

**Multilinguisme  
et traitement  
de l'information**

*sous la direction de*  
Frédérique Segond

**hermes**  
**Science**  
— publications —

*Il a été tiré de cet ouvrage  
25 exemplaires hors commerce réservés  
aux membres du comité scientifique,  
aux auteurs et à l'éditeur  
numérotés de 1 à 25*

7677

***IST 2761***

Multilinguisme et traitement de l'information

BIBLIOTHEQUE DU CERIST

© LAVOISIER, 2002

LAVOISIER

11, rue Lavoisier

75008 Paris

Serveur web : [www.hermes-science.com](http://www.hermes-science.com)

ISBN 2-7462-0523-8

---

Catalogage Electre-Bibliographie

Segond, Frédérique (sous la direction de)

Multilinguisme et traitement de l'information

Paris, Hermès Science Publications, 2002

ISBN 2-7462-0523-8

RAMEAU : analyse automatique (linguistique)

multilinguisme : ressources internet

DEWEY : 418 : Linguistique appliquée. Traduction

Généralités

---

Le Code de la propriété intellectuelle n'autorisant, aux termes de l'article L. 122-5, d'une part, que les "copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective" et, d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, "toute représentation ou reproduction intégrale, ou partielle, faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause, est illicite" (article L. 122-4). Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles L. 335-2 et suivants du Code de la propriété intellectuelle.

---

COMITÉ SCIENTIFIQUE DU  
TRAITÉ DES SCIENCES ET TECHNIQUES DE L'INFORMATION

Jacques Rouault  
Hubert Fondin  
Brigitte Guyot

---

Le Traité des Sciences et Techniques de l'Information répond au besoin de disposer d'un fonds commun de connaissances dans les domaines où les technologies ont transformé la production et la gestion de l'information. Absorbant et transformant des disciplines traditionnelles, comme la documentation et la bibliothéconomie, les sciences de l'information occupent en effet un champ à part, où se rencontrent l'informatique et les sciences humaines et sociales.

Les ouvrages du Traité des Sciences et Techniques de l'Information analysent les différents aspects d'une discipline étendue et diverse : les documents et leur gestion, l'information dans la société, la communication homme-machine, la représentation et l'instrumentalisation des connaissances. Ils abordent les technologies de base, les institutions, les grandes options méthodologiques et les savoir-faire.

Chaque volume étudie aussi bien les aspects fondamentaux qu'expérimentaux. Une classification des différents chapitres contenus dans chacun, une bibliographie et un index détaillé orientent le lecteur vers ses points d'intérêt immédiats : celui-ci dispose ainsi d'un guide pour ses réflexions ou pour ses choix.

BIBLIOTHEQUE DU CERIST

## Liste des auteurs

Elisabeth ANDRÉ  
Institut für Informatik  
Universität d'Augsbourg  
Allemagne

Caroline BRUN  
Xerox Research Centre Europe  
Meylan

Stéphane CHAUDIRON  
CRIS  
Université Paris X  
Nanterre

Yun-Chuang CHIAO  
DIAM  
CHU Pitié-Salpêtrière  
Université Paris VI

Eva DAUPHIN  
Centre commun de recherche  
EADS  
Toulouse

Thierry DECLERCK  
DFKI  
Sarrebruck, Allemagne

Luca DINI  
CELI  
Turin, Italie

Marc DYMETMAN  
Xerox Research Centre Europe  
Meylan

Gregory GREFENSTETTE  
Clairvoyance Corporation  
Pittsburgh, Etats-Unis

Geoffrey NUNBERG  
CSLI  
Université de Stanford  
Etats-Unis

Thibault PARMENTIER  
Objet Direct  
Veurey

Sylvie REGNIER  
Centre commun de recherche  
EADS  
Suresnes

Ágnes SÁNDOR  
Xerox Research Centre Europe  
Meylan

Frédérique SEGOND  
Xerox Research Centre Europe  
Meylan

Monique SLODZIAN  
Centre de recherche en ingénierie  
multilingue  
Inalco  
Paris

Jean-David STA  
EDF R&D  
Clamart

BIBLIOTHEQUE DU CERIST

## Table des matières

<b>Introduction</b> . . . . .	17
<b>PREMIÈRE PARTIE. INTERNET ET MULTILINGUISME</b> . . . . .	25
<b>Chapitre 1. Internet et ses enjeux linguistiques</b> . . . . .	27
Geoffrey NUNBERG	
1.1. Introduction . . . . .	27
1.2. Une enquête . . . . .	34
1.3. Effets d'Internet sur les communautés linguistiques . . . . .	40
1.4. Conclusion . . . . .	45
1.5. Bibliographie . . . . .	46
<b>Chapitre 2. Présence des langues sur le WWW et construction des ressources linguistiques</b> . . . . .	47
Gregory GREFENSTETTE	
2.1. Introduction . . . . .	47
2.2. L'accès à l'information . . . . .	48
2.3. Estimation de nombre de mots dans une langue sur le web . . . . .	49
2.3.1. Algorithme de prédiction de nombre de mots . . . . .	51
2.3.2. Utilisation d'un portail pour estimer la présence d'une langue . . . . .	51
2.4. Les outils nécessaires à un accès multilingue . . . . .	56
2.4.1. Ensemble minimal de ressources . . . . .	56
2.4.2. Exemple d'extension automatique des ressources . . . . .	57
2.4.3. Gestion des ressources . . . . .	58
2.5. Conclusions . . . . .	60
2.6. Bibliographie . . . . .	61

**Chapitre 3. La question du multilinguisme en contexte**

**de veille automatisée sur Internet . . . . . 63**

Stéphane CHAUDIRON

- 3.1. Introduction . . . . . 63
- 3.2. Le contexte de veille sur Internet . . . . . 64
  - 3.2.1. Les différentes facettes de la veille . . . . . 64
  - 3.2.2. Caractéristiques du contexte de veille . . . . . 65
- 3.3. Spécificité de la veille sur Internet . . . . . 66
  - 3.3.1. L'existence de différentes sphères informationnelles . . . . . 66
  - 3.3.2. Risque de surinformation . . . . . 67
  - 3.3.3. Fiabilité et pertinence . . . . . 67
  - 3.3.4. Renouvellement continu . . . . . 68
  - 3.3.5. Instabilité de la localisation . . . . . 68
  - 3.3.6. Fragmentation . . . . . 68
  - 3.3.7. Information non structurée . . . . . 69
- 3.4. L'information multilingue sur Internet . . . . . 69
  - 3.4.1. Répartition des langues . . . . . 69
  - 3.4.2. Contraintes techniques . . . . . 70
  - 3.4.3. Codage des jeux de caractères . . . . . 70
  - 3.4.4. Balisage de l'information . . . . . 73
  - 3.4.5. Classement de l'information . . . . . 74
- 3.5. Analyse de l'offre en matière d'outils  
de veille stratégique automatisée . . . . . 75
  - 3.5.1. Recherche et collecte de l'information . . . . . 75
    - 3.5.1.1. Navigateurs . . . . . 75
    - 3.5.1.2. Outils de recherche . . . . . 76
  - 3.5.2. Analyse et traitement de l'information . . . . . 78
    - 3.5.2.1. Analyse et visualisation . . . . . 78
    - 3.5.2.2. Résumé automatique . . . . . 80
    - 3.5.2.3. Traduction automatique . . . . . 81
  - 3.5.3. Plates-formes de veille . . . . . 81
- 3.6. Perspectives . . . . . 82
- 3.7. Bibliographie . . . . . 82

**DEUXIÈME PARTIE. PRODUCTION D'INFORMATION MULTILINGUE . . . . . 87**

**Chapitre 4. Terminologie et multilinguisme :**

**des principes à l'application . . . . . 89**

Monique SLODZIAN

- 4.1. Introduction . . . . . 89
- 4.2. La doctrine terminologique . . . . . 91
  - 4.2.1. Une sémiotique du signe « pur » . . . . . 91
  - 4.2.2. Le refoulé linguistique . . . . . 93
  - 4.2.3. La langue de spécialité (LSP) . . . . . 94

4.2.4. Le découpage du terme . . . . .	96
4.2.5. L'autorité de la norme . . . . .	96
4.3. Le nouveau paysage de la terminologie . . . . .	97
4.3.1. La question du domaine . . . . .	97
4.3.2. L'explosion textuelle . . . . .	98
4.3.3. Le terme dans tous ses états . . . . .	99
4.4. Le renouveau de la terminologie . . . . .	99
4.4.1. Retour au contexte . . . . .	99
4.4.2. La terminologie textuelle . . . . .	100
4.4.3. Des méthodes et des outils . . . . .	101
4.4.4. Le primat de la tâche . . . . .	102
4.4.5. De la terminologie à l'ontologie . . . . .	103
4.5. Le temps du multilinguisme . . . . .	104
4.5.1. De quelques ambiguïtés . . . . .	104
4.5.2. La variabilité des langues . . . . .	104
4.5.3. Constitution de ressources multilingues hors-texte . . . . .	105
4.5.4. Constitution de ressources multilingues non taxinomiques . . . . .	106
4.6. Conclusion . . . . .	106
4.7. Bibliographie . . . . .	108

## **Chapitre 5. Accès à l'information multilingue et terminologie . . . . . 111**

Yun-Chuang CHIAO et Jean-David STA

5.1. Introduction . . . . .	111
5.2. Construction d'une terminologie multilingue . . . . .	113
5.2.1. Incomplétude des systèmes terminologiques . . . . .	113
5.2.2. Constitution d'un corpus pour l'acquisition terminologique . . . . .	114
5.2.3. Extraction de termes à partir de corpus . . . . .	115
5.2.4. Extraction de relations entre termes à partir de corpus . . . . .	116
5.2.5. Construction d'une terminologie multilingue à partir de corpus . . . . .	117
5.2.6. Terminologie et multilinguisme . . . . .	117
5.3. Reformulation par traduction de la requête . . . . .	118
5.3.1. Introduction . . . . .	118
5.3.2. Différentes méthodes de traduction . . . . .	119
5.3.3. Désambiguïsation de la traduction d'une requête . . . . .	120
5.4. Reformulation par extension de la requête . . . . .	121
5.4.1. Introduction . . . . .	121
5.4.2. Différentes approches pour étendre une requête . . . . .	122
5.4.2.1. Extension à partir des textes . . . . .	122
5.4.2.2. Extension à partir d'un réseau sémantique . . . . .	123
5.4.3. Extension de requête et multilinguisme . . . . .	123
5.4. Conclusion . . . . .	124
5.5. Bibliographie . . . . .	125

**Chapitre 6. Rédaction multilingue assistée dans le modèle MDA . . . . . 129**

Caroline BRUN et Marc DYMETMAN

- 6.1. Introduction. . . . . 129
  - 6.1.1. Notre approche. . . . . 130
  - 6.1.2. Structure de ce chapitre . . . . . 131
- 6.2. Caractéristiques fonctionnelles du système MDA . . . . . 131
- 6.3. Le modèle algorithmique . . . . . 133
  - 6.3.1. Grammaires hors-contexte et arbres de choix . . . . . 133
  - 6.3.2. Explicitation des choix . . . . . 134
  - 6.3.3. Grammaires abstraites . . . . . 135
  - 6.3.4. Types dépendants . . . . . 136
  - 6.3.5. Grammaires parallèles et compositionnalité pilotée par la sémantique . . . . . 137
  - 6.3.6. Arbres hétérogènes et interactivité . . . . . 138
- 6.4. Applications . . . . . 139
  - 6.4.1. Choix du domaine d'application . . . . . 139
  - 6.4.2. Analyse de corpus . . . . . 139
  - 6.4.3. Structure des notices . . . . . 140
  - 6.4.4. Dépendances sémantiques . . . . . 142
  - 6.4.5. Grammaticalité . . . . . 144
  - 6.4.6. Interface utilisateur . . . . . 146
  - 6.4.7. Multilinguisme. . . . . 148
- 6.5. Perspectives. . . . . 149
- 6.6. Bibliographie . . . . . 151

**Chapitre 7. Aide à la production de documentation technique multilingue . . . . . 153**

Sylvie REGNIER et Eva DAUPHIN

- 7.1. EADS : le multilinguisme en question . . . . . 153
  - 7.1.1. Typologie des documents . . . . . 154
  - 7.1.2. Contraintes et exigences contractuelles . . . . . 155
    - 7.1.2.1. Cadre de normalisation . . . . . 155
    - 7.1.2.2. Langue de rédaction et multilinguisme . . . . . 155
    - 7.1.2.3. L'anglais simplifié : *simplified english* (SE) . . . . . 155
    - 7.1.2.4. Le français rationalisé (FR) . . . . . 156
    - 7.1.2.5. Normes et autres contraintes . . . . . 157
  - 7.1.3. Processus de rédaction et traduction . . . . . 157
    - 7.1.3.1. Langue de production . . . . . 157
    - 7.1.3.2. Cycle de vie . . . . . 158
    - 7.1.3.3. Scenarii de traduction . . . . . 158
- 7.2. Méthodes et outils expérimentés – Retours d'expérience . . . . . 158
  - 7.2.1. Phases de spécifications . . . . . 159
    - 7.2.1.1. Etudes linguistiques et recommandations de rédaction. . . . . 159

7.2.1.2. Spécification de l'environnement et des contraintes de production documentaire . . . . .	160
7.2.1.3. Construction d'un référentiel de ressources linguistiques . . . . .	160
7.2.2. Outils de rédaction . . . . .	161
7.2.2.1. Base terminologique . . . . .	161
7.2.2.2. Contrôleur de conformité à l'anglais simplifié . . . . .	163
7.2.2.3. Génération . . . . .	165
7.2.3. Outils de traduction . . . . .	168
7.2.3.1. Aptitude des documents à un traitement par TAO . . . . .	168
7.2.3.2. Types d'outils et de traitements adéquats : TA ou TAO . . . . .	169
7.2.3.3. Stratégie de mise en place d'un outil de traduction . . . . .	171
7.2.4. Retour d'expérience . . . . .	175
7.2.4.1. Forte interaction homme-machine-processus . . . . .	175
7.2.4.2. Rentabilité des projets linguistiques sur le long terme . . . . .	175
7.2.4.3. Marche de niche et instable . . . . .	175
7.3. Vers des outils d'assistance à la production documentaire multilingue . . . . .	176
7.3.1. Changement d'optique . . . . .	176
7.3.1.1. Aide/facteur humain . . . . .	176
7.3.1.2. Indépendance vis-à-vis des formats/outils propriétaires . . . . .	177
7.3.1.3. Partage des ressources/travail collaboratif . . . . .	178
7.3.2. Outils d'assistance . . . . .	178
7.3.2.1. Mémoires de rédaction multilingues . . . . .	178
7.3.2.2. Aide à la saisie . . . . .	178
7.3.2.3. Personnalisation interactive (traduction à la demande, <i>profiling</i> , etc.) . . . . .	179
7.3.2.4. Supports sémiotiques (intégration de nouveaux supports de sens – vidéos, images, photos, sons, etc.) . . . . .	179
7.4. Conclusion . . . . .	179
7.5. Bibliographie . . . . .	180

## TROISIÈME PARTIE. APPLICATIONS MULTILINGUES . . . . . 181

### Chapitre 8. Compréhension multilingue et extraction de l'information . . . . . 183

Luca DINI

8.1. Introduction . . . . .	183
8.2. Le modèle d'extraction d'informations multilingues (EIM) . . . . .	185
8.3. Première étape : l'extraction des données à partir des textes . . . . .	188
8.3.1. L'extraction d'informations . . . . .	189
8.3.2. Buts et difficultés en extraction d'informations . . . . .	190
8.3.2.1. Les buts de l'extraction d'informations . . . . .	190
8.3.2.2. Points techniques . . . . .	191
8.3.3. L'outil choisi : Sophia 2.1 . . . . .	193
8.3.3.1. Une IDE pour l'extraction d'informations . . . . .	193
8.3.3.2. Stratégies d'extraction dépendant du domaine d'application . . . . .	194

8.3.3.3. Modalité template multiple . . . . .	195
8.3.4. Exemple d'une configuration pour des nouvelles financières . . . . .	196
8.4. Deuxième étape : la génération en langage naturel. . . . .	198
8.4.1. Génération dans une langue . . . . .	198
8.4.2. Génération en différentes langues . . . . .	199
8.5. Troisième étape : tout regrouper . . . . .	201
8.6. Conclusion . . . . .	203
8.7. Bibliographie . . . . .	203

**Chapitre 9. L'indexation conceptuelle de documents multilingues**

<b>et multimédias . . . . .</b>	<b>205</b>
---------------------------------	------------

Thierry DECLERCK et Elisabeth ANDRÉ

9.1. Introduction . . . . .	205
9.2. L'extraction d'information . . . . .	207
9.2.1. Recherche d'information et extraction d'information . . . . .	207
9.2.2. Les tâches spécifiques de l'extraction d'information . . . . .	209
9.2.3. L'extraction d'information et la génération d'annotations . . . . .	209
9.3. SMES – Un système d'extraction d'information pour l'allemand. . . . .	210
9.3.1. Les outils d'analyse linguistique . . . . .	211
9.3.2. La modélisation du domaine d'application . . . . .	211
9.4. Rôle du TAL dans les applications multimédias . . . . .	213
9.4.1. Systèmes multimédias et multimodaux . . . . .	213
9.4.2. Intégration de modalités . . . . .	213
9.4.3. Coordination de médias . . . . .	214
9.4.4. Accès en langage naturel aux archives digitales multimédias . . . . .	214
9.5. Challenges pour le TAL . . . . .	215
9.6. MUMIS – Un environnement pour l'indexation et la recherche de données multimédias . . . . .	216
9.6.1. Données multilingues, multisources et multimédias. . . . .	217
9.6.2. Extension des technologies d'extraction de l'information . . . . .	219
9.6.3. Synchronisation de la séquence vidéo et des annotations formelles. . . . .	220
9.7. Conclusion . . . . .	220
9.8. Bibliographie . . . . .	221

**Chapitre 10. Les outils de TAL au service de la e-formation en langues . . . . . 223**

Caroline BRUN, Thibault PARMENTIER, Ágnes SÁNDOR, Frédérique SEGOND

10.1. Introduction . . . . .	223
10.2. Outils de TAL. . . . .	225
10.2.1. Outils de base. . . . .	225
10.2.1.1. Automates et transducteurs à états finis . . . . .	225
10.2.1.2. Segmentation. . . . .	226
10.2.1.3. Normalisation . . . . .	226

10.2.1.4. Analyse morphologique . . . . .	226
10.2.1.5. Désambiguïsation des parties du discours. . . . .	227
10.2.1.6. Analyseurs syntaxiques robustes . . . . .	228
10.2.2. Désambiguïsation lexicale sémantique. . . . .	229
10.2.2.1. Extraction des règles . . . . .	229
10.2.2.2. Application des règles pour la désambiguïsation lexicale . . . . .	231
10.3. Intégration des composants. . . . .	232
10.3.1. Extraction de terminologie : monolingue et bilingue . . . . .	232
10.3.2. Outil de recherche d'information multilingue : LIRIX . . . . .	232
10.3.3. Locolex . . . . .	233
10.3.4. Dictionnaire sémantique . . . . .	234
10.4. Intégration à la e-formation en langues . . . . .	236
10.5. Edutainment for Internet Language Learning Solution (Exills) : un logiciel d'apprentissage des langues utilisant le TALN. . . . .	240
10.5.1. Un identificateur de langue . . . . .	243
10.5.2. Un analyseur morphologique . . . . .	243
10.5.3. Un outil de consultation de dictionnaire intelligent . . . . .	244
10.5.4. Un correcteur orthographique . . . . .	245
10.5.5. Accès libre aux traitements linguistiques . . . . .	246
10.5.6. askOnce . . . . .	247
10.6. Conclusion . . . . .	248
10.7. Bibliographie . . . . .	249
<b>Index . . . . .</b>	<b>251</b>