

TABLE DES MATIÈRES

INTRODUCTION	7
1. Le regain d'intérêt pour les corpus	7
2. À quoi servent les corpus annotés?	8
2.1. <i>La linguistique descriptive anglo-saxonne et ses questions</i>	8
2.2. <i>Le changement de cap en TALN</i>	10
3. Choix terminologiques	11
4. Notations	12
5. Orientation de l'ouvrage	13
5.1. <i>L'écrit au travers de corpus enrichis de langues vivantes</i>	13
5.2. <i>Les corpus, les ressources et les recherches de langue anglaise</i>	14
5.3. <i>Un point de vue aux frontières de la linguistique</i>	14
5.4. <i>La diversité des publics concernés</i>	15
6. Démarche suivie	15
6.1. <i>Les corpus annotés et leurs utilisations</i>	15
6.2. <i>Dimensions transversales</i>	16
6.3. <i>Méthodologies et techniques</i>	16
7. Principaux corpus cités	17
7.1. <i>Corpus anglais ou américains</i>	17
7.2. <i>Corpus français</i>	18

PREMIÈRE PARTIE

LES CORPUS ANNOTÉS ET LEURS UTILISATIONS

CHAPITRE I. LES CORPUS ÉTIQUETÉS	21
1. Définitions	21
1.1. <i>Exemples</i>	21
1.2. <i>L'inévitable éparpillement des étiquetages</i>	23
1.3. <i>Une représentation canonique</i>	24
1.4. <i>Types d'étiquetage</i>	26
2. Étiquetage partiel et typologie de textes	28
2.1. <i>Circularité des démarches typologiques habituelles</i>	28
2.2. <i>Dégager les corrélations de traits linguistiques : D. Biber</i>	29
2.3. <i>Généralité des typologies induites</i>	30
3. Étiquetage intégral et socio-stylistique	32
3.1. <i>Caractériser les types de locuteurs</i>	32
3.2. <i>Varié le jeu d'étiquettes selon les phénomènes observés</i>	32
3.3. <i>Une première opposition : style nominal et style verbal</i>	34
3.4. <i>Patrons syntaxiques caractéristiques pour chaque type de locuteur</i>	35
3.5. <i>Préciser l'emploi des adjectifs : qualificatifs et relationnels</i>	36
3.6. <i>Évaluation et perspectives</i>	37
4. Utiliser étiqueteurs et corpus étiquetés	37
4.1. <i>Adapter l'étiquetage aux objectifs de recherche</i>	37
4.2. <i>Environnements de catégorisation et de manipulation de texte étiqueté</i>	39
5. Enjeux théoriques	40
5.1. <i>Le dit est le dire</i>	40
5.2. <i>Linguistique et textualité</i>	41
5.3. <i>Analyses multidimensionnelles</i>	41

CHAPITRE II. LES CORPUS ARBORÉS.....	43
1. Diversité des corpus arborés.....	43
1.1. Noter des relations syntaxiques.....	44
1.2. Obtenir des analyses.....	48
1.3. Types d'analyse.....	48
1.4. Analyseurs de texte « tout-venant ».....	50
1.5. Niveaux d'analyse.....	52
2. Une réalisation exemplaire : Suzanne.....	54
2.1. Une annotation « exhaustive ».....	54
2.2. Informations fournies dans Suzanne.....	55
3. Phraséologie et traitements syntaxiques.....	55
3.1. Le renouveau des études linguistiques de la phraséologie.....	56
3.2. La flexibilité en corpus d'expressions polylexicales.....	58
3.3. La variation de termes en langue de spécialité.....	60
3.4. La recherche de candidats termes.....	65
3.5. Enjeux pratiques et théoriques.....	69
4. Utiliser des parseurs et des corpus arborés.....	71
4.1. Utiliser des parseurs.....	71
4.2. Utiliser des corpus arborés.....	71
CHAPITRE III. LES RESSOURCES LEXICALES POUR L'ÉTIQUETAGE	
SÉMANTIQUE.....	73
1. Un objectif : la désambiguïsation lexicale.....	74
2. Une opposition fondamentale : construction lexicale ou conceptuelle.....	75
2.1. Bases de connaissances lexicales.....	76
2.2. Bases de connaissances conceptuelles.....	79
2.3. Une opposition réelle mais floue.....	80
3. Une grande diversité de ressources lexicales.....	81
3.1. Des distinctions de sens plus ou moins fines.....	81
3.2. Des ressources générales ou spécialisées.....	82
3.3. Des sources plus ou moins informatisées.....	83
4. Un réseau lexical : WordNet.....	85
4.1. Un projet ambitieux.....	85
4.2. Une structure riche et différenciée.....	88
5. Tabler sur l'existant.....	90
DEUXIÈME PARTIE	
DIMENSIONS TRANSVERSALES	
CHAPITRE IV. DES MOTS AUX SENS : SÉMANTIQUE EN CORPUS.....	95
1. Définitions et enjeux.....	95
1.1. Un objectif commun : accéder au sens.....	95
1.2. Des applications variées.....	96
2. Construire automatiquement des entrées de dictionnaire.....	98
2.1. Des ébauches d'entrées de dictionnaires.....	99
2.2. Une méthode entièrement automatique.....	102
2.3. Les limites d'une approche empirique.....	104
3. Distinguer des sens pour la recherche documentaire.....	105
3.1. Retrouver des textes dans une base documentaire.....	106
3.2. Désambiguïser des corpus à l'aide de WordNet.....	107
3.3. De la désambiguïsation lexicale à la recherche documentaire.....	113
4. Un même parti pris d'empirisme.....	114
4.1. Fonder une sémantique sur les corpus.....	114
4.2. Exploiter des résultats approximatifs.....	115
4.3. Combiner des techniques simples.....	116
4.4. Modéliser par ajustements successifs.....	117
4.5. Expérimenter pour mieux expliquer.....	118

CHAPITRE V. LE LANGAGE AU FIL DU TEMPS : CORPUS ET DIACHRONIE	121
1. Définitions et enjeux.....	121
2. Un corpus pour l'étude de la diachronie : <i>Archer</i>	122
2.1. <i>L'anglais et l'américain de 1650 à aujourd'hui</i>	122
2.2. <i>Échantillonnage de registres</i>	123
2.3. <i>Structuration temporelle</i>	124
2.4. <i>Représenter les états de langue ou des idiolectes ?</i>	124
3. Études de la diachronie.....	125
3.1. <i>La courte durée</i>	125
3.2. <i>La moyenne durée</i>	126
3.3. <i>La longue durée</i>	127
4. Problèmes méthodologiques.....	130
4.1. <i>Des corpus « petits » et peu annotés</i>	131
4.2. <i>Vérifier et préciser les évolutions</i>	132
4.3. <i>Acceptabilité et fréquence</i>	132
4.4. <i>Affiner les explications</i>	133
CHAPITRE VI. D'UNE LANGUE À L'AUTRE : LES CORPUS ALIGNÉS	135
1. Définition et exemples.....	135
2. Utilisation des textes alignés.....	137
3. Méthodes d'alignement.....	138
4. Problèmes et enjeux.....	139
TROISIÈME PARTIE	
MÉTHODES ET TECHNIQUES	
CHAPITRE VII. CONSTITUER UN CORPUS	143
1. Définitions et typologie des corpus.....	143
2. Langue générale.....	145
2.1. <i>Étudier une dimension particulière</i>	145
2.2. <i>Constituer un corpus de référence</i>	146
2.3. <i>Peut-on constituer des échantillons représentatifs ?</i>	148
3. Langues de spécialité et sous-langages.....	148
3.1. <i>Les hypothèses de Z. Harris</i>	148
3.2. <i>Analyses de sous-langages</i>	149
3.3. <i>Évaluation et perspectives</i>	150
4. Articuler typologie interne et typologie externe.....	152
4.1. <i>Typologie des textes, genres et registres</i>	152
4.2. <i>Typologie des paramètres situationnels</i>	152
5. Normaliser un corpus.....	153
5.1. <i>Représentations logiques : SGML</i>	153
5.2. <i>Les types de textes : TEI</i>	155
6. Documenter un corpus.....	156
6.1. <i>Origine et histoire du corpus</i>	156
6.2. <i>Jurisprudence d'annotation</i>	157
7. Contraintes et conditions institutionnelles.....	158
7.1. <i>Assises institutionnelles</i>	158
7.2. <i>Problèmes juridiques</i>	159
CHAPITRE VIII. ANNOTER UN CORPUS	161
1. Nettoyage et homogénéisation.....	161
2. Segmentation.....	162
2.1. <i>Repérer les unités</i>	162
2.2. <i>Techniques</i>	163
2.3. <i>Difficultés</i>	164
3. Étiquetage morpho-syntaxique.....	165
3.1. <i>Taux d'ambiguïté</i>	165

3.2. Désambiguïsation par règles	166
3.3. Désambiguïsation probabiliste	167
3.4. Performances	168
3.5. Post-traitement et coûts	169
3.6. Évaluation et nouvelles tendances	169
4. Analyse syntaxique	170
4.1. Structuration par règles	170
4.2. Structuration probabiliste	171
4.3. Performances et évaluation	172
4.4. Post-traitement	173
4.5. Coûts	176
4.6. Difficultés	176
5. Étiquetage sémantique	177
5.1. Construire des catégories sémantiques	177
5.2. Projeter des catégories sur un corpus	181
CHAPITRE IX. QUANTIFIER LES FAITS LANGAGIERS.....	183
1. Pourquoi quantifier ?	184
1.1. Étudier la variation de traits linguistiques dans un corpus	184
1.2. Réaliser des typologies de textes et de documents	185
1.3. Déceler des corrélations entre phénomènes	185
2. Les unités.....	186
2.1. Normes de dépouillement	187
2.2. Décomptes automatisés	188
2.3. Incidence de la norme sur les décomptes	189
2.4. Exemple : l'accroissement du vocabulaire	190
3. Mesures de récurrence sur l'axe syntagmatique	191
3.1. Séquences d'unités	191
3.2. Quasi-segments	192
3.3. Cooccurrences	192
3.4. Filtrage des résultats	193
4. Comparer des décomptes au sein d'un corpus partitionné.....	194
4.1. Organiser la partition du corpus	195
4.2. Repérer les faits saillants	196
5. Approches multidimensionnelles.....	198
5.1. Classer les unités et les textes	198
5.2. L'approche factorielle	201
6. Articuler des constats sur des unités différentes.....	204
6.1. Articuler unités isolées et séquences d'unités	204
6.2. Articuler différents systèmes d'unités	205
7. Temps lexical.....	207
7.1. Accroissements spécifiques	208
7.2. Formes chrono-homogènes	209
8. Bilan.....	211
CONCLUSION.....	213
BIBLIOGRAPHIE	219
INDEX.....	230

Masson & Armand Colin Éditeurs
 34 bis, rue de l'Université - 75007 Paris
 N° 1775/01
 Dépôt légal : décembre 1997

Achévé d'imprimer sur les presses de la
 SNEL S.A.
 rue Saint-Vincent 12 - B-4020 Liège
 tél. 32(0)4 343 76 91 - fax 32(0)4 343 77 50
 novembre 1997
 8472