

Des corpus représentatifs : de quoi, pour quoi, comment ?

B. Habert

UMR 8503 — ENS Fontenay/Saint-Cloud — bh@ens-fcl.fr

...il n'y a pas de caractérisation globale du langage dans son ensemble qui soit satisfaisante... (Biber, 1993, p. 220)

La francophonie cède à l'engouement pour les corpus, avec quelque retard par rapport aux initiatives et aux recherches anglo-saxonnes. Les rencontres et les projets s'enchaînent, non sans quelque confusion : le mot *corpus* est tiraillé dans des directions parfois bien éloignées. La réalité même des corpus a en outre beaucoup évolué. La vieille question de la représentativité des corpus resurgit. Il importe d'évaluer si les termes mêmes dans lesquels elle se posait se sont ou non déplacés.

La section 1 est consacrée aux corpus « nouveaux » mais aussi au changement d'acteurs : créateurs ou utilisateurs de corpus. La seconde section rappelle deux positions classiques : l'accroissement de la taille des données disponibles sous forme électronique fait de ces données des échantillons de plus en plus représentatifs des usages langagiers vs la diversité langagière est encore mal connue et suppose de constituer des corpus visant à rendre compte de la variation linguistique. Cette section oppose également la tradition anglo-saxonne des *corpus de référence* aux regroupements plus opportunistes qui ont aussi cours. Deux directions doivent en fait être explorées simultanément pour améliorer l'adéquation d'un corpus aux utilisations qui en sont faites : l'analyse précise et la mémorisation des conditions de production du corpus et de ses composants (section 3) ainsi que la mesure de l'hétérogénéité interne, en termes linguistiques, du corpus (section 4). La section 5 indique les tâches et les contraintes concrètes qui découlent des deux directions esquissées dans les deux sections précédentes. La section 6 aborde enfin les nouvelles conditions de l'articulation entre intuition et attestation.

1 Les corpus ont changé, leurs « facteurs » et leurs utilisateurs aussi

C'est sous l'angle de la représentativité que nous dépeignons les renouvellements. Nous n'abordons donc ni le niveau d'annotation des corpus¹ ni les mises en relation de corpus².

1.1 Les corpus nouveaux sont arrivés

1.1.1 Géométrie variable des corpus

« Réservoirs » à corpus

À côté des corpus « fermés », mis au point une fois pour toute, existent désormais des « réservoirs à corpus ». Les données signalétiques attachées à chaque composant permettent de réaliser « à façon » un corpus répondant à une recherche particulière. Le BNC (*British National Corpus*) qui est présenté *infra* constitue le meilleur représentant de ces regroupements de textes. Les textes littéraires du moyen français à nos jours rassemblés par l'INaLF (*Institut national de la langue française*) dans *Frantext* fournissent un autre exemple. On parlera alors plutôt de *base textuelle* que de corpus : c'est l'opération de choix raisonné parmi les composants disponibles qui crée un corpus³. Nous proposons d'ailleurs une définition de *corpus* encore plus restrictive que celle de (Sinclair, 1996, p. 4)⁴ : un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et *extra-linguistiques* explicites pour servir d'échantillon d'emplois déterminés d'une langue⁵.

Corpus « ouverts »

Beaucoup de corpus constituent des ressources achevées, dès lors immuables (sauf à en extraire des sous-corpus). À l'inverse, avec la possibilité de « capter » en continu des données dans certains secteurs (les fichiers de composition de grands journaux, par exemple) est apparue la notion de *corpus de suivi* (*monitor corpus*) (Renouf, 1993). Par définition, un tel corpus ne cesse de croître. Et il devient alors possible d'étudier l'évolution de certains phénomènes langagiers : néologismes, emplois privilégiés à un moment donné de certains suffixes ou préfixes, etc., un peu comme les éditions papier de certains dictionnaires d'usage (*Le Petit Larousse*, *Le Petit Robert*) servent de « sonde » sur le lexique et ses changements. Les CD-ROMs du journal *Le Monde* permettent aujourd'hui de de telles analyses pour le français. D'autres corpus accueillent sans cesse de nouveaux composants. C'est le cas du corpus de points de vue des acteurs sociaux constitué à des fins de veille sociale interne à la Direction des Études et Recherches d'EDF (projet *Scriptorium*)⁶.

¹Aux corpus « nus » des années 70 se sont ajoutés les corpus assortis d'étiquettes morpho-syntaxiques, puis au début des années 90, les corpus munis d'arbres syntaxiques partiels ou complets. La période actuelle est consacrée à des annotations plus ambitieuses : étiquetage sémantique ; marques de co-références ; identification d'entités comme les noms propres de personnes, de sociétés, de lieux ; transcription phonétique alignée avec le signal sonore... On se reportera à (Habert *et al.*, 1997) pour une présentation détaillée – mais déjà datée – des niveaux d'annotation et de leur utilisation en linguistique.

²*Corpus comparables* : les textes, dans des langues ou des états de langue différents, sont rassemblés selon des critères similaires, en ce qui concerne le domaine, le « genre »... ; *corpus parallèles* : les textes sont en relation de traduction ; *corpus alignés* : on indique la correspondance exacte entre des traductions : phrase à phrase ou constituant à constituant, etc.

³Le fait de pouvoir sélectionner la base entière ne menace pas ce choix terminologique : c'est en fin de compte valider les choix qui ont conduit à la sélection des composants.

⁴« Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage. »

⁵J'ajoute, pour des raisons que ces pages essaient de préciser, les critères extra-linguistiques : les critères purement linguistiques ne sont pas dans l'immédiat suffisants, ne serait-ce parce que nous manquons à la fois de typologies linguistiques de textes et d'outils de typologisation (cf. sections 3 et 4). Par ailleurs, Sinclair parle d'« échantillon du langage ». Notre ignorance de la population d'événements que constitue un langage dans son ensemble (cf. section 2) m'amène à vouloir des objectifs plus limités et plus spécifiques.

⁶Ce projet, qui fait l'objet d'un contrat pour la période 1997-2000 entre la Direction des Etudes et Recherches d'EDF et l'ENS de Fontenay/Saint-Cloud, a pour objectif la constitution d'un corpus de 20 millions de mots. Les documents rassemblés sont extrêmement variés tant pour le format que pour le type

Corpus éphémères vs « persistants »

Certains corpus sont constitués pour une recherche ponctuelle. S'il est envisageable parfois de les intégrer dans une base de textes, leur réutilisation dans une autre perspective ne s'impose pas toujours. Certains glanages de données textuelles n'ont même pas vocation à être conservés au delà de l'étude qui les a nécessités. Par exemple, l'étude de la structuration des interactions sur la Toile suppose de prendre des échantillons⁷ dans différents cadres : forums électroniques, pages personnelles, etc. Mais, la Toile évoluant rapidement, il faut constituer régulièrement de telles « carottes » de sondage (même si les précédentes peuvent offrir des points de comparaison). Au rebours, d'autres bases de textes sont patiemment complétées pour fournir un corpus aussi représentatif que possible d'un emploi déterminé du langage.

1.1.2 Changements de paramètres

Butinage, glannage et pillage

Les corpus étaient naguère protégés par la complexité même de leur constitution et de leur gestion. L'usage en était sinon privé du moins restreint à une petite communauté d'utilisateurs. La facilité actuelle d'accès aux ressources (par numérisation ou par simple copie) rend cruciale la résolution en amont des problèmes juridiques tant par rapport aux ayants droits sur les documents primaires que par rapport aux institutions qui ont ajouté de la valeur en fournissant des versions électroniques et des annotations. Le statut juridique des données est en effet souvent incertain, ce qui obère leur réutilisation. Les pillages s'accompagnent en effet d'un « oubli », bien compréhensible, de l'édition (électronique ou papier dans le cas d'une numérisation) mise à contribution (involontaire).

Homogénéisation *a priori* / *a posteriori*

Le coût important de la mise sous forme électronique d'un corpus allait auparavant de pair avec l'édition de normes de saisie⁸. C'est maintenant l'abondance de données et la multiplicité des formats qui prévalent : pages HTML issues de la Toile, fichiers provenant de traitements de texte ou de reconnaissance optique, etc. L'homogénéisation s'opère désormais *a posteriori*⁹.

Gigantisme

Une dizaine d'années du journal *Le Monde* sur CD-ROM représentent une masse textuelle dépassant ce qui a été engrangé dans *Frantext* en 40 ans.

« Laïcisation »

Disposer de texte sous forme électronique, même en grande quantité, n'est plus l'apanage des institutions. De nombreux logiciels sont également « à portée de la main » tant par leur prix que par leur facilité de maniement. Les ressources et outils permettant de travailler sur corpus se multiplient¹⁰. Plusieurs étiqueteurs sont accessibles pour le français. Il en sera bientôt de même pour les analyseurs syntaxiques automatiques ou *parseurs*. C'est déjà le cas pour les outils statistiques (sans compter les facilités offertes par les suites bureautiques).

Les micro-ordinateurs offrent désormais l'environnement matériel et logiciel nécessaire. Le temps des traitements (et des sauvegardes) sur gros systèmes est révolu. Celui des informaticiens assurant les « passages » et les ajustements logiciels aussi. Il en résulte la nécessité pour le « corpiste » d'assurer tout ou partie des tâches découlant du recours au corpus : normalisation, correction, choix de logiciels, traitements, etc., au prix parfois d'incohérences et de « bricolages » pas toujours documentés.

1.2 Nouveaux « facteurs » de corpus

La tradition anglo-saxonne de linguistique descriptive s'appuyant sur les corpus électroniques, qui s'est maintenue obstinément malgré la disqualification apriorique du recours aux corpus dans le paradigme chomskyen, a reçu ces dernières années un appui vigoureux et inattendu de la communauté du traitement automatique du langage (TAL). Cet appui découle de la prise de conscience progressive d'une inadéquation relative des paradigmes utilisés pour le TAL. En effet, la sophistication des formalismes utilisés ne débouche pas toujours sur des systèmes de traitement fiables et efficaces. Deux explications sont généralement avancées. Tout d'abord, un système de TAL a besoin de ressources (dictionnaires, grammaires) à la fois très vastes (en nombre d'entrées lexicales et de règles) et très détaillées (concernant les conditions syntaxiques d'emploi des mots, par exemple). Les ressources actuelles sont notoirement insuffisantes, surtout pour ce qui est de la finesse de la description. En second lieu, leur amélioration, semble-t-il, n'est ni uniquement ni même principalement à chercher dans des nouvelles études « en chambre » mais plutôt dans l'observation des larges ensembles de données textuelles qui sont maintenant disponibles.

1.3 Des « textuaires » multiples

Les travailleurs du texte électronique, les « textuaires »¹¹ sont désormais légion. Aux spécialistes d'analyse de discours des années 70, aux sociologues et ethnométriciens, aux linguistes « de terrain », se sont adjoints, en force, les spécialistes du TAL et ceux de la recherche d'information (*information retrieval* – cf. (Sparck-Jones & Willett, 1997)). Autant dire que les corpus requis ne sont pas les mêmes, ni en taille, ni par leur structure et leur format. La convergence apparente des intérêts ne doit pas masquer les divergences théoriques et pratiques. C'est ce qui nous a amenés à parler des linguistiques de corpus dans (Habert *et al.*, 1997).

de données langagières : tracts, extraits de livre, presse syndicale, presse d'entreprise, comptes-rendus de comités d'entreprise, transcriptions de messages syndicaux enregistrés, etc.

⁷ Il existe des « aspirateurs à Web » qui permettent de telles récoltes.

⁸ Comme (Labbe, 1990) ou (Lafon *et al.*, 1985) pour le traitement statistique du lexique ou *lexicométrie*.

⁹ (Habert *et al.*, 1998) présente les standards et les savoir faire nécessaires. (Heiden, 1999) montre les conséquences de cette situation pour la lexicométrie.

¹⁰ Pour le projet *Scriptorium*, les logiciels utilisés en aval sont également variés : associations thématiques et lexicales privilégiés, via ALCESTE (Reinert, 1993)(Lebart & Salem, 1994) ; traitements statistiques divers avec SAS ; traitements linguistiques avec LEXTER (Bourigault, 1993), etc.

¹¹ Le mot est de Lucien Fevre dans sa leçon inaugurale au Collège de France du 13 décembre 1933 (Cerquiglini, 1989, p. 17) et il y est négatif.

2 Des corpus représentatifs : de quoi ?

Curieusement, l'expression *corpus représentatif* se rencontre parfois sans que l'on précise quelle population langagière le corpus en cause est censé représenter : le français dans son ensemble, la langue littéraire, la langue familière, un langage spécialisé... D'un point de vue statistique, on peut considérer un corpus comme un échantillon d'une population (d'événements langagiers). Comme tout échantillon, un corpus est passible de deux types d'erreurs statistiques qui menacent les généralisations à partir de lui (Biber, 1993, p. 219–220) : « l'incertitude » (*random error*) et la « déformation » (*bias error*). L'incertitude survient quand un échantillon est trop petit pour représenter avec précision la population réelle. Une déformation se produit quand les caractéristiques d'un échantillon sont systématiquement différentes de celles de la population que cet échantillon a pour objectif de refléter. Un extrait de 2 000 mots d'une interview de F. Mitterrand par Y. Mourousi ne permet guère d'extrapoler et d'en tirer des conclusions sur le français mitterrandien ou sur l'interaction journaliste-homme politique. Utiliser les articles de la seule rubrique *Économie* du *Monde*, quel que soit le volume textuel rassemblé, risque fort de déboucher sur une image déformée du français employé par ce journal.

2.1 « Gros, c'est beau » vs « l'insécurité dans les grands ensembles »

Deux positions constituent les pôles entre lesquels se répartissent les créateurs de corpus (Pery-Woodley, 1995). « Gros, c'est beau » pourrait être le slogan de la première. La conviction sous-jacente est que l'élargissement mécanique des données mémorisables (les centaines de millions de mots actuelles deviendront à terme des milliards) produit inévitablement un échantillon de plus en plus représentatif de la langue traitée. Si l'on n'arrive pas à cerner précisément les caractéristiques de l'ensemble des productions langagières, il ne reste qu'à englober le maximum d'énoncés possibles. À terme, la nécessité de choisir finirait par s'estomper. La seconde approche, plus sensible aux variations propres aux données textuelles et à « l'insécurité dans les grands ensembles »¹², constitue des ensembles aux conditions de production et de réception plus nettement définies et corrélées à leurs caractéristiques langagières. C'est cette logique qui inspire les « facteurs » de corpus de référence.

2.2 L'héritage des corpus de référence

2.2.1 La tradition des corpus « panachés »

Plusieurs projets de constitution de *corpus de référence* ont été menés à bien aux États-Unis ou en Angleterre : Brown¹³, LOB¹⁴... « Un corpus de référence est conçu pour fournir une information en profondeur sur une langue. Il vise à être suffisamment grand¹⁵ pour représenter toutes les variétés pertinentes de cette langue et son vocabulaire caractéristique, de manière à pouvoir servir de base à des grammaires, des dictionnaires et d'autres usuels fiables » (Sinclair, 1996, p. 10). Il s'agit donc d'associer étroitement deux caractéristiques : une taille suffisante et la diversité des usages représentés. Le British National Corpus les allie effectivement.

2.2.2 Une réalisation exemplaire : le British National Corpus (BNC)

Ce corpus de 100 millions de mots étiquetés¹⁶ répond à l'objectif de constituer un ensemble de données textuelles aux conditions de production et de réception définies avec précision et qui soient représentatives d'une grande variété de situations de communication.

Il mêle oral (10 %) et écrit (textes de fiction à partir de 1960 et textes « informatifs » à partir de 1975). À ce titre, c'est, sous réserve d'inventaire, le plus gros corpus d'oral au monde.

En ce qui concerne l'écrit, les variables prises en compte sont le domaine (textes informatifs et textes de fiction), le support (livres, périodiques, discours), la datation et la diffusion (sélection parmi les listes des meilleures ventes, celles de prix littéraires, les indications de prêts en bibliothèque, etc.). L'accent mis sur la diffusion effective certifie la représentation d'usages majeurs de l'anglais.

Pour l'oral, des conversations spontanées ont été recueillies à partir d'un échantillonnage démographique en termes d'âge, de sexe, de groupe social et de région. Ont été également intégrées des transcriptions d'interactions orales typiques dans divers domaines : affaires (réunions, prises de parole syndicales, consultations médicales ou légales) ; éducation et information (cours et conférences, informations radio-télévisées) ; prises de parole publiques (sermons, discours politiques, discours parlementaires et légaux) ; loisirs (commentaires sportifs, réunions de clubs). Ces interactions institutionnelles ont été jusqu'à présent relativement mal représentées dans les corpus d'oral.

2.2.3 Panachage et échantillonnage

La volonté de « représenter » une diversité maximale de situations de communication dans un corpus de référence s'accompagne souvent dans les travaux anglo-saxons d'une démarche d'échantillonnage. La logique de cette position conduit à « équilibrer » en taille les échantillons retenus – de taille limitée (de 2 000 à 40 000 mots) –, voire à ne pas retenir des empanes de texte continus, de manière à ne pas risquer de sur-représenter des « lieux » du texte aux caractéristiques particulières (l'introduction par exemple). L'échantillonnage touche donc à la fois le choix des documents à intégrer et la partie de ces documents à conserver¹⁷. Ce « saucissonnage » rend par contre impossible l'étude des changements de corrélations de traits linguistiques au fil des textes.

¹² C'est le titre de (Geffroy & Lafon, 1982).

¹³ Ce corpus mis au point par Francis et Kucera, de l'Université Brown, aux États-Unis, comprend un million de mots (500 extraits de 2 000 mots) et regroupe des textes américains produits en 1961 et relevant de 15 « genres » différents. Il a été soigneusement étiqueté : une catégorie morpho-syntaxique est attachée à chaque mot. Par sa mise dans le domaine public, il a joué un rôle moteur dans le renouveau des recherches sur corpus.

¹⁴ Acronyme de Lancaster-Oslo-Bergen, les universités qui l'ont mis au point. Ce corpus a été explicitement conçu pour être l'équivalent anglais du corpus de Brown. Il comprend également 500 extraits de 2 000 mots relevant des mêmes genres, mais de textes anglais cette fois-ci, produits aussi en 1961.

¹⁵ On notera que la taille correspondant à un « gros corpus » ne cesse de croître. Au début des années 80, un million de mots étiquetés correspondait à une réalisation d'ampleur. Quinze ans après, ce sont cent millions de mots étiquetés que le BNC rend accessibles.

¹⁶ <http://info.ox.ac.uk/bnc/>

¹⁷ Cf. (Bronckart *et al.*, 1985, p. 69). Biber consacre plusieurs articles à la taille des échantillons à utiliser en fonction des phénomènes à étudier : (Biber, 1990)(Biber, 1994).

2.3 Misère de la philologie...

Les récents projets de constitution de corpus en France ne s'inscrivent pas vraiment dans cette perspective de représentation planifiée de la diversité langagière. Ils reposent plutôt sur l'« assemblage » ou le recyclage de données préexistantes.

C'est le cas par exemple du corpus réalisé¹⁸ dans le cadre du projet européen *Parole* (1996–1998). L'objectif était pour 12 langues, dont le français, de fournir un corpus de 20 millions de mots, datant pour l'essentiel (80 % au moins) d'après 1980. 250 000 mots devaient avoir été étiquetés et vérifiés quant à la partie du discours, et 50 000 mots dans ce sous-ensemble vérifiés quant à l'ensemble des traits attribués. Le corpus devait provenir pour 60 % de journaux, pour 30 % de livres, pour 10 % de périodiques (les 10 % restant pouvant relever de diverses provenances). Les 20 093 099 mots obtenus se répartissent à l'issue du projet en 2 025 964 mots de transcriptions de débats au parlement européen, de 3 267 409 mots issus d'une trentaine d'ouvrages de disciplines variées (en sciences humaines) fournis par CNRS-Éditions, de 942 963 mots provenant des notes de vulgarisation de la revue CNRS Info et d'articles sur la communication de la revue Hermès et enfin de 13 856 763 mots correspondant à 25 654 articles provenant du choix aléatoire de numéros entiers parmi ceux des années 1987, 1989, 1991, 1993 et 1995 du journal *Le Monde* (Naulleau, 1998). Les données rassemblées sont certes variées, mais sans pouvoir prétendre pour autant à représenter de manière cohérente les emplois principaux du français. La presse n'est présente que par un seul journal, quotidien. La presse régionale, les hebdomadaires, la presse spécialisée sont absents. Les langages techniques et scientifiques également (hormis les sciences humaines et des notes très brèves de vulgarisation dans des domaines extrêmement variés). Cette perspective se rapprochant d'un « vide-grenier » est explicite dans le projet SILFIDE (*Serveur Interactif sur la Langue Française, son Identité, sa Diffusion et son Étude*) de l'AUFPEL-UREF (pour 1996–2000)¹⁹ où il s'agit de rendre accessibles des ressources et des outils linguistiques pré-existants dans un cadre logiciel unifié.

Dans tous les cas, se trouvent agrégés des documents avant tout parce qu'ils sont faciles d'accès : leur mise en relation n'a pas été réellement pensée. On part non pas des emplois du français que l'on souhaite représenter mais des données disponibles et des annotations dont on veut les enrichir.

2.4 ...philologie de la misère

Dans ce qu'on pourrait appeler des regroupements « opportunistes », certains emplois du français sont privilégiés tandis d'autres restent dans l'ombre.

2.4.1 Emplois du français majorés de facto

le français d'hier Les difficultés juridiques et la frilosité des éditeurs ralentissent ou bloquent l'accès aux textes contemporains, en particulier littéraires. Cet obstacle pousse à recourir aux textes « libres de droits ». Dans la pratique, il s'agit souvent de textes datant au mieux du siècle dernier²⁰. Or le français, comme toute langue, évolue (Marchello-Nizia, 1999). Représenter le français actuel et le français de jadis et naguère n'est pas la même chose.

le français écrit Le coût de la transcription manuelle précise de l'oral spontané²¹ aboutit à une sous-représentation manifeste de l'oral, et en particulier de certains types d'interaction (le cours ou la conférence, les discours politiques...).

le « bon usage » Le recours aux écrits littéraires pour étayer les entrées des dictionnaires comme *Le Littré*, *Le Robert* ou *Le Trésor de la Langue Française* a fait place ces dernières années au curieux privilège accordé aux versions électroniques des journaux *Le Monde* et *Le Monde diplomatique*. Cette situation tient sans doute à la politique offensive et novatrice de cette société de presse en matière d'accès électronique à l'information. Elle s'explique probablement aussi par le magistère qui est plus ou moins accordé implicitement à cet organe en matière de langage. Irait-il aussi aisément de soi de s'appuyer sur *La Montagne*, *Sud-Ouest* ou *L'Yonne Républicaine* ?

2.4.2 Emplois du français sous-représentés

le français non-hexagonal Français parlé au Québec, en Suisse, en Belgique, en Afrique et au Moyen-Orient ...

le français scientifique et technique Il peut différer de la langue « générale » non seulement par le lexique, mais éventuellement par des constructions ou des modes d'organisation globales distincts.

les français non-standard À côté des emplois « policés » de la langue, qui ont bénéficié de relectures et de corrections, comme les livres, les journaux, se rencontrent les textes peu ou pas ou mal révisés. C'est le cas des énoncés dictés (comme les comptes rendus d'hospitalisation, les réponses à des enquêtes), du courrier électronique, des thèses et rapports techniques. La généralisation du « tout-électronique » accentue la place de ce « français ordinaire » (Gadet, 1997), par exemple lors de la fourniture d'exemplaires prêts à cliquer (*camera-ready*) de livres ou d'articles.

2.5 Un corpus représentatif : pour quoi

Sinclair, on l'a vu, fixe comme but à un corpus de référence de représenter « toutes les variétés pertinentes [d'une] langue ». À partir d'un tel objectif, louable, peut-on dresser réellement la liste des types d'énoncés à intégrer ? Dans le cadre d'un projet de corpus échantillonné pour le français, *Corpus CLEF*²², P. Zweigenbaum a repéré les registres suivants dans le domaine médical :

1. Dossier patient (hospitalier et médecine de ville)
 - compte rendu d'hospitalisation (CRH), opératoire (intervention chirurgicale), d'examen : d'imagerie (radiologie, scanner, échographie, endographie...); d'exploration fonctionnelle (respiratoire, neurologique, anatomopathologie, ECG, EEG, électromyogramme); biologique, biochimique (ex. : antibiogramme)...; notes de suivi.
 - lettre de correspondant : de sortie (version plus synthétique du CRH); pour adresser un patient à un autre médecin.
 - prescription : ordonnance; demande d'examen.
2. Enseignement : livre de cours; polycopié; QCM, question de cours.
3. Ressources : monographie sur médicament (Vidal); notice de médicament; références médicales opposables; guide de bonne conduite; protocole d'essai clinique; dictionnaire; encyclopédie
4. Publications (principalement scientifiques): mémoire de thèse; article (de type scientifique; dépêche AFP santé; communication institutionnelle, d'entreprise); péri-article (résumé d'article scientifique; notice bibliographique); forums de discussion, listes de diffusion électroniques en médecine; autorisation de mise sur le marché d'un médicament.

¹⁸ Par Georges Vignaux (INaLF) et moi.

¹⁹ Participants : CLIPS–GETA, CRIN, INaLF, LIMSI, LPL.

²⁰ Comme ceux mis en ligne par ABU – *Association des bibliophiles universels* (<http://cedric.cnam.fr/ABU/>).

²¹ De l'ordre d'1/2 heure pour une minute d'ora.

²² <http://www.biomath.jussieu.fr/CLEF/>

5. Oral : cours ; staff – réunion d'équipe soignante (voire staff à distance : vidéoconférence) ; entretien médecin / patient ; exposé de conférence ; film documentaire.

P. Zweigenbaum fournit ainsi un inventaire – probablement encore à compléter – des types d'énoncés en circulation dans cette sphère d'activité. Si l'on souhaite faire un corpus représentatif de ce domaine, faut-il faire figurer un échantillon de chacun des « genres » qui ont été isolés ? Ou bien, peut-on considérer (mais sur quelles bases objectives ?) que certains « genres » partagent suffisamment de caractéristiques linguistiques pour qu'il suffise de représenter l'un d'entre eux (le livre de cours et le polycopié par exemple) ?

On le voit, au moins trois visions distinctes de la représentativité peuvent conduire les choix :

privilégier les conditions de réception C'est le choix du BNC ;

favoriser les conditions de production Ce serait le cas d'un corpus de langage médical qui rassemblerait des échantillons de tous les types d'énoncés repérés par P. Zweigenbaum ;

retenir les types de textes On regroupe alors des énoncés dont on postule qu'ils sont similaires sur le plan linguistique.

Quel que soit le choix (ou la combinaison de choix) fait, nous ne disposons pas actuellement de données empiriques nécessaires pour pouvoir être confiants dans la validité du corpus résultant. Notre connaissance de la « population » des données langagières est encore extrêmement fragmentaire. Pour l'oral, par exemple, il n'est pas évident de spécifier les « genres » les plus produits ou les plus importants en réception.

Par ailleurs, on peut se demander si « les variétés de langage représentées correspondent à ce pour quoi est fait le corpus en cause » (Biber, 1993, p. 220). Pour constituer une terminologie médicale, il est probable par exemple que certains des « genres » recensés par P. Zweigenbaum ne sont pas pertinents (l'entretien médecin / patient, voire l'ordonnance et la demande d'examen).

Dans tous les cas, améliorer la représentativité d'un corpus consiste à préciser la production et la réception de chacun de ses composants, en lien avec les motifs qui ont conduit à la création du corpus, mais aussi à pouvoir déterminer sur des bases objectivables les différents emplois du langage auxquels on s'intéresse. Ce sont ces deux dimensions, externe et interne, de la représentativité, qui sont abordées dans les deux sections suivantes.

3 Caractérisations *a priori* de textes

On peut opposer sommairement deux types de classification :

1. La première opère *a priori*. Elle repose sur les conditions de production des textes (section 3.1), sur les buts visés par les textes (section 3.2), sur l'inscription dans des « genres » (section 3.3) – type de classification qui sera particulièrement détaillé –, sur l'emploi ou non de certaines marques linguistiques (section 3.4) ou bien sur une combinaison de ces critères.
2. La seconde, développée section 4, procède *a posteriori* : les types obtenus reposent sur les propensions d'un groupe de textes à recourir à un ensemble de traits linguistiques et à en éviter d'autres.

3.1 Définition de la situation de communication

Une des difficultés à ce stade est que font partie de la situation de communication non seulement les données objectives sur le destinataire et le destinataire mais également les représentations que possède le destinataire de lui-même et de son « public » (Kerbrat-Orecchioni, 1980, p. 22–26).

3.2 Précision de la fonction visée

Le classement porte sur la fonction du texte : distraire, informer, convaincre, etc., dans la tradition par exemple de Condillac qui distinguait le *didactique*, le *normatif* et le *descriptif* (Chiss, 1987, p. 24) ou dans la lignée de décomposition de la communication en six facteurs par Jakobson (Jakobson, 1963) qui associe à chaque facteur une fonction linguistique déterminée²³. Les textes se distinguaient dans cette perspective par la domination de telle ou telle fonction, même si chacun d'eux fait appel à toutes.

À la suite de Werlich (1975), qui distinguait les textes descriptifs, narratifs, expositifs, argumentatifs et instructifs, J.-M. Adam (Adam, 1985) proposait sept types textuels de base²⁴.

3.3 Rattachement à des thèmes et à des domaines

Les classifications en termes de sujets et de domaines sont sujettes à caution²⁵. Trop raffinées, elles se trouvent vite battues en brèche par l'évolution des sociétés, des techniques et des mentalités. Grossières, elles sont trop floues pour être utiles. Il faut alors s'en servir comme d'un débroussaillage imparfait mais commode.

²³ Cf. (Kerbrat-Orecchioni, 1980, p. 11–29) pour une critique détaillée de cette analyse qui a fait florès.

²⁴ Narratif (reportage, fait divers, roman, nouvelle, conte, récit historique, parabole, publicité narrative, film, bande dessinée), descriptif (description, inventaire, guide touristique), explicatif (discours didactique ou scientifique), argumentatif (essai, publicité), prédictif (prophétie, bulletin météorologique, horoscope), conversationnel (interview, dialogue) et rhétorique ou poétique (poème, chanson, slogan, proverbe, dicton, maxime). Le partage narratif/descriptif oppose énoncés de faire/énoncés d'état.

²⁵ J. Sinclair (Sinclair, 1996) fournit les thèmes et domaines utilisés dans 20 corpus différents. Voici les catégories recensées (chaque catégorie est suivie du nombre de corpus qui y ont recours) : religion (14), techniques et technologie (12), droit (11), sports (11), belles lettres (10), politique (9), histoire (8), médecine (8), philosophie (7), économie (8), éducation (7), psychologie (7), sciences (8), sociologie (8), loisirs (8), civilisations (6), physique (6), biologie (6), mathématiques (5), vie domestique (5), voyages (5), anthropologie (5), affaires militaires (5), médias et communication (5), langage (5), littérature (4), architecture (4), mode et vêtements (4), informatique (4), agriculture (4), géographie (4), écologie et environnement (3), transports (3), chimie (3), finances (3). Même si chacun des corpus concernés ne couvre pas l'ensemble des thèmes mentionnés, on ne peut qu'être frappé par le « bricolage » que manifeste cette répartition. Les catégories ne s'excluent pas forcément. En effet certaines entrées en recouvrent d'autres : *sciences* rassemble *physique*, *chimie*, *biologie*, etc. (sans compter que l'on ne sait pas si cette vedette très générale comprend ou non les sciences dites humaines comme l'anthropologie, la sociologie ou l'économie). En outre, certaines catégories semblent être en intersection : *vie domestique* et *mode*, ou encore *médias et langage*, voire *religion* et *philosophie*.

3.4 Inscription dans des « genres » ou registres

3.4.1 La catégorisation « ordinaire » des textes

En français au moins, la notion de genre reste souvent liée aux textes littéraires et à leurs subdivisions traditionnelles (comédie/tragédie/épopée...) ²⁶. Elle a pourtant depuis une quinzaine d'année pris une extension plus large, sous l'influence en particulier des travaux en analyse du discours et des écrits de Bakhtine. (Biber, 1989, p. 5–6) définit ainsi la typologie de sens commun que sont les genres (*the folk-typology of 'genres'*) : « Les genres sont les catégories de textes distinguées spontanément par les locuteurs confirmés (*mature*) d'une langue ; par exemple, les genres de l'anglais incluent les romans, les articles de journaux, les éditoriaux, les articles de recherche (*academic articles*), les discours en public, les nouvelles radiophoniques, et la conversation de tous les jours. » Pour D. Maingueneau (Maingueneau, 1996, p. 43), il s'agit de « dispositifs de communication socio-historiquement définis », comme « le fait divers, l'éditorial, la consultation médicale, l'interrogatoire policier, les petites annonces, la conférence universitaire, le rapport de stage, etc. » ²⁷. S. Branca indique (Branca-Rosoff, 1999, p. 5) : « Les usagers de la langue classifient spontanément leurs productions discursives. Par exemple, dans les médias, les journalistes, et leurs lecteurs emploient *fait divers, reportage, débats*. De même, *notes de synthèse, compte rendu...* s'entendent dans les bureaux et dans les entreprises ; *dissertation, thèse, compte rendu de lecture...* à l'Université. Au demeurant, les locuteurs s'en tiennent aux noms d'espèce et ne semblent pas utiliser souvent les termes englobants de "genre", de "modes" ou de "types". » Elle souligne d'ailleurs (*ibid.*, p. 17) qu'il s'agit de regroupements *a posteriori*, sans critères systématiques.

Dans ses travaux plus récents (Biber, 1995), Biber utilise *registre* pour cette conception élargie des genres. C'est ce terme englobant qui sera utilisé par la suite.

3.4.2 Les registres : un inventaire à géométrie variable

L'inventaire des registres est destiné à rester ouvert ²⁸. L'éventail disponible évolue en effet au fil du temps ²⁹. L'émergence d'un nouveau médium entraîne la naissance de nouveaux registres ou la transformation de registres existants (Maingueneau, 1996, p. 43). C'est le cas avec Internet (Beaudoin & Velkova, 1999) : courrier électronique, forums, « news », listes de diffusion électroniques... À l'inverse, d'autres registres reculent ou disparaissent : le sermon et la confession ont pâti de la baisse de la pratique religieuse.

L'ancrage institutionnel des registres varie en outre. Les registres majeurs sont clairement identifiés et éventuellement objets de normes. Ils sont enseignés ³⁰. L'apprentissage de certains de ces registres tient une place importante dans la scolarité (dissertation, oraux de concours, etc.). La plupart des secteurs d'activité possèdent leurs registres majeurs spécifiques dont la maîtrise est nécessaire ³¹, comme la thèse ou l'article, etc., dans le champ universitaire. Ces registres « reconnus », au nombre limité dans chaque domaine, rejettent dans l'ombre l'existence du grand nombre d'autres classes d'énoncés qui ne font pas l'objet d'une transmission explicite mais qui régulent néanmoins les discours écrits ou oraux effectivement produits. La lettre de recommandation, l'intervention à un jury de thèse, la question pendant un séminaire ou un colloque en sont des exemples pour le champ universitaire. Notons d'ailleurs, comme (Bakhtine, 1984, p. 286), que les registres oraux sont encore moins bien connus que les registres écrits. On peut supposer enfin une certaine tendance à sous-estimer la place des registres dans la production et l'interprétation des énoncés : il est difficile d'imaginer voire d'accepter que les discours les plus ordinaires soient contraints dans leur thématique, leur style, leur structure, sans que nous en ayons le plus souvent conscience.

3.4.3 Les contraintes liées à un registre

Si certains registres laissent plus de « jeu » que d'autres (Dolinine, 1999, p. 36)(Vion, 1999, p. 97) ³², les contraintes à l'œuvre portent sur le volume global « de bon aloi » (une lettre de recommandation peut-elle dépasser le recto-verso, voire le simple recto ?), l'organisation d'ensemble (les parties requises, celles qui sont optionnelles, et leur structuration) – un compte rendu d'hospitalisation rappelle l'histoire clinique du patient, ses antécédents, évoque ce qui justifie la présente hospitalisation et les traitements effectués et s'achève par les traitements nécessaires après sortie de l'hôpital ³³ –, mais aussi la « palette » linguistique disponible : temps des verbes, connecteurs ³⁴, ancrage spatio-temporel, etc. C'est le deuxième volet de contraintes qu'une DTD SGML permet de modéliser.

²⁶(Chiss, 1987, p. 12). (Branca-Rosoff, 1996, p. 193) : « Le mot *genre* est ... issu du métalangage de la rhétorique et de la poétique. Le syntagme *genres de discours (genera dicendi)* appartient à la tradition rhétorique gréco-latine qui distinguait, en fonction de situations sociales codifiées, le genre judiciaire qui s'exerce au tribunal, le genre délibératif à l'assemblée, et le démonstratif ou épictique dans les fêtes publiques. À ces lieux d'énonciation institutionnels, correspondaient des actes de langage au service d'une finalité pragmatique. »

²⁷C'est en fait la même intuition qui est à l'origine de la TEI (*Text Encoding Initiative*), la proposition de normes d'encodage pour les principaux types de textes utilisés en sciences humaines : le découpage des énoncés en grandes classes aux régularités formelles identifiables (Burnard & Sperberg Mc Queen, 1996)(Ide & Veronis, 1995). Une DTD (Définition de Type de Document – *Document Type Definition*) est dans cette optique une manière de formaliser un genre.

²⁸T. Todorov (Todorov, 1978) étudie comme « genres non littéraires » : la devinette, le discours de la magie, le mot d'esprit, les jeux de mots. P. Charaudeau (Charaudeau, 1983) étudie les genres suivants : information, genre publicitaire, instructions officielles, genre littéraire. (Petitjean, 1987) est consacré au fait-divers. D. Maingueneau (Maingueneau, 1996, p. 43) évoque également la recette de cuisine, la prière et le journal télévisé.

²⁹Voir l'apparition de la conversation comme genre au 18^e siècle et sa non-reconnaissance par la rhétorique (Branca-Rosoff, 1996, p. 194).

³⁰Voir pastichés. Voir par exemple *La réaction yellante chez la cantatrice soprano (Experimental demonstration of the tomatotopic organization in the Soprano (Cantatrix sopranica L.)*, pastiche d'article scientifique par Georges Pérec qui exerçait les fonctions de documentaliste à l'INSERM (Pérec, 1991, p. 11-32). Ou encore ce détournement des remerciements par J.-C. Anscombe « Voulez-vous dériver avec moi ? », *Communications*, n°32, 1980, p. 123 : « Je remercie de leurs critiques, conseils et suggestions les personnes suivantes : A.-M. Diller, O. Ducrot, B. Fradin, F. Récanati. J'en ai scrupuleusement tenu compte, ne tenant pas à être seul responsable des bêtises disséminées ça et là dans ce texte. »

³¹Outre la nécessité d'émettre des phrases bien formées et bien « enchaînées » : une suite de phrases correctes ne respecte pas forcément les contraintes de *cohésion* et de *cohérence* qui caractérisent un texte *réussi*. Ces contraintes ont été étudiées par le courant des grammaires textuelles (Combettes, 1988).

³²La marge de manœuvre est très faible pour les énoncés ritualisés (mariage, collation de grade).

³³Voir aussi (Petitjean, 1987) pour la structure prototypique d'un fait-divers.

³⁴Voir (Branca-Rosoff, 1999, p. 119) dans l'utilisation de certaines locutions prépositionnelles comme marqueurs d'un registre (administratif) : « compte tenu », « suite à »...

P. Charaudeau (Charaudeau, 1983, p. 50) invite à ajouter, pour chaque registre, un *contrat* implicite qui précise les droits et les devoirs du destinataire et du destinataire des textes qui en relèvent³⁵.

3.4.4 Prolifération ou regroupement de registres ?

On peut accroître presque à l'infini la liste des registres. A contrario, n'existe-t-il pas des « méta-registres » qui regrouperaient des types de texte relevant globalement des mêmes contraintes mais présentant néanmoins des variantes sensibles ? Le méta-registre *mode d'emploi* subsumerait alors des registres comme *guide de l'utilisateur*, *notice d'utilisation* (d'un appareil) et *recette de cuisine*. Le corpus de thèses fourni dans (Habert *et al.*, 1998) rassemble ainsi 10 thèses de doctorat « nouveau régime » – soutenues entre 1991 et 1997 – relevant de quatre disciplines (linguistique, informatique, biologie et économie). Au sein de ce registre bien défini, chaque discipline pourtant entraîne des contraintes particulières. Par exemple, les renvois bibliographiques dans le fil du texte sont assortis de citations, parfois longues, en linguistique. Ce n'est pas le cas en informatique où la référence bibliographique au fil du texte peut se réduire au numéro d'ordre dans la bibliographie fournie *in fine*³⁶.

3.5 Types linguistiques postulés

Jakobson (Jakobson, 1963, ch. 9) a caractérisé les *embrayeurs*³⁷ (*shifters*) comme les unités linguistiques dont la valeur référentielle nécessite de connaître les conditions de leur *énonciation*, c'est-à-dire le moment, le lieu et l'identité des co-locuteurs : dans *je viens ici demain*, l'interprétation de *je*, *ici* et *demain* suppose de connaître l'identité du locuteur ainsi que la localisation de l'énonciation dans le temps et dans l'espace.

On peut alors opposer les énoncés qui organisent leurs repérages par rapport à la situation d'énonciation, et qui recourent donc aux embrayeurs, et ceux dont les repérages reposent sur l'énoncé lui-même. Cette opposition a donné lieu à la distinction par Benveniste (Benveniste, 1966) entre *discours* et *histoire*³⁸. Dans le discours, « quelqu'un s'adresse à quelqu'un, s'énonce comme locuteur et organise ce qu'il dit dans la catégorie de la personne » (*ibid.*, p. 242) tandis que dans l'histoire « les événements semblent se raconter eux-mêmes ». Tant César dans *La guerre des Gaules* que De Gaulle dans ses *Mémoires de guerre* ont utilisé l'histoire, en parlant d'eux-mêmes à la troisième personne : ils ont ainsi changé le mode de présentation de leurs faits de guerre et de leur rôle.

À la suite de D. Maingueneau (Maingueneau, 1996, p. 34), on utilisera plutôt les termes de *plan embrayé* (\approx discours) et *plan non embrayé* (\approx histoire), moins ambigus que les dénominations choisies par Benveniste.

3.6 Malaise dans la classification ?

Ce titre, emprunté à J.-L. Chiss (Chiss, 1987), rend compte de l'éparpillement des travaux typologiques. Ce manque de convergence a même pu conduire à des tentatives de typologie des typologies. (Petitjean, 1989), en fonction des critères utilisés, distingue ainsi d'abord les *typologies de textes* ou les *typologies de séquences* (travaux de J.-M. Adam), où les critères sont homogènes et où les textes constituent le domaine de validité, ensuite les *typologies de discours* (E. Benveniste et J.-P. Bronckart), où des critères hétérogènes sont articulés dans une même perspective, la mise en situation des textes, et enfin les *typologies de genres*, aux critères totalement hétérogènes (dimensions discursive, communicationnelle et textuelle).

4 Typologies inductives de textes

Une autre optique consiste à faire émerger les types de textes – considérés comme des agglomérats de traits linguistiques – grâce à un traitement statistique de textes (étiquetés ou non). Cette démarche inductive, présentée section 4.1, peut se centrer sur des textes spécialisés (section 4.3) ou au contraire chercher à traiter la « langue générale » (section 4.2). Elle peut également s'attacher aux sous-régularités perceptibles au sein des textes, c'est-à-dire aux types qui s'y succèdent (section 4.4).

4.1 Dégager des constellations de marques linguistiques

Dans cette optique, un type de texte est défini par la cooccurrence d'un certain nombre de traits linguistiques (et éventuellement par l'évitement systématique d'autres traits). Un corpus est constitué pour examiner la répartition de traits considérés (préalablement ou *a posteriori*) comme discriminants et significatifs.

4.1.1 Constitution d'un corpus

La démarche inductive peut avoir pour objectif de confirmer/amender une typologie pré-existante et de caractériser les emplois correspondant à chaque type présumé. C'est la logique de (Bronckart *et al.*, 1985). Le postulat de départ est l'existence de trois pôles textuels ou *architypes*, liés à des situations de communication distinctes : le discours en situation, le discours théorique et la narration (*ibid.*, p. 43–44). Cinquante textes ont été recueillis pour chaque architype.

La démarche inductive peut au contraire être purement exploratoire : elle vise à *révéler* des types sans correspondance obligatoire avec des genres repertoriés.

- Elle peut opérer sur des textes relevant d'un même genre (la *résolution générale de congrès confédéral* dans (Bergounioux *et al.*, 1982) comme dans (Habert, 1983)).
- La démarche inductive peut explorer les régularités de textes relevant de genres aussi variés que possibles. C'est la ligne directrice des travaux de D. Biber (Biber, 1988)(Biber, 1989). Il examine les cooccurrences entre 67 traits linguistiques dans les 1 000 premiers mots de 481 textes d'anglais britannique contemporain écrit et oral. Ces textes proviennent de LOB et de London-Lund³⁹ et deux ensembles de lettres. Les « genres » représentés sont les quinze

³⁵Un texte peut jouer sur la violation de ce pacte. C'est le cas du *Meurtre de Roger Acroyd* d'A. Christie : le narrateur n'est autre que le criminel, mais cette identité n'est révélée qu'à la fin du roman qui viole les conventions tacites du roman policier : soit le narrateur omniscient n'est pas coupable soit son crime est donné dès le départ.

³⁶Voir aussi (Beacco, 1992, p.15–17)(Regent, 1992) pour l'étude de la variation au sein d'un même registre en fonction de différentes langues.

³⁷«...on parle aussi d'éléments *déictiques* (Bühler), d'*expressions sui-référentielles*, d'*éléments indiciels*, de *symboles indexicaux*... » (Maingueneau, 1996, p. 33).

³⁸Amendée dans (Simonin-Grumbach, 1975).

³⁹Ce corpus étiqueté (Svartvik *et al.*, 1982) totalise 435 000 mots d'anglais parlé, répartis en 87 extraits de 5 000 occurrences de locuteurs adultes ayant fait des études. Il comprend de nombreuses informations prosodiques (pauses, limites, etc.).

de LOB, deux de lettres (personnelles et professionnelles), six de London-Lund : conversations face à face ; conversations téléphoniques ; conversations publiques, débats et interviews ; (*broadcast*) ; discours improvisés (*spontaneous speeches*) ; discours préparés (*planned speeches*).

Le corpus peut être constitué de textes complets (Bergounioux *et al.*, 1982)(Sueur, 1982)(Habert, 1983)(Biber & Finegan, 1994) ou d'extraits. (Bronckart *et al.*, 1985) comme (Biber, 1988) se limitent à des fragments de 1 000 mots choisis au hasard, donnés comme suffisants pour dégager les cooccurrences souhaitées. Les motivations de cette restriction diffèrent. Dans le premier cas, c'est la lourdeur de l'étiquetage manuel des traits retenus. Dans le second, la volonté de faciliter les comparaisons entre les fréquences des traits.

4.1.2 Sélection et examen de marqueurs linguistiques

Le passage de certains mots (ou suites de mots) à des traits linguistiques (de *je et j'* à *embrayeur* par exemple) peut s'effectuer *a priori* : il s'agit alors d'un étiquetage, manuel (Bronckart *et al.*, 1985) ou automatique (Biber, 1988). Chaque texte analysé est remplacé par les étiquettes correspondant aux items de la grille employée. L'étiquetage mis en œuvre s'éloigne de l'étiquetage morpho-syntaxique « traditionnel ». Il est partiel (une partie seulement de la surface du texte est marquée) et partial. À géométrie variable, « inéquitable », il s'intéresse à des fonctionnements linguistiques très spécifiques qu'il analyse en détail tandis qu'il en laisse d'autres dans l'ombre.

Dans d'autres cas, il n'existe pas de grille pré-existante : c'est l'examen des formes employées dans le corpus qui suggère des regroupements, dont certains rejoignent néanmoins des catégories établies. Cette approche « montante » est suivie dans (Bergounioux *et al.*, 1982) et dans (Habert, 1983).

4.1.3 Mise en évidence de constellations de traits

Dans (Bergounioux *et al.*, 1982) et (Habert, 1983), un programme qui isole les éléments significativement sur-employés dans une partie d'un corpus au regard de leur emploi dans le corpus entier (Lebart & Salem, 1994, p. 172-180) est utilisé pour évaluer les phénomènes étudiés. Ce programme dégage en même temps les sous-emplois significatifs d'une partie au regard du tout⁴⁰.

Dans (Sueur, 1982)(Bronckart *et al.*, 1985)(Biber, 1988), la statistique multidimensionnelle (Bouroche & Saporta, 1980)(Saporta, 1990, ch. 8)(Lebart & Salem, 1994, ch. 3) est mise à contribution pour repérer les oppositions majeures entre associations de traits linguistiques. Elle rassemble les traits qui ont tendance à apparaître ensemble. Elle constitue dans le même temps les configurations de traits qui sont systématiquement évités par les mêmes rassemblements. Cette démarche permet d'obtenir des pôles multiples, positifs et négatifs, correspondant à ces constellations.

Seule la statistique multidimensionnelle, dont relève la seconde série de travaux, permet de certifier l'existence de corrélations, positives ou négatives, entre traits⁴¹. L'emploi du programme des spécificités n'a d'ailleurs pas pour objectif de mettre en évidence de telles corrélations. Il permet d'en percevoir certaines mais laisse une trop grande latitude à l'interprète humain.

4.1.4 Des cooccurrences de marques aux types de textes

Une interprétation relativement immédiate des regroupements opérés à partir des sur-emplois et des sous-emplois conduit dans (Bergounioux *et al.*, 1982) et dans (Habert, 1983) à opposer à chaque fois deux types. Dans (Bronckart *et al.*, 1985), l'examen des premiers facteurs est jugé valider les trois architypes postulés mais elle conduit également à proposer des types intermédiaires (*ibid.*, ch. VI, p 103-137).

Biber procède en deux temps. Les facteurs les plus significatifs sont considérés comme autant de *dimensions* pouvant caractériser un texte donné. Biber souligne que ces dimensions sont en fait des prototypes. Chacune des dimensions obtenues (cf. tableau *Dimensions de Biber et traits correspondants*) oppose deux pôles de fonctionnements textuels, mais les textes concrets se situent en des points variés des « échelles » ainsi définies. Biber calcule alors pour chaque texte ses coordonnées sur chacun de ces axes⁴². La classification automatique est alors requise pour rapprocher les textes en fonction de leurs coordonnées sur ces cinq axes. Les huit regroupements obtenus sont alors considérés comme des types de textes, au terme d'une nouvelle phase d'interprétation qui s'appuie en particulier sur l'examen des textes les plus proches du centre de chacune des classes définies.

1 – <i>Production impliquée</i>	<i>Production informationnelle</i>
Verbes “privés”, omission de <i>that</i> , négation analytique, subordinants de cause, pronoms indéfinis, relatives, questions en <i>WH</i> , modaux de possibilité, <i>do</i> comme pro-verbe, <i>be</i> comme verbe principal, présent, démonstratifs, contractions du type <i>don't</i> , 1 ^{ère} et 2 ^{ème} personne du singulier, pronom <i>it</i>	Noms, mots longs, adjectifs attributs, prépositions, adverbiaux de lieu
2 – <i>Orientation narrative</i>	<i>Orientation non-narrative</i>
Verbes au passé, pronoms de la 3 ^{ème} personne, verbes “publics”, négation synthétique, propositions participes	Verbes au présent, adjectifs attributs
3 – <i>Référence explicite</i>	<i>Référence dépendante de la situation d'énonciation</i>
Propositions relatives en position objet et en position sujet, coordination de syntagmes, nominalisations	Adverbes de temps et de lieu, adverbes
4 – <i>Visée persuasive explicite</i>	
Infinitifs, modaux de prédiction, de nécessité et de possibilité, verbes de persuasion, conditionnelles, auxiliaires discontinus	
5 – <i>Style abstrait</i>	<i>Style non-abstrait</i>
Conjonctions, passifs sans agent, propositions au participe passé, passif avec agent en <i>by</i> , <i>past-participial WHIZ</i> deletions, autres subordinants adverbiaux	

Tableau *Dimensions de Biber et traits correspondants*

⁴⁰Soulignons deux apports de ce programme. La simple lecture ne perçoit qu'une partie limitée des sur-emplois effectifs. Elle est bien en peine de juger s'ils sont significatifs ou non. Les sous-emplois, le « creux » d'une partie au regard de l'ensemble, échappent le plus souvent à la conscience. Ils sont ici dégagés.

⁴¹Voir néanmoins (Karlgrén, 1999) pour les problèmes que pose l'étude de traits qui ne relèvent pas forcément tous de la même loi de probabilité.

⁴²Il additionne les fréquences des traits correspondant aux traits positifs de l'axe et soustrait celles des marques ressortissant aux traits négatifs de l'axe. Dans les deux cas, il s'en tient aux traits *significatifs*, qui sont constitutifs de l'axe.

4.2 Typologies inductives « générales »

4.2.1 Les travaux de J.-P. Bronckart

J.-P. Bronckart croise deux paramètres qui aboutissent à quatre architypes discursifs (cf. tableau *Architypes discursifs de Bronckart*, repris de (Bronckart, 1996, p. 42)). Le premier est le rapport à la situation d'énonciation : l'émetteur peut intégrer dans son texte des renvois explicites aux paramètres de l'énonciation (locuteur, interlocuteur, temps et lieu), paramètres alors *impliqués* dans le texte, ou bien il peut éliminer ces indications : le texte, *autonome* par rapport à sa situation de production, ne requiert aucune connaissance de celle-ci pour son interprétation. La représentation du monde fournie dans l'énoncé peut être présentée comme mise à distance de l'interaction sociale en cours (les faits sont racontés comme s'ils étaient passés), c'est la disjonction. Elle peut inversement se situer dans le monde de l'interaction sociale en cours, c'est la conjonction : le texte montre des états, des actions, des événements accessibles dans le monde des protagonistes de l'interaction.

		Rapport au	monde
Rapport interactif à la situation		Conjonction	Disjonction
	Implication	<i>Discours interactif</i>	<i>Récit</i>
	Autonomie	<i>Discours théorique</i>	<i>Narration</i>

tableau *Architypes discursifs de Bronckart*

4.2.2 Les travaux de D. Biber

Les 67 traits étudiés ressortissent à 16 catégories distinctes comme marqueurs de temps et d'aspect, adverbess et locutions adverbess de temps et de lieu, pronoms et pro-verbess, questions, passifs, modaux, coordination, négation...L'objectif est l'« inclusion d'un grand nombre de caractéristiques linguistiques représentant l'éventail des possibilités fonctionnelles de l'anglais » (Biber, 1988, p. 211).

À partir des cinq dimensions issues de l'analyse factorielle, en utilisant la classification automatique, Biber aboutit à huit types de textes (tableau *Types de textes postulés par Biber*), en fonction de leur place sur chacune de ces dimensions.

Ces types ne correspondent pas forcément aux intuitions communes. C'est ainsi qu'on ne débouche pas sur un type unique interaction ou dialogue, mais deux : l'interaction à visée informationnelle et l'interaction à visée interpersonnelle. De la même manière, Biber distingue plusieurs types de textes « expositifs » et de textes narratifs (Biber, 1989, p. 38). Comme le souligne R. Sigley (Sigley, 1997, p. 231–232), ces catégories permettent de questionner les catégories qui ont été utilisées pour constituer les corpus concernés et éventuellement de réorganiser les constituants de ces corpus : certaines catégories doivent être subdivisées tandis que des similarités amènent à rapprocher des textes relevant de catégories distinctes.

Dénomination	traduction
Intimate interpersonal interaction	Interaction interpersonnelle intime
Informational interaction	Interaction informationnelle
“scientific” exposition	Exposé scientifique
Learned exposition	Exposé savant
Imaginative fiction	Fiction narrative
General narrative exposition	Récit
Situated reportage	Reportage situé
Involved persuasion	Argumentation impliquée

Tableau *Types de textes postulés par Biber*

4.3 Typologies inductives spécialisées

Dans (Bergounioux *et al.*, 1982, p. 169-186), l'étude de la répartition précise d'un certain nombre de formes (marques d'énonciation, détermination, coordination, pronoms, prépositions, etc.) dans les résolutions générales des congrès des confédérations CFDT, CFTC, CGT et FO de 1971 à 1976 a pour objectif de dégager le fonctionnement d'ensemble de ces résolutions. Les convergences des sur-emplois et des sous-emplois permettent d'opposer (*ibid.*, p. 175) une structure dite *analytique*, utilisée par la CFDT et la CGT à une structure dite *déclarative*, préférée par FO et la CFTC ⁴³. Le premier type de résolution sur-emploi en particulier le verbe *être* à la troisième personne de l'indicatif présent, les modaux, les pronoms à la première personne du pluriel et les possessifs de même personne, les pronoms de troisième personne. Le deuxième type sur-emploi les verbes déclaratifs (*appelle, considère, estime, exige...*), ayant pour sujet le congrès ou le sigle (*la CFTC*), suivis d'une complétive en *que*. La CGT et la CFDT sur-emploient par exemple la préposition *de* et ses variantes (*du, des*), en raison du recours à des syntagmes nominaux complexes ⁴⁴. Dans ces syntagmes nominaux, se rencontrent des unités complexes « en langue », comme *pouvoir d'achat, lutte de classes*, mais aussi des dénominations propres à ce type de syndicalisme ou à un certain discours politique (*unité d'action, union de la gauche, capitalisme monopoliste d'État*) voire à la CGT ou à la CFDT (*union des forces populaires*).

4.4 Typologies inductives et séquentialité

Certains travaux examinent également les changements de corrélations de traits linguistiques au fil de textes et les types qui y correspondent.

J.-P. Sueur soumet dans (Sueur, 1982) la Résolution Générale du congrès de la CFDT de 1976 à un étiquetage manuel fin suivi d'une analyse factorielle destinée à « décrire les régularités existant dans le texte dans l'occurrence de formes lexicales au sein de cadres syntaxiques et énonciatifs spécifiques, et à des places spécifiques à l'intérieur de ces cadres » (*ibid.*, p. 148). Pour les groupes nominaux, par exemple, sont codés la nature de la préposition, celle du déterminant et

⁴³ L'étude consacrée aux résolutions générales des congrès de la CFTC de 1945 à 1964 et de la CFDT de 1965 à 1979 (Habert, 1983) trouve une opposition similaire

⁴⁴ La résolution générale du congrès CGT de 1975 fournit l'exemple suivant : *le droit et les moyens de discuter du bien-fondé des décisions de licenciement et de fermetures d'entreprises avec la possibilité d'instances de recours*.

son interprétation sémantique⁴⁵, la classe « sémantique » du nom-tête (1 : locuteurs collectifs : *la CFDT, le congrès* ; 2 : individus ou groupes inclus dans la classe 1 ; ceux au nom de qui s'expriment les locuteurs de la classe 1, etc.). Pour les verbes, sont codés la forme positive ou négative, le statut actif/passif/pseudo-passif (*se* moyen), le mode et le temps, le trait contrôlable si le sujet superficiel peut contrôler l'action exprimée par le verbe, la classe sémantique (performatifs ; obligation et nécessité ; possibilité ; définition ; verbes d'analyse ; verbes marquant l'extension dans le temps et l'espace ; verbes psychologiques ; verbes marquant un processus). Sur la Résolution Générale du congrès de la CFDT de 1976, l'analyse factorielle met en évidence par exemple des corrélations entre certaines classes de sujets et certains types de verbes (*la CFDT, le congrès* / verbes performatifs ; *action(s), lutte(s), combat* / verbes d'obligation, etc.). Par ailleurs la Résolution étudiée se divise en quatre grandes parties. La première est consacrée à l'analyse, la seconde aux principes stratégiques, les deux dernières sont tournées vers l'action. Les corrélations de traits linguistiques manifestent l'existence au sein de ce texte unique de types distincts (discours analytique/théorique, discours stratégique/injonctif, etc.). Par exemple, la première partie est corrélée avec les verbes non contrôlables à sujet animé, la seconde avec les verbes non contrôlables sans sujet animé, tandis que les deux dernières, qui fixent des tâches précises pour les différents niveaux de l'organisation, sont corrélées avec des verbes contrôlables à sujet animé.

D. Biber et E. Finegan (Biber & Finegan, 1994) sur un corpus d'articles du *New England Journal of Medicine* et du *Scottish Medical Journal* montrent que les parties canoniques d'un article scientifique (introduction, méthodes, résultats, conclusion) comportent des différences linguistiques sensibles entre elles. Les pronoms de la première personne, les complétives en *that* marquent Introduction et Conclusion. La section Méthodes se caractérise par un emploi très fréquent des passifs sans agents et par un privilège donné au passé sur le présent. La partie Discussion fait un appel important aux modaux de possibilité et au présent.

Ces travaux raffinent l'idée de types de texte. Un texte donné n'est pas forcément homogène sur le plan des types de texte auxquels il recourt⁴⁶. Il peut inclure des « sous-types » ou faire appel pour telle ou telle part à un autre type que celui qui y prédomine. Le « grain » pour l'étude des types de textes n'est donc pas forcément un document dans son ensemble, même s'il est bref, ce qui est le cas des articles de médecine étudiés par Biber et Finegan.

4.5 Évaluation

4.5.1 Fiabilité des typologies dégagées

Les traits employés par les travaux analysés sont à l'évidence « fragiles » :

arbitraire relatif On ne saurait constituer une liste fermée de traits. Les limites fixées tiennent à la fois à la formation linguistique du chercheur et aux tendances linguistiques dominantes au moment de la recherche⁴⁷, à la stratégie d'étiquetage choisie (l'étiquetage manuel permet un grain plus fin) et aux outils statistiques convoqués.

fiabilité de l'étiquetage L'étiquetage morphosyntaxique « simple » est moins fiable, moins homogène quand il est manuel plutôt qu'automatique (Marcus *et al.*, 1993). L'incohérence risque d'être encore plus grande pour des distinctions fines comme celles de (Sueur, 1982). À l'inverse, un étiquetage automatique fin portant sur des données volumineuses pose la question de la validation du résultat.

La dispersion évidente des typologies obtenues n'invalide pas l'approche. Les travaux analysés n'ont en effet ni les mêmes objectifs (validation/exploration) ni les mêmes domaines d'étude (genre unique/diversité de genres). C'est pourquoi ce ne sont pas les mêmes aspects qui prêtent le flanc à la critique :

Réification d'oppositions « fabriquées »

Des techniques comme l'analyse factorielle révèlent des corrélations avant tout relatives aux données et aux variables qui sont utilisées. Extrapoler à l'extrême conduit à *réifier* ces dimensions, à croire saisir l'essence même des types de texte existants. D'autres jeux de données ainsi que le recours à d'autres techniques de statistique multidimensionnelle pourraient conduire à d'autres constats, et partant, à d'autres interprétations.

Représentation de la diversité effective de registres

Pour les études se limitant à un genre unique, la question ne se pose pas. Pour celles de Biber, le problème n'est pas celui de la représentation *effective* de l'ensemble des registres reconnus comme tels par les locuteurs : à l'évidence, la liste, ouverte, dépasse très largement la vingtaine effectivement retenue par Biber. Il s'agit plutôt de savoir si les registres retenus donnent accès à l'intégralité des dimensions sous-jacentes aux textes et énoncés en circulation dans une communauté langagière. La multiplication des corpus et des outils d'étiquetage devrait permettre de répondre à cette interrogation.

4.5.2 Généralité des typologies dégagées

Biber (Biber, 1995) a appliqué la même démarche à quatre corpus, le corpus anglais initial et trois ensembles de textes en coréen, somali⁴⁸ et nukulaelae tuvalu⁴⁹. Malgré des différences nettes, liées en particulier au degré d'alphabetisation et à la place des traditions orales dans les langues considérées, Biber (*ibid.*, p 359) pense pouvoir émettre l'hypothèse que les types textuels qu'il dégage sont communs à plusieurs langues, même si leurs réalisations linguistiques diffèrent d'une langue à l'autre.

4.5.3 Valider les types proposés

Biber souligne à plusieurs reprises la place de l'interprétation dans la dénomination des dimensions⁵⁰ et des types de textes et sur les précautions à reprendre pour contraindre au mieux cette interprétation. Il montre les va-et-vients nécessaires entre les caractérisations globales opérées par l'analyse factorielle, par la

⁴⁵ Défini spécifique, défini non spécifique ou générique, indéfini spécifique, indéfini non spécifique, générique (*tout*, etc.), possessif, démonstratif, divers, pronom anaphorique (dans ce cas, sont indiquées les caractéristiques du SN auquel ce pronom se réfère).

⁴⁶ Cf. les travaux de J.-M. Adam (Adam, 1992)(Adam & Revaz, 1996) qui visent à dégager au sein des textes, conçus comme hétérogènes, des composantes homogènes – les séquences – relevant du récit, de la description, de l'explication, de l'argumentation et du dialogue.

⁴⁷ L'insistance sur les embrayeurs dans (Bronckart *et al.*, 1985) renvoie ainsi à la place des problématiques énonciatives dans la linguistique française des années 70.

⁴⁸ Langue parlée par environ 5 millions de personnes en Somalie, à Djibouti, en Éthiopie et au Kenya.

⁴⁹ Langue parlée par 350 personnes sur l'atoll Nukulaelae du groupe Tuvalu (Pacifique).

⁵⁰ « Les dimensions textuelles sont des constructions théoriques, issues de l'interprétation des résultats d'une procédure statistique connue sous le nom d'analyse factorielle. C'est-à-dire que l'analyse factorielle identifie les traits linguistiques qui cooccurrent très fréquemment dans les textes, et chaque groupe de traits cooccurrents peut être interprété pour déterminer la fonction de communication sous-jacente la plus largement partagée par ces traits. » (Biber, 1985, p. 338)

classification automatique et l'examen des traits linguistiques correspondants et des textes liés. C'est la complémentarité qu'il souligne dans (Biber, 1985) entre approche « macroscopique » et approche « microscopique ».

4.5.4 Grammaires de discours ou restrictions sur les discours

J.-P. Sueur (Sueur, 1982, p. 148) se donne comme projet l'élaboration d'une *grammaire de discours* : « ...le but de cette grammaire est de définir le plus précisément possible les caractéristiques qui permettent d'identifier spontanément un discours : nous reconnaissons le discours de tel individu, de tel groupe, de tel parti, de tel syndicat, etc. Or si la spécificité du lexique joue un rôle dans ce processus de reconnaissance, d'autres facteurs interviennent : la syntaxe, les faits d'énonciation, mais surtout la connexion entre ces divers types de faits. Tel mot apparaît de manière privilégiée, mais surtout, il apparaît de manière privilégiée à telle place et dans tel cadre. Tout cela fait partie de la "compétence" propre au locuteur – et vient se combiner avec les traits qui définissent les diverses formes d'énonciation (une résolution n'est pas un discours ni la réponse à une interview, etc.) »

En fait, les types de textes dégagés par les travaux qui viennent d'être présentés ne débouchent pas sur de véritables grammaires. Une constellation de traits n'indique pas les restrictions fines à l'œuvre, mais des restrictions préliminaires sur les « matériaux » linguistiques utilisables et leurs proportions licites d'emploi. Par contre, elle ne permet pas de connaître les schémas de phrase probables ou la structure d'ensemble licite. En ce sens, elle constitue une abstraction éloignée de la perception des sujets parlants.

4.5.5 Pôles et continuum

Les travaux examinés convergent pour concevoir les types de texte comme des constructions théoriques, éventuellement jamais réalisées entièrement, comme des pôles multiples entre lesquels se situent les textes effectifs (Bronckart *et al.*, 1985, p. 137)(Biber, 1989, p. 3).

5 Sur le métier, remettre les tâches

Les deux dimensions de la représentativité, externes et internes, qui ont été abordées dans les deux sections précédentes, se traduisent en deux volets de tâches concrètes : la documentation des composants de bases textuelles, des corpus et de leurs traitements d'une part, la mise au point d'outils de profilage de textes d'autre part.

5.1 Documenter un corpus, ses composants et leur histoire

5.1.1 Mieux de mémoire et savoirs de mémoire : la vie éternelle pour les corpus ?

Pour qu'une base textuelle permette l'extraction à la demande des « documents » en fonction d'une utilisation donnée⁵¹, il importe que chacune des unités élémentaires⁵² qui la constituent soit « autonomisable ». On doit posséder suffisamment d'informations fines sur elle pour pouvoir l'extraire de la base et l'assembler avec d'autres éléments de la même base ou d'autres bases sans perdre ces renseignements qui sont indispensables pour interpréter les contrastes et les convergences manifestés dans le corpus qui vient d'être rassemblé. Ces renseignements doivent couvrir à la fois la description précise du contexte de production du composant et une caractérisation en termes de domaine (thématique) et de « genre » (au sens indiqué *supra*).

Comme chaque « document » présent dans le BNC est assorti d'informations précises sur la situation de communication dans laquelle il s'insère et sur les données originelles dont cette version électronique dérive, le BNC fonctionne effectivement comme une « réserve à corpus ». En fonction d'une recherche ou d'une application déterminée, on peut extraire les documents qui correspondent le mieux à ce qu'on veut étudier. On peut ainsi ne retenir que l'écrit. Ou s'en tenir à l'oral, voire être plus précis : les locuteurs d'un certain âge, ou d'une certaine région du Royaume-Uni, ou s'attacher à une certaine situation d'interlocution (conférence, par exemple).

5.1.2 Normaliser les corpus et leur documentation

Séparer représentation physique et représentation logique des documents : SGML et XML

L'échange des corpus et leur réutilisation ont buté jusque récemment sur l'éclatement des codages pratiqués. Un travail de normalisation est en cours pour y remédier. Cette normalisation sépare représentation physique et représentation logique des documents. Elle propose des conventions générales pour les différents types de textes. Le balisage logique d'un document revient à indiquer sa structure : ses subdivisions et leurs relations. Il se réalise en deux étapes. La première est l'identification des éléments possibles pour un texte donné et de leurs relations. C'est en quelque sorte écrire une « grammaire de texte ». C'est ce qu'on appelle une *Définition de Type de Document* (DTD). La deuxième étape est l'introduction des balises choisies dans le document relevant de cette DTD, en respectant les règles éditées pour leur combinaison. Le balisage employé rend explicite les éléments du texte et leur agencement. Il obéit au langage standard de balisage SGM qui est maintenant présent dans pratiquement tout logiciel de gestion de document.

S'entendre sur les types de textes majeurs : la TEI

Ce premier niveau de normalisation s'avère cependant insuffisant. Rien n'empêche en effet plusieurs groupes ou individus de se donner des conventions différentes pour un même type de document, ce qui entrave la comparaison et l'échange des résultats. Un deuxième niveau est donc nécessaire : s'entendre sur des descriptions génériques pour les grands types de documents utilisés (dictionnaires, poésie, théâtre, oral, textes alignés, documents historiques), ainsi que pour les niveaux d'annotation qui peuvent les décorer (étiquettes, arbres, appareil critique, références croisées). Une initiative de grande ampleur, la TEI (*Text Encoding Initiative*) rassemble depuis plus de dix ans des chercheurs de différentes disciplines et de toutes nationalités pour proposer des conventions sur ces types de

⁵¹La collection de documents réunie à des fins de veille sociale pour le projet *Scriptorium* peut ainsi faire l'objet d'extractions multiples (par acteur, en fonction d'une sous-période, par genre de document, autour de mots-pivots...).

⁵²Il peut s'agir de textes « complets », comme dans *Frantext*, mais aussi de fragments ou d'extraits, comme dans le BNC, voire de paragraphes ou de phrases (par exemple concernant un mot-clé donné ou une construction syntaxique déterminée). Il peut même s'agir de segments textuels discontinus, comme les prises de parole successives d'un même orateur lors d'un débat public (Heiden, 1999). Plus le « grain », c'est-à-dire la taille moyenne des composants de la base est fin, plus il permet de constituer des sous-ensembles distincts et adaptés à une tâche donnée. Par exemple, si l'on dispose des indications signalétiques attachées à chacun des articles des CD-Roms du *Monde*, on peut extraire des corpus en fonction d'un type de document donné (nécrologie, interview, portrait...) ou d'une thématique (pages financières, sport, politique internationale).

documents. Elle a débouché sur des Recommandations en 1994. De nombreux projets de constitution de corpus et de ressources linguistiques ont adopté la TEI (le BNC par exemple). Pour reprendre les termes de J. André (André, 1996, p. 17), la TEI constitue un « inventaire – une sorte de flore, au sens de Buffon – des divers éléments pouvant constituer un document littéraire », et elle représente en ce sens une avancée dans la description et la formalisation des types de documents en circulation dans les diverses communautés langagières. Elle fournit ainsi indirectement des éléments pour les typologies de textes et les études sur les genres discursifs.

Lier la documentation au corpus

La TEI fait obligation aux concepteurs de corpus de faire figurer au tout début du corpus un *en-tête (header)* ou encore *cartouche*⁵³. Ce cartouche documente quatre aspects du corpus :

1. le rapport entre les sources utilisées et la version électronique ;
2. les choix d'annotation effectués ;
3. des renseignements sur le contexte du corpus : langues et dialectes représentés, type de textes retenus, etc. ;
4. le détail des révisions subies par le corpus, en quelque sorte, le livre de bord des responsables du corpus.

On distingue le cartouche dominant l'ensemble du corpus et les cartouches des composants du corpus. Le cartouche du corpus met en facteur les choix qui valent pour toutes les données textuelles englobées dans le corpus. Chaque composant du corpus peut comporter son propre cartouche, qui détaille les informations qui lui sont propres. C'est cette répartition entre informations partagées et renseignements particuliers qui permet en particulier, à partir d'un corpus donné, d'en extraire un sous-ensemble sur des critères tout à fait précis.

La version électronique et ses sources

Ce premier volet distingue nettement la version électronique et ses sources :

1. le titre. On distingue, par exemple en ajoutant « version électronique », ce titre de celui de la source utilisée, pour souligner les écarts éventuels entre les deux états du document. On note d'une part l'auteur, qui correspond à celui de la source et le responsable du contenu intellectuel de l'édition électronique ;
2. la mention d'édition, en particulier le numéro de version de l'édition électronique de ce document ;
3. la taille approximative, par exemple en caractères et mots, du document. Ces indications permettent en particulier de prévoir la place nécessaire pour le document ;
4. les indications sur la diffusion possible. Elles regroupent la mention de l'éditeur, l'adresse, mais surtout des précisions sur la disponibilité du document (son usage est-il soumis à des restrictions comme un *copyright* ?) ;
5. la description bibliographique de la source utilisée.

Choix d'annotation

Les objectifs sous-jacents au corpus s'accompagnent d'une description des méthodes choisies pour sélectionner les documents retenus. Le type d'annotation effectué est décrit globalement : choix opérés pour la segmentation, traitement des citations, corrections apportées au texte de départ, etc. C'est aussi dans cette section que sont déclarées les catégories utilisées pour classer les composants regroupés (par exemple pour *Le Monde*, ce pourrait être la rubrique, les mots-clés fournis par la documentation du journal, le type d'article : brève, interview...). Un même composant peut être classé simultanément sur plusieurs axes, ce qui permet ensuite des extractions fines.

Contexte

Dans ce volet sont mémorisées les informations concernant la création du corpus ou d'un de ses composants, en particulier ses dates et lieu de mise au point, mais aussi les langues, les registres et les dialectes représentés. C'est à cet endroit que figure, pour un composant, sa place dans les classifications signalétiques choisies pour le corpus.

Historique des révisions

C'est le journal des modifications apportées au corpus, qui sont notées séquentiellement (date, personne responsable de la modification et une description détaillée de la révision effectuée).

Le cartouche proposé par la TEI peut paraître trop détaillé, voire verbeux. Il permet néanmoins de s'assurer que les renseignements fondamentaux pour donner sens aux analyses issues de ce corpus ont effectivement été rassemblés.

5.1.3 Documenter les analyses faites sur un corpus

Une collection de textes permet d'engendrer de multiples corpus distincts. Chacun de ces corpus peut donner lieu à des annotations variées qui constituent autant d'« interprétations », au sens large : étiquetage morpho-syntaxique, projection de catégories sémantiques, lemmatisation, etc. Chacun de ces traitements produit une version différente du corpus. L'écart entre les versions peut être plus ou moins important.

Il importe donc, pour une analyse donnée, de mémoriser non seulement la sélection de textes sur laquelle elle opère mais aussi les traitements auxquels ces textes ont été soumis. C'est la condition *sine qua non* pour que l'analyse en question soit reproductible et pour qu'on puisse relier de manière sûre les constats effectués et les caractéristiques du corpus traité. L'expérience prouve que ce lien se perd rapidement : restent des analyses dont on ne sait pas toujours précisément sur quoi elles ont porté. Elles se vident alors de sens.

De la même manière que la TEI enjoint d'attacher à chaque corpus son cartouche, de sorte que la description du corpus, de son origine, de l'annotation réalisée, des révisions faites, ne puisse être dissociée du corpus lui-même, il convient d'« amarrer » chaque analyse au corpus sur lequel elle porte. La solution minimale est le renvoi aux cartouches du corpus et des composants extraits pour l'analyse. Une solution intermédiaire, l'inclusion de ces cartouches. La solution optimale, plus coûteuse en place (mais les moyens de stockage croissent continuellement), revient à mémoriser en même temps que l'analyse les choix qui ont présidé à la création du corpus, le cartouche du corpus, ceux des composants et les composants eux-mêmes.

⁵³ « Emplacement réservé à la légende ou au titre, situé au bas d'un tableau, d'une carte géographique, etc. » (*Le Petit Robert*).

5.2 Mesurer/maîtriser l'hétérogénéité langagière : « profilage » de corpus

5.2.1 De l'hétérogénéité subie à l'hétérogénéité visée

On serait tenté de voir dans les nouveaux corpus à géométrie variable « du texte », texte dont on ne sait pas toujours très bien de quels usages langagiers il est représentatif. Les données du journal *Le Monde* disponibles sous forme électronique rassemblent ainsi des textes de longueur très différentes (de quelques dizaines de mots dans les « brèves » à des milliers de mots pour les articles de dossiers), relevant de domaines distincts – les *rubriques* ou *sections*⁵⁴ – et de *genres* multiples : biographie, chronique, chronologie, encadré, correspondance, entretien, opinion, portrait, rectificatif, revue de presse... L'étude (Illouz *et al.*, 1999) menée sur les 6 rubriques principales⁵⁵ de ce journal montrait ainsi des écarts significatifs entre ces rubriques à la fois pour le vocabulaire utilisé et pour les catégories syntaxiques qui y sont privilégiées⁵⁶.

Les nouveaux corpus nécessitent donc des outils de profilage pour évaluer leur homogénéité interne, pour pouvoir dégager des sous-parties homogènes, pour déterminer les conséquences de l'ajout ou de l'élimination d'une partie de leurs composants, ou encore pour assembler tout ou partie de leurs composants avec des éléments provenant d'autres corpus. La maîtrise des caractéristiques du corpus utilisé détermine en effet partiellement la qualité des connaissances acquises à partir de lui. Pour préparer un dictionnaire de français langue étrangère, on peut souhaiter par exemple disposer dans un domaine donné des textes les moins techniques, employant le vocabulaire le plus central du domaine. Il en va de même si l'on veut rassembler des données pertinentes sur la variation du français dans le temps ou en fonction des conditions sociales. Il faut donc pouvoir « profiler » les corpus et les textes.

Nous appelons profilage de textes l'utilisation d'outils de calibrage donnant des indications sur l'emploi du vocabulaire, mais aussi de catégories morpho-syntaxiques et de patrons, dans les parties d'un corpus⁵⁷, pour en déterminer l'homogénéité ou l'hétérogénéité. Ces outils doivent également permettre de positionner un nouveau texte par rapport aux regroupements obtenus sur un corpus pré-existant. Nous proposons de développer une méthodologie de profilage⁵⁸ qui prolonge les travaux de D. Biber.

5.2.2 Choix de marques linguistiques

Démarches

On peut partir d'un type d'énoncé et chercher les marqueurs linguistiques qui y correspondent. Par exemple, associer à un style informationnel, décontextualisé les nominalisations, les passifs sans agent, etc., comme le fait Biber (Biber, 1985, p. 344–345). On peut à l'inverse partir d'éléments relevant de niveaux différents de l'analyse linguistique et examiner quels sont leurs usages en discours et en quoi ils marquent tel ou tel type de texte. Par exemple, certains suffixes évaluatifs (-ard dans *chauffard*, -âtre dans *verdâtre*) indiquent une certaine implication personnelle de l'énonciateur et paraissent peu compatibles avec un style informationnel.

Trait et fonction

Un trait linguistique isolé ne donne pas, le plus souvent, d'indications sur le type de texte dans lequel il s'insère. C'est dans son alliance avec d'autres traits qu'il prend sens. Ainsi, le pronom indéfini *on* associé à des pronoms de la première et de la deuxième personne du singulier peut renvoyer à un discours familier tandis qu'accompagné du présent de l'indicatif, du passif, et sans première et deuxième personne du singulier, il peut pointer un discours factuel (vulgarisation, communication scientifique...).

Examen de deux palettes de traits

Comparer les choix faits par Bronckart (Bronckart *et al.*, 1985, p. 147–167) et Biber (Biber, 1988, p. 211–245) pour leurs typologies généralistes révèle des convergences et des divergences qui aident à mettre au point une grille d'analyse pour le français.

5.2.3 Architectures : étiquetage et sur-étiquetage

De l'étiquetage manuel à l'étiquetage automatique

L'absence d'étiqueteurs morphosyntaxiques pour le français au début des années 80 explique l'utilisation des convergences dans l'emploi d'un certain nombre de mots, en particulier « outils », dans (Bergounioux *et al.*, 1982) et dans (Habert, 1983).

Le travail de Biber, qui a été réalisé au milieu des années 80, il y a un peu plus d'une dizaine d'années donc, est parti de versions « nues » des corpus choisis. Si une version étiquetée du LOB est devenue disponible au moment de la recherche rapportée, l'absence de son équivalent pour London-Lund a conduit D. Biber à développer un ensemble unique de programmes en PL/1 pour traiter ces textes non étiquetés. D. Biber a commencé par réaliser lui-même un étiquetage morpho-syntaxique des textes⁵⁹. La recherche de « patrons » plus spécifiques et marqueurs d'un trait donné constitue la deuxième étape. La vérification manuelle est limitée

⁵⁴ ETR(ranger), ECO(nomie), POL(itique), ING (? information générale : sport, faits divers), ART (médias, spectacles), EMS (? éducation, médecine, société), etc. Ce sont les classifications utilisées par la rédaction du journal *Le Monde* qui sont reprises dans les champs signalétiques de la version électronique distribuée par ELRA. On ne dispose pas toujours de la signification des libellés (par exemple pour ING et EMS), d'où le point d'interrogation.

⁵⁵ ETR, ECO, POL, ING, ART, EMS.

⁵⁶ Pour le vocabulaire, ont été étudiés les 15 438 articles totalisant 7 millions de mots extraits des 14 millions de mots provenant par choix aléatoire de numéros entiers parmi ceux des années 1987, 1989, 1991, 1993 et 1995 (Naulleau, 1998) qui constituent la partie *Presse* du corpus réalisé dans le cadre du projet européen PAROLE. Pour les catégories syntaxiques, ont été examinés 241 484 mots, provenant de 7 numéros de septembre 1987, qui ont été extraits de l'ensemble précédent, étiquetés automatiquement et corrigés manuellement pour la partie du discours, toujours dans le cadre de PAROLE.

⁵⁷ Dans une perspective donc plus large que (Kilgariff & Rose, 1998)(Kilgariff, 1997)

⁵⁸ Dans le cadre du projet TyPTex (*Typage et Profilage de Textes*) commun au LIMSI et à l'UMR 8503 et soutenu financièrement par ELRA (*European Language Resources Association*).

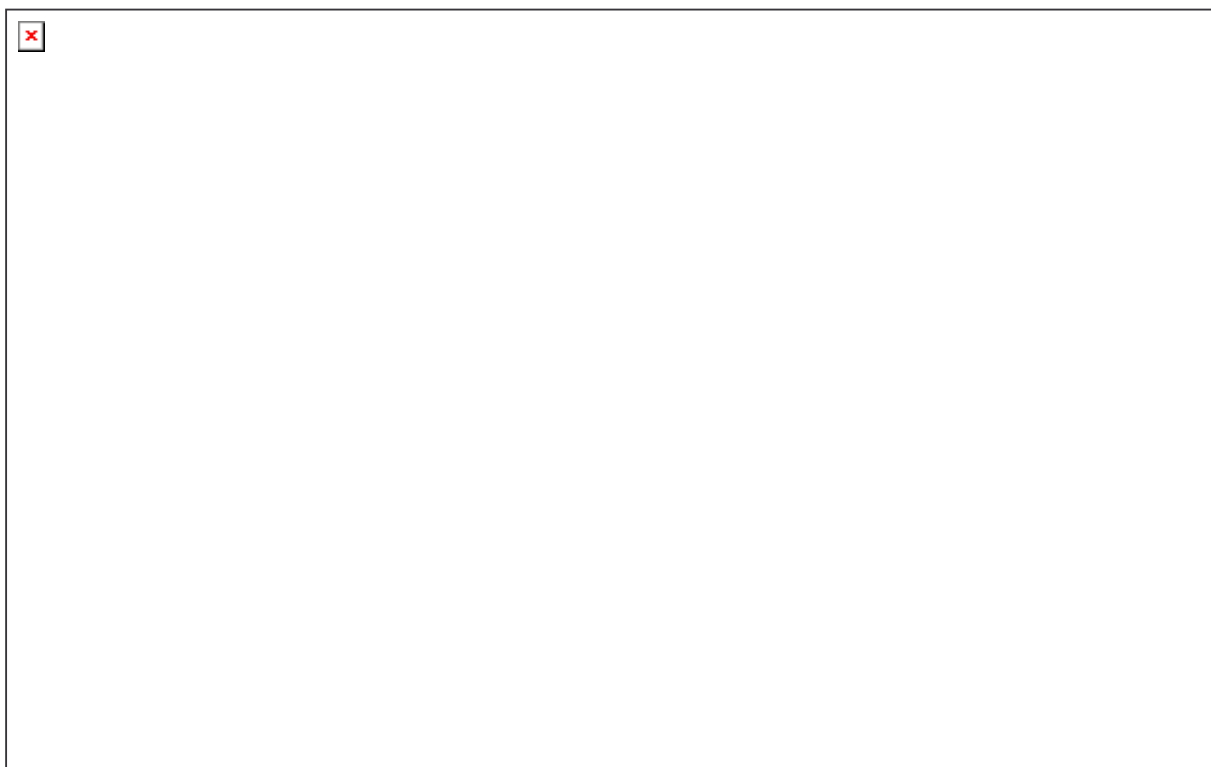
⁵⁹ En s'appuyant sur Brown pour l'inventaire des couples ⟨forme, catégorie⟩ (*ibid.*, p. 212) et sur (Quirk *et al.*, 1985) pour l'écriture de règles de désambiguïsation. Certains mots trop polymorphes sont laissés de côté : *as, that...*

au maximum. D. Biber fournit pour chacun des traits l'algorithme pour son repérage automatique⁶⁰, une description des rôles joués par ce trait⁶¹ et un renvoi aux études préexistantes sur lui (*ibid.*, p. 221–245). Un certain nombre de traits ne pouvant être identifiés automatiquement ont été écartés de l'étude.

Pour certains traits de (Sueur, 1982), la finesse des distinctions opérées (pour les emplois des déterminants, par exemple) exclut clairement l'automatisation, tandis que le repérage d'autres marques relèverait aujourd'hui soit d'étiqueteurs ordinaires soit de programmes à développer de manière spécifique.

Organisation d'ensemble

Comme le montre la figure *Architecture de profilage de textes*, on dispose au départ d'une base de textes. Chacun comprend un cartouche suivant les recommandations de la TEI (Dunlop, 1995). Les critères d'une requête ou d'une sélection aboutissent à un corpus, c'est-à-dire un ensemble de textes rassemblés en fonction d'une recherche ou d'une application déterminée. Chacun de ces textes est soumis à un étiquetage morpho-syntaxique, qui permet d'associer à chaque mot ou unité polylexicale un lemme, une partie du discours et des indications morphosyntaxiques plus fines. Le marquage typologique utilise l'ensemble de ces informations et les remplace par de nouvelles catégories, correspondant aux traits linguistiques dont on veut étudier la distribution (embrayeurs, modalités, présentatifs, usage des temps, passif, classes d'adverbes – négation, degré...– et de déterminants, etc.). Le corpus marqué (et éventuellement corrigé par le biais de CORTECS (Heiden *et al.*, 1998)) est alors soumis à des logiciels d'analyse textuelle. En particulier, on construit la matrice des fréquences de chaque trait dans chaque texte. Cette matrice sert tant à la recherche optimale de traits pertinents à une opposition, qu'à la classification inductive ou supervisée.



6 Articuler autrement intuition et attestation

Nous avons détaillé les conditions externes et internes qui aident à sélectionner des données langagières en fonction des emplois d'une langue que l'on souhaite représenter. Nous avons essayé de montrer les tâches qui en résultent en matière de documentation de corpus et de « profilage » des composants de corpus.

Ces « principes de précaution » n'empêchent pas de réexaminer la question de l'articulation entre les régularités constatées dans un corpus et les règles postulées⁶². J.-C. Milner affirme (Milner, 1989, p. 55) : « [...] l'activité grammaticale ne consiste pas à enregistrer les données de langue ; elle consiste à émettre sur ces données un jugement différentiel », c'est-à-dire à isoler « l'impossible de langue » (*ibid.*). Dans le même esprit, S. Auroux (Auroux, 1998, p. 197) rappelle : « La règle est une hypothèse sur les faits, les faits contiennent aussi bien du possible que de l'impossible. » Il continue (*ibid.*, p. 240) : « Il y a hétérogénéité essentielle entre l'étude statistique de la régularité, et l'existence de règles. Supposons en effet qu'existe une règle, la régularité des actions connectées avec une règle contient à la fois des actions correctes et incorrectes ; pour les distinguer, il faut connaître la règle. »

Le recours renouvelé aux corpus confronte effectivement à des énoncés que l'on juge *a priori* impossibles⁶³. On peut répartir ces « existants impossibles » au moins dans quatre catégories. La première relève du simple lapsus, de l'erreur. La seconde du viol intentionné d'une règle par le locuteur. Une telle transgression

⁶⁰Par exemple, pour le *perfect*, les « patrons » (a) HAVE + (Adverbe) + (Adverbe) + Participe Passé (b) HAVE + Nom/Pronom + Participe Passé (questions) (*ibid.*, p. 223).

⁶¹Par exemple, « Les verbes au présent traitent de sujets et d'actions immédiatement pertinentes. Ils servent également dans le style universitaire à centrer l'attention sur l'information présentée et à écarter le déroulement dans le temps » (*ibid.*, p. 224).

⁶²Cf. (Laks, 1996, ch. 6) pour une discussion approfondie du rapport règle/régularité.

⁶³Ou nettement improbables. Ainsi, dans (Corbin, 1997), D. Corbin examine certaines contraintes – linguistiques et pragmatiques – qui empêchent l'intégration dans le lexique (*la lexicalisation*) de certaines séquences pour le reste « bien formées ». Ce sont par exemple des conventions (des tabous) qui font obstacle, à la lexicalisation de *le bras gauche de* Nom Propre (parallèle à *le bras droit de* Nom Propre) : la gauche est toujours *sinistre* alors que

(c'est l'exemple de Le Carré cité en note) ne bouscule pas la lisière entre possible et impossible. Ce genre de « fusée langagière » en joue et s'en joue, ce qui est précisément une manière de la confirmer. La troisième catégorie correspond à la variation interne à la langue (dans une perspective proche de celle de Labov). La quatrième, étroitement liée, témoigne de l'évolution des règles. Ce qui sépare les deux premiers types des deux derniers, c'est la fréquence. Les premiers sont négligeables. Le nombre d'occurrences des seconds est significatif⁶⁴.

On peut en outre refuser de suivre S. Auroux quand il affirme (*ibid.*, p. 183) : « La recherche d'attestations dans des textes (quelles que soient sa sophistication et l'utilisation de moyens techniques coûteux, voire informatiques), la constitution d'un *corpus* ...ne relèvent pas directement des protocoles expérimentaux. À cela deux raisons : i) elles ne sont pas en relation directe avec une hypothèse explicite à tester ; ii) elles ne correspondent pas à la production d'un phénomène. » L'existence de corpus annotés permet au contraire de tester des hypothèses explicites. C'est la démarche défendue dans (Aarts, 1990). La mesure par Barkema (Barkema, 1993)(Barkema, 1994)(Habert *et al.*, 1997, p. 58–60) du degré de figement de séquences en relève. Par ailleurs, un certain nombre de phénomènes langagiers qui affleurent dans les corpus échappent à la perception des locuteurs et sont difficilement explicites par eux. C'est le cas des corrélations entre traits linguistiques constitutives des types de textes mis en évidence par les traitements statistiques multidimensionnels. En ce sens, les corpus actuels et les outils de traitement qui les accompagnent donnent à voir, représentent des dimensions du langage relativement mal explorées.

Références

- AARTS, J. (1990). Corpus linguistics : an appraisal. In J. HAMESSE & A. ZAMPOLLI, Eds., *Computers in Literary and Linguistic research*, pp. 13–28. Paris-Genève: Champion-Slatkine.
- ADAM, J.-M. (1985). Quels types de textes ? *Le français dans le monde*, (192).
- ADAM, J.-M. (1992). *Les textes : types et prototypes*. Paris: Nathan.
- ADAM, J.-M. & REVAZ, F. (1996). *L'analyse des récits*. N° 22 in *Mémo*. Paris: Seuil.
- ANDRÉ, J. (1996). Balises, structures et tei. *Cahiers Gutenberg*, (24), 11–22.
- AUROUX, S. (1998). *La raison, le langage et les normes*. Sciences, modernités, philosophies. Paris: Presses Universitaires de France.
- BAKHTINE, M. (1984). *Esthétique de la création verbale*. Bibliothèque des idées. Paris: Gallimard. Traduction d'Alfreda Aucouturier. Préface de Tzvetan Todorov.
- BARCKEMA, H. (1993). Idiomaticity in English NPs. In J. AARTS, P. DE HAAN & N. OOSTDIJK, Eds., *English language corpora : design, analysis and exploitation*, pp. 257–278. Amsterdam: Rodopi.
- BARCKEMA, H. (1994). Determining the syntactic flexibility of idioms. In U. FRIES, G. TOTTIE & P. SCHNEIDER, Eds., *Creating and using English language corpora*, pp. 39–52. Amsterdam: Rodopi.
- BEACCO, J.-C. (1992). Les genres textuels dans l'analyse du discours : écriture légitime et communautés translangagières. *Langages*, (105), 8–27. Ethnolinguistique de l'écrit, Jean-Claude Beacco (éd.).
- BEAUDOIN, V. & VELKOVSKA, J. (1999). Constitution d'un espace de communication sur Internet (forums, pages personnelles, courrier électronique...). *Réseaux*.
- BENVENISTE, E. (1966). *Problèmes de linguistique générale*, volume 1 of *Coll. TEL*. Gallimard.
- BERGOUNIOUX, A., LAUNAY, M.-F., MOURIAUX, R., SUEUR, J.-P. & TOURNIER, M. (1982). *La parole syndicale*. Politique d'aujourd'hui. Paris: Presses Universitaires de France.
- BIBER, D. (1985). Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics*, (23), 337–360.
- BIBER, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- BIBER, D. (1989). A typology of English texts. *Linguistics*, (27), 3–43.
- BIBER, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, 5(4), 257–270.
- BIBER, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, 19(2), 243–258.
- BIBER, D. (1994). Representativeness in corpus design. *Linguistica Computazionale*, IX-X, 377–408. Current Issues in Computational Linguistics : in honor of Don Walker.
- BIBER, D. (1995). *Dimensions of register variation : a cross-linguistic comparison*. Cambridge: Cambridge University Press.
- BIBER, D. & FINEGAN, E. (1994). Intra-textual variation within medical research articles. In N. OOSTDIJK & P. DE HAAN, Eds., *Corpus-based research into language*, N° 12 in *Language and computers : studies in practical linguistics*, pp. 201–222. Amsterdam: Rodopi.
- BOURIGAU, D. (1993). Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *TAL*, 34(2).
- BOUROCHE, J.-M. & SAPORTA, G. (1980). *L'analyse des données*. Que sais-je ? Paris: Presses Universitaires de France.
- BRANCA-ROSOFF, S. (1996). Retour aux genres. In S. AUROUX, S. DELESALLE & H. MESCHONNIC, Eds., *Histoire et grammaire du sens*, U Linguistique, chapter 14, pp. 189–203. Paris: Armand Colin. Hommage à Jean-Claude Chevalier.
- BRANCA-ROSOFF, S. (1999a). Des innovations et des fonctionnements de langue rapportés à des genres. *Langage et société*, (87), 115–129.
- BRANCA-ROSOFF, S. (1999b). Types, modes et genres : entre langue et discours. *Langage et société*, (87), 5–24.
- BRONCKART, J.-P. (1996). Genres de textes, types de discours et opérations discursives. *Enjeux*, (37–38), 31–47. Namur.
- BRONCKART, J.-P., BAIN, D., SCHNEUWLY, B., DAVAUD, C. & PASQUIER, A. (1985). *Le fonctionnement des discours : un modèle psychologique et une méthode d'analyse*. Lausanne: Delachaux & Niestlé.
- BURNARD, L. & SPERBERG-MCQUEEN, C. M. (1996). La tei simplifiée : une introduction au codage des textes électroniques en vue de leur échange. *Cahiers Gutenberg*, (24), 23–151. Traduction de François Role.
- CERQUIGLIANI, B. (1999). *Éloge de la variante. Histoire critique de la philologie*. Des travaux. Paris: Seuil.
- CHARAUDEAU, P. (1983). *Langage et discours. Éléments de sémiolinguistique (théorie et pratique)*. Langue Linguistique Communication. Paris: Hachette.
- CHISS, J.-L. (1987). Malaise dans la classification. *Langue française*, (74), 10–28. La typologie des discours, J.-L. Chiss, J. Filliolet (eds).
- COMBETTES, B. (1988). *Pour une grammaire textuelle: la progression thématique*. De Boeck-Duculot.
- CORBIN, D. (1997). Entre les mots possibles et les mots existants : les unités lexicales à faible probabilité d'actualisation. In D. CORBIN, B. FRADIN, B. HABERT, F. KERLEROUX & M. PLÉNAT, Eds., *Mots possibles et mots existants*, pp. 79–90. Lille.
- DOLININE, C. (1999). Le problème des genres du discours quarante-cinq ans après bakhtine. *Langage et société*, (87), 25–40.
- DUNLOP, D. (1995). Practical considerations in the use of TEI headers in large corpora. *Computers and the Humanities*, (29), 85–98. Text Encoding Initiative. Background and Context, edited by Nancy Ide and Jean Véronis.
- GADET, F. (1997). *Le français ordinaire*. U Linguistique. Paris: Armand Colin/Masson, 2ème édition.

la droite est adroite...D. Corbin prédit donc que *le bras gauche de Nom Propre* est possible mais peu probable. Dans *Le directeur de nuit*, de John Le Carré (Robert Laffont, 1993), traduction de *The hight manager*, on rencontre cependant la phrase suivante : « Seul Amato, le bras droit américano-vénézuélien de Strelski, restait de marbre... **L'étonnant bras gauche** de Strelski était un Irlandais obèse au visage empâté nommé Pat Flynn... ».

⁶⁴ Même s'il reste à déterminer la manière de mesurer cette significativité, selon les niveaux linguistiques en cause

- GEFFROY, A. & LAFON, P. (1982). L'insécurité dans les grands ensembles. Aperçu critique sur *le vocabulaire français de 1789 à nos jours* d'Etienne brunet. *MOTS*, (5), 129–141.
- HABERT, B. (1983). Études des formes spécifiques et typologie des énoncés (les résolutions générales des congrès de la CFTC-CFDT de 1945 à 1979). *MOTS, Presses de la Fondation Nationale des Sciences Politiques*, (7), 97–124.
- HABERT, B., FABRE, C. & ISSAC, F. (1998). *De l'écrit au numérique : constituer, normaliser, exploiter les corpus électroniques*. Informatiques. Paris: InterÉditions/Masson.
- HABERT, B., NAZARENKO, A. & SALEM, A. (1997). *Les linguistiques de corpus*. U Linguistique. Paris: Armand Colin/Masson.
- HEIDEN, S. (1999). Encodage uniforme et normalisé de corpus. application à l'étude d'un débat parlementaire. *Mots*, (60), 113–132. Presses de Sciences Po.
- HEIDEN, S., CUQ, A., DUCOUT, D., HORLAVILLE, P., ROBERT, J.-P., PRIEUR, V. & DOHM, B. (1998). *CorTeCs – 1.0β : Manuel de l'utilisateur*. Laboratoire de Lexicométrie et Textes Politiques – UMR 9952, CNRS – ENS Fontenay/Saint-Cloud.
- IDE, N. & VÉRONIS, J. (1995). *The Text Encoding Initiative: Background and context*. Dordrecht: Kluwer Academic Publishers.
- ILLOUC, G., HABERT, B., FLEURY, S., HEIDEN, S. & LAFON, P. (1999). Maîtriser les déluges de données hétérogènes. In A. CONDAMINES, C. FABRE & M.-P. PÉRY-WOODLEY, Eds., *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, pp. 37–46, Cargèse.
- JAKOBSON, R. (1963). *Essais de linguistique générale*. Paris: Edition de Minuit.
- KARLGREN, J. (1999). Stylistic experiments in information retrieval. In T. STRZALKOWSKI, Ed., *Natural language information retrieval*, Text, speech and language technology, chapter 6, pp. 147–166. Dordrecht: Kluwer.
- KERBRAT-ORECCHIONI, C. (1980). *L'énonciation de la subjectivité dans le langage*. Coll. Linguistique. Paris: Armand Colin.
- KILGARIFF, A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Fifth ACL Workshop on Very Large Corpora*, Beijing.
- KILGARIFF, A. & ROSE, T. (1998). Measures for corpus similarity and homogeneity. In *3rd Conference on Empirical Methods in Natural Language Processing*, pp. 46–52, Granada.
- LABBÉ, D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Cahier 7, CERAT – Institut d'Études Politiques de Grenoble, Saint-Martin d'Hères.
- LAFON, P., LEFEVRE, J. & SALEM, A. (1985). *Le machinal. Principes d'enregistrement informatique des textes*. Saint-Cloud. Klincksieck.
- LAKS, B. (1996). *Langage et cognition. L'approche connexionniste*. Langue, raisonnement, calcul. Paris: Hermès.
- LEBART, L. & SALEM, A. (1994). *Statistique textuelle*. Paris: Dunod.
- MAINGUENEAU, D. (1996). *Les termes clés de l'analyse du discours*. N° 20 in Mémo. Paris: Seuil.
- MARCHELLO-NIZIA, C. (1999). *Le français en diachronie : douze siècles d'évolution*. L'essentiel français. Paris: Ophrys.
- MARCUS, M., SANTORINI, B. & MARCINKIEWICZ, M. A. (1993). Building a large annotated corpus of english : The Penn Treebank. *Computational Linguistics*, 19(2), 313–330.
- MILNER, J.-C. (1989). *Introduction à une science du langage*. Des Travaux. Paris: Seuil, 1ère édition.
- NAULLEAU, E. (1998). *Transformation of Le Monde data to obtain PAROLE DTD conformance*. Technical report, INaLF – CNRS, Saint-Cloud.
- PEREC, G. (1991). *Cantatrix sopránica L. et autres écrits scientifiques*. La librairie du XX^e siècle. Paris: Seuil.
- PETITJEAN, A. (1987). Les faits divers : polyphonie énonciative et hétérogénéité textuelle. *Langue française*, (74), 73–96. La typologie des discours, J.-L. Chiss, J. Filliolet (eds).
- PETITJEAN, A. (1989). Les typologies textuelles. *Pratiques*.
- PÉRY-WOODLEY, M.-P. (1995). Quels corpus pour quels traitements automatiques ? *TAL*, 36(1–2), 213–232. Traitements probabilistes et corpus, Benoît Habert (resp.).
- QUIRK, R., GREENBAUM, S., LEECH, G. & SVARTVIK, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.
- REINERT, M. (1993). Les « mondes lexicaux » et leur « logique » à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, (66), 5–39.
- RENOUF, A. (1993). A word in time : first findings from the investigation of dynamic text. In J. AARTS, P. DE HAAN & N. OOSTDIJK, Eds., *English language corpora : design, analysis and exploitation*, pp. 279–288. Amsterdam: Rodopi.
- RÉGENT, O. (1992). Pratiques de communication en médecine : contextes anglais et français. *Langages*, (105), 66–75. Ethnolinguistique de l'écrit, Jean-Claude Beacco (éd.).
- SAPORTA, G. (1990). *Probabilités analyse des données et statistique*. Paris: Technip.
- SIGLEY, R. (1997). Text categories and where you can stick them: a crude formality index. *International Journal of Corpus Linguistics*, 2(2), 199–237.
- SIMONIN-GRUMBACH, J. (1975). Pour une typologie des discours. In *Langue, discours, société (pour Emile Benveniste)*, pp.85–121. Paris: Seuil.
- SINCLAIR, J. (1996). *Preliminary recommendations on Corpus Typology*. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards).
- SPARCK-JONES, K. & P. WILLETT, Eds. (1997). *Readings in Information Retrieval*. San Francisco, California: Morgan Kaufmann.
- SUEUR, J.-P. (1982). Pour une grammaire du discours : élaboration d'une méthode; exemples d'application. *MOTS*, (5), 145–185.
- SVARTVIK, J., EEG-OLOFSSON, M., FORSHEDEN, O., ORESTRÖM, B. & THAVENIUS, C. (1982). *Survey of Spoken English*. Lund: Lund University Press.
- T. TODOROV, Ed. (1978). *Les genres du discours*. Paris: Le Seuil.
- VION, R. (1999). Pour une approche relationnelle des interactions verbales et des discours. *Langage et société*, (87), 95–114.
- WERLICH, E. (1975). *Typologie der texte. Entwurf eines textlinguistischen Modells zur Grundlegung einer Textgrammatik*. Heidelberg: Quelle und Meyer.