

Université Paris 7
Centre d'Etude et de Recherche en Informatique Linguistique
Laboratoire d'Automatique Documentaire et Linguistique

Thèse de doctorat en Informatique Fondamentale

ALARIC
Un système d'Analyse Linguistique Automatique des
Résumés d'Incidents en Centrale

par Tarek Chellali

Jury :

Jean Pierre Desclés (Rapporteur)
Michel Gondran (Rapporteur)
Maurice Gross (Président)
Philippe Lefèvre
Dominique Perrin

23 Juin 1993

SOMMAIRE

	SOMMAIRE.....	3
		3
Chapitre 1	Introduction	7
1.1	Le problème	7
1.2	Description des fiches.....	7
1.3	Nature des expressions linguistiques utilisées dans les résumés	11
1.4	Modélisation simplifiée des résumés.....	11
1.5	Contenu de l'étude.....	12
1.5.1	L'étude du lexique	13
1.5.2	L'étude syntaxique	13
1.5.3	L'étude sémantique	13
1.5.4	Le formalisme	14
1.5.5	Les outils.....	14
Chapitre 2	Etat de l'art	15
2.1	Généralités sur le traitement du langage naturel	15
2.2	Rappel de quelques supports de connaissances lexicales.....	16
2.2.1	Dictionnaires.....	16
2.2.2	Les outils d'extraction d'unités terminologiques complexes	17
2.2.3	Les outils d'analyse lexicale.....	18
2.3	Les formalismes grammaticaux.....	19
2.3.1	La grammaire en chaîne	21
2.3.2	Grammaires syntagmatiques généralisées	22
2.3.3	Grammaires logiques.....	25
2.3.4	Grammaires de cas	26
2.3.5	Grammaires lexicales fonctionnelles	28
2.3.6	Conclusion	31
2.4	La représentation du sens	32
2.4.1	La représentation du sens des mots	32
2.4.2	Le modèle conceptuel de Schank	33
2.4.3	Les graphes conceptuels	35
2.4.4	Conclusion	38

Chapitre 3	Étude linguistique du corpus	39
3.1	Particularités du lexique	39
3.1.1	Style télégraphique	39
3.1.2	Néologismes	40
3.1.3	Nominalisation des verbes.....	40
3.1.4	Quelques aspects morphologiques de la nominalisation	41
3.1.5	Formes nominalisées avec négation	42
3.1.6	Formes impersonnelles	42
3.1.7	Noms	43
3.1.8	Adjectifs.....	44
3.2	Mots composés	45
3.2.1	Nécessité de prendre en compte les mots composés.....	45
3.2.2	Classes des mots composés.....	46
3.2.3	Les classes des noms composés	47
3.2.4	Classes des adverbes composés	48
3.2.6	Conjonctions composées.....	49
3.2.7	Extraction des mots composés	49
3.3	Particularités syntaxiques	53
3.3.1	Introduction	53
3.3.2	Les verbes.....	53
3.3.3	Transformations sur les verbes	55
3.3.4	Les verbes opérateurs	55
3.3.5	Normalisation des constructions impersonnelles des verbes opérateurs	56
3.3.6	Nominalisation des verbes opérateurs	58
3.3.7	Les verbes de mouvement	58
3.3.8	Les verbes à trait attributif.....	59
3.3.9	Réductions de déterminants et de prépositions	60
3.3.10	Réduction de la copule <i>être</i>	60
3.3.11	Réduction par l'adjectif ou par abréviation	60
3.3.12	Une syntaxe irrégulière	61
3.4	Particularités sémantiques	62
3.4.1	Ambiguïtés sémantiques de ce corpus	62
3.5	Systématisation du lexique	65
3.5.1	Les éléments porteurs de sens	65
3.5.2	Méthode de classement	66
3.5.3	Les principales classes des noms évoquant des notions.....	68
3.5.4	Les classes de noms de notion complémentaires	72
3.5.5	Classes d'adjectifs.....	74
3.5.6	Classes d'adverbes	75
3.5.7	Classes des mots outils.....	75

3.5.8	Prépositions et conjonctions.....	76
3.6	La théorie de Fillmore et la classification des verbes.....	77
3.6.1	Verbes	77
3.6.2	Verbes supports.....	79
3.6.3	Verbes de présentation	79
3.6.4	Verbes de changement	80
3.6.5	Verbes causatifs d'action	81
3.6.6	Verbes d'état.....	82
3.6.7	Verbes d'action	82
Chapitre 4	Grammaires sémantiques	84
4.1	Grammaires sémantiques.....	84
4.1.1	Notion de cas sémantique.....	84
4.1.2	Définition	85
4.1.3	Intérêt pour l'analyse des résumés d'incidents.....	87
4.2	Mise en oeuvre de la grammaire sémantique	88
4.2.1	Options choisies	88
4.2.2	Éléments porteurs de sens : la notion d'îlot.....	89
4.2.3	Les éléments perturbateurs : traitement du bruit.....	93
4.2.4	Les grammaires d'extrapositions et leurs variantes : formalisme utilisé pour traiter le bruit	93
4.3	L'analyse ciblée	94
4.3.1	Analyse par niveaux.....	94
4.3.2	Arbres de dérivation.....	95
4.4	Avantages et inconvénients de la méthode choisie	96
4.4.1	Avantages de la méthode	96
4.4.2	Inconvénients	97
Chapitre 5	Grammaires attribuées	100
5.1	Grammaires attribuées.....	100
5.1.1	Nécessité d'étendre le formalisme des grammaires sémantiques.....	100
5.1.2	Définition d'une grammaire attribuée	100
5.1.3	Un formalisme adapté à l'utilisation des grammaires attribuées.....	101
5.1.4	Hypothèses syntaxiques et sémantiques	102
5.2	Les caractéristiques de la grammaire attribuée.....	102
5.2.1	Le mécanisme d'abstraction (instanciation)	102
5.2.2	Calcul d'attributs (Eq-lex)	103
5.2.3	Contextes syntaxiques et sémantiques intégrés dans le calcul d'attributs.....	105
5.2.4	Traitement des chaînes spéciales (Eq-syn)	106
5.2.5	Des exceptions pour éviter des règles ambiguës (Eqse).....	108

5.3	Avantages des grammaires attribuées comme modèle de représentation.....	114
5.4	Comment utiliser ce formalisme.....	115
5.4.1	Avantages d'un tel formalisme	115
5.4.2	Une bonne construction du lexique	117
5.4.3	Les règles de grammaire	117
5.5	Conclusion.....	119
Chapitre 6	Analyseur lexical	120
6.1	Reconnaissance lexicale des mots composés par l'application de la stratégie du regard en avant.....	120
6.1.1	L'algorithme du regard en avant appliqué à la reconnaissance des mots composés.....	122
6.1.2	Discussion sur la stratégie utilisée	124
6.2	Implémentation en prolog.....	124
Chapitre 7	Implémentation du prototype	128
7.1	L'architecture générale du prototype.....	128
7.1.1	Description de l'architecture du modèle.....	128
7.1.2	Le gestionnaire de programme	130
7.1.3	Le découpeur	130
7.1.4	Dictionnaire.....	131
7.1.5	La grammaire	132
7.1.6	L'analyseur syntaxico-sémantique.....	132
7.1.7	Les automates.....	132
7.2	Rappel de quelques notions de Prolog.....	134
7.2.1	Le langage	134
7.2.2	Le mécanisme d'unification	134
7.2.3	Grammaire à clause définie (DCG)	135
7.3	Description de l'analyseur syntaxico-sémantique en Prolog.....	138
7.3.1	Implémentation des règles de la grammaire attribuée.....	138
7.3.2	Exceptions	141
7.4	Résultats et commentaires	141
7.4.1	Le comportement de l'analyseur lexical.....	141
7.4.2	Comportement de l'analyseur syntaxico-sémantique	143
7.4.3	Saut des expressions non analysables.....	145
	Conclusion	147
	Bibliographie	149