Revue québécoise de linguistique

Génération automatique de rapports boursiers français et anglais

Chantal Contant

Volume 17, numéro 1, 1988

Psychomécanique du langage

URI : https://id.erudit.org/iderudit/602620ar
DOI : https://doi.org/10.7202/602620ar

Aller au sommaire du numéro

Éditeur(s)

Université du Québec à Montréal

ISSN

0710-0167 (imprimé)
1705-4591 (numérique)

Découvrir la revue

Citer cet article

Contant, C. (1988). Génération automatique de rapports boursiers français et anglais. *Revue québécoise de linguistique*, *17*(1), 197–221. https://doi.org/10.7202/602620ar

Résumé de l'article

Depuis peu de temps, il est possible, dans un sous-langage technique bien délimité, de créer des systèmes automatiques capables de générer, à partir d'une représentation sémantique, des textes linguistiquement bien formés. Un tel système existe pour le sous-langage boursier. En effet, à partir des données de la Bourse de New York, ce logiciel produit de façon automatique des résumés boursiers en anglais et en français. Le présent article présente le système anglais et français de génération automatique de texte et décrit brièvement les particularités du sous-langage boursier.

GÉNÉRATION AUTOMATIQUE DE RAPPORTS BOURSIERS FRANÇAIS ET ANGLAIS

Chantal Contant

1. Génération de texte

De nos jours, plusieurs domaines de l'intelligence artificielle nécessitent la collaboration des linguistes. Le traitement des langues naturelles par ordinateur, qui semblait simple à première vue, fascine de plus en plus les linguistes et les informaticiens par sa complexité. Tel est le cas, entre autres, de la génération automatique de texte (appelée aussi synthèse de texte) faite à partir de représentations sémantiques. En effet, depuis peu de temps il est possible, dans un sous-langage technique bien délimité, de créer des systèmes automatiques produisant des textes linguistiquement bien formés.

1.1 Représentation des diverses connaissances impliquées

La génération automatique nécessite diverses connaissances, principalement des connaissances sémantiques et des connaissances linguistiques (Danlos 1985a et 1985b). Les connaissances sémantiques ne peuvent actuellement correspondre aux connaissances du monde en général car nous ne savons pas encore exactement comment les représenter. C'est pourquoi les connaissances des systèmes de génération automatique se limitent à un sous-domaine bien délimité. Ce qui nous amène à générer du texte aux caractéristiques d'un sous-langage (sémantique restreinte).

La représentation de ces connaissances peut être indépendante de la langue ou non, selon les théories et les buts visés. Chose certaine, les représentations conceptuelles des logiciels actuels sont insuffisantes et imprécises en comparaison avec la complexité et la richesse de la langue. Une simple phrase signifie tant de choses et évoque de multiples détails implicites sur la connaissance du monde.

Les connaissances proprement linguistiques employées dans la génération de texte concernent la syntaxe, la grammaire: comment combiner les mots et les propositions. Elles concernent aussi la morphologie, le lexique, la stylistique et la rhétorique. Ces deux dernières sont parfois traitées au niveau de la sémantique (et parfois même pas du tout). Malgré tout, il est évident que la structure de texte, le bon ordonnancement des idées doivent faire partie des connaissances d'un système de génération. Mais comment intégrer toutes ces connaissances? Comment programmer l'ordinateur pour qu'il sache produire un texte élégant? Nous avons besoin de théories linguistiques formelles et de techniques de représentation des connaissances pour mener à bien cette tâche.

1.2 Quoi dire et comment le dire

Différents travaux ont été réalisés sous l'étiquette de génération automatique et il en ressort deux principales préoccupations: *QUOI dire* et *COMMENT le dire*.

> «Computer generation of natural language requires the ability to make reasoned choices from a large number of possible things to say as well as from a large number of expressive possibilities.» (McKeown 1983, p. 582)

On remarque dans la littérature de linguistique computationnelle que l'aspect QUOI dire semble privilégié (Mann 1981, 1982; Winograd 1983; McDonald et al. 1984). Les chercheurs et chercheuses construisent souvent des systèmes dont la composante sémantique qui sert à retirer l'information pertinente et à l'organiser en une structure de texte cohérente est bien développée mais dont la composante linguistique est plus rapidement esquissée. La génération est plus souvent basée sur le contenu que sur la forme.

> «In the past, research systems [...] have been limited by the weakness of their linguistic bases, especially their grammars...» (Mann 1983, p. 261)

Il reste tout de même que les deux questions sont présentes. Et parmi les systèmes qui ont déjà été conçus, il existe deux tendances face à ces deux grandes questions:

1) Aborder les deux questions de façon séquentielle: d'abord décider QUOI dire puis dans un deuxième temps décider COMMENT le dire.

2) Aborder les deux questions de façon parallèle: QUOI dire et COMMENT le dire sont étroitement reliés et interdépendants (Danlos 1983 et 1984).

Il y a plus d'adeptes pour la tendance à séparer les deux questions en des modules distincts. On détermine dans un premier temps la représentation de sens et ensuite on la traduit en mots. C'est l'approche qui a été adoptée dans le système de génération de rapports boursiers anglais et français que nous décrivons plus loin.

1.3 Sous-langage

1.3.1 Pourquoi un sous-langage

Les systèmes de génération automatique de textes linguistiquement bien formés ne sont possibles que s'ils se situent dans un sous-langage technique bien délimité. Il a été démontré que les langues en général sont trop vastes pour être traitées correctement (au niveau syntaxique et sémantique) par les systèmes actuels. Il existe dans le monde un seul système compétent et performant: le cerveau humain. Puisque tout être humain en bonne santé est capable de parler, cet acte semble trivial et simple à reproduire. Et c'est ce qu'on a cru dans les années 50, au début des premiers travaux en traduction automatique. Or, on s'est vite aperçu que tel n'est pas le cas.

Aujourd'hui, les systèmes fonctionnels d'interrogation de bases de données, de génération, d'analyse ou de traduction automatique sont limités à un sous-domaine particulier. Citons en exemple l'expérience de l'équipe TAUM (Traduction Automatique Université de Montréal) dans les années 70. Essayant d'abord de créer un système de traduction automatique pour l'anglais en général vers le français (TAUM-71 - Colmerauer et al. 1971; TAUM-73 - Kittredge et al. 1973), le groupe a ensuite créé le système TAUM-MÉTÉO, où les phrases traitées se limitaient à celles que l'on retrouvait dans les bulletins météorologiques d'Environnement Canada. On s'est donc restreint à un sous-ensemble de l'anglais, à un **sous-langage**. Dès lors, un système satisfaisant de traduction automatique a été possible puisque le vocabulaire, la syntaxe et la sémantique étaient bien délimités et restreints à un domaine particulier.

En ne retenant qu'un sous-ensemble d'une langue, en se limitant à un sous-langage, on restreint le nombre de mots et le nombre de sens pour chaque mot, facilitant ainsi le traitement automatique.

1.3.2 Le domaine boursier

Le monde de la Bourse est associé à un domaine de références relativement restreint et concerne une communauté d'individus particuliers partageant des connaissances communes et spécialisées. Pourquoi d'ailleurs utilisons-nous l'expression *le «monde» de la Bourse* si ce n'est parce qu'il s'agit d'un univers particulier, un sous-monde de notre monde.

Les résumés du marché boursier que l'on retrouve dans les grands journaux quotidiens et qui décrivent l'activité de la Bourse d'une ville en particulier, sont très structurés et bien délimités quant à leur contenu. Ils constituent un mode de communication restreint au niveau de la forme et ont un but particulier: informer une population cible (les investisseurs) au sujet d'un domaine précis (l'activité boursière). Leur structure de texte est également régulière d'une journée à l'autre.

Il semble donc que nous sommes en présence ici d'un sous-langage: le sous-langage boursier. Nous avons étudié ces rapports quotidiens (Contant et Gauthier 1983; Contant 1985) qui font état de l'activité du marché boursier et que l'on retrouve dans les grands journaux. Par ces études, nous avons constaté qu'effectivement, ce sous-langage a un vocabulaire particulier et restreint: des verbes de mouvement vers le haut ou vers le bas, des verbes désignant la clôture, etc. Il a également une syntaxe assez stéréotypée et une sémantique bien délimitée. Nous verrons plus loin les caractéristiques de ce sous-langage. Ces caractéristiques sont avantageuses pour réaliser un traitement automatique.

2. Application

2.1 Le système Ana

Le système de génération automatique de rapports boursiers dont nous parlerons ici a d'abord été conçu en anglais par Karen Kukich (1983a) à l'Université de Pittsburgh et il se nomme ANA. La fonction de ce système est de convertir les données traitables par la machine (tableaux de chiffres et d'abréviations) en information traitable de façon naturelle par l'humain, c'est-à-dire en un texte linguistiquement adéquat.

Dans un tel système, deux principaux domaines de connaissances entrent en jeu. La connaissance sémantique reliée au domaine dont on parle (ici le domaine boursier) et la connaissance linguistique, c'est-à-dire la compétence à rédiger un texte avec l'organisation structurelle (rhétorique), la syntaxe, la morphologie et les choix lexicaux appropriés. Il y a donc les deux aspects dont nous avons parlé: QUOI dire et COMMENT le dire. C'est ce deuxième aspect que nous avons développé (Contant 1985) afin d'obtenir également la rédaction de rapports boursiers en français, le premier étant commun aux deux langues.

QUOI dire se réalise par la formulation de messages sémantiques pertinents concernant une base de données particulière. Dans le cas du système Ana, la base de données est constituée de colonnes de chiffres et d'abréviations provenant de la Bourse de New York, illustrant l'activité boursière de la journée. La courbe de l'indice Dow Jones sous forme de tableau numérique est interprétée par Ana qui décide quels sont les points importants à souligner à propos de l'activité du marché, des indices boursiers, du volume de transactions, etc.

COMMENT le dire se réalise par un rapport boursier en langue naturelle. C'est le passage de la représentation sémantique au texte final. Ana regroupe d'abord ses messages en une suite d'idées qui s'enchaîneront de façon cohérente et logique. C'est l'organisation du discours en paragraphes. Puis, pour chaque message sémantique (chaque sens ou idée à véhiculer), elle choisit une entrée lexicale sous forme de syntagme. Elle choisit ensuite la forme syntaxique adéquate pour permettre un texte naturel, aux structures variées, non monotone ni télégraphique. Après le choix du sujet, une étape de morphologie est nécessaire. Et sur le tout se greffent de multiples

contraintes de style et de rhétorique: longueur maximale de syllabes par phrase, alternance de phrases longues et courtes, nombre maximal de messages par phrase, ponctuation, préférence pour certaines structures et certaines entrées syntagmatiques (fréquence d'apparition), interrogation sur le contenu sémantique à venir ou sur le contenu linguistique et sémantique de ce qui a déjà été rédigé précédemment, etc.

La structure de Ana est constituée de quatre modules séquentiels. Dans chaque cas, la sortie (output) de l'un sert d'entrée (input) au module suivant. Le premier module prend pour entrée les données provenant de la Bourse de New York et le dernier module produit comme sortie le rapport boursier anglais correspondant aux données reçues (voir fig. 1).

Le premier module exécute certains calculs pour transformer en «faits» les données qui proviennent de New York et qui seront nécessaires au générateur de messages.

Le deuxième module constitue le module sémantique du système. Il prend pour entrée la série de faits générés par le premier module. Ce module de génération de messages est le deuxième en importance au niveau du volume et de la complexité. Le but de ce module est de générer des *messages* pertinents à partir des *faits* qu'il reçoit. C'est lui qui décide QUOI dire.

Dix classes de messages sémantiques sont envisagées tour à tour afin de créer zéro, un ou plusieurs messages de chaque classe, selon le cas. Il s'agit de: fluctuations intéressantes du marché; situation du marché à la fermeture; marché mitigé ou non; qualification de l'activité boursière; fluctuations intéressantes de l'indice Dow Jones des industrielles; situation de cet indice à la fermeture; situation de l'indice des transports à la fermeture; situation de l'indice des services publics à la fermeture; volume des transactions à la Bourse de New York; et nombre des titres à la hausse et à la baisse.

Du point de vue informatique, ce module sémantique, ainsi que les deux modules suivants, est programmé en langage de productions OPS-5 (Forgy 1981). Une règle de productions est de la forme SI *condition(s)* ALORS *action(s)* et les éléments manipulés sont composés d'une liste ↑ *attribut valeur*.

Voici un exemple de message sémantique:

(make message ↑repdate 01/10 ↑top DOW ↑subtop DOWPT
↑subsubtop HILO ↑subjclass DOW ↑dir up ↑deg hi
↑vardeg 19.13 ↑varlev 1095.2 ↑tim late ↑vartim 3:30pm)

qui se traduira grâce aux règles de la grammaire du module 4 français par: *L'indice Dow Jones a enregistré son plus gros gain à 3h30, soit 19.13 points, à 1095.2.*

Le troisième module de Ana prend les messages non ordonnés du module 2 et les regroupe en paragraphes selon leur *topique* (thème) et ordonne chacune des idées à l'intérieur des paragraphes. La sortie est composée des mêmes messages sémantiques reçus en entrée mais avec un attribut additionnel pour chacun: un numéro de priorité illustrant la structure du texte, c'est-à-dire l'ordre d'apparition des messages dans le rapport boursier final.

Le quatrième module constitue le module linguistique anglais du système Ana. C'est le plus volumineux et le plus complexe des quatre modules. Il contient 110 règles de productions et près de 450 entrées syntagmatiques.

Les 110 règles forment la grammaire du module linguistique anglais. Une grammaire, nous le verrons, qui est dépendante du contexte relatif à l'état du système, à connaissances multiples intégrées et dont les règles combinent des syntagmes et des propositions (macro-niveau).

Ce module linguistique reçoit les messages sémantiques sous forme organisée. En fouillant le dictionnaire syntagmatique, il associe les entrées (COMMENT le dire) avec les messages et combine le tout à l'aide de sa grammaire pour produire un texte élégant grâce à diverses contraintes.

2.2 Ajout d'un module français: Frana

Le système Ana de Karen Kukich (1983a) dont nous venons d'esquisser la structure sépare donc QUOI dire de COMMENT le dire. Nous avons vu qu'il est constitué entre autres d'un module sémantique (module 2), d'un module d'organisation du discours (module 3) et d'un module linguistique anglais (module 4). Comme le montre l'exemple plus haut, la représentation sémantique de Ana est sous forme de

messages avec chacun une liste *attribut-valeur* qui se concrétisera en une proposition dans le module linguistique.

La sémantique étant indépendante de la langue dans Ana, un nouveau module linguistique compatible peut prendre comme entrée les messages sémantiques sortant du module 2 ou 3 et générer le texte dans sa langue de choix. C'est ce que nous avons fait en français. En effet, la modularité de Ana nous a permis de concevoir un module linguistique français, que nous nommons FRANA, et qui peut se substituer au module 4 anglais de Ana.

Résumons-nous:

Ana est un système qui, partant de données numériques de la Bourse de New York, produit des messages sémantiques, les ordonne, et de là génère un texte en langue anglaise. Pour traduire ce texte dans une autre langue, disons le français, nous aurions pu bien sûr en analyser le contenu, faire les transferts d'une langue à l'autre puis générer un deuxième texte, cette fois-ci en français. Mais Ana possède déjà une représentation sémantique du texte (module 2) disposée de façon organisée pour l'élaboration du discours (module 3). Notre idée a été de profiter de la présence de cette représentation sémantique pour épargner temps et travail à analyser le texte source et à faire les transferts.

Frana a donc été élaboré pour se substituer au quatrième module de Ana. Plutôt que d'être soumis au bloc linguistique anglais, les messages ordonnés sortant du module 3 sont soumis à un bloc linguistique français nommé Frana. Le texte français sera indépendant du genre de choix syntaxiques et lexicaux faits lors de la génération du texte anglais car la représentation sémantique est indépendante de la langue. Mais le contenu sémantique sera évidemment le même puisque la représentation du sens est la base commune aux deux textes.

Le système n'a donc «réfléchi» qu'une fois à ce qu'il veut dire (QUOI dire) et, selon qu'il soumet ses messages au module linguistique anglais ou français ou les deux, il obtient un texte dans la ou les langues désirées. La modularité de Ana rend possible la substitution de modules linguistiques différents, permettant économie de travail et production accrue. On peut imaginer de tels modules pour n'importe quelle langue.

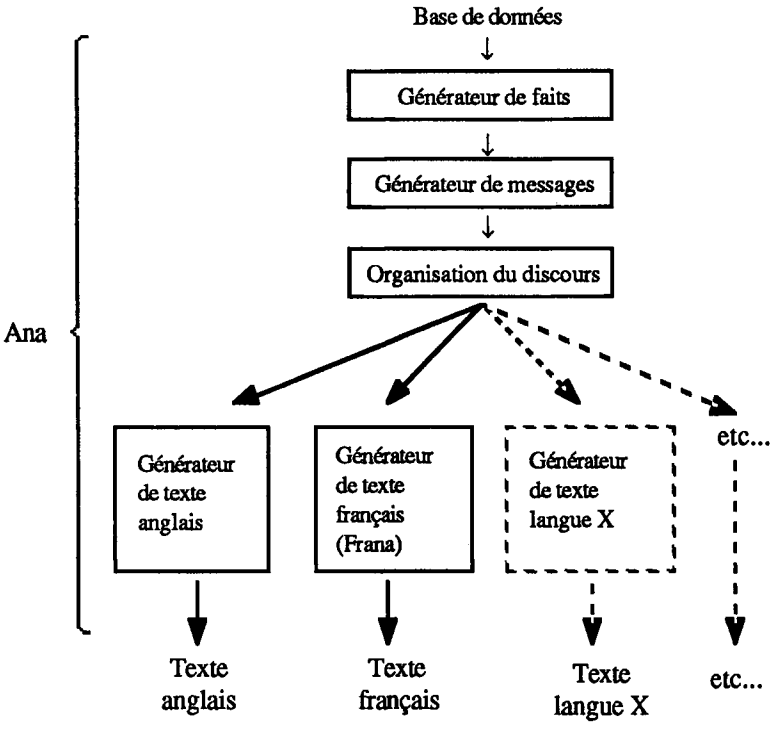

Figure 1 Structure du système bilingue de génération automatique de rapports boursiers

Il ne s'agit pas d'une traduction mais d'une francisation du système. Bien sûr, le module linguistique anglais a été une ressource inestimable, un point de départ précieux à l'élaboration de Frana. C'est ce module que nous avons modifié et transformé pour répondre aux besoins du français. D'abord au niveau des entrées syntagmatiques où les attributs sémantiques se devaient de rester les mêmes mais où les attributs syntaxiques, lexicaux et rhétoriques étaient fonction de la langue française. Ensuite au niveau des règles, où contraintes, paramètres et règles ont été enlevés, remplacés, ajoutés, modifiés. Les plus gros changements, mis à part le lexique, sont sans doute au niveau des formes syntaxiques et au niveau de la morphologie, le français étant plus complexe que l'anglais avec ses nombreux accords.

2.3 Grammaire dépendante du contexte et macro-niveau

Frana est écrit en langage de productions OPS-5 (Forgy 1981) pour respecter le formalisme déjà établi dans Ana et contient 143 règles. Les règles de productions étant de type SI → ALORS, l'application d'une règle est fonction de la valeur de vérité de la conjonction de ses conditions. Les conditions dans les règles de Ana et Frana peuvent porter sur des connaissances sémantiques, lexicales, syntaxiques, rhétoriques ou sur des éléments de contrôle. Elles tiennent compte de l'état du système et de ses connaissances qui varient au cours du traitement. À cause de ces conditions, l'application d'une règle dépend donc fortement du contexte. Précisons qu'il ne s'agit pas ici du contexte linéaire (précédent, suivant) des règles de réécriture des langages de type 1 de la hiérarchie de Chomsky (Baudot 1981) mais du contexte relatif à l'état du système durant l'exécution.

Ana et Frana utilisent des structures de connaissances à un macro-niveau, ce qui permet une représentation et un traitement efficaces pour la génération automatique de rapports en langue naturelle.

Plutôt que de manipuler des primitifs sémantiques, Ana traite des messages sémantiques entiers, c'est-à-dire des unités complexes. Et plutôt que d'être des unités linguistiques minimales, les catégories lexicales se retrouvent au niveau des syntagmes, plus ou moins complexes. Alors, contrairement aux grammaires formelles de génération de phrases traditionnelles de la théorie chomskienne (Baudot 1981), Ana et Frana combinent des syntagmes entiers (Becker 1975) et des propositions.

Bien sûr en opérant à un macro-niveau nous perdons de la flexibilité linguistique et sémantique. Mais ceci n'est pas un réel handicap puisque nous travaillons à l'intérieur d'un sous-langage. Le champ sémantique est limité car les événements à exprimer dans un rapport sont sémantiquement restreints puisqu'ils dépendent du domaine. Par le fait même, le contenu lexical est également limité. Les expressions figées ou semi-figées, les collocations, les tournures de phrases particulières sont chose courante pour le rédacteur ou la rédactrice de rapports boursiers (Kittredge 1982a et 1982b; Cohen 1980 et 1981). La validité psychologique de ce principe n'est donc

pas à négliger. De plus, la manipulation à un macro-niveau permet un travail informatique plus rapide, efficace et moins coûteux.

2.4 Fonctionnement général et choix syntaxiques

Pour qu'une entrée syntagmatique soit choisie pour exprimer un message dans le texte final, il faut que les attributs sémantiques de cette entrée correspondent à ceux du message à exprimer et que différentes contraintes de rhétorique soient respectées: longueur des syllabes (len), fréquence d'apparition (rand), niveau d'hyponymie pour le sujet. Voici un exemple d'entrée syntagmatique prédicative de Frana:

(make phraselex ↑ptype pred ↑top DOW ↑subtop DOWPT ↑subsubtop FIRST
↑subjclass DOW ↑classespec DOW ↑dir down ↑deg great ↑tim first
↑vardeg vardeg ↑varlev varlev
↑verbe accuser ↑sppre 2
↑predrem un recul de <vardeg> points dès la première heure de transactions
↑len 15 ↑rand 1)

(make sppre ↑cle 2 ↑terme en baisse de <vardeg> points à l'ouverture
↑len 9)

Une fois l'entrée adéquate choisie, il faut faire le choix syntaxique, c'est-à-dire décider sous quelle forme on va produire l'entrée. Ce choix est soumis à une multitude de contraintes reliées aux connaissances variées d'Ana ou de Frana, selon le cas, et exprimées dans les règles.

En comparant Frana avec le module linguistique anglais de Ana, nous pouvons dresser le tableau suivant concernant les divers choix syntaxiques disponibles dans chacun des modules.

Module anglais (Kukich 1983a)	Module français (Frana - Contant 1985)
sentence	phrase simple (indépendante ou principale)
coordinate sentence (and - but)	proposition coordonnée (et - mais)
subordinate sentence	
After	—
Before	—
As	subordonnée conjonctive
subordinate participial clause	
After	complément de temps antéposé
Before	complément de temps postposé
subordinate participial sentence	(?) subordonnée conjonctive
prepositional phrase	syntagmes prépositionnels antéposé et postposé
adverbial clause	adverbe
—	infinitive avec *pour*
—	relative en *où*
—	épithète détachée
—	nominalisation avec préposition

On peut souligner que le complément de temps en français, selon qu'il soit antéposé ou postposé, fait appel à deux temps de verbe différents et peut être l'objet d'un accord avec le sujet, alors qu'en anglais, il s'agit du même temps et la forme est invariable:

after plunging	- après s'être dégonflé (-e-s)
before plunging	- avant de se dégonfler

Frana, contrairement au module linguistique anglais de Ana, ne contient pas dans ses entrées syntagmatiques toutes les formes temporelles du verbe dont elle pourrait faire usage mais contient plutôt l'attribut ↑*verbe* qui est une clé d'accès à une table de conjugaison.

De façon générale, les principales étapes pour rédiger un texte dans le module linguistique français sont: mettre le focus sur le message suivant, choisir une entrée syntagmatique prédicative adéquate, choisir la forme syntaxique et la conjonction appropriée lorsque requise, choisir un sujet, accorder le verbe selon le sujet, accorder les autres éléments nécessaires, écrire la partie de texte générée. Parmi ces étapes se

greffent diverses contraintes qui s'assurent de mettre un point, une virgule ou un syntagme adverbial au bon moment, de changer de paragraphe lorsque requis, etc.

3. Analyse du corpus au préalable

3.1 Frana face au corpus

Frana possède trois types d'entrées syntagmatiques à l'intérieur de son vocabulaire: 281 prédicats (syntagmes verbaux, nominaux, adjectivaux et/ou prépositionnels), 73 sujets (syntagmes nominaux) et 17 adverbes (syntagmes prépositionnels). Ces entrées couvrent bien l'ensemble de la sémantique de Ana. À chaque message sémantique potentiel correspond une ou plusieurs entrées disponibles.

C'est à partir de l'analyse de notre corpus du domaine boursier (Contant 1985) que nous avons inventorié le vocabulaire de Frana et construit les 371 entrées syntagmatiques. Dans Contant (1985), notre corpus est constitué de 119 rapports boursiers (40 concernant la Bourse de Montréal, 40 pour la Bourse de Toronto et 39 décrivant les activités de la Bourse de New York) et comprend près de 25 000 mots, si on inclut les chiffres. En effet, ceux-ci sont très présents dans le vocabulaire boursier puisqu'ils occupent à eux seuls 20% du corpus. Ces 119 rapports ont été soumis au logiciel JEUDEMO de l'Université de Montréal (Ouellette 1983).

Donc, grâce à l'étude de ce corpus, nous avons permis à Frana de générer des textes ressemblant à ceux rédigés manuellement. Lors de l'élaboration du lexique de Frana, nous avons respecté les termes utilisés ainsi que leur fréquence d'apparition. Ensuite, nous nous sommes attaquée aux règles de la grammaire, la partie la plus difficile à programmer étant les choix syntaxiques possibles en français.

3.2 Structure de texte des rapports boursiers

Un rapport boursier comprend habituellement de six à huit paragraphes dont le contenu est assez standardisé. On y parle d'abord des fluctuations intéressantes du marché au cours de la journée ainsi que de la situation à la fermeture. On qualifie également l'activité du marché (faible à forte). Viennent ensuite les fluctuations de

l'indice composé et son niveau de clôture. Il s'y glisse souvent un paragraphe concernant l'opinion des analystes qui touche les nouvelles d'ordre économique ou politique. Puis on donne le volume des transactions ainsi que le nombre de titres à la hausse et à la baisse. Enfin, on termine en décrivant la situation des différents indices à la fermeture puis celle de certains titres particuliers (normalement les plus actifs). R. Kittredge (1983) parle de deux domaines de référence à l'intérieur de ces rapports:

> 1) le domaine primaire qui serait le «*coeur*» de l'information boursière: évolution du marché au long de la journée, volume d'actions échangées, valeur des indices et de certains titres à la clôture...
>
> 2) le domaine secondaire qui concerne les nouvelles d'ordre économique ou politique pouvant influencer le marché et intéresser les investisseurs (guerre, annonce du bilan financier d'une compagnie importante, etc.)

Le paragraphe concernant l'opinion des analystes fait partie du domaine secondaire puisque les informations qu'il contient ont trait davantage au domaine économique en général qu'au domaine boursier. Nous avons d'ailleurs analysé notre corpus (Contant 1985) sous trois versions différentes: la version originale d'abord, une seconde version n'incluant que le texte relatif au domaine primaire et une troisième version se limitant au contenu sémantique que couvre le système de génération automatique (Ana) décrit précédemment.

3.3 Structure de phrases

On observe dans les rapports boursiers certains stéréotypes de structures. Ces structures sont reliées au contenu sémantique qu'elles véhiculent et semblent donc être le propre de certains types d'informations. Ainsi, pour décrire la situation à la clôture de titres particuliers ou celle des différents secteurs, le gapping (constructions scindées) est très souvent utilisé.

> Mais Consolidated Bathurst gagne 1 1/2 à $56 1/4, Bell 1/2 à $30 1/8, Bombardier 13 1/4 à $15 3/4 et Gulf Canada 1/8 à $18 1/2. (Le Devoir, 18 oct. 1983)

De plus, la structure:

$$\left\{\begin{array}{l} \text{X à (\$)Y} \\ \text{X cents à \$Y} \end{array}\right\} \quad \text{pour les titres}$$

ou X points (,) à Y pour les indices

est tellement fréquente qu'elle se retrouve 2062 fois sur 2086 (98,8%) occurrences où l'on décrit la situation à la clôture. Parmi les autres structures possibles (1,2%), on retrouve par exemple: *L'indice préliminaire des industrielles s'est établi à Y, en baisse de X points*. Dans toutes ces expressions, X équivaut à la fluctuation (écart absolu entre l'ouverture et la fermeture) et Y équivaut à la valeur du titre ou de l'indice à la fermeture. Supposons Z la valeur du titre ou de l'indice à l'ouverture. Nous obtenons:

Hausse: $Z + X = Y$

Baisse: $Z - X = Y$

Ainsi, dans la phrase: *Dome Pete a ajouté 10 cents à $4.25* (Le Devoir, 20 oct. 1983), le sens de *ajouter* n'est pas

$$4.25 + .10 = 4.35$$

comme on a l'habitude de l'interpréter, mais bien

$$4.15 + .10 = 4.25$$

3.4 Vocabulaire

3.4.1 Généralités

Puisque le contenu sémantique des rapports boursiers est assez restreint, les rédacteurs et les rédactrices doivent varier les expressions qu'ils emploient en utilisant des paraphrases (phrases distinctes syntaxiquement et/ou lexicalement mais à contenu

sémantique équivalent). Ils doivent user d'imagination pour ne pas rendre leurs rapports ennuyeux et redondants.

> «L'écoute ou la lecture des bulletins de la Bourse, en anglais ou en français, ne manque pas d'étonner par la richesse lexicale dont les deux langues font preuve pour rendre compte des fluctuations des cours.» (Dominique 1971, p. 55)

Ainsi, pour mieux exprimer le mouvement de baisse, nous retrouvons dans notre corpus une gamme de verbes et de noms tels: *céder, perdre, abandonner, chuter, reculer, fléchir, baisser, diminuer, dégringoler, retraiter, baisse, recul, pertes, chute, déficit, revers, repli,* etc.

Cependant, le vocabulaire est varié mais restreint. En effet, si on double (triple ou quadruple) la grosseur du corpus en nombre de mots, on ne doublera (triplera, quadruplera) pas le nombre de formes différentes car les mêmes mots reviennent après un certain temps. La figure 2 représente la courbe du vocabulaire (nombre de mots et nombre de formes en excluant les chiffres) que l'on obtient lorsqu'on divise la troisième version du corpus (celle limitée au contenu sémantique couvert par le système Ana) en quatre parties égales de 30 rapports boursiers chacune.

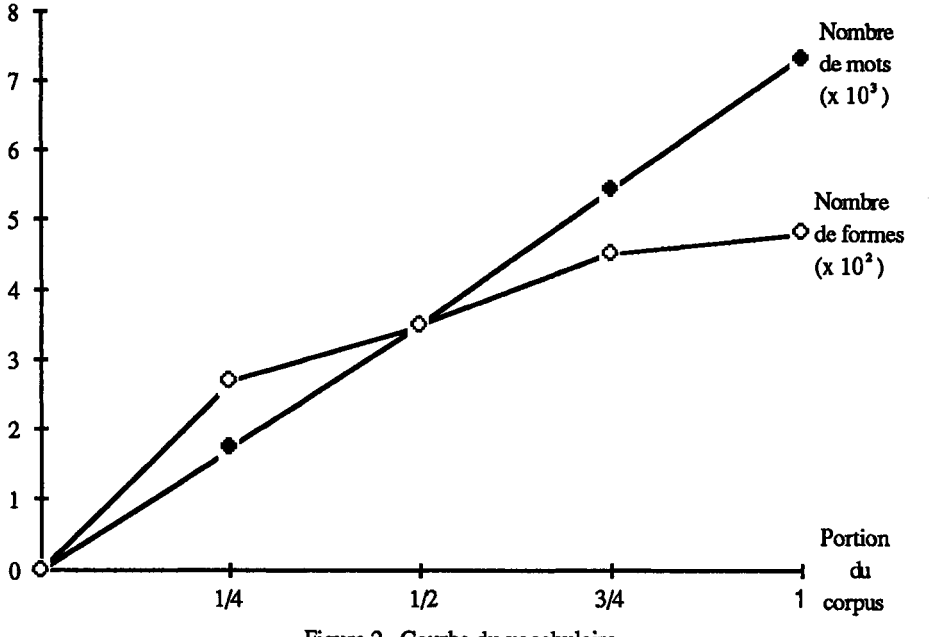

Figure 2 Courbe du vocabulaire

Portion du corpus	Nombre de mots	Augmentation de mots	Nombre de formes	Augmentation de formes	Nombre de mots/forme
1/4	1794	1794	276	276	6,5
1/2	3665	1871	381	105	9,6
3/4	5428	1763	448	67	12,1
1	7040	1612	499	51	14,1

On remarque dans le corpus que les mots utilisés ne sont pas techniques. Ce sont des mots connus de tous, des mots du vocabulaire général, sauf qu'ils prennent un sens particulier. Voici quelques exemples tirés de notre corpus. Nous avons mis volontairement certains mots en caractère italique pour indiquer au lecteur les extensions de sens, les réductions, etc. Certains termes sont même appliqués à des objets comme si ceux-ci étaient animés ou ressentaient des émotions. On utilise l'anthropomorphisme, i.e. qu'on attribue des propriétés humaines ou intelligentes à des objets.

> «... mais les *automobiles*, les *pétroles* et les compagnies financières ont été *déprimées*» (Le Devoir - 9 nov. 1983)
>
> «*Walt Disney* a *chuté* de 1 1/2 à 58 7/8.» (Le Devoir - 10 nov. 1983)
>
> «*Faible* à l'ouverture, l'indice des industrielles s'est ensuite *raffermi* pour *se détendre* à nouveau ...» (Le Devoir, 1er déc. 1983)

3.4.2 Cooccurrences

Dans un sous-langage, la sémantique des mots est restreinte. En conséquence, les mots ne peuvent être agencés entre eux de façon libre, comme on le fait dans la langue en général. Ainsi en est-il dans le sous-langage boursier. En étudiant les cooccurrences (mots avec leur contexte), on s'aperçoit que certains mots sont quasi indissociables. Ils forment des paires ou des unités à part entière.

Par exemple, chaque fois que le mot *main* ou le mot *changer* apparaissent dans notre corpus, ils sont toujours regroupés ensemble sous l'expression *quelque W millions d'actions ont changé de main*. Le mot *quelque* revient d'ailleurs presque toujours dans ce contexte. Sur 234 occurrences, le verbe *gagner* sélectionne toujours

comme complément la forme: *gagner X à Y* sauf un seul cas où l'on retrouve l'expression *gagner du terrain.* Le verbe *dominer* sélectionne toujours un complément du type *la séance, la journée.* Le verbe *s'établir* pour sa part se retrouve chaque fois dans la structure: *s'établir à Y, en hausse (baisse) de X points.* Le verbe *dépasser* sous-catégorise comme sujet et comme complément les formes *le nombre des baisses* ou *le nombre des hausses* (ex: *le nombre des baisses a dépassé celui des hausses).*

Au niveau des noms, on parle toujours dans les rapports boursiers de *l'ouverture* . On ne dira jamais *une ouverture* . Le nom *la veille* est toujours suivi d'une virgule ou d'un point. Il ne sous-catégorisera jamais un complément comme il est permis de le faire dans le langage courant (ex: *la veille de Noël*). L'adjectif *composé* accompagne seulement le nom *indice* . La préposition *parmi* apparaît toujours en début de phrase (45 fois sur 45) dans le contexte: *Parmi les valeurs canadiennes, Parmi les industrielles* . D'ailleurs, l'adjectif *canadienne* se retrouve toujours à côté du mot *valeur*, tout comme le mot *traité* ne s'applique qu'aux *actions, valeurs* ou *volume d'affaire* . Les verbes et les noms ont donc une sous-catégorisation restreinte. Ainsi, beaucoup d'agencements de mots parmi ceux du vocabulaire de ce sous-langage sont impossibles même si dans le langage courant on pourrait leur donner un sens.

3.4.3 Fréquences

Les dix mots les plus fréquents du corpus sont, dans l'ordre: à, de, les, et, le, a, la, cent(s), ont, en.

Environ 44% des mots du vocabulaire linguistique (excluant les chiffres) contenus dans notre corpus font partie de ce groupe de dix ! Les déterminants *le, la* et *les* s'occupent d'accompagner les noms, tout comme la préposition *en* qui précède un nom dans 96% des cas (*en hausse, en recul, en clôture...*).

Les auxiliaires *a* et *ont* du verbe *avoir* sont fréquents étant donnée l'utilisation plus ou moins généralisée du passé composé dans les rapports boursiers.

Les prépositions *à* et *de* sont les deux mots les plus fréquents du vocabulaire. Ceci est justifié par le fait que les rapports décrivent les fluctuations des indices et des

cours en mentionnant l'écart absolu entre l'ouverture et la fermeture (fluctuation de la journée) et la valeur à la clôture.

L'indice a monté de X, à Y points.
Bell a perdu X à $Y.

La présence de la préposition *à* est chaque fois nécessaire alors que celle du *de* dépend du type de complément que requiert le verbe. Les verbes transitifs sont beaucoup plus fréquents que les verbes introduisant leur complément à l'aide de la préposition *de* (546 contre 329 occurrences relevées pour les verbes de fluctuation), d'où, entre autres, la fréquence plus grande de *à* que de *de* . Il ne faut pas oublier que ces prépositions, très fréquentes en général en français, ont aussi d'autres contextes d'utilisation dans le corpus.

Dans les rapports boursiers, les noms de compagnies occupent une large place puisqu'on y décrit les fluctuations des titres importants. Dans notre corpus, ils constituent 9,5% du vocabulaire linguistique total! Mais dans notre troisième version, ils ont complètement disparu puisque Ana ne couvre pas les titres particuliers.

Parmi les autres noms, on retrouve des noms de fluctuation (*hausse, baisse, recul* , etc.), des noms de secteurs qui constituent des indices (*indice des industrielles, les transports, les services publics, les pétrolières...*). Les noms *actions, titres, valeurs* et *cours* , qui sont quatre synonymes, apparaissent 557 fois dans les 119 rapports et le mot *indice* apparaît 122 fois. Voici la liste des noms les plus fréquents dans le corpus original: cent(s), action(s), valeur(s), hausse(s), titre(s), baisse(s), indice, million(s), marché(s), volume, Bourse, point(s). Dans la version du corpus où on se limite à la sémantique couverte par le système de génération automatique de rapports boursiers anglais et français, les noms énumérés précédemment occupent près de 46% du total de noms.

C'est donc toutes ces particularités du sous-langage boursier que nous avons étudiées, ce qui nous a permis d'élaborer le module linguistique français Frana de façon à produire des textes ressemblant à ceux rédigés manuellement.

4. Conclusion

4.1 Résultats

Les résultats de Ana et Frana sont très satisfaisants. La sortie (output) comprend un texte de trois paragraphes, avec en-tête. Voici un exemple de ce que produit le module Frana lorsque combiné avec les trois premiers modules de Ana:

Rapport boursier

Mardi, le 11 janvier 1983

le marché des valeurs boursières est demeuré soutenu tout au long de la journée hier, à Wall Street, où les titres ont fermé sur une forte hausse. les échanges se sont suivis à un rythme frénétique.

l'indice Dow Jones des industrielles a enregistré son plus gros gain en fin d'après-midi pour clôturer en hausse de 16.28 points, se hissant à 1092.35. le groupe des transports a été en avance de 9.06, à 469.43 et celui des services publics a inscrit un gain de 0.83 points à 124.49.

le volume des transactions s'est dégonflé à 101890000, comparativement à 127290000 vendredi dernier. les valeurs à la hausse ont enterré celles à la baisse par 1152 contre 519.

Dans cet exemple, on voit que le texte est linguistiquement adéquat. Il est fidèle au style des rapports boursiers rédigés par des humains francophones (vocabulaire et syntaxe) et les accords sont bien faits, à l'exception du mot *points* qui devrait être au singulier lorsque la valeur numérique est inférieure à 2 mais dont Frana ne s'est pas souciée. De plus, le choix des entrées correspond bien à la sémantique sous-jacente. Le texte n'a pas besoin de révision humaine (mis à part l'exception dont on vient de parler) mais d'un simple traitement de texte pour une disposition

plus attrayante. Le seul accroc orthographique reste l'utilisation de la minuscule en début de phrase. Donc Ana et Frana génèrent du texte linguistiquement bien formé, cohérent, naturel, reflétant bien la réalité des faits qu'il décrit.

4.2 Limites sémantiques

Ana est limitée au contenu sémantique qu'elle possède. Mais grâce à la modularité du système, on peut assez facilement étendre ou modifier ses connaissances. Pour l'instant, il existe diverses lacunes pour générer des rapports aussi complets que ceux présentés dans les grands quotidiens. Il s'agit entre autres des résultats de titres particuliers et des données historiques. Ces informations supplémentaires se laisseraient facilement intégrer au système déjà existant. En effet, la description des cours particuliers respecte une syntaxe rigoureuse et un vocabulaire semblable à celui décrivant la situation des indices à la fermeture. Il faudrait cependant élaborer des règles de sémantique pour sélectionner les titres à mentionner dans le rapport. Les données historiques, quant à elles, supposent la mise à jour d'une base de données des records établis et de la tendance constante à la hausse ou à la baisse durant plusieurs jours.

4.3 Limites linguistiques

Si les sorties (output) du système sont impeccables, c'est en partie grâce au principe de macro-niveau. En effet, le fait de combiner des syntagmes entiers en propositions et de combiner ces propositions en phrases diminue les risques d'erreurs dans le texte final.

Donc, au niveau du vocabulaire, nous avons dit que Ana et Frana utilisent des entrées syntagmatiques et non lexicales. Ce choix est justifié étant donné que nous l'appliquons à un sous-langage, où les expressions toutes faites sont courantes.

Mais ce choix nous apporte des contraintes linguistiques. Ainsi, si on veut modifier un verbe par un adverbe ou un nom par un adjectif pour amplifier ou minimiser un aspect (ex: *se détendre fortement, un maigre gain)*, il faut écrire une nouvelle entrée presque identique à l'exception de cette petite modification. Pourtant, ces processus de modifications sont généraux et s'appliquent de façon automatique par

les humains pour tout nom ou tout verbe compatible. De même, l'utilisation de mots synonymes nécessite une copie de l'entrée. Au lieu d'avoir un patron du style:

FERMER en BAISSE

où FERMER peut prendre différentes valeurs de verbes (*fermer, clôturer, terminer, finir*) et BAISSE différentes valeurs de noms (*baisse, déficit, recul, retrait*), il faut multiplier les possibilités (4 x 4 = 16 entrées). Et on fait un raccourci sur certains problèmes linguistiques: accords internes, cooccurrences permises, choix de l'article, etc., en se situant à un niveau supérieur. Cependant, il reste à accorder le nom sujet avec le verbe et, en français plus spécialement, avec aussi les adjectifs attributs et les participes passés.

Au niveau de la syntaxe, il n'y a pas de structure syntaxique profonde ni de transformations. On passe directement de la représentation sémantique à une structure de surface correspondante. Cependant, tous les types de structures syntaxiques ne sont pas représentés. Certaines structures syntaxiques sont volontairement absentes car l'étude du sous-langage boursier a montré qu'elles n'apparaissaient pas dans la partie des rapports couverts par Ana (impératives, interrogatives, nominalisations propres, relatives autres que avec *où* , etc.). D'autres structures ne sont pas présentes dans les règles de Frana mais une extension du module linguistique pourrait les intégrer (clivées, constructions scindées, etc.). Les règles de syntaxe sont tantôt générales, tantôt spécifiques aux rapports boursiers.

Au niveau de la structure de texte, la cohérence dans la description des faits et la structuration en paragraphes sont adéquats. Mais à cause des entrées syntagmatiques reliées au macro-niveau, il est difficile de prévoir si un mot contenu à l'intérieur d'une entrée ne réapparaîtra pas dans une entrée utilisée subséquemment dans le texte. Quelques mécanismes telle la contrainte d'hyponymie des sujets permettent d'éviter la redondance lexicale mais le problème reste présent. Ces considérations sont plus globales car elles se situent au niveau du texte et elles nous poussent à poser la question: suffit-il de juxtaposer dans un ordre bien établi des propositions bien formées pour obtenir un texte convenable? Il semble que cela suffit dans notre cas puisque le vocabulaire, propre à chaque type d'informations faisant l'objet d'une proposition, est à la fois varié et restreint. En effet, chaque type d'informations contient plusieurs entrées synonymes (donc un lexique varié), et chaque type

d'informations contient des entrées à contenu restreint qui lui est propre. Là où il peut y avoir intersection sémantique, il faut s'assurer que deux types différents d'informations utilisent des entrées dont le lexique est relativement différent. C'est, entre autres, ce dont se charge la contrainte d'hyponymie.

4.4 Avantages

Nous terminons en soulignant quelques avantages de ce système bilingue de génération automatique de rapports boursiers. D'abord, le système génère un texte final qui est fidèle aux événements décrits. Les structures syntaxiques sont bien choisies et les termes employés sont justes, reflétant par leur fréquence le style des textes écrits manuellement. Le texte n'a pas besoin de révision humaine car il est sans erreur. Au niveau de la conception d'Ana, sa modularité permet diverses extensions comme celle du module linguistique Frana que nous avons construit.

Chantal Contant
Université de Montréal

Remerciements

Ce travail a été subventionné en partie par le CRSHC (No 458-84-3064) et par le Fonds F.C.A.R. à titre de bourses de maîtrise. Je remercie particulièrement Richard Kittredge, mon directeur de maîtrise, pour m'avoir initiée au domaine boursier et à la génération automatique de texte. Merci également à Karen Kukich, auteure de Ana, pour sa précieuse collaboration et pour m'avoir ainsi permis de réaliser Frana. Merci à Guy Lapalme pour ses remarques éclairées et à Michel Boyer pour son support technique.

Références

BAUDOT, J. A. (1987) *Introduction aux grammaires formelles*, Sodilis, Montréal.

BECKER, J. D. (1975) «The Phrasal Lexicon» dans *Theoretical Issues in Natural Language Processing* , Cambridge, Mass., pp. 60-63.

COHEN, Betty (1980) *Description préliminaire du langage de la Bourse (I)*, projet de recherche sur les sous-langages, Université de Montréal.

COHEN, Betty (1981) *Description préliminaire du langage de la Bourse (II)* , projet de recherche sur les sous-langages, Université de Montréal.

COLMERAUER, Alain et al. (1971) *TAUM-71*, Groupe TAUM, Université de Montréal.

CONTANT, Chantal (1985) *Génération automatique de texte: Application au sous-langage boursier français*, mémoire de maîtrise, Université de Montréal.

CONTANT, Chantal et M.-H. Gauthier, (1983) *Manipulation du corpus, grammaires de textes, paraphrases* , projet de recherche sur les sous-langages, Département de linguistique et philologie, Université de Montréal.

DANLOS, Laurence (1983) «Some Issues in Generation from a Semantic Representation» dans *Proceedings of the 18th International Joint Conference on Artificial Intelligence* , volume 1, Karlsruhe, pp. 606-608.

DANLOS, Laurence (1984) «Conceptual and Linguistic Decisions in Generation» dans *Proceedings of Coling 84* , Standford University, pp. 501-504.

DANLOS, Laurence (1985a) *Génération automatique de textes en langues naturelles* , Masson.

DANLOS, Laurence (1985b) «Un survol des recherches en génération automatique» dans *Revue québécoise de linguistique* , volume 14, n°2, Montréal, pp. 65-102.

Journal *LE DEVOIR*, rapports boursiers parus dans les pages économiques du 18 octobre 1983 au 10 décembre 1983 (Bourses de New York, Montréal et Toronto).

DOMINIQUE, Philippe (1971) «Vocabulaire boursier de la hausse et de la baisse» dans *La linguistique* , volume 7, n°1, pp. 55-72.

FORGY, C. L. (1981) *OPS5 User's Manual* , Department of Computer Science, Carnegie-Mellon University.

KITTREDGE, Richard I. (1982a) «Variation and Homogeneity of Sublanguages» dans *Sublanguage: Studies of Language in Restricted Semantic Domains*, Walter de Gruyter, pp. 107-137.

KITTREDGE, Richard I. (1982b) «Sublanguages» dans *American Journal of Computational Linguistics*, volume 8, n°2, pp. 79-84.

KITTREDGE, Richard I. (1983) «Semantic Processing of Texts in Restricted Sublanguages» dans *Computers & Mathematics with Applications*, volume 9, n°1, pp. 45-58.

KITTREDGE, Richard et al. (1973) *TAUM-73*, Groupe TAUM, Université de Montréal.

KUKICH, K. (1983a), *Knowledge-Based Report Generation: A Knowledge-Engineering Approach to Natural Language Report Generation*, thèse de Ph. D., Department of Information Science, University of Pittsburgh.

KUKICH, K. (1983b), «Design of Knowledge-Based Report Generator» dans *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, MIT, Cambridge, pp. 145-150.

KUKICH, K. (1983c), «Knowledge-Based Report Generation: A technique for Automatically Generating Natural Language Reports From Numeric Databases» dans *Proceedings of the ACM-SIGIR Meeting*, Washington.

MANN, William C. (1982) «Text Generation» dans *American Journal of Computational Linguistics*, volume 8, n°2, pp. 62-69.

MANN, William C. (1983) «An Overview of the Nigel Generation Grammar» dans *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, MIT, Cambridge, pp. 79-84.

MANN, William C. et J.A. Moore (1981) «Computer Generation of Multiparagraph English Text» dans *American Journal of Computational Linguistics*, volume 7, n°1, pp. 17-29.

McDONALD, D.D., M.E. Cook et W.G. Lehnert (1984) «Conveying Implicit Content in Narrative Summaries» dans *Proceedings of Coling84*, Stanford University, pp. 5-7.

McKEOWN, Kathleen R. (1983) «Focus Constraints on Language Generation» dans *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, volume 1, Karlsruhe, pp. 582-587.

OUELLETTE, Francine (1983) *JEUDEMO*, bulletin 96-01b, Centre de Calcul, Université de Montréal.

WINOGRAD, Terry (1983) *Language as a Cognitive Process: Syntax*, Addison-Wesley.