 48,17 % et 51,83 % du corpus. Du point de vue général, le corpus possède donc un équilibre entre les fragments narratifs et les fragments interlocutifs.

Pour calculer la dimension du Récit (86 683 occ. soit 48,17 % du corpus) nous avons regroupé le Récit syntaxiquement dominant (Rdo) et le Récit syntaxiquement dépendant (Rdp). Ils totalisent respectivement 84 233 et 2 450 occurrences soit 97,17 % et 2,83 % du bloc narratif. Pour l'Interlocution (93 261 occ. soit 51,83 % du corpus) nous avons regroupé l'ensemble des personnages individuels et collectifs des nouvelles.

Nous présentons dans le tableau suivant la distribution Récit/Interlocution pour chacune des douze « Nouvelles Exemplaires » :

Nom de la Nouvelle	Récit	Interlocution	Total

La Gitanilla (Gita)	11 394	(48,54%)	12 078	(51,46%)	23 472
El Amante Liberal (Aman)	7 255	(39,18%)	11 264	(60,82%)	18 519
Rinconete y Cortadillo (Rinc)	6 065	(44,33%)	7 617	(55,67%)	13 682
La Española Inglesa (Espa)	11 573	(72,17%)	4 463	(27,83%)	16 036
El Licenciado Vidriera (Vidr)	6 529	(68,77%)	2 965	(31,23%)	9 494
La Fuerza de la Sangre (Sang)	5 576	(72,54%)	2 111	(27,46%)	7 687
El Celoso Extremeño (Celo)	9 845	(71,95%)	3 839	(28,05%)	13 684
La Ilustre Fregona (Freg)	12 180	(60,82%)	7 846	(39,18%)	20 026
Las dos Doncellas (Donc)	8 280	(55,41%)	6 663	(44,59%)	14 943
La Señora Cornelia (Corn)	7 572	(52,80%)	6 768	(47,20%)	14 340
El Casamiento Engañoso (Casa)	377	(7,50%)	4 650	(92,50%)	5 027
El Coloquio de los Perros (Colo)	37	(0,16%)	22 997	(99,84%)	23 034
TOTAL	86683	(48,17%)	93261	(51,83%)	179 944

Bien que l'ensemble du corpus montre un équilibre entre le récit et l'interlocution, nous ne retrouvons pas cette répartition dans chaque nouvelle. En d'autres termes, si certaines nouvelles possèdent une part à peu près égale de récit et d'interlocution (comme « La Gitanilla » et « La Señora Cornelia »), d'autres sont à dominante interlocutive (telles que « El Coloquio de los Perros » et « El Casamiento Engañoso »), et d'autres sont à dominante narrative (« La Fuerza de la Sangre » et « La Española Inglesa », par exemple).

Cette variabilité structurelle pour chacune des douze « Nouvelles Exemplaires » est encore plus nette dans la représentation graphique suivante :

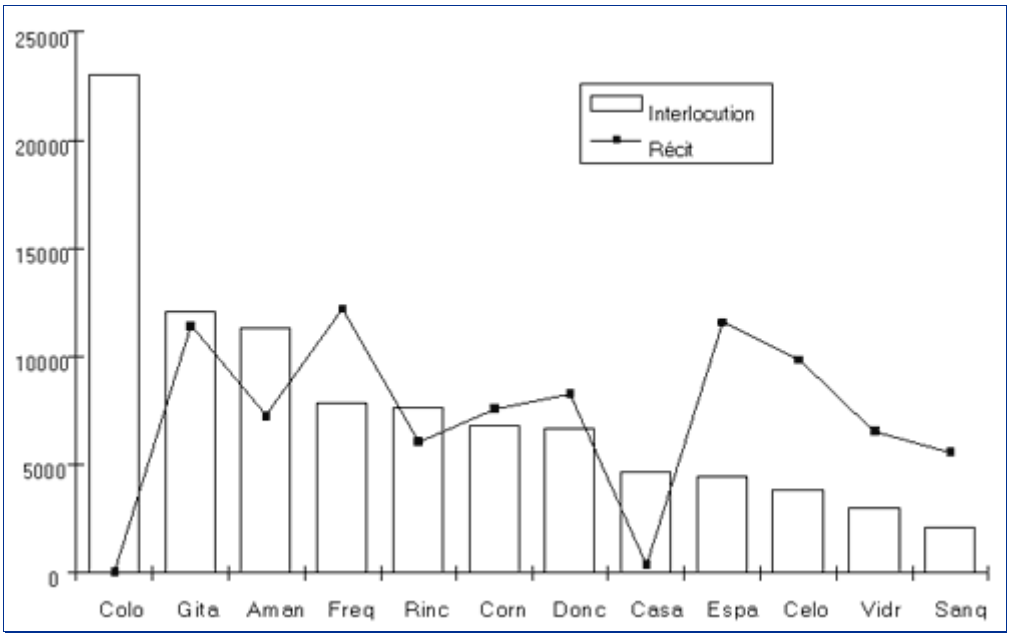

Figure 2 : Graphique de la distribution du Récit et de l'Interlocution dans le Nouvelles.

À partir de ce constat et comme nous l'avons mentionné plus haut, peut-on s'interroger sur le fait que l'équilibre des types de discours puisse plus ou moins moduler la richesse lexicale ? Peut-on dire que l'interlocution génère plus de richesse lexicale que le récit, ou inversement ?

Toutefois, cette hypothèse de recherche basée sur la définition intrinsèque du discours doit être reliée à une autre notion, celle d'inventaires fermés et

d'inventaires ouverts [8] et leur typologie narrative ou interlocutive.

Pour comprendre cela, regardons l'index hiérarchique global des douze Nouvelles Exemplaires :

que	9 843	sus	689	estaba	307	padres	196
de	8 785	yo	682	te	301	Preciosa1	195
y	8 603	tan	568	qué	293	aquella	192
la	5 782	sin	558	esta	288	dar	192
a	5 101	ni	548	quien	288	gran	191
en	3 992	él	540	os	284	entre	190
el	3 829	ella	524	sino	275	son	189
no	2 731	ser	519	hacer	268	aquí	188
los	2 428	porque	514	les	267	mis	186
con	2 376	pero	499	este	260	aquel	185
se	2 250	dos	490	tenía	251	verdad	184
por	1 978	o	489	otro	247	cosa	183
su	1 849	esto	485	otra	245	ojos	179
le	1 681	todo	483	Luego	236	hay	178
las	1 558	así	458	vida	234	mismo	178
lo	1 517	era	450	decir	233	dicho	176
me	1 087	bien	439	hasta	230	mí	176
un	1 049	pues	392	señora	222	dio	175
del	993	cuando	380	he	220	Dios	174
como	956	ha	379	ellos	219	noche	171
más	911	don	363	aunque	216	tal	169
si	887	respondió	362	allí	213	uno	168
es	822	todos	354	muy	213	Isabela	167
al	805	señor	346	ver	213	otros	167
mi	759	cual	339	todas	205	Juan	165
una	751	ya	331	día	203	menos	165
había	746	casa	330	nos	200	tiempo	165
dijo	723	donde	326	tanto	200	(...)	
para	695	fue	321	poco	199		

En effet, si nous observons ce classement hiérarchique, nous constatons que la majeure partie des formes à très forte occurrence correspondent aux inventaires fermés. À titre d'exemple, en examinant les 30 premiers rangs, nous notons que les deux premières formes ouvertes, « había » (746 occ.) et « dijo » (723 occ.), n'apparaissent qu'à la 27e et à la 28e position, alors que les 26 premières formes fermés totalisent à elles seules 73 323 occurrences, soit 40,7 % de l'ensemble du corpus. Cela souligne le poids important des inventaires fermés dans la construction du discours.

Par ailleurs, pour mettre en relation ce poids par rapport à la dimension des oeuvres, observons le graphique suivant [9], qui présente la distribution de la forme « que » (au 1er rang de l'index hiérarchique avec 9 843 occ.) dans chacune des nouvelles :

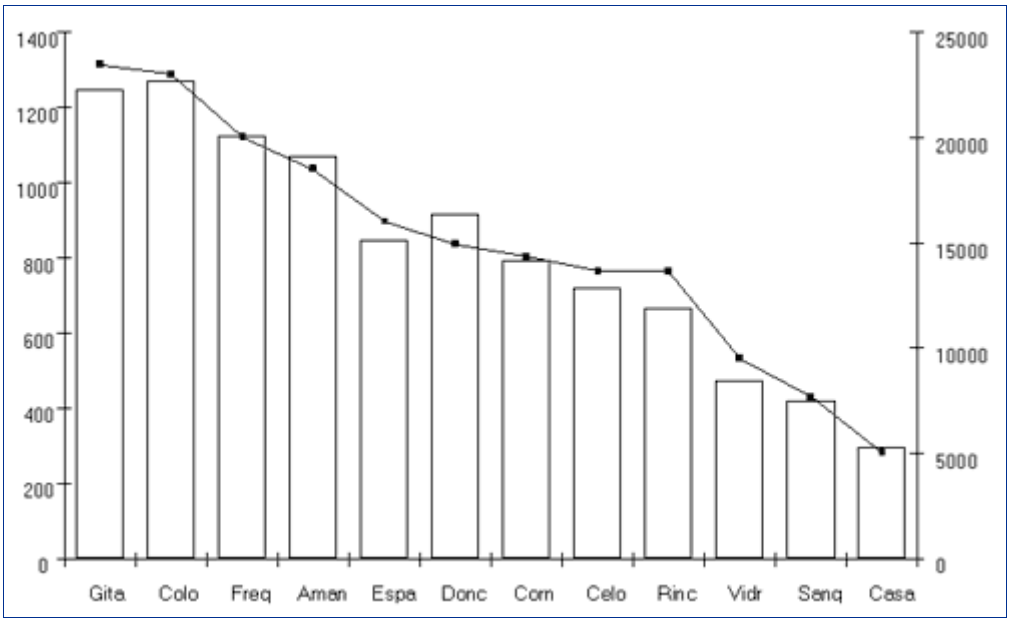

Figure 3 : Graphique sur la distribution des occurrences de la forme « QUE » (courbe) par rapport à la dimension des nouvelles (histogramme)

La représentation graphique nous montre que le nombre d'occurrences de la forme « que » est proportionnel à la taille des nouvelles. Les autres formes « fermées » fonctionnant de façon similaire, on se rend compte de leur contribution importante à la dimension d'une oeuvre, et elles modulent ainsi le coefficient de richesse lexicale. Si deux textes sont de même taille (avec la même typologie de discours) et disposent d'un coefficient de richesse lexicale différent, on pourra alors le considérer comme un indicateur d'une différence significative de leur vocabulaire.

La solution idéale pour l'étude du lexique serait de disposer de filtres puissants permettant d'ignorer les formes fermées lors des différents traitements lexicométriques afin d'inventorier uniquement les formes ouvertes et ainsi pouvoir comparer plusieurs oeuvres indépendamment de leur taille et de leur typologie discursive [10].

3.2. Typologie du système démonstratif Cervantin

Notre objectif était d'examiner la relation entre les trois familles de formes démonstratives (este / ese / aquel) et les différents types du discours (narratif / interlocutif). Nous avons recensé les formes démonstratives existant dans l'ensemble des douze Nouvelles, puis nous avons déterminé leur distribution pour chacune des Nouvelles. L'ensemble du corpus compte 39 formes graphiques qui totalisent 2 285 occurrences. La répartition hiérarchique est la suivante [11] :

Nouvelle	**(Code)**	**Occurrences**
La Gitanilla	(Git)	309
Coloquio de los Perros	(Col)	298
La Ilustre Fregona	(Fre)	271
La Española Inglesa	(Esp)	222

El Amante Liberal	(Ama)	220
Rinconete y Cortadillo	(Rin)	195
La Señora Cornelia	(Cor)	185
Las dos Doncellas	(Don)	177
El Celoso Extremeño	(Cel)	155
El Licenciado Vidriera	(Vid)	97
La Fuerza de la Sangre	(San)	87
El Casamiento Engañoso	(Cas)	69
TOTAL		2 285

En ce qui concerne la dimension en nombre d'occurrences de chaque forme démonstrative, nous avons obtenu le classement hiérachique suivant :

esto	485	deste	50	aquél	14	deso	4
esta	283	aquello	46	desa	11	déste	3
este	254	ese	45	esos	9	desos	3
aquella	192	éste	43	éstos	9	déstos	3
aquel	185	ésta	34	aquélla	8	aquesta	2
eso	118	esa	33	dese	8	aquesas	1
estas	105	aquellas	30	désta	7	aqueso	1
desta	95	desto	19	ésa	7	désos	1
estos	71	destos	19	ése	7	ésos	1
aquellos	56	destas	18	esas	5		

Bien que ces deux tableaux nous renseignent sur les oeuvres les plus privilégiées par la présence des démonstratifs, et sur les formes dominantes pour l'ensemble des 12 Nouvelles, cela reste insuffisant en termes d'analyse. Ces premiers résultats ont donc donné lieu à la création de tableaux lexicaux – qu'il nous est impossible de présenter ici étant donné leur très grande dimension – dans lesquels nous avons mis en relation la distribution des formes démonstratives dans chaque Nouvelle et leur appartenance aux types de discours narratif ou interlocutif.

Ces tableaux étant trop complexes pour pouvoir être analysés à partir d'une lecture directe, ils ont donné lieu à une étude statistique [12] pouvant nous aider à l'interprétation. Sans entrer dans le détail, nous rappelons que ce type d'analyse [13] (analyse factorielle des correspondances) permet de transformer des tableaux linguistiques complexes en graphiques constitués de points qui illustrent les variables étudiées autour de deux axes factoriels, ce qui offre une représentation synthétique lisible. L'interprétation de chaque graphique ou plan factoriel se réalise en fonction de la distribution (voisinage) des mots représentés (lignes du tableau) et des différents textes et/ou types de discours (colonnes du tableau). En ce qui concerne notre application, les graphiques factoriels représentent les formes démonstratives et les types de discours par nouvelle.

La première analyse factorielle permet de mettre en évidence une différence de typologie entre les fragments interlocutifs et narratifs. En effet, ce plan montre que les blocs narratifs et blocs interlocutifs se trouvent dans des zones

différentes du plan factoriel et que leur opposition traduit des caractéristiques linguistiques spécifiques à chaque type de discours [14]. La disposition de ce premier graphique est importante pour analyser la distribution des formes démonstratives contenues dans le second plan factoriel, car en effet les mots ont soit une typologie interlocutive soit une typologie narrative [15].

Le deuxième plan factoriel permet d'observer que les trois familles de démonstratifs (este, ese, aquel) se regroupent entre elles marquant ainsi trois zones de spécificité linguistique. Nous notons, en comparant ce deuxième graphique au premier, que la famille démonstrative « ese » est associée aux blocs interlocutifs et que la famille « aquel » est associée aux blocs narratifs. Ainsi la famille « ese » est typique (ou dominante) des blocs interlocutifs et la famille « aquel » est typique des blocs narratifs.

Quant à la famille « este » elle montre un usage plus général de ses formes appartenant ainsi aux deux types de discours. Mais il faut nuancer cette analyse car les démonstratifs « éste, este, esta, estos, éstos, ésta » se situent à proximité de la zone interlocutive (proche de la famille « ese ») traduisant un usage un peu plus dense dans l'interlocution. En revanche, « esto, desta et estas » mettent en évidence une utilisation spécifique du récit.

Si la place importante de la famille « aquel » dans le récit s'explique par l'une de ces fonctions qui est de présenter des événements qui ont eu lieu dans le passé (localisation temporelle) [16], les démonstratifs « esto, desta et estas » sont attirés par le récit pour des caractéristiques du discours narratif autres que la localisation temporelle (ou spatiale). En effet, le récit fait appel à différentes valeurs de ces trois formes démonstratives. D'une part, afin d'assurer la cohérence textuelle du discours, la fonction anaphorique [17] des démonstratifs répond au principe d'économie en évitant ainsi les répétitions lors des reformulations, c'est-à-dire, le renvoi à un mot ou à une série de mots apparus antérieurement dans le discours [18]. D'autre part la fonction cataphorique [19] de ces déictiques permet d'introduire le discours direct des personnages (mais dans une moindre mesure). Ces deux types de fonctions observées dans le corpus, expliquent la forte liaison des formes « esto, desta et estas » avec le discours narratif.

Dans l'interlocution, les démonstratifs peuvent posséder la valeur localisatrice spatio-temporelle du fait que les interlocuteurs se situent dans le même espace-temps. La valeur spatiale est ainsi liée à la position des deux interlocuteurs (émetteur/récepteur) par rapport à l'objet : « este » (proximité), « aquel » (éloignement) et « ese » (position intermédiaire). La valeur temporelle de la famille « aquel » existe dans l'interlocution principalement dans le cas du personnage narrateur, et par conséquent il s'agit d'une fonction relativement limitée, réservée surtout au récit, comme le met en évidence le graphique.

Bien que nous puissions étendre ce type d'approche à l'ensemble des formes

démonstratives, nous limitons ici notre propos à ces quelques réflexions qui nous montrent l'intérêt de disposer de banques de données textuelles et d'instruments d'observation et d'analyse nous permettant de relier les statistiques des formes à leur appartenance structurelle (types de discours) dans l'optique de dégager des typologies spécifiques, en l'occurrence celle de la langue cervantine.

4. Conclusion

Les outils informatiques et d'analyse de statistique lexicale, dont nous avons ici donné une illustration d'application sur corpus de très grande dimension, font donc partie de notre méthodologie de recherche au sein de notre Laboratoire. Ces ainsi que l'exploration actuelle des corpus spécialisés scientifiques, économiques et techniques se situent dans le continuum de l'application que nous venons de présenter sur les aspects lexicaux, syntaxiques et la typologie du discours de la langue cervantine. Nous recherches sont en accord avec les orientations de notre laboratoire partenaire CIEL de l'Université de Paris 7 travaille sur le problème de classement des unités lexicales et de la construction du discours. Notre méthodologie et les principes linguistiques sous-jacents à ce type d'approche sur corpus permettront donc de développer des projets de recherche communs entre l'université d'Évry et de Paris 7.

Si les questions fondamentales de la recherche linguistique demeurent, elles s'enrichissent depuis quelques années des progrès informatiques et statistiques qui offrent ainsi de nouvelles perspectives d'investigation, en augmentant les potentialités d'observation et d'analyse, grâce notamment à la recherche méthodologique et à la création des grands corpus électroniques. Ainsi pour étudier les différentes typologies des formes linguistiques il était nécessaire d'établir, de quantifier et de comparer les formes et les structures morpho-syntaxiques préférentielles dans chaque type de texte ou de discours. Notre méthodologie nous a permis d'atteindre cet objectif. Cette première série d'analyses empiriques effectuées sur l'écriture cervantine, grâce à la construction de la banque de données sur les douze « Nouvelles Exemplaires », met en évidence l'existence de micro-systèmes structurels et syntaxiques des formes et par-là même celle des typologies des discours. La double contextualité des formes structurelle et syntaxique pèse donc sur les occurrences constitutives des énoncés. Cela exige l'utilisation de l'informatique afin de pouvoir réaliser les analyses linguistiques des grands échantillons des données issues des méga-corpus dont nous travaillons également sur les discours scientifiques, économiques et techniques nécessaires au développement des filières universitaires de L.E.A.

Javier Sanchez et Collaborateurs
Département de Langues Étrangères Appliquées
Directeur du Laboratoire de Traduction spécialisée

et nouvelles technologies
Université d'Évry-Val d'Essonne
et membre du CIEL de l'Université Paris 7.

NOTES

[1] Collaborateurs : Baldo Sabrina, Boely Christian, Bonneville Christine, Galiana Philippe, Zamoun Slimane.

[2] Sanchez J., *Méthodologie et outils de l'analyse relationnelle informatique des textes*, Collection « Analyse Textuelle et Nouvelles Technologies », publié par l'Université de Paris-VIII, Saint-Denis, Juillet 1992 (350 pages), (épuisé). [« Book Review » par Simone Monsonégo (CNRS-INaLF), dans la Revue spécialisée L.L.C. (Literary and Linguistic Computing), vol. 8, n° 3, p. 186, Oxford University Press, 1992.]

[3] Sanchez J., Qu'est-ce que l'analyse relationnelle informatique des textes ?, Revue *Informatique et Statistique dans les sciences humaines*, Université de liège, n^{os} 1 à 4, 1993, p. 135-165.

[4] Pour cette phase du protocole, nous avons retenu le logiciel qui nous paraissait le plus adapté à tout point de vue (convivialité, rapidité et performances). Il s'agit du programme Omnipage dont il faut préciser qu'il existe sous deux versions, pour compatible IBM et pour Macintosh.

[5] Le logiciel que nous utilisons actuellement est Micro-OCP du centre de calcul de l'Université d'Oxford et qui a été développé sous Ms-Dos.

[6] L'édition utilisée pour l'élaboration de cette banque est : « Novelas Ejemplares » (2 Vol.), Ediciones Cátedra, Madrid, 1989.

[7] Sanchez J., Techniques informatiques de désambiguïsation et d'organisation par blocs des données textuelles et procédures initiales pour l'étude de l'interlocution, *Cahiers de Linguistique Relationnelle Informatique*, Centre de Recherche de l'Université de Paris VIII, 1989 (70 pages).
Sanchez J., Mesure et dynamique interlocutive en analyse relationnelle informatique des textes, dans *Revue Literary and Linguistic Computing*, Oxford University Press, Vol. 6, n° 2, p. 104-108, 1990.

[8] Nous appelons inventaires fermés les catégories grammaticales existantes dans tous les textes telles que les prépositions, les articles, les conjonctions, etc. qui ne portent pas de sens lexical mais qui constituent des éléments de relation ou de détermination. Par opposition, nous appelons inventaires ouverts l'ensemble des mots porteurs de sens et qui sont propres à chaque texte (substantifs, verbes, adjectifs).

[9] Dans ce graphique, on lit à gauche le nombre d'occurrences de la forme « que » (représentée par la courbe) et à droite la dimension en nombre

d'occurrences (représentée par l'histogramme).

[10] En effet, rappelons qu'il existe un lien entre les formes et les types de discours. Donnons l'exemple, dont nous avons déjà parlé, des formes verbales « había » (746 occ.) et « dijo » (723 occ.) qui appartiennent respectivement à 75 % et 88 % au récit, donc à dominante narrative. De façon identique, les inventaires fermés peuvent comporter du point de vue statistique des formes typiques ou plus au moins liées à un type de discours. Nous allons en rendre compte à partir de la distribution des occurrences des démonstratifs dans les douze nouvelles et ce pour les deux types de discours.

[11] Si nous comparons les occurrences des démonstratifs avec la dimension des nouvelles, nous constatons que ces déictiques gardent la même proportion. Cela prouve encore, comme pour d'autres inventaires fermés, que le système démonstratif est utilisé en fonction du temps de parole de chaque nouvelle.

[12] Cette analyse à été effectuée avec le logiciel SPAD-N (Système Portable d'Analyse des Données Numériques) développé par le CISIA, Saint-Mandé (France).

[13] L'analyse factorielle des correspondances a été créée par J.-P. Benzécri et est utilisée actuellement dans pratiquement tous les domaines des sciences humaines où l'on nécessite une aide à l'interprétation des grands échantillons de données. Il existe des nombreuses publications à ce sujet, mais nous donnons celle qui correspond à sa première présentation par l'auteur : *Pratique de l'analyse des données : analyse des correspondances, exposé élémentaire*, Dunod, 1979 (3e édition).

[14] Par contre, si les fragments narratifs et interlocutifs avaient formé un seul ensemble compact de points, ce voisinage aurait traduit des caractéristiques linguistiques communes. Pour citer un exemple concret, l'analyse factorielle des correspondances des prépositions espagnoles montre ce type de schéma puisque aussi bien le récit et l'interlocution comportent cette catégorie d'inventaire fermé.

[15] Quand nous parlons de typologie nous faisons référence à un concept de dominance (statistique) d'une ou plusieurs formes dans un type de discours mais sans pour autant impliquer une exclusivité d'emploi.

[16] Voici un exemple : « ... al punto declaró la una a la otra su determinación amorosa, y desde *aquella* noche determinaron de dar principio a la conquista de sus dos desapasionados amantes. » (*La ilustre fregona*).

[17] Pour les valeurs des démonstratifs espagnols voir : Asenjo Oriva M.-R., *Los demostrativos*, Publicationes del Colegio de España, 1990.

[18] Voici un exemple : « ... pero ahora os suplico con todo encarecimiento que

os vais y me dejéis, que me importa. Hablando *esto* se tentó la cabeza, y vio que estaba sin sombrero, y volviéndose a los que habían venido pidió que... » (*La señora Cornelia*).

[19] Voici un exemple : « Hiciéronlo así, y quedando solos los cinco, sin esperar que otro hablase, con sosegada voz, limpiándose los ojos, *desta manera* dijo Carrizales : -Bien seguro estoy, padres y señores míos... » (*El celoso Extremeño*).

Association EPI
Février 2006