Atelier Thématique TALN 1999, Cargèse, 12–17 juillet 1999

Maîtriser les déluges de données hétérogènes

G. Illouz*, B. Habert‡, S. Fleury‡, H. Folch‡, S. Heiden‡, P. Lafon‡

*LIMSI – Université Paris XI
‡UMR 8503 — ENS Fontenay/Saint-Cloud
gabrieli@limsi.fr, {bh, fleury, floch, slh, lafon}@ens-fcl.fr

Résumé

Le traitement automatique des langues fait de plus en plus appel à de volumineux corpus textuels pour l'acquisition des connaissances. L'obstacle actuel n'est plus la disponibilité de corpus, ni même leur taille, mais l'hétérogénéité des données qui sont rassemblées sous ce nom. Dans cet article, nous examinons l'hétérogénéité que manifestent les articles du *Monde* quand on les regroupe selon les rubriques de la rédaction du journal. Les conséquences d'une telle hétérogénéité pour l'étiquetage et le parsage sont soulignées. Partant de ce constat, nous définissons la notion de « profilage de corpus » par le biais d'outils permettant d'évaluer l'homogénéité d'un corpus (sur-emploi du vocabulaire, de catégories morpho-syntaxiques, ou de patrons) et l'utilisation qui peut en être faite.

1. La nouvelle donne : de grands corpus hétérogènes

Le traitement automatique des langues fait de plus en plus appel à de volumineux corpus textuels pour l'acquisition des connaissances qui lui sont nécessaires : cadres de sous-catégorisation des verbes, collocations, enchaînement de catégories... L'obstacle actuel n'est plus la disponibilité de corpus, ni même leur taille, mais l'hétérogénéité des données qui sont rassemblées sous ce nom. C'est le cas par exemple des données de presse qui, par leur accessibilité, sont souvent mises à contribution.

Les données du journal *Le Monde* disponibles 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* : ETR(anger), ECO(nomie), POL(itique), ING (?information générale : sport, faits divers), ART (médias, spectacles), EMS (?education, médecine, société), etc.[1]– et de *genres* multiples : biographie, chronique, chronologie, encadré, correspondance, entretien, opinion, portrait, rectificatif, revue de presse...

Dans la partie 2, nous examinons l'hétérogénéité que manifestent les articles du *Monde* quand on les regroupe selon les rubriques de la rédaction du journal. Dans la partie 3, nous

1. 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.

Rubrique	*articles*		*mots*		*moyenne*		*minimum*		*maximum*	
ETR(anger)	**5 464**	149	**2 366 055**	77 347	**433**	519	**13**	33	**3 624**	2 585
ECO(nomie)	**3 478**	108	**144 3923**	38 540	**415**	356	**15**	52	**3 058**	1 473
POL(itique)	**2 305**	83	**1 326 576**	36 703	**575**	442	**36**	74	**5 202**	1 604
ING(info. g^{ale})	**838**	80	**364 590**	34 284	**435**	427	**28**	44	**3 168**	2 109
ART(médias)	**2 261**	76	**1 080 620**	37 220	**477**	489	**1**	46	**2 990**	2 087
EMS(?Edu., Méd., Soc.)	**1 092**	42	**45 7626**	17 390	**419**	414	**25**	58	**3 551**	3 127
Total	**15 438**	538	**7 039 390**	241 484	**455**	448	**1**	33	**5 202**	3 127

TAB. 1 – *Caractéristiques des parties « **nue** » (extraits de 1987; 89; 91; 93; 95) et* catégorisée *(extraits de septembre 1987) du* Monde *par rubriques.*

soulignons les conséquences d'une telle hétérogénéité pour l'étiquetage et le parsage. Dans la partie 4, nous dégageons les caractéristiques des « profileurs de corpus »[2] qu'implique le recours actuel aux corpus en TALN.

2. Hétérogénéité dans *Le Monde* : un corpus peut en cacher un autre

2.1. Données : 14 millions de mots « nus » et 250 000 mots étiquetés

14 millions de mots provenant par choix aléatoire de numéros entiers parmi ceux des années 1987, 1989, 1991, 1993 et 1995 du journal *Le Monde* (Naulleau, 1998) constituent la partie *Presse* du corpus réalisé dans le cadre du projet européen `PAROLE`[3]. Nous dénommerons ce corpus `LeMondeNu`.

241 484 mots, provenant de 7 numéros du Monde de septembre 1987, ont été extraits de `LeMondeNu`, étiquetés automatiquement et corrigés manuellement pour la partie du discours, toujours dans le cadre de `PAROLE`. Nous dénommerons ce sous-corpus `LeMondePOS`.

Dans `LeMondeNu` et `LeMondePOS`, chaque article constitue un composant à part entière, pourvu d'un cartouche documentaire – *header* (Dunlop, 1995) – suivant les propositions de la TEI (*Text Encoding Initiative*). Les champs signalétiques fournis par la documentation du *Monde* ont été transformés en catégories classificatoires dans les cartouches. Il est ainsi possible d'extraire les articles relevant de différentes *rubriques* ou de différents *genres*. C'est l'homogénéité des rubriques principales représentées qui est examinée. Le tableau 1 fournit la taille, la longueur moyenne des articles de chacune des rubriques retenues (les rubriques qui occupaient une place trop faible dans `LeMondePOS` ont été éliminées des expériences présentes).

La partition de `LeMondeNu` en rubriques permet un premier examen, au niveau des mots, de l'hétérogénéité que manifestent les parties correspondantes. `LeMondePOS` permet d'affiner ce constat en s'attachant à la répartition des étiquettes ainsi qu'aux sur et sous-emplois de certains mots fonctionnels.

2. Nous définissons les « profileurs de corpus » comme des outils de calibrage donnant des indications sur l'emploi du vocabulaire, de catégories morpho-syntaxiques et de patrons dans les parties d'un corpus, pour en déterminer l'homogénéité ou l'hétérogénéité.

3. Ce projet a abouti à la constitution pour chacune des langues de la communauté d'un corpus de 20 millions de mots issus de textes récents (après 1980), se répartissant dans des proportions déterminées entre journaux, périodiques, livres, etc.

2.2. *La structuration par les rubriques de 14 millions de mots « nus »*

Méthode utilisée À chaque objet (ici, une partie du corpus), on associe un vecteur représentatif de celui-ci. L'analyse de ces vecteurs nous permet alors d'obtenir une classification des types de textes présents. Le choix des traits[4] est capital, et nombre de méthodes sont pertinentes pour représenter un texte, partant de traits surfaciques (longueur des paragraphes et des mots, jeu de caractères…), ou bien linguistiques (fréquence du vocabulaire présent dans le texte, proportion d'adjectifs, fréquence de certains motifs linguistiques…).

Les résultats présentés sont obtenus en utilisant des vecteurs dont les traits correspondent au nombre d'occurrences des formes lexicales les plus fréquentes (supérieures à 500 occurrences, soit 108 formes) sur l'ensemble de `LeMondeNu` (cf. parties 5 et 2.4). Le choix de la fréquence des formes les plus présentes est classiquement un bon moyen de contraster les corpus ou leurs parties ((Lebart & Salem, 1994)). C'est la donnée la plus immédiatement observable, à condition de se donner une segmentation simple et de ne pas lemmatiser les formes.

Notre but n'est pas de classer les données mais simplement d'observer si ces données se prêtent à une classification. Les vecteurs représentatifs sont explorés à l'aide de la méthode de Sammon (Sammon, 1969) qui, partant d'un nuage de points de dimension n, projette les données dans un espace de dimension moindre k ($k < n$), avec la propriété de conserver au mieux les distances existantes dans l'espace de départ.

Hypothèse à infirmer : les rubriques ne structurent pas le corpus Si les rubriques ne sont pas structurantes, des tranches de 10 000 mots consécutifs regroupées en ensembles de mêmes tailles que ceux obtenus à partir des rubriques (pour se retrouver dans une situation comparable à la partition en rubriques) devraient se disperser de façon aléatoire, ne permettant pas de retrouver des proximités entre textes d'une même rubrique. Les graphiques de la figure 1, construits à partir de l'application de la méthode de Sammon à une telle partition, manifestent de fait une dispersion où il est impossible de distinguer des regroupements, mis à part quelques fragments singuliers regroupés en haut de la partie *a* de cette première figure.

Contrastes entre rubriques En revanche, dans la figure 2, si les frontières entre classes sont floues, des proximités entre textes de même rubrique émergent. Dans le nuage principal, que l'on peut observer plus précisément dans la partie droite *b* de la figure, on distingue nettement les regroupements de certaines rubriques : ART en haut, ECO en bas à droite, ETR en bas au centre, POL en bas à gauche. Les rubriques EMS et ING, probablement plus hétérogènes vu leur intitulé, sont plus dispersées. Dans la partie *a*, on remarque par ailleurs la présence d'un « bruit », sans doute causé par des textes atypiques : il s'agit en fait du rassemblement de quelques textes de la rubrique POL qui correspondent à des résultats d'élections législatives (mars 1993)[5]. Ce sous-ensemble de POL se détache nettement du nuage principal et même du reste de POL. Une rubrique définie *a priori* et relativement homogène peut donc comprendre des sous-ensembles qui s'en éloignent.

4. En classification, un trait est une coordonnée du vecteur représentant les valeurs possibles d'une caractéristique jugée pertinente dans la prise de décision.

5. Une ligne de résultat est de la forme: `Ball. : Jean-Yves Haby, UDF-PR, d.s., adj. m. de Courbevoie, 20 327(47,56) ; Pierre Roussel, PS, c.m. de La Garenne-Colombes, 6 836 (15,99)`

2.3. *La structuration par les rubriques de 250 000 mots étiquetés*

Méthode Nous faisons l'hypothèse que la nature (ici la rubrique) des textes traités détermine partiellement la qualité de la plupart des traitements élaborés (comme l'étiquetage ou le parsage) : les expériences de la sections 3.1 confirment d'ailleurs cette hypothèse. Il ne nous était donc pas possible de partir des résultats non corrigés d'un étiqueteur quelconque pour comparer les répartitions des catégories d'une rubrique à l'autre. C'est pourquoi nous n'avons pas utilisé sur `LeMondeNu` les étiqueteurs dont nous disposions (`WinBrill`[6] et `Sylex`(Constant, 1991)) et que nous nous sommes volontairement limités à l'examen au sein de la partie corrigée manuellement (`LeMondePOS`) de l'influence des rubriques sur l'emploi des catégories.

Nous avons utilisé la méthode des spécificités (Lafon, 1980) pour dégager les sur et sous-emplois significatifs d'une catégorie dans une partie par rapport à sa répartition dans l'ensemble du corpus.

Contrastes `LeMondePOS` emploie les catégories[7] suivantes : N(om), V(erbe), A(djectif), D(éterminant), P(ronom), C(onjonction), R (adveRbe), S (prépoSition), I(nterjection), X (inconnu). Le tableau 2 résume les contrastes : le niveau d'abstraction des entités manipulées rend leur interprétation délicate (celle-ci serait facilitée par l'utilisation de catégories plus fines comme temps et mode des verbes, type de pronoms. . .). C'est pourquoi on a complété le tableau par la répartition des formes fonctionnelles les plus fréquentes (non lemmatisées) dans tout le corpus. On notera tout de même l'opposition entre ART qui valorise le nominal au détriment du verbal et POL qui manifeste la répartition inverse. Les mots fonctionnels sur-employés par POL soulignent la place des adverbes et des conjonctions (*que, si, qu', Si, soit, comme, lors*), en particulier de la négation (*ni, ne, n'*), tandis que les adverbes de degré marquent ART (*peu, très, trop*), ainsi que les pronoms personnels, tout particulièrement la « non-personne » (*on, On*) et un embrayeur (*je, Je*). Le sur-emploi de *M* dans POL correspond à l'abréviation de *monsieur*.

Rubrique	*Sous-emplois*	*Sur-emplois*	*Sur-emplois fonctionnels*
ART	D, S, V	N, P, R, C	Mais, Et, si, ou, comme, sa, Un, De, qui, on, y, lui, elle, c', C', On, je, Elle, Je, où, peu, très, trop, avec, sans
ECO	V, P, R, C	D, S	Cette, les, de, des, La, celui, plus, ainsi, pour, Selon, aux
EMS	N, R	C, X	et, Mais, ni, ou, cette, L', les, leur, leurs, tous, ils, Ils, aussi, aux
ETR	P, R, C, X	A	ces, la, le, l', Le, d', des, dont, où, lors, notamment, également, depuis, contre, entre, dans, par, selon, Selon, au, du, devant
ING	A, X	V, P, R	Et, son, Les, Une, il, lui, elle, C', je, Je, cela, En, n', encore, donc, après, sur, sans, sous, Au, A
POL	N, D, S, A	V, P, R, C	ni, que, si, qu', Si, soit, comme, ce, tous, il, se, Il, on, ils, nous, -t-il, ceux, Nous, Je, ne, n', lors, pour, au, du, M

TAB. 2 – *Sur-emplois et sous-emplois significatifs de catégories dans* `LeMondePOS` *(probabilité d'apparition inférieure à 5%).*

6. http://jupiter.inalf.cnrs.fr/WinBrill/

7. Les catégories utilisées sont celles qui ont été corrigées à la main. Seules les catégories principales sont utilisées. Ainsi, les connecteurs sont inclus dans les conjonctions ou les adverbes et les numéraux dans les déterminants.

FIG. 1 – *Projections de Sammon (108 dimensions vers 2 dimensions) de 6 groupes (de taille équivalente aux rubriques de la fig. 2) de tranches de 10 000 mots consécutifs (*a. en haut : *Toutes les tranches.* b. en bas : *Zoom sur le nuage principal).*

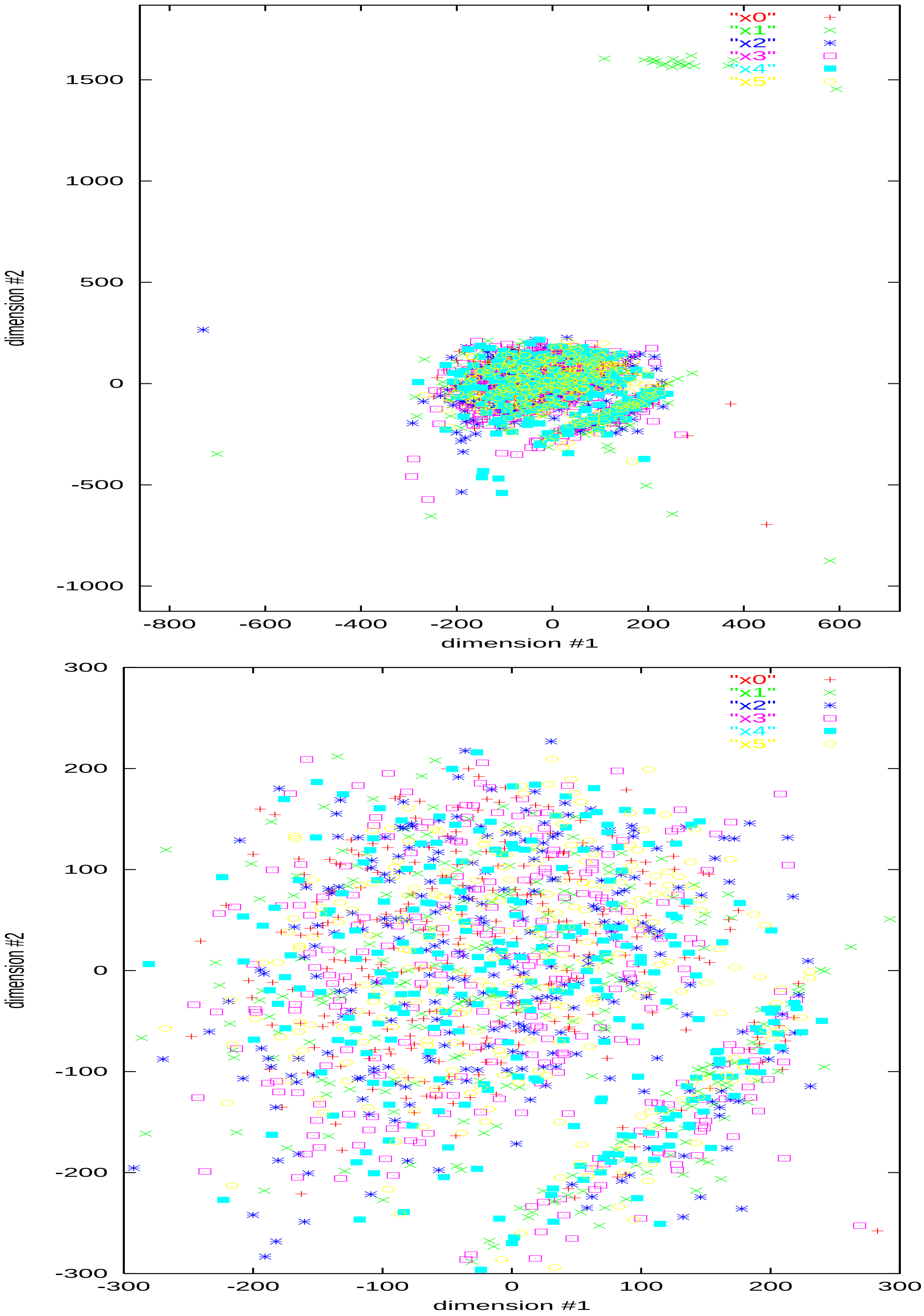
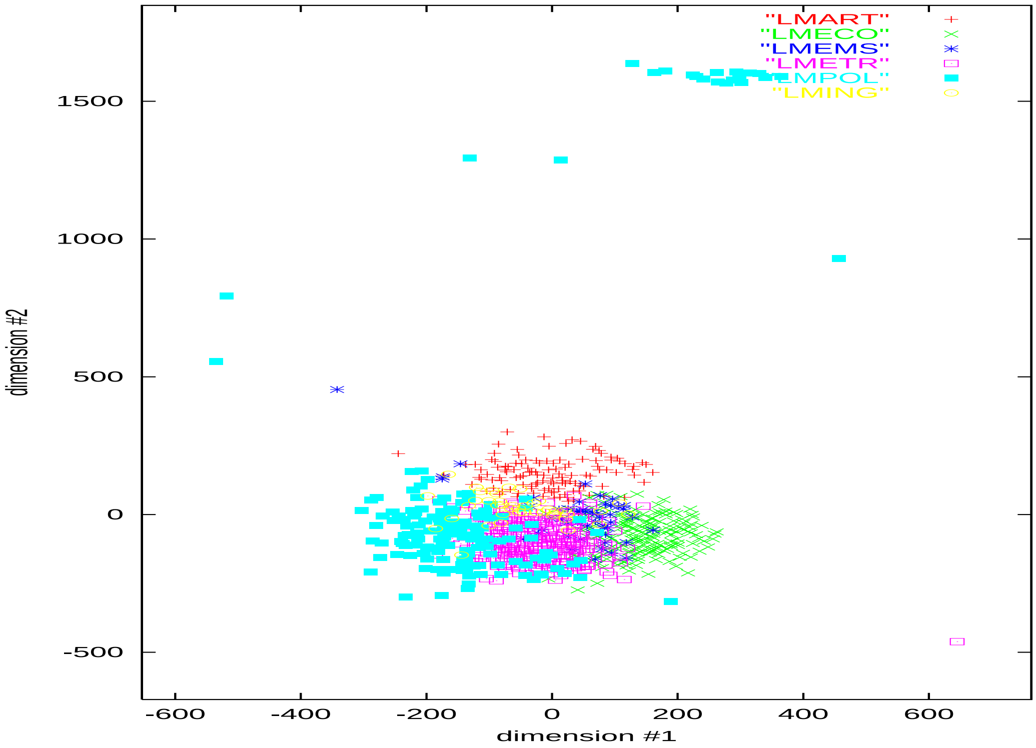
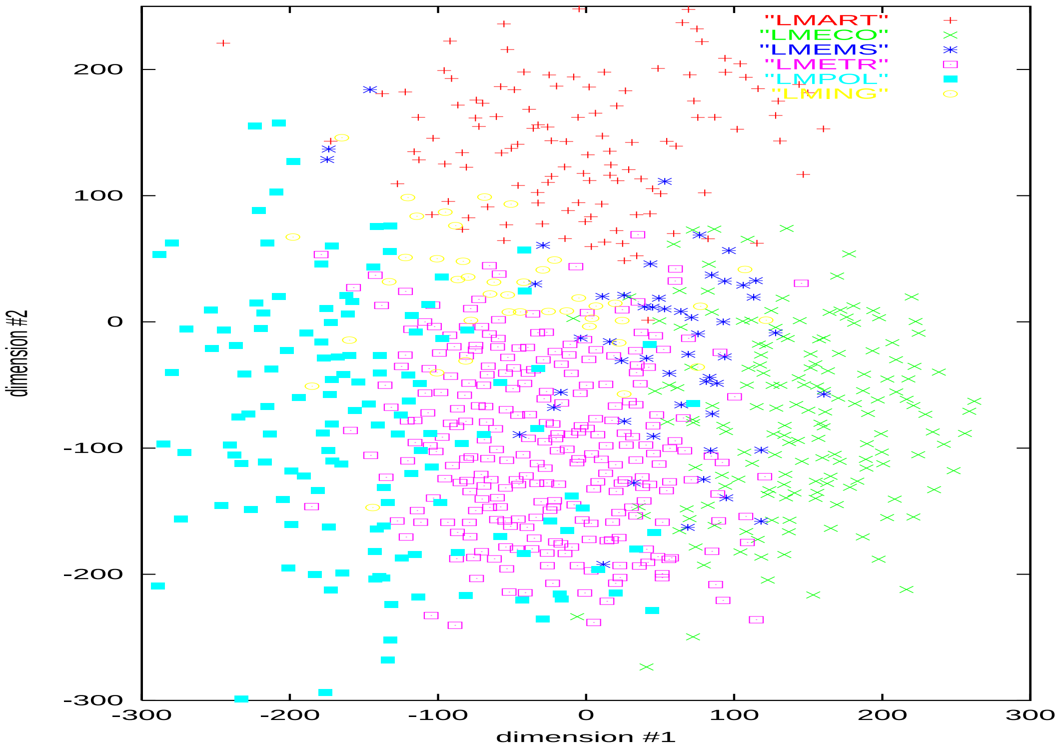


FIG. 2 – *Projections de Sammon (108 dimensions vers 2 dimensions) de tranches de 10 000 mots au sein de 6 rubriques (*a. en haut : *Toutes les tranches.* b. en bas : *Zoom sur le nuage principal).*





2.4. Bilan

L'effet des rubriques sur l'organisation du nuage peut être observé tant pour `LeMondeNu` que pour `LeMondePOS`. D'autres analyses avec d'autres vecteurs représentatifs (jeux de caractères, bigrammes d'étiquettes…), non intégrés ici faute de place, mènent à des résultats similaires. La partition des articles du *Monde* selon les rubriques manifeste en outre une relative stabilité quel que soit l'ensemble de traits employé. On peut par conséquent se contenter d'approximations du type de celles offertes par le vocabulaire le plus fréquent, la longueur des mots, etc.

Cependant, diminuer les tailles des tranches de textes relevant d'une même rubrique (10 000, 5 000, 2 500, 1 000 mots) dilue progressivement l'effet des rubriques qui devient peu perceptible pour 1 000 mots. Ce brouillage progressif dépend toutefois des traits utilisés : il est moins sensible pour les caractères que pour les mots et *a fortiori* pour les étiquettes et les bigrammes d'étiquettes.

3. La fiabilité des traitements dépend de l'homogénéité des données

La présente partie est consacrée aux conséquences pour deux tâches du TALN – l'étiquetage et le parsage – de l'hétérogénéité interne des corpus du type de celle que nous venons de constater.

3.1. Étiquetage

D. Biber a montré (Biber, 1993, p. 223) sur des corpus subdivisés en domaines que la probabilité d'une catégorie morpho-syntaxique donnée est fonction du domaine. Il a souligné également (*ibid.*, p. 225) les différences dans l'enchaînement des probabilités des catégories morpho-syntaxiques d'un domaine à l'autre, ainsi que les différences dans les collocations.

La précision des différents systèmes d'étiquetage dans le corpus *MULTEXT* (résultant de l'action `GRACE`), mesurée par rapport au corpus de référence corrigé manuellement, manifeste des variations significatives en fonction de la partie de ce corpus concernée. Ce corpus de 100 000 mots rassemble en effet des extraits du *Monde* (2) et de textes littéraires : mémoires (2), romans (6), essais (2). Ainsi, les mémoires font chuter les performances de la plupart des étiqueteurs (Illouz, 1999).

Dans la perspective de la mise au point d'un étiqueteur, se limiter à un seul domaine ou à un seul genre biaiserait donc singulièrement les apprentissages. « Additionner » les textes de provenance variée sans avoir vérifié leur homogénéité aboutirait à des résultats sans doute peu utilisables.

3.2. Parsage

Slocum (Slocum, 1986) montre l'existence de règles syntaxiques différentes à utiliser selon deux types de textes, en allemand. Ces deux types de textes sont composés de deux manuels écrits par des ingénieurs et de deux brochures écrites par des commerciaux. Il propose aussi un moyen de caractériser le type *manuel* (impératifs, acronymes, suppressions de déterminants) par rapport au type *brochure commerciale* (phrases longues, utilisation des pronoms, syntaxe plus riche).

Sekine (Sekine, 1998) utilise 8 sections du corpus Brown qu'il sépare en deux « classes » :

fiction (fiction, western, romans sentimentaux) et *non-fiction* (reportages, éditoriaux, loisirs, sciences). Il examine les performances, mesurées en rappel/précision, d'un analyseur syntaxique probabiliste selon que l'apprentissage de la grammaire s'effectue sur la même section ou la même classe que celle du test, sur toutes les sections confondues. Les performances vont en général dans l'ordre décroissant suivant : identité section d'apprentissage/de test, appartenance des sections d'apprentissage/de test à la même « classe », apprentissage sur un corpus relevant de toutes les sections à la fois. Entraîner l'analyseur sur une classe (*fiction* par exemple) et l'utiliser sur l'autre classe (*non-fiction*) donne les résultats les plus mauvais.

4. « Profilage » de corpus : tâches et conditions

4.1. Une nécessité pour les corpus à géométrie variable

À côté des corpus « fermés », mis au point une fois pour toute, existent désormais des « réservoirs de corpus » ou des « corpus ouverts ». Le BNC[8] (*British National Corpus*), qui rassemble 100 millions de mots étiquetés (dont 10 % d'oral) d'anglais contemporain, constitue par exemple un réservoir de corpus : les données signalétiques attachées à chaque composant permettent de réaliser « à façon » un corpus répondant à une recherche particulière (par exemple, l'oral des personnes d'une certaine classe d'âge). D'autres corpus accueillent sans cesse de nouveaux composants (corpus de suivi ou de veille – *monitor corpus*). 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`). C'est le cas aussi de la textothèque politique de l'UMR 8503 : elle accueille des collections de textes très variés et sur une longue période (1750-1999) : pétitions, tracts, discours parlementaires, éditoriaux de presse… Ces corpus à géométrie variable nécessitent 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.

4.2. Profilages : apports d'une démarche typologique inductive

Les traits retenus permettent de percevoir des écarts importants entre des parties définies *a priori* (ici les rubriques du *Monde*, mais de très nombreux autres champs signalétiques sont utilisables, comme les « genres »). Cependant, les composants de corpus ne sont pas toujours assortis de tels champs signalétiques. On a constaté en outre qu'une partie, homogène par rapport aux autres, peut comporter des sous-ensembles qui s'en distinguent fortement (POL et résultats d'élections). Par ailleurs, les classifications *a priori* foisonnent, sans converger pour autant et sans que le choix de telle ou telle s'impose. Il n'est pas sûr enfin qu'une classification d'un ensemble de textes « en aveugle » déboucherait forcément sur des regroupements correspondant aux catégories attribuées par ailleurs à ces textes. Pour ces trois raisons, nous proposons de développer en parallèle une typologie inductive des textes dans l'optique proposée par D. Biber.

Dégager les corrélations de traits linguistiques pour aboutir à des types de textes, c'est la ligne directrice des travaux de D. Biber (*ibid.*). Il examine les cooccurrences entre 67 traits linguistiques dans les 1 000 premiers mots de textes d'anglais contemporain écrit et oral relevant de « genres » divers : articles de recherche, reportages, conversations, nouvelles radiopho-

8. `http://info.ox.ac.uk/bnc/`.

niques... Les traits étudiés ressortissent à 16 catégories distinctes comme marqueurs de temps et d'aspect, adverbes et locutions adverbiales de temps et de lieu, pronoms et pro-verbes, questions, passifs, modaux, coordination, négation... Ils sont identifiés automatiquement (en limitant au maximum la vérification manuelle). La statistique multidimensionnelle permet alors d'obtenir des pôles multiples, positifs et négatifs, correspondant à des constellations de traits linguistiques corrélés. Ces pôles deux à deux constituent des dimensions. Chaque texte, par son emploi des traits linguistiques étudiés, se situe en un point déterminé de l'espace à n dimensions déterminé par cette analyse. À partir de ces dimensions, en utilisant des techniques de classification automatique, Biber aboutit à différents types de textes, en fonction de leur place sur chacune de ces dimensions. Ces types de textes ne recoupent pas directement les « genres » intuitivement distingués. Par exemple, l'oral se trouve réparti entre deux types de textes.

5. Vers des « profileurs » à géométrie variable

L'objectif de Biber est l'inclusion d'un grand nombre de caractéristiques linguistiques représentant l'éventail des possibilités fonctionnelles de l'anglais. Il s'intéresse à des fonctionnements linguistiques très spécifiques. La recherche de « patrons » fins et de marqueurs d'un trait donné opère en aval d'un étiquetage morpho-syntaxique préalable. Un certain nombre de traits ne pouvant être identifiés automatiquement sont écartés.

L'intégration du profilage dans la constitution et l'utilisation des vastes corpus hétérogènes actuels entraîne des contraintes différentes. En premier lieu, la taille des données à profiler conditionne l'éventail de traits examinés. Ce sont ainsi les traits de surface, aisément détectables par des analyseurs lexicaux et moins sujets à discussion (Kilgariff, 1997), qui seront privilégiés pour des données volumineuses. En second lieu, la nature de la tâche envisagée conduit à des priorités distinctes. Un étiqueteur stochastique sera peut-être plus sensible à l'homogénéité des données d'apprentissage qu'un outil d'extraction d'information. Enfin, il s'agit de rendre utilisables les profils mis en évidence, en fournissant des indicateurs à la fois robustes et compréhensibles. En effet, les expériences de la partie 2 manifestent les divergences de comportement des rubriques, elles ne permettent pas dans l'immédiat de comprendre les différences sous-jacentes d'usages langagiers.

La mise au point d'un « profileur » de corpus apparaît comme une étape nécessaire pour la maîtrise de l'hétérogénéité interne d'un corpus et pour son utilisation optimale en TALN. Plus généralement, elle conditionne la validité des généralisations qui peuvent être faites à partir d'un corpus.

Références

BIBER D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, **19**(2), 243–258.

CONSTANT P. (1991). *Analyse syntaxique par couche*. Doctorat de l'enst, École Nationale Supèrieure des Télécommunications, Paris.

DUNLOP D. (1995). Practical considerations in the use of TEI headers in large corpora. *Computers and the Humanities*, (29), 85–98.

ILLOUZ G. (1999). Méta-étiqueteur adaptatif : Vers une utilisation pragmatique des ressources linguistiques. In *TALN99*, Cargèse.

KILGARIFF A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Fifth ACL Workshop on Very Large Corpora*, Pékin.

LAFON P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *MOTS*, (1), 128–165. Presses de la Fondation Nationale des Sciences Politiques.

LEBART L. & SALEM A. (1994). *Statistique textuelle*. Paris: Dunod.

NAULLEAU E. (1998). *Tranformation of* Le Monde *data to obtain PAROLE DTD conformance*. Rapport interne, INaLF – CNRS, Saint-Cloud.

SAMMON J. (1969). A non linear mapping for data structure analysis. *IEEE Transactions on Computing*, (18), 401–409.

SEKINE S. (1998). The domain dependence of parsing. In *Fifth Conference on Applied Natural Language Processing*, p. 96–102, Washington.

SLOCUM J. (1986). How one might automatically identify and adapt to a sublanguage. In R. GRISHMAN & R. KITTREDGE, Eds., *Analyzing Language in Restricted Domains*, chapter 11, p. 195–210. Hillsdale, NJ: Lawrence Erlbaum Ass.