CONSTITUER UN CORPUS DE TEXTES DE SPECIALITE

M. Teresa CABRÉ
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra (Barcelone)

Il ne peut y avoir de doute que le développement des corpus textuels a permis à la linguistique descriptive de faire un saut qualitatif très important. Ce progrès a permis aux linguistes de tenir compte de façon plus adéquate du fonctionnement des langues à partir du moment où les analyses ont pu se fonder pour la première fois sur des exemples abondants et représentatifs de production langagière qui n'étaient ni limités ni biaisé subjectivement comme auparavant. En outre, la linguistique de corpus permet d'explorer de manière exhaustive les productions langagières et d'offrir au linguiste des échantillons de données qu'une analyse manuelle n'est pas en mesure d'exploiter de façon aussi approfondie.

Dans cet article nous poursuivons trois objectifs. Tout d'abord nous exposerons quelques caractéristiques de ce qu'on appelle les "langues de spécialité", qui sont à l'origine des textes spécialisés. Ensuite nous présenterons brièvement le corpus textuel spécialisé de l'Institut universitaire de linguistique appliquée comme échantillon d'adaptation aux conditions mentionnées dans notre premier objectif. Nous monterons enfin à titre d'exemple une étude sur les différences entre les textes de spécialité et les textes non spécialisés au regard de leurs caractères linguistico-grammaticaux.

1 LA CONSTITUTION D'UN CORPUS DE SPECIALITE : QUESTIONS ET CRITERES

Pour construire un corpus textuel de spécialité la première question qui se pose est de savoir ce qu'on entend par texte spécialisé ou comment nous identifions les textes spécialisés. Sans réponse précise à cette question il n'est évidemment pas possible de commencer à sélectionner le matériel.

Une fois établis les critères qui permettent d'identifier les textes produits par des spécialistes dans des situations professionnelles, il convient de déterminer quels types de texte nous devons retenir pour que le corpus qui en résulte soit suffisamment équilibré.

En troisième lieu, nous devons déterminer la quantité de productions qui feront partie de ce corpus, pour savoir s'il sera suffisamment représentatif de chaque spécialité ou bien seulement pour analyser un thème préalablement choisi. Il convient à ce propos d'avoir préalablement déterminé afin de pouvoir définir sa dimension. Dans quel but constituons-nous ce corpus? Quelle en est la finalité que nous comptons atteindre grâce à lui? A quelles études linguistiques souhaitons-nous qu'il donne lieu?"

En répondant à ces trois questions nous pouvons déjà commencer le travail, qui logiquement devra comporter d'autres questions de type plus techniques, qu'il s'agisse de questions linguistiques ou d'informatiques ?

Une fois constitué le corpus sous forme numérique enfin, nous devons factoriser toutes les possibilités d'exploration, possibilités que nous avons dû établir dans l'étape préliminaire de la caractérisation du corpus à constituer. Nous répondrons point par point à chacune de ces questions.

1.1 Qu'est-ce un texte spécialisé? Comment reconnaître parmi tous les textes ceux qui sont spécialisés?

Les textes spécialisés sont les productions linguistiques, orales ou écrites, qui se manifestent dans le cadre des communications professionnelles et dont la finalité est exclusivement professionnelle. On reconnaît les situations professionnelles par les interlocuteurs qui interagissent, par le sujet évoqué qui relève du domaine ou des domaines concernés par la profession, et par la finalité essentielle de rechercher l'information auprès du récepteur, bien que pour ce faire on utilise des stratégies discursives différentes.

D'un point de vue analytique, on peut dire que les textes spécialisés se définissent par trois types de conditions :

- conditions discursives : les propriétés de la situation spécialisée de ce type de communication
- conditions cognitives : le thème qui est traité et la façon dont il est traité
- conditions linguistiques : les conditions textuelles générales (précision, concision et systématicité, les deux dernières à des degrés différents suivant les conditions discursives), la forme macro et micro textuelle, et surtout les unités lexicales propres au domaine dont il est question dans le texte.

1.2 Quels sont les variables que nous pouvons prendre en considération dans un corpus spécialisé?

Les textes de spécialité ne sont pas homogènes, mais ils sont organisés en différents types en fonction des critères de classification qui sont pris en considération. Les critères qui sont à notre avis les plus pertinents pour organiser les textes de spécialité dans un corpus sont les suivants :

- le thème/sujet
- la perspective ou dimension disciplinaire
- le niveau de spécialisation
- les sources
- le genre textuel
- la classe de texte d'après la stratégie discursive
- les langues
- la relation entre les textes des langues du corpus dans le cas de textes plurilingues (bilingues, trilingues, etc.)

En matière de thème ou de sujet nous distinguons entre corpus monodisciplinaire et pluridisciplinaire. Un exemple de ce dernier : la banque de droit de l'environnement pour le groupe TERMISUL de l'Université de Porto Alegre (Brésil).

Quant au niveau de spécialisation, un corpus peut comporter des textes d'un seul niveau de spécialité (par exemple : des textes d'articles scientifiques

provenant de périodiques de même type) ou encore comporter des textes de différents niveaux de spécialité[1].

Par le canal de transfert, les textes du corpus peuvent émaner d'un seul type de source ou de plusieurs types de sources. La diversité des sources peut résulter aussi d'une grande diversité de critères parmi lesquels nous nous intéressons au critère du mode de transmission ici car les textes d'un corpus peuvent être exclusivement oraux ou écrits ou audiovisuels ou encore comporter des occurrences de toutes les possibilités.

En ce qui concerne le genre textuel, un corpus peut être homogène et ne comporter que des textes d'un seul genre (par exemple des abstracts de périodiques scientifiques), ou bien il peut inclure des textes de différents genres textuels.

En ce qui concerne le type de texte d'après la stratégie discursive, les corpus peuvent être homogènes ou hétérogènes en matière de genre textuel. Par exemple, un corpus homogène ne comportera que des textes argumentatifs, ou narratifs, etc.

En application du critère des langues, les corpus peuvent être monolingues, bilingues, trilingues, etc. Les textes qui comportent plus d'une langue peuvent être mélangés au sein d'un seul sujet ou bien comporter des textes dans une langue donnée et la traduction correspondante dans la deuxième ou la troisième langue. Dans ce dernier cas on parle de corpus parallèles.

1.3 Quelle est la bonne taille d'un corpus spécialisé?

La réponse à cette question ne peut être que la suivante : ceci dépend de la finalité du corpus. A quoi va servir un corpus? Pour extraire des données représentatives de l'usage d'une langue dans son ensemble? Dans ce cas-là nous devons constituer un type de corpus que l'on appelle corpus de référence, qui comporte un échantillon d'usage représentatif de la totalité de la langue, y compris toute la variation interne et externe. En revanche, s'il s'agit de constituer un corpus pour étudier un problème particulier, la taille du corpus doit être en adéquation avec les finalités proposées. Par exemple, le corpus que nous devons constituer pour analyser l'usage d'un pronom en position enclytique sera de taille moins importante que celui qu'il faudrait

[1] La pertinence d'un texte à un niveau haut, moyen ou bas se détermine par les caractéristiques des destinataires, son support et ses finalités. Ainsi, un texte produit par un spécialiste pour des étudiants peut se définir comme de niveau moyen. Pour de plus amples renseignements, voir Cabré (1998) et Ciapuscio (2003).

pour extraire la terminologie d'un domaine de spécialité. L'extraction de collocations nécessitera un corpus encore plus important.

1.4 Le processus de constitution de corpus

La constitution effective, une fois les critères établis, se déroule en phases distinctes :

a. la sélection des sources

b. les critères de sélection des textes et la décision de savoir s'il faut prendre le texte complet ou des fragments du même texte[2]

c. les décisions quant à l'architecture de base

d. les décisions quant à l'infrastructure logicielle et matérielle (système de gestion de corpus textuels)

e. la sélection des conventions pour la représentation des textes

f. les critères, langage et système de balisage structurel

1.5 Outils d'interrogation

Les textes d'un corpus peuvent être utilisés sous forme brute ou déjà traités linguistiquement. Si l'on utilise des textes déjà traités, il paraît logique de tenir compte des ressources et des outils de traitement automatique de l'information :

- outils de marquage structural et linguistique
- dictionnaire initial de traitement
- système d'analyse morphologique
- système de lemmatisation
- système de désambiguïsation
- système de gestion de dictionnaires
- système de structuration syntaxique (« chunker »), etc.

[2] Cette décision dépend des études que nous souhaitons réaliser grâce au corpus. Pour l'analyse textuelle (connecteurs, structuration informative, genres textuels, etc.), il faut des textes complets.

1.6 Possibilités d'exploitation

Les possibilités d'exploitation linguistique d'un corpus dépendent enfin du traitement que les données ont subies pendant cette phase. Les possibilités d'application des données du corpus se réalisent dans les secteurs suivants :

- en ingénierie linguistique, pour la mise au point d'outils et de robots
- en extraction d'information pour des besoins de recherche, d'enseignement, d'exploitation industrielle, de publication, etc.
- en récupération d'information pour des besoins documentaires et bibliographiques.

Les linguistes s'intéressent aux corpus de spécialité surtout pour les applications suivantes :

- la recherche sur le discours spécialisé, la terminologie et la phraséologie spécialisés
- l'élaboration de dictionnaires spécialisés
- l'enseignement des langues de spécialité ou de langues sur objectifs spécialisés.

Pour l'enseignement des langues de spécialité, les corpus donnent la possibilité de mieux préparer les programmes (en rapport avec les besoins et le niveau de connaissances des étudiants), d'élaborer des exercices et d'alimenter des systèmes d'auto-apprentissage des langues.

Dans le domaine de la documentation, et plus concrètement pour la gestion de l'information, les corpus fournissent de l'information pour la construction automatique ou assistée par ordinateur de thésaurus, pour l'indexation automatique et pour élaborer des systèmes de classification de documents ou pour mieux orienter la consultation selon le profil de l'utilisateur individuel.

2 LE CORPUS TECHNIQUE PLURILINGUE DE L'IULA

L'Institut universitaire de linguistique appliquée (IULA) est un centre de l'Université Pompeu Fabra, de Barcelone, consacré à la recherche et à la formation doctorales. Il fut créé en 1993 par Maria Teresa Cabré. L'IULA est sous-divisé en groupes de recherche : Lexique, terminologie et discours spécialisé (Groupe IULATERM, qui héberge la Linguistique Informatique), Lexicographie (Groupe INFOLEX), Variation linguistique (Groupe UVAL), Documentation et publication numérique (Groupe DIGIDOC), ainsi que trois laboratoires : OBNEO (Observatoire de néologie), LATEL (Laboratoire de

technologie linguistique) et le Laboratoire de linguistique judiciaire. Depuis 1993 jusqu'aujourd'hui le projet Corpus est le projet de recherche commun auquel participent tous les membres de l'IULA. Il comporte des textes écrits dans cinq langues (catalan, castillan, anglais, français et allemand) des domaines de l'économie, du droit, de l'environnement, de la médecine et de l'informatique. Le corpus comporte en plus des documents parallèles, facilitant ainsi l'étude de la traduction. Le corpus multilingue de l'IULA est constitué d'un sous-corpus de la langue générale, extrait de la presse de grande diffusion, qui représente un corpus contrastif.

L'objectif de ce corpus est de faciliter l'analyse de données linguistiques afin de pouvoir établir les lois qui régissent le comportement de chaque langue dans chaque domaine. Il est ouvert aux chercheurs et à tous ceux qui ont besoin de consulter dans les domaines de spécialité concernés. L'exploitation du corpus a débouché sur des études de caractère terminologique, discursif, morphologique, syntaxique, néologique ou traductologique. Afin de faciliter l'exploitation des données, l'IULA a mis au point une série d'outils d'interrogation. Parmi ceux-ci on peut signaler un extracteur automatique de néologie, un détecteur automatique de terminologie, un aligneur de textes, un outil permettant l'alimentation des dictionnaires. De fait, ce corpus est le principal support des activités de recherche et d'enseignement de notre institut.

L'outil qui permet d'accéder aux données du corpus par Internet est BwanaNet, qui peut être consulté sur la page principale du site de l'IULA (http ://bwananet.iula.upf.edu/), dans la rubrique intitulée " Recursos IULA".

Le corpus de l'IULA comporte des textes écrits dans cinq langues (catalan, castillan, anglais, français et allemand) des domaines de l'économie, du droit, de l'environnement, de la médecine et de l'informatique, ainsi que des documents parallèles sur ces sujets. Chacun des domaines a été structuré par un spécialiste en différents sous-domaines de telle sorte que les textes puissent être récupérés avec une grande précision thématique.

Voici comment est structuré le domaine de la médecine :

Anatomie	(AN)
Organismes	(OR)
Maladies	(MA)
Produits chimiques et pharmaceutiques	(PQ)
Techniques et équipements analytiques, diagnostiques et thérapeutiques	(TE)
Psychiatrie et psychologie	(PS)
Sciences biologiques	(CB)
Sciences physiques	(CF)
Anthropologie, éducation, sociologie et phénomènes sociaux	(FS)
Technologie, industrie, agriculture	(TI)
Sciences humaines	(HU)
Information scientifique	(IC)
Groupes nominaux	(GN)
Planification et gestion sanitaires	(GS)

Le traitement des textes du corpus suit les étapes suivantes.

2.1 Phase de sélection des textes

Les spécialistes de chaque matière sélectionnent les textes qu'ils considèrent comme pertinents et les classent par thème dans une structuration du domaine préalablement conçu par des spécialistes.

2.2 Phase d'annotation et d'enregistrement de l'information du document

Les documents sont balisés selon la norme SGML et les conventions établies par la norme Corpus Encoding Standards (CES) du projet EAGLES. Ensuite l'information de type documentaire est enregistrée (auteur, titre, édition, pages retenues, sous-domaine auquel il appartient, langues qu'un document unique peut comporter dans le corpus).

2.3 Phase de traitement linguistique

Le traitement linguistique de la documentation est automatisé et comporte un prétraitement afin de traiter linguistiquement les entités susceptibles d'une détection automatique avant l'analyse morphologique (dates, chiffres, locutions, noms propres, sigles et abréviations), une analyse morphologique, par laquelle tous les mots du documents sont lemmatisés et pourvus d'une ou de plusieurs étiquettes morphologiques, en accord avec le système d'étiquetage morphosyntaxique conçu à l'IULA, ainsi qu'une désambiguïsation linguistique et statistique de sorte que chaque mot ne se voie attribué qu'un seul lemme et une seule étiquette.

2.4 Stockage dans une base de données textuelles

Finalement quand chaque mot est associé à un lemme et à une catégorie grammaticale, les textes sont stockés dans une base de données textuelles, qui comporte toute l'information générée sur ce document. Le résultat de tout ce processus de traitement des textes peut être consulté en ligne à l'adresse suivante : *http :// brangaene.upf.es/bwananet/index.htm*.

Domaine	catalan	espagnol	anglais	français	allemand	total
Droit	1 463 000	2 085 000	431 000	44 000	16 000	4 039 000
Economie	1 776 000	1 091 000	274 000	78 000	27 000	3 246 000
Environnement	1 506 000	1 062 000	599 000	230 000	429 000	3 826 000
Informatique	655 000	1 227 000	338 000	194 000	83 000	2 497 000
Médecine	2 619 000	4 077 000	1 555 000	27 000	198 000	8 476 000
Total	8 019 000	9 542 000	3 197 000	573 000	753 000	22 084 000

Figure 1 : nombre de mots par langue et par domaine

Le corpus de médecine comporte un sous-corpus de textes sur le génome humain, élaboré par le groupe Iulaterm, qui comporte 945 000 mots en catalan, 1 447 000 en espagnol et 1 119 000 en anglais.

Les donnés en relation avec le corpus parallèle pour les paires linguistiques les plus significatives catalan-espagnol, catalan-anglais, espagnol-anglais, sont présentées dans la figure 2.

Domaine	catalan-espagnol	catalan-anglais	espagnol-anglais
Droit	460 000	12 000	57 000
Economie	600 000	250 000	283 000
Environnement	214 000	213 000	144 000
Informatique	28 000	–	300 000
Médecine	118 000	40 000	640 000
Total	420 000	515 000	1 424 000

Figure 2 : Nombre de mots dans les corpus parallèles par domaine et par langue

Les données du corpus témoin sont indiquées dans la figure 3.

Domaine	catalan	espagnol	total
Général	1 526 000	3 230 000	4 756 000

Figure 3 : nombre de mots dans le corpus de la langue générale

Le corpus technique de l'IULA (CT-IULA) est indexé grâce à un ensemble d'outils mis au point par l'*Institut für Maschinelle Sprachverarbeitung*, de l'Université de Stuttgart (Corpus Workbench). L'IULA a mis au point l'outil qui permet l'interrogation en ligne de CT-IULA (brangaene.upf.es/bwananet/index htm).

2.5 Une application de linguistique de corpus : comparaison grammaticale entre textes spécialisés et textes non spécialisés

Grâce à ce corpus plus de vingt thèses de doctorat ont pu être réalisées. En plus des thèses, le corpus a permis de mettre au point une base de connaissances (GENOMA) qui peut être consultée à www.iula.upof.edu/genoma.

En ce moment, un projet de recherche sur les caractéristiques spécifiques des textes spécialisés par rapport aux textes non spécialisés est sur le point d'être achevé. Une brève synthèse de ce projet et certains de ses résultats sont présentés ci-dessous.

Le projet ESPETEX, qui fait partie d'un projet plus vaste financé par le Ministère de l'Education et de la culture espagnol (TEXTERM-2. *Fondements, stratégies et outils pour le traitement et l'extraction automatiques de l'information spécialisée* Nº REFERENCIA : BFF2003-02111) auquel participent une vingtaine de chercherurs et collaborateurs, comporte deux objectifs :

Vérifier si les caractéristiques grammaticales que les manuels de langues de spécialité attribuent aux langues de spécialité sont confirmées par rapport à un corpus suffisamment représentatif.

Au cas où ceci ne serait pas confirmé en totalité ou en partie, tenter de relever et d'établir quelques-uns de facteurs grammaticaux spécifiques qui caractérisent les textes spécialisés.

Pour mener à bien ce projet nous sommes partie de la liste des caractéristiques des textes spécialisés exposés dans les deux manuels suivants :

- Kocourek, R. (1991) *La langue française de la technique et de la science. Vers une linguistique de la langue savante*, Wiesbaden, Oscar Brandstetter Verlag.
- Sager, J.C. ; Dungworth, D. ; McDonald, P. (1980) *English Special Languages*. Wiesbaden, Oscar Brandstetter Verlag.

Ces manuels sont basés sur des corpus de taille modeste. Pour le projet ESPETEX nous avons constitué un corpus double : un premier corpus de textes spécialisés et un second de textes de caractère général.

Le corpus de la langue générale, issue de la presse, comporte 5.002.121 mots, répartis dans 155 documents du Corpus de l'IULA.

Le corpus de spécialité est composé de 5.018.193 mots répartis dans 251 documents du Corpus de l'IULA (droit, économie, informatique, environnement, médecine : 1.000.000 mots par domaine).

Les caractéristiques grammaticales non lexicales que les manuels attribuent aux textes de spécialité peuvent être répartis, selon Kocourek (1991), en quatre groupes[3]

1. sélection des catégories grammaticales
2. complexité de la structure
3. condensation syntaxique
4. impersonnalité de la phrase

En ce qui concerne la sélection des catégories grammaticales on relève les phénomènes suivants :

- prédominance des substantifs
- emploi particulier de certaines catégories grammaticales, surtout en relation avec le verbe (ainsi qu'avec les pronoms personnels) :

[3] En plus de Kocourek, d'autres linguistes ont publié sur le sujet. Parmi ceux-ci nous signalons en particulier : Phal (1968), Vigner et Martin (1976), Loffler-Laurian (1980, 1982, 1983, 1985, 1986), D. Candel (1984), Hoffmann (1985) et L'Homme (2005).

    - absence de la deuxième personne du singulier comme du pluriel
    - usage rare de la première personne du singulier, auquel on préfère *nous*.

- absence de certains mots ou morphèmes grammaticaux de la morphologie verbale :
    - prédominance de la troisième personne du singulier
    - prédominance du présent de l'indicatif
    - fréquence de la deuxième personne du pluriel à l'impératif
    - prédominance de phrases déclaratives
    - emploi limité de phrases interrogatives directes.

Par rapport à la complexité structurelle, on distingue comme spécifiques des textes spécialisés les traits grammaticaux suivants :

- faible longueur de la phrase
- nominalisation des verbes
- fréquence d'expansions de noms et de syntagmes nominaux
- abondance de propositions relatives
- constructions réalisées avec des participes et l'infinitif
- diversité des conjonctions circonstancielles
- constructions insérées dans la phrase

Comme exemples de condensation syntaxique, nous relevons les phénomènes suivants :

- emploi abondant de pronominalisation
- emploi de propositions infinitives et participiales
- nominalisation de formes verbales

Le caractère impersonnel de la phrase dans les textes de spécialité est exprimé par les phénomènes suivants :

- pronom de modestie : *nous*
- emploi de l'indéfini : *un*
- tournures impersonnelles comme *est* + adjectif (*probable*, *certain*, *surprenant*, etc.), *il en résulte que*, etc.
- emploi fréquent de la voix passive.

En plus de toutes ces caractéristiques grammaticales, il convient de souligner sur le plan textuel :

- absence de certains genres (cartes, pièces de théâtre, etc.)
- abondance de certains genres : suivant le domaine (droit, médecine, génomique, etc.)
- contrôle de la structuration de l'information (marqueurs discursifs et méta-discursifs, tables, listes, etc.).

Sur le plan lexical :

- présence abondante de terminologie
- absence relative d'unités polysémiques
- tendance à employer systématiquement la même unité pour un concept, évitant ainsi l'emploi de synonymes.

Sur le plan graphique, enfin :

- présence de symboles, de formules
- représentations iconiques
- unités lexicales : *commande –c*, etc.

L'analyse réalisée sur notre double corpus se limite aux phénomènes suivants :

- classes grammaticales : N, V, Adj, Adv, Prép, Conj,
- noms propres et noms communs
- genre et nombre des substantifs
- nom précédé du déterminant défini
- adjectifs qualificatifs
- pronoms relatifs
- personne, mode et temps des verbes
- formes verbales impersonnelles
- prépositions
- conjonctions

Parmi les substantifs et pronoms :

- N + Adj
- N + SP
- Pronoms 1°, 2°, 3° personne du singular et pluriel

- Forme impersonnelle se
- Pronoms relatifs : que, qui, dont

Quant aux formes verbales, nous avons analysé

- temps : présent/passé
  - personne : 1°, 2°, 3°
  - nom : singulier/pluriel
  - formes de la 1°, 2°, 3° personne actives et passives
- mode
  - indicatif/subjonctif/impératif/ conditionnel

Nous avons observé en outre certaines prépositions, des conjonctions simples et complexes, à savoir :

- préposition *de*
- conjonctions de coordination : *et, ou, ni, mais*
- conjonctions subordonnées : *parce que*, etc.
- conjonctions subordonnées complexes : *par conséquent, de sorte que, à moins que...*

Ainsi que certaines marques métadiscursives

- lemme : *définir, désigner, appeler, sous-entendre*
- lemme : *connaître, définir, entendre + comme*
- lemme : *entendre par*
- lemme : + lemme : *vouloir dire*
- lemme : *recevoir + le nom de*
- *c'est à dire*
- *c'est*
- *ou bien*

Les résultats auxquels nous sommes arrivés peuvent être résumés dans les tableaux suivants.

	Langue générale	**LSP**
Noms	1.218.815	1.302.211
Adj qualificatifs	381.813	430.576
Verbs	684.530	624.766
Determinants	612.499	659.823
Préposition de	366.827	457.584
Conjonctions	239.865	235.434
Adverbes	231.341	202.956

	TG	TE
Adj qualificat.	*381.813*	*430.576*
N+Adj	***150.386*** ***(38,07%)***	***225.856*** ***(42,68%)***
N+SP	*244.635* *(61,93%)*	*303.469* *(57,33%)*
N+participe	--	--

	TG	TE
Formes personnelles	497.278	454.947
Formes non personnelles	187.252	169.819

	TG	TE
Indicatif	313.992	219.648
Subjonctif	9.437	8.315
Ambigues Imperatif-Indicatif	115.917	120.258
Ambigues Imperatif-Sbjonctif	29.614 (0,72%)	41.202 (0,88%)
Conditionnel	9.378	7.612

	TG	TE
présent	287.983	312.423
passé	***148.318***	***40.079***

	TG	TE
1ère personne	36.243 (12,47%)	26.190 (11,61%)
2ème personne	4.525 (1,56%)	3.316 (1,47%)
3ème personne	249.989 (85,9 %)	196.049 (86,9 %)
1ª singulier/pluriel	***23.270***/12.973	***12.472***/13.718
2ª singulier/ pluriel	4.214/311	3.210/106
Total formes sing	***174.904 (63,08%)***	***102.389 (36,92%)***
Total formes plur.	115.853 (48,48%)	123.166 (51,52%)

	TG	TE
PASSIVE	3.469	3.562
ACTIVE		
1r sing/plu	16/17	0/0
2e sing/plur	0/0	1/0
3esing/plur	1.8292/1.544	1.570/1.991

	TG	TE
Total	120.453	105.222
que	114.204	97.391
cual, cuales	***1.216***	***3.948***
quien, quienes	***1.103***	***387***
cuyo,-a, cuyos, -as	***1.743***	***2.973***
se impersonnel	***69.867***	***97.418***

2.6 En guise de conclusion

Nous partons du principe que ce que l'on appelle les *langues* de spécialité font partie de l'ensemble de la langue en générale et qu'elles peuvent constituer des ensembles uniquement virtuels. Si nous acceptons ce principe, les langues de spécialité sont alors uniquement des variétés ou des styles de la langue toute entière. C'est sur la base des textes produits dans des situations de communication spécialisée que nous pourrons extraitre leur caractéristiques discriminantes par rapport à ceux qui sont issus de situations non spécialisés. Ces caractéristiques comportent des ressources lexicales aussi bien que morphologiques, syntaxiques et graphiques.

De tous les phénomènes que les linguistes ont considérés comme discriminants, dans cette étude empirique portant sur un assez vaste corpus nous avons pu vérifier que seuls certains traits apparaissent assez souvent dans les textes de spécialité tandis que d'autres ne peuvent pas être considérés

comme représentatifs, car d'occurrences trop peu fréquentes. En revanche nous avons pu observer des phénomènes qui n'ont pas été relevés dans les ouvrages sur les langues de spécialité.

Parmi ceux-ci nous pouvons signaler en particulier :

- Noms propres moins représentés en langue de spécialité
- Predominance N+Adj en langue de spécialité
- Pronoms de 1ª personne du singular et du plural plus présentes en langue générale
- Distribution complementaire des formes du pronom relatif (sauf que en espagnol)
- Conjonctions complexes en langue de spécialité
- Que completif en langue générale
- Conjonction o en langue de spécialité
- Conjunctions pero, porque, ni (mais, parce que, ni) en langue générale
- Marqueurs métadiscursifs en langue de spécialité, etc.

En revanche les données confirment que les traits suivants apparaissent comme significatifs dans les textes de spécialité :

- La prédominance des substantifs (par rapport à d'autres catégories ; pas plus qu'en langue générale)
- Emploi particulier de catégories grammaticales, surtout en relation avec les verbes (surtout des pronoms personnels)
- absence de la 2° personne du singulier comme du pluriel
- emploi rare de la 1° personne du singulier en faveur du *nous*
- emploi considérable de la 3° personne du singulier
- prédominance du présent de l'indicatif (par rapport aux temps passé)
- expansion adjectivale des substantifs
- nominalisation des formes verbales
- *nosotros*
- *uno*

Grâce à ces résultats nous pensons pouvoir contribuer à la caractérisation grammaticale des textes spécialisés et faciliter ainsi leur traitement automatique

BIBLIOGRAPHIE

Beaugrande, R. de, Dressler, W. (1997) Introducción a la lingüística del texto. Barcelona, Ariel.

Cabré, M.T. (1998) VariacIó pel tema. El discurs especialitzat o la variació funcional determinada per la temática : noves perspectives. En : Caplletra, Revista Internacional de Filología, Tardor, 1998, pp. 137-194.

Cajolet-Laganière, H., N. Maillet (1995) « Caractérisation des textes techniques québécois », Présence francophone 47, pp. 113-147.

Ciapuscio, G. (2003) Textos especializados y terminología. Barcelona, IULA.

Coulon, R. (1972) « French as it is written by French sociologists », Bulletin pédagogique des IUT18, p. 11-25.

Harris, Z. (1952) Discourse Analysis, Language, 28, 1-30, p. 474-494.

Hoffmann, L. (1976) Kommunikationsmittel Fachsprache – Eine Einführung, Berlin, Sammlung Akademie Verlag.

Kocourek, R. (1991) La langue française de la technique et de la science. Vers une linguistique de la langue savante. Wiesbaden, Oscar Branstetter.

L'Homme, M.C. (1993) Contribution à l'analyse grammaticale de la langue de spécialité : le mode, le temps et la personne du verbe dans quelques textes,scientifiques écrits à vocation pédagogique. Québec, Université Laval.

L'Homme, M.C. (1995) « Formes verbales de temps et texte scientifique », Le langage et l'homme, 31(2-3), p. 107-123.

Lauffler-Laurian, A.M. (1983) Typologie des discours scientifiques : deux approches, *Études de Linguistique Appliquée*, 51.

Lauffler-Laurian, A.M. (1984) Vulgarisation scientifique : formulation, reformulation, traduction, *Langue Française*, 64, p. 109-125.

Opitz, K. (1980) "Language for Special Purposes. An intractable presence", Fachsprache 2(2), p. 21-27.

Sager, J.C., Dungworth,D. (1980) *English Special Languages*. Wiesbaden, Oscar Brandstetter Verlag.

ANNEXE : THESES ET MEMOIRES

Les thèses et mémoires suivants ont été réalisés en exploitant les données du corpus :

Araceli Alonso : *Descripción y análisis de los sufijos nominalizadores en el área del medio ambiente* / Description et analyse des suffixes de nominalisation dans le domaine de l'environnement

Rosanna Folguerà : *Adjectius en el discurs especialitzat : una primera descripció deis adjectius en els textos del genoma humá* / Adjectifs en discours spécialisé : une première description des adjectifs dans les textes sur le génome humain

Vanesa Vidal : *Aproximación al fenómeno de la combinatoria verbo-nominal en el discurso especializado en Genoma Humano* / Une approche du phénomène de la combinaison verbe-nom dans le discours spécialisé sur le génome humain

Gabriel Quiroz : *Las unidades sintagmáticas extensas especializadas en inglés y en español : descripción y clasificación en un corpus de genoma* / Les unités syntagmatiques développées spécialisées en anglais et en espagnol : description et classification dans un corpus de génomique

John Jairo Giraldo : *Análisis y descripción de las siglas en el discurso especializado de Genoma humano y Medio ambiente* / Analyse et description des sigles en discours spécialisés du génome humain et en environnement

Iria de Cunha : *Hacia un modelo lingüístico de resumen automático de artículos médicos en español* / Vers un modèle linguistique du résumé automatique des articles de médecine en espagnol

Rogelio Nazar : *Aproximación cuantitativa al mapeo conceptual* / Approche quantitative de la carte conceptuelle

Carles Tebé : *La representació conceptual en terminologia : l'atribució temàtica en els bancs de dades terminològiques* / La représentation conceptuelle en terminologie : l'attribution de domaine dans les banques de données terminologiques.

Ricardo Guantiva : *Terminología y variación vertical : clasificación de textos en niveles de especialización a partir del análisis del tipo y la densidad de las unidades terminológicas* / Terminologie et variation verticale : classification de textes en niveaux de spécialisation à partir d'analyses de type et la densité d'unités terminologiques.

Ona Domènech : *Textos especialitzats i variació vertical : la diversitat terminològica com a factor discriminant del nivell d'especialització d'un text* / Textes spécialisés et variation verticale : la diversité terminologique comme facteur discriminant du niveau de spécialisation d'un texte.