

Génération automatique de textes en langues naturelles

Laurence Danlos

Préface du Pr. M. GROSS

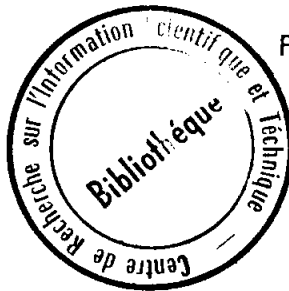
CT135 CQ34 PG845.180/D/4. ((<...La marquise sortit à 5 heures.
Non. Erreur : à 16 h 59 mn 43/100. Il faisait beau ; 23 C° à l'ombre.
Une brise de force 2 soulevait sa jupe à 47,29 cm du sol. Une mèche
de 623 cheveux flottait, légère, sur son front pâle. Soudain, son
sans (5 l 135 cl) ne fit qu'un tour 3/4 ; devant elle, à 7,28 m se
tenait Xavier, 1,773 m, 71,257 kg, 38 ans et 17 jours, 18 615 Francs
par mois, brut. Cela faisait 568 jours qu'elle ne l'avait revu, mais
elle se souvint avec émotion de cette brève rencontre qui lui avait
valu, à 23 h 42, un orgasme d'amplitude S.M. 349.>)) 87CT145 CQ30.

BIBLIOTHEQUE DU CERIST

Génération automatique de textes en langues naturelles

Laurence Danlos

*Chargé de recherche au CNRS
Laboratoire d'Automatique Documentaire et Linguistique (LADL)*

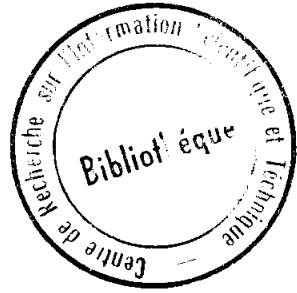


Préface du Professeur M. Gross

MASSON

Paris New York
Barcelone Milan Mexico
São Paulo

1985



Préface

L'Intelligence Artificielle est souvent plus un art qu'une méthode de simulation informatique de comportements "intellectuels". Mais dans le domaine de la communication intelligente homme-machine, certaines questions sont susceptibles d'une approche rigoureuse. En effet, un ensemble d'études linguistiques a permis de formaliser des grammaires de langues naturelles avec la cohérence et la rigueur de l'approche scientifique. Le modèle développé par Laurence Danlos et présenté dans ce volume est l'un des premiers construits sur des méthodes nouvelles qui exploitent les ressources de la linguistique formalisée. Il constitue donc une contribution importante à l'explicitation de mécanismes linguistiques dont l'étude est souvent négligée.

Le système de génération automatique proposé par Laurence Danlos part d'une représentation sémantique des informations qui doivent être véhiculées dans le texte à engendrer. Le problème consiste donc à traduire une représentation sémantique en un texte syntaxiquement cohérent et de bonne lisibilité tant du point de vue de la clarté que du style. La cohérence syntaxique des textes s'obtient sans trop de difficulté et de manière systématique grâce au lexique-grammaire général du français. C'est surtout dans ce domaine que les progrès en linguistique des vingt dernières années ont été les plus solides, rendant les applications correspondantes possibles. La qualité stylistique a toujours été l'objet de remarques scolaires ou de conseils formulés dans des manuels de rédaction. Laurence Danlos a systématisé un niveau de représentation stylistique. Elle a montré que le passage de représentations sémantiques à des textes bien formés requérait l'utilisation d'une "grammaire de discours", notion originale qui permet d'intégrer décisions conceptuelles, syntaxiques et stylistiques.

Le système présenté ici n'utilise donc que des connaissances linguistiques formelles. Il est illustré par un programme qui construit des textes de qualité dans un domaine restreint. Le développement des connaissances linguistiques formelles devrait permettre la réalisation de programmes engendrant des textes dans différents domaines par application des méthodes développées dans cet ouvrage. Les exemples abondent, les références sont complètes, le livre de Laurence Danlos est donc un véritable manuel de construction de systèmes de génération.

Maurice Gross

Professeur à l'Université de Paris 7

Table des matières

Préface du Pr. Gross	5
Introduction	11
1. Exemple introductif : production de messages d'erreur	21
1.1 Présentation du problème	21
1.2 Données de base du générateur	24
1.3 Algorithme de génération	31
1.4 Connaissances linguistiques/Notions abstraites	43
2. Survol des recherches en génération automatique .	47
2.1 Quoi dire ? / Comment le dire ?	47
2.2 Premiers travaux	52
2.3 Représentations sémantiques	54
2.3.1 Difficultés avec les représentations calquées sur une langue	54
2.3.1.1 Paraphrases et inférences	54
2.3.1.2 Passage à une autre langue	54
2.3.1.3 Variations du contenu informatif des textes produits	55
2.3.2 Degré d'imprécision dans les représentations abstraites	56
2.3.3 Choix d'une représentation en fonction des objectifs visés	58
2.3.4 Entrées des générateurs	62
2.4 Travaux récents	63
2.4.1 Générateur de McKeown	63
2.4.1.1 Présentation	63
2.4.1.2 Commentaires	69
2.4.2 Modèles "psychologiquement plausibles"	71

3. Les bases linguistiques d'un système de génération 75

3.1	Sémantique et formulations des relations causales	75
3.1.1	Subordonnées en <i>parce que</i> / formes verbales à sujet causatif	75
3.1.2	Nature de la cause pour un résultat donné	78
3.1.3	Relations causales et relations temporelles	81
3.1.4	Relations causales et relations locatives	85
3.1.5	Relations causales et lexique	86
3.1.6	Ebauche des relations sens/forme pour les relations causales	86
3.2	Choix lexicaux	87
3.2.1	Position du problème	87
3.2.2	Recherche d'un élément dans un lexique propre au domaine	90
3.2.2.1	Premiers facteurs pertinents	90
3.2.2.2	Informations sur l'auteur de l'acte	92
3.2.2.3	Phénomènes de coordination	94
3.2.3	Décisions locales/non locales	96
3.3	Linéarisation en phrases	97
3.3.1	Combinaisons où l'acte et le résultat sont à l'actif	98
3.3.1.1	Présentation	98
3.3.1.2	Structures avec un gérondif/participe présent	102
3.3.2	Combinaisons où l'acte est à l'actif, le résultat au passif	109
3.3.3	Combinaisons où l'acte et le résultat sont au passif	111
3.3.4	Combinaisons où l'acte est au passif et le résultat à l'actif	112
3.3.5	Grammaire de discours	112
3.4	Système de génération	118
3.4.1	Données linguistiques du générateur	118
3.4.2	Sélection d'une structure de discours	118
3.4.3	Sélection d'une structure de discours et choix lexicaux	119
3.4.4	Algorithme de génération	119

4. Grammaire de discours 125

4.1	Contraintes formelles	125
4.2	Contraintes non formelles	131
4.3	Combinaisons à une seule phrase simple	134
4.3.1	Formulations exprimant simultanément la cause et le résultat	134
4.3.2	Les nominalisations et les compléments instrumentaux	137
4.4	Questions stylistiques et sélection d'une structure de discours	141

5. Le programme	147
5.1 Entrées du générateur	149
5.1.1 Représentation des connaissances par formulaire	149
5.1.2 Implémentation de la représentation par formulaire	152
5.1.3 Domaine des attentats	157
5.1.4 Menu	160
5.2 Composant stratégique	163
5.2.1 Sélection d'une structure de discours et choix lexicaux pour les attentats	165
5.2.2 Choix lexicaux	166
5.2.2.1 Représentation des informations syntaxiques	167
5.2.2.2 Représentation des propriétés syntaxiques	169
5.2.2.3 Sélection des phrases simples	170
5.2.2.4 Instanciation des phrases du lexique	174
5.2.3 Sélection d'une structure de discours	176
5.2.4 Construction des schémas de phrases	177
5.2.5 Propositions relatives	184
5.3 Composant syntaxique	188
5.3.1 Traitement de la phrase syntaxique	189
5.3.2 Synthèse des compléments. Pronominalisation	191
5.3.3 Synthèse des étiquettes	208
5.3.3.1 Synthèse de la première occurrence d'une étiquette	209
5.3.3.2 Synthèse de la n ^{ième} occurrence d'une entité	211
5.3.3.3 Synthèse des modifieurs	212
5.3.3.4 Synthèse des déterminants	214
5.3.4 Coordination	215
5.3.5 Dernières opérations	217
Conclusion	219
Appendice	221
Bibliographie	227
Index	237