Université de Montréal

Sélection de l'information pour la génération d'un texte associé à un graphique statistique

par
Marc Corio

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M. Sc.)
en informatique

Décembre 1998

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:
Sélection de l'information pour la génération d'un texte associé à un graphique statistique

présenté par:
Marc Corio

a été évalué par un jury composé des personnes suivantes:

Yoshua Bengio	(président-rapporteur)
Guy Lapalme	(directeur de recherche)
Esma Aïmeur	(membre du jury)

Mémoire accepté le: …….

Sommaire

Ce mémoire s'intéresse à un aspect de la génération automatique de texte en langue naturelle: la sélection de l'information pour un texte associé à un graphique dans un rapport statistique.

Dans sa thèse de doctorat, Massimo Fasciano a décrit un modèle de génération intégrée de textes et de graphiques statistiques dont il a tiré une implantation, **Postgraphe**. Il a obtenu des résultats satisfaisants pour la partie graphique, mais il aurait été préférable d'améliorer la partie texte. Nous avons raffiné la partie texte, particulièrement l'aspect "QUOI DIRE", en établissant un modèle pour sélectionner l'information à divulguer et ainsi, diriger l'attention du lecteur en énonçant les informations les plus pertinentes.

Une étude de corpus de 411 extraits de texte associés à des graphiques statistiques constitue la base de ce projet. Cette étude nous a permis d'observer comment les professionnels écrivent des légendes ou des textes qui accompagnent un graphique. Nous avons ensuite déterminé comment choisir les informations les plus pertinentes et les plus intéressantes à communiquer dans un rapport statistique.

L'analyse du corpus nous a permis d'identifier 55 codes de classification qui indiquent la façon dont l'information est exprimée. Nous avons distingué sept thèmes principaux dans la nature des messages de nos extraits de texte: les messages descriptifs, quantitatifs, de domination, déductifs, discriminants, qualitatifs et justificatifs.

Nous avons dégagé certaines observations sur la façon de sélectionner l'information et sur les types de messages qu'il sera possible ou impossible à générer de façon automatique. Nous avons développé des techniques de sélection pour générer des extraits de texte associés aux codes de classification les plus fréquents qui couvrent 71% des

extraits de notre corpus. Pour implanter ces techniques, nous avons développé **SelTex** un module PROLOG que nous avons intégré à Postgraphe.

Table des matières

Liste des tableaux

Liste des figures

Remerciements

Je tiens d'abord à remercier Guy Lapalme, mon directeur de recherche, pour ses conseils pertinents, ses interventions encourageantes, sa grande patience, sa disponibilité et ses sympathiques courriers électroniques!

Pour leur précieuse collaboration, je remercie toute l'équipe de RALI / INCOGNITO ainsi que Massimo Fasciano, Esma Aïmeur, Michel Boyer et Richard Kittredge.

Je remercie le FCAR pour son support financier, grâce au programme de réintégration à la recherche qui encourage le retour aux études après une longue période sur le marché du travail.

Enfin, j'aimerais remercier mes parents, mes amis de PRIMUS ainsi que mes collègues de GIRO pour leurs encouragements et leur support.

Chapitre 1 - Introduction

1.1 Problématique

Dans le domaine de l'Intelligence Artificielle, on cherche constamment à satisfaire le désir de l'être humain de réaliser des machines qui reproduisent ses facultés. On veut rendre l'ordinateur capable de reproduire les processus cérébraux de l'être humain mais surtout, on essaie de faire accomplir par l'ordinateur ce qu'il n'est pas encore capable de faire.

À l'époque où l'on rédigeait nos travaux avec une dactylo manuelle, le cerveau envoyait un signal, à chaque fois qu'on approchait la fin d'une ligne, qu'il était temps de cesser de taper et d'effectuer un retour de chariot. Ce "signal du cerveau", l'être humain a réussi à le transposer sur une machine, dans les logiciels de traitement de texte. Si cet "exploit" ne nous impressionne pas aujourd'hui, il en sera sans doute de même dans quelques décennies pour certaines tâches qui nous paraissent complexes actuellement.

Ce défi de vouloir transposer sur une machine le processus de réflexion humain peut apporter des bénéfices économiques évidents au sein de l'industrie tout en nous aidant à réfléchir sur le fonctionnement du cerveau, et à mieux comprendre comment on réfléchit.

Notre projet de recherche se rattache à une branche importante de l'Intelligence Artificielle: le traitement de la langue naturelle et de façon plus précise, la génération de texte. Un objectif général de cette branche est de déterminer comment on utilise la langue pour communiquer.

La génération de texte: une sous-discipline de l'Intelligence Artificielle

La génération automatique de texte consiste également à reproduire un processus cérébral: lorsque vous écrivez une phrase, votre main et votre stylo sont guidés par votre cerveau! Mais quelles sont les étapes à franchir pour décider quoi écrire? Quel "algorithme" utilise-t-on pour choisir les mots justes et l'ordre dans lesquels ils se suivent?

Une approche pour décrire les problèmes en génération de texte consiste à poser les 2 questions QUOI DIRE? et COMMENT LE DIRE? Pour répondre à la question QUOI DIRE?, il faut déterminer la nature de l'information contenue dans le message: la sémantique. Pour répondre à la question COMMENT LE DIRE?, il faut se préoccuper de la forme: syntaxe, morphologie, mise en page, etc.

Génération intégrée de textes et de graphiques: l'ère multimédia

Suite à l'évolution rapide de la puissance des ordinateurs au cours des dernières années, le mot multimédia est devenu à la mode: transmettre et diffuser des fichiers graphiques et sonores de plusieurs centaines de Kilo-octets est devenu aujourd'hui une opération commune. Depuis une dizaine d'années, plusieurs systèmes importants de génération intégrée ont vu le jour. Nous énumérons ici les étapes importantes d'un générateur de rapports statistiques multimédias:

- La sélection du contenu (quoi dire?)
- La sélection du médium (par quels outils l'exprimer?)
- L'organisation du discours (comment ordonner et agencer l'information?)
- Réalisation finale des éléments textuels et graphiques du rapport (comment le dire?)

Notre projet se concentre sur la génération d'un rapport statistique, en particulier la sélection du contenu textuel.

1.2 QUOI DIRE?

L'objectif d'un rapport statistique est de faire parler les chiffres. Plus la quantité de données brutes dont on dispose est grande, plus il y a de façons différentes de faire parler les chiffres et plus la capacité d'extraire les informations pertinentes et d'en déduire les conclusions intéressantes devient un mécanisme complexe.

Par exemple, tous les cinq ans, on fait un recensement au Canada, le dernier ayant eu lieu en 1996. Quelques mois peuvent suffire pour compiler et publier les données brutes sous forme de tableaux, sans essayer d'interpréter les données. Par contre, dans les années qui suivent, Statistique Canada publie des centaines de publications comportant des rapports statistiques où l'on établit différentes comparaisons, évolutions et corrélations basées sur les données brutes du recensement.

Lorsqu'une personne analyse des statistiques dans le but de rédiger un rapport, elle fait appel à son intelligence pour décider quelles informations sont les plus intéressantes et les plus pertinentes à communiquer. On veut cibler les détails intéressants afin de s'assurer que le lecteur retient les détails essentiels. C'est ce processus "intelligent" qu'on a voulu cerner et transposer dans une machine, plus précisément dans l'application Postgraphe, un système de génération intégrée de textes et graphiques statistiques. Les résultats obtenus par Postgraphe dans [FAS96] sont satisfaisants au niveau des graphiques mais le traitement du texte est demeuré superficiel. À travers notre projet, nous tentons d'établir les principaux critères de décisions qui peuvent amener une personne à sélectionner les faits pertinents à exprimer explicitement sous forme de texte, pour accompagner un graphique.

"Idéalement, la sélection du contenu doit se faire en connaissant et en utilisant l'intention du rédacteur pour mieux cibler la nature du message à transmettre." [FAS96]

Les intentions du rédacteur sont l'élément directeur du processus de génération, mais la sélection de l'information est aussi basée sur les valeurs des données, les types de données et les relations entre les données; le type de graphique choisi pourra aussi influencer la sélection du texte.

1.3 Description du projet

De façon concrète, notre projet consiste à améliorer la génération de texte dans Postgraphe**.** Pour ce faire, nous développons un nouveau module, SelTex, qui s'intègre à Postgraphe pour choisir des modèles de texte pertinents pour chaque type d'intention et en établissant des critères de sélection pour déterminer le "QUOI DIRE". La figure 1.1 montre un exemple de graphique et de texte généré par Postgraphe et SelTex.

Au chapitre suivant, nous décrivons les travaux antérieurs en génération intégrée de textes et de graphiques, en insistant principalement sur le texte. Nous présentons le système d'intentions du rédacteur ainsi que l'application Postgraphe. Nous montrons un exemple de rapport produit par Postgraphe et décrivons nos objectifs quant aux améliorations à apporter au système, via notre module SelTex.

Le chapitre 3 est consacré à notre corpus: nous avons relevé 411 extraits de textes accompagnant des graphiques statistiques. Nous expliquons la méthodologie utilisée et la classification effectuée. Nous identifions les types d'extraits de texte les plus fréquents pour chaque type d'intention du rédacteur. Ensuite, nous évaluons les résultats obtenus afin de déterminer les principaux critères de décision pour la sélection du contenu.

<insérer fig11.ps>

L'Alberta, la Colombie-Britannique et l'Ontario ont un taux plus élevé de ménages ayant un ordinateur alors que le Québec se situe au septième rang avec 24.0%.

FIG. 1.1 - Exemple de sortie générée par Postgraphe et SelTex.

Les éléments essentiels de notre analyse de corpus ont fait l'objet d'une publication [COR98] dans le cadre d'un *workshop* sur les représentations multimédia à la conférence COLING-ACL '98.

Les techniques de sélection utilisées par SelTex sont présentées au chapitre 4, pour chaque type d'intention. Ces techniques ont été développées principalement en fonction des conclusions tirées lors de notre étude de corpus.

1 Rapport généré automatiquement par POSTGRAPHE 4.1

Nouvelle section (1 intentions à traiter).

nouvelles intentions : comparaison de ordinateur entre province.

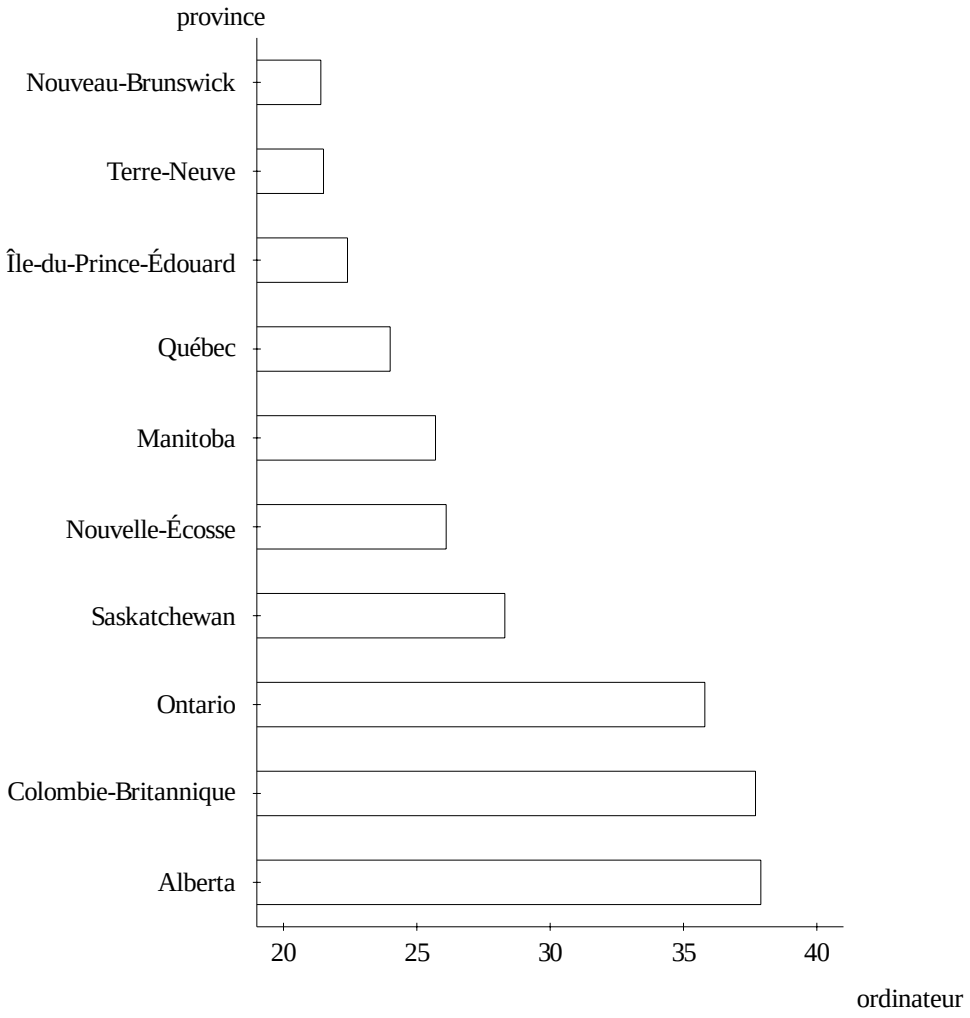


FIG. 1: [Schéma : *barres1*]. comparaison de ordinateur entre province (100).

[Schéma : *position*]. comparaison de ordinateur entre province (100).

L'Alberta , la Colombie-Britannique et l'Ontario ont un taux plus élevé de ménages ayant un ordinateur , tandis que le Québec se situe au septième rang avec 24.0 % .

Au chapitre 5, nous exposons de façon technique comment nos techniques de sélection ont été intégrées à Postgraphe.

Au dernier chapitre, nous terminons avec une évaluation de notre projet ainsi que diverses suggestions pour de futurs travaux.

Chapitre 2 - Travaux antérieurs

Dans ce chapitre, nous décrivons certains résultats de travaux antérieurs sur la génération intégrée de textes et de graphiques en insistant sur les principaux concepts utilisés dans le cadre de notre projet. Nous présentons aussi quelques systèmes reliés à notre domaine de recherche, mais qui ont eu peu d'influence directe sur nos travaux. Puis, nous terminons en exposant brièvement nos objectifs pour faire avancer les travaux réalisés dans le cadre du système Postgraphe.

2.1 Les intentions du rédacteur

Le concept d'intentions du rédacteur est à la base des travaux qui ont mené à Postgraphe et SelTex. L'origine du système d'intentions provient des travaux de Gene Zelazny [ZEL89] qui a proposé un modèle pour choisir le type de graphique le plus adéquat dans un rapport statistique.

Comme point de départ, Zelazny a déterminé que le choix d'un graphique quantitatif se limite à 5 figures de base, illustrées à la figure 2.1.

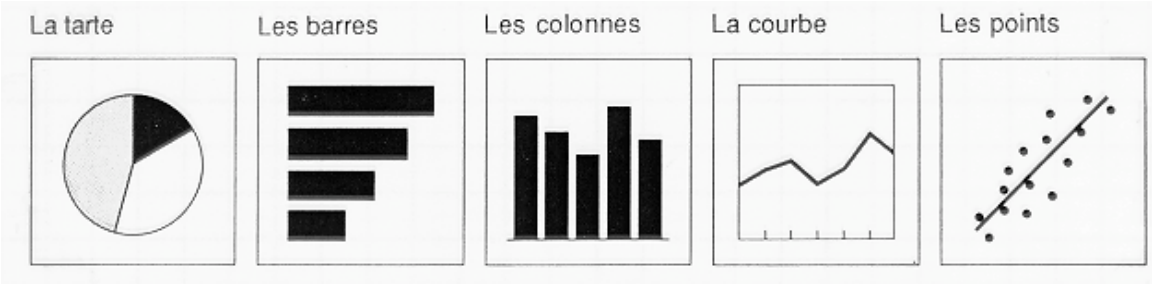


FIG. 2.1 - Les 5 graphiques de base de Zelazny

Avant de faire son choix, le concepteur doit d'abord avoir une idée claire du message qu'il veut transmettre et Zelazny encourage fortement à utiliser le message comme titre du graphique. Par exemple, un titre générique tel "Évolution des ventes de la société" donne le sujet du graphique sans préciser l'élément important à retenir, tandis qu'un titre comme "Les ventes de la société ont doublé" fera en sorte que le lecteur s'attachera à l'aspect que le concepteur veut mettre en évidence.

Une fois le message défini, il faut ensuite identifier l'intention (que Zelazny appelle type de comparaison) sous-jacente à ce message: décomposition, position, évolution, répartition ou corrélation. Le tableau 2.1 fournit les éléments nécessaires pour déterminer l'intention associée à un message: pour chaque intention, on donne la définition, un exemple et quelques mots clés qu'on retrouve typiquement dans les messages associés à l'intention.

Intention	**Définition**	**Exemple**	**Mots clés**
Décomposition	Montrer la taille de chaque fraction (pourcentage) d'un total.	La part de marché de notre client en 1989 est inférieure à 10% du secteur.	• fraction • pourcentage • représente X%
Position	Comparer comment les éléments se classent les uns par rapport aux autres.	La marge bénéficiaire de ce client se situe au quatrième rang.	• plus grand que • plus petit que • égal à
Évolution	Montrer la façon dont les éléments varient dans le temps.	Les ventes ont régulièrement augmenté depuis le mois de janvier.	• changement • croissance, déclin • hausse, baisse • augmentation • diminution • variation
Répartition	Montrer combien d'éléments se répartissent dans chaque intervalle (ou catégorie) d'une série numérique continue.	Au mois de mai, la plupart des ventes se situaient dans une fourchette de mille à deux mille francs.	• catégorie de x à y • concentration • fréquence • distribution • répartition
Corrélation	Montrer si une relation entre deux variables se comporte ou non comme on pourrait s'y attendre.	La rémunération des directeurs généraux ne dépend pas de la taille de l'entreprise.	• lié à • augmente avec • diminue avec • varie avec • évolue en même temps que

TAB. 2.1 - Critères d'identification des intentions selon Zelazny

La dernière étape consiste à sélectionner le graphique le plus approprié correspondant à l'intention. La matrice de la figure 2.2 indique le choix de base pour chaque intention. La justification de ces choix est discutée en détail avec de nombreux exemples dans [ZEL89]. Cette matrice donne une ligne directrice pour le choix du graphique mais d'autres critères peuvent influencer le choix, comme le nombre de données à représenter.

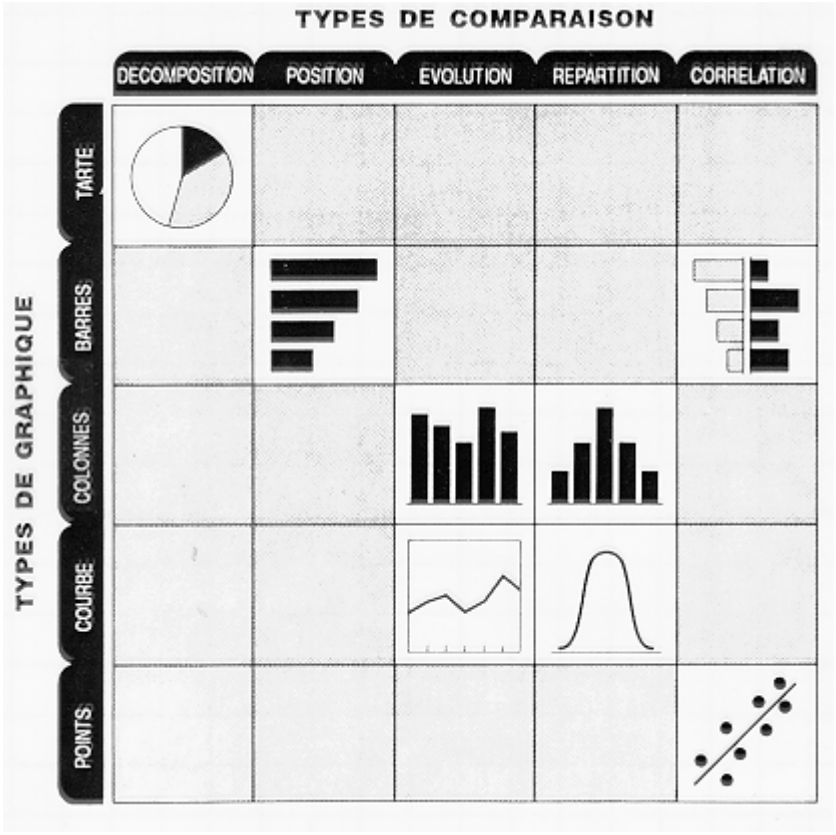


FIG. 2.2 - Matrice de sélection du graphique selon Zelazny

2.2 Intégration graphique et texte: un départ

À l'Université de Montréal, les travaux sur la génération intégrée de graphiques et de texte ont débuté avec le mémoire de maîtrise de Micheline Bélanger [BEL90] qui s'est intéressée particulièrement à des rapports statistiques associés à des données relationnelles. En faisant la synthèse de modèles de générateurs de graphiques et de techniques de générations de texte, elle a retenu les éléments nécessaires pour proposer une méthodologie d'intégration texte/graphique.

Elle discute des principaux problèmes rencontrés dans les étapes de l'élaboration du discours ainsi que des méthodes de présentation des informations. Deux méthodes sont présentées pour diviser l'information en paragraphe.

Dans la première, chaque paragraphe est accompagné d'un graphique représentant les relations qui y sont abordées, en traitant l'information la plus importante en priorité. Cette méthode avait l'inconvénient de faire perdre une notion de globalité souhaitée.

La seconde méthode consiste à séparer les informations qui comparent plusieurs relations de celles qui ne traitent que des particularités d'une seule relation. Cette méthode a été retenue pour l'implantation d'un système de génération intégrée, après avoir été raffinée pour obtenir un résultat acceptable. Le système obtenu peut produire des graphiques en barres ou en points ainsi que du texte mentionnant des valeurs moyennes, des valeurs extrêmes ou des corrélations entre deux variables. La figure 2.3 illustre un exemple de résultat obtenu.

Les textes produits constituent un premier pas intéressant car ils sont sémantiquement et syntaxiquement valables. Mais son système se contentait d'extraire quelques données quantitatives pour les insérer dans des phrases "à trous" sans se soucier de la pertinence du message ou de l'intention du rédacteur.

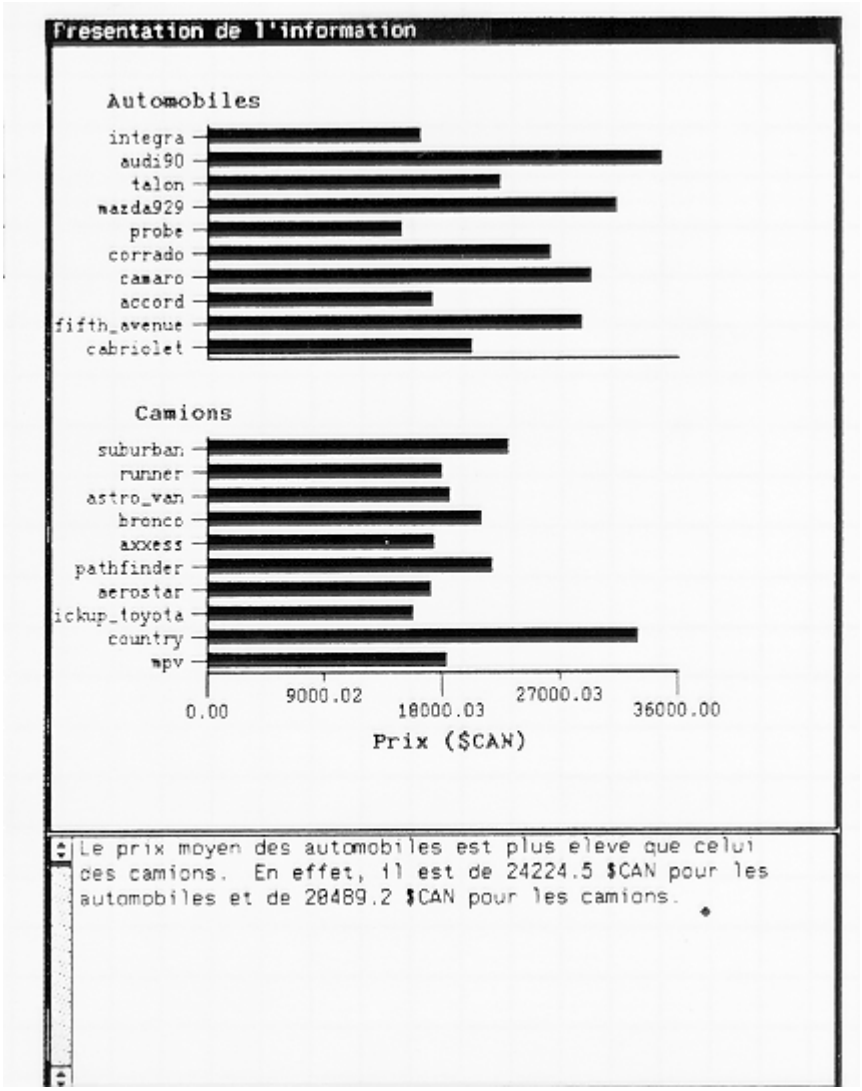


FIG. 2.3 - Exemple de sortie du système [BEL90]

2.3 Postgraphe

Dans le cadre de son projet de doctorat, Massimo Fasciano a poursuivi les travaux de Micheline Bélanger en créant Postgraphe, un système de génération intégrée de graphiques et de textes écrit entièrement en Prolog qui génère des fichiers textes en format Latex et des figures Postscript. Les principaux avancements à la recherche apportés par ce projet sont les suivants:

- Enrichissement du système de types à partir du modèle de Mackinlay [MAC86].
- Développement et intégration d'un modèle de l'intention du rédacteur à partir des travaux de Zelazny (voir section 2.1).
- Diversification importante de la banque de schémas graphiques .

Nous présentons dans cette section les concepts de la thèse de Massimo Fasciano qui nous ont été utiles pour notre projet: le système de types, le modèle d'intention du rédacteur et les éléments théoriques sur les différentes formes d'expression que peuvent prendre le texte dans les rapports statistiques. La section se termine par un exemple d'entrée et de sortie du système Postgraphe.

2.3.1 Système de types

Les informations sur le typage des données permettent de contrôler le choix et la réalisation du texte et des graphiques. Le système de types de Postgraphe est organisé en un graphe d'héritage multiple, composé de 6 facettes principales:

- **Organisation**: variable nominale, ordinale ou quantitative. Inspirée de la classification de Bertin [BER83].

- **Format:** utile pour déterminer comment imprimer les valeurs: étiquette, intervalle, nombre entier, nombre réel.
- **Domaine**: énumération ou bornes inférieures et supérieures. Utile pour préparer les échelles dans les graphiques.
- **Temps**: années, mois. Utile pour savoir comment traiter et exprimer les données, surtout dans un texte.
- **Mesures**: notions de position et de longueur: date, durée, distance.
- **Entités spécifiques**: classes plus concrètes utiles pour résoudre des problèmes particuliers: pays, province, pourcentage, dollar. Ces classes ont été définies spécifiquement à cause de leur fréquence importante dans différentes publications.

2.3.2 Modèle de l'intention du rédacteur

Pour les besoins du projet de recherche, Massimo Fasciano a ajouté et modifié quelques éléments au modèle de Zelazny présenté à la section 2.1.

Une nouvelle intention de base est introduite: l'intention de **présentation** (aussi appelée lecture) définie comme le besoin de lire des valeurs individuelles plutôt que de regarder uniquement les tendances globales, comme la mention d'une moyenne ou d'un total. Cette intention se distingue des autres par le fait qu'elle n'est pas porteuse d'un message particulier. Pour cette raison, dans notre étude de corpus (voir chapitre 3), nous avons inclus les titres génériques parmi les extraits de texte associés à une intention de présentation.

On a également introduit le concept de modificateur**:** celui-ci sert à définir le contenu de la transmission (le message) alors que le terme intention est utilisé pour décrire le but communicatif du rédacteur. On a établi que la décomposition et la position dans le modèle de Zelazny ont le même but communicatif: celui de comparer.

Le tableau 2.2 établit le parallèle entre la classification de Zelazny et celle de Fasciano. Nous avons choisi d'utiliser le mot distribution plutôt que répartition (le terme répartition était trop facilement confondu avec décomposition).

Fasciano		**Zelazny**
Message de base	Modificateur	
Présentation		
Évolution		Évolution
Évolution	augmentation	
Évolution	diminution	
Évolution	stagnation	
Évolution	récapitulation	
Corrélation		Corrélation
Comparaison		Position
Comparaison	décomposition	Décomposition
Distribution		Répartition
Distribution	proportion	

TAB. 2.2 - Classification des intentions (tableau comparatif)

2.3.3 Formes d'expression du texte dans les rapports statistiques

Dans la partie théorique de la thèse, on y décrit les différentes formes d'expression que peuvent prendre le graphique et le texte. Dans le domaine plus spécifique du texte dans les rapports statistiques (le centre d'intérêt de notre projet), il y a 2 formes d'expressions principales: les légendes textuelles et le texte continu.

a) Principales caractéristiques des légendes textuelles

Les légendes textuelles sont généralement limitées à une ou deux phrases qui décrivent les points essentiels qui caractérisent le graphique. Elles visent à guider le lecteur dans son inspection du graphique soit de façon très générale, soit par un commentaire plus précis, ou une combinaison des deux. La légende peut aussi orienter la perception du lecteur de façon subjective.

Contrairement au texte continu, la légende ne situe pas le graphique dans l'ensemble du rapport. Aussi, la proximité de la figure rend les références externes inutiles mais des références à des éléments internes peuvent être utiles pour diriger l'attention du lecteur.

b) Principales caractéristiques du texte continu

Le texte continu contient un plus grand nombre de phrases que les légendes. Habituellement, il situe chaque figure dans l'ensemble du rapport et fait le lien entre les figures. Il inclut les transitions entre les différents thèmes.

Également, le texte continu doit permettre au lecteur de parcourir le texte sans avoir à consulter les figures. Il peut contenir des références externes utiles (le texte n'est pas toujours juxtaposé aux figures).

La description d'un texte continu peut être semblable à la légende mais l'impact est moins direct sur le lecteur.

Notre travail se concentre exclusivement sur les légendes textuelles. Les caractéristiques énumérées ci-dessus nous ont servi de guide dans nos choix d'implantation pour SelTex, d'autant plus que plusieurs d'entre elles ont été confirmées par notre étude de corpus (chapitre 3).

2.3.4 Exemple d'utilisation de Postgraphe

La figure 2.4 reproduit un exemple de fichier d'entrée du système Postgraphe. La dernière portion du fichier contient les données brutes représentant les profits réalisés par 3 compagnies (nommées simplement A, B et C) au cours des années 1994 à 1997. Les autres informations contenues dans le fichier sont dans l'ordre: les noms des variables,

les types des variables, les clés pour construire les relations entre les variables et les intentions du rédacteur.

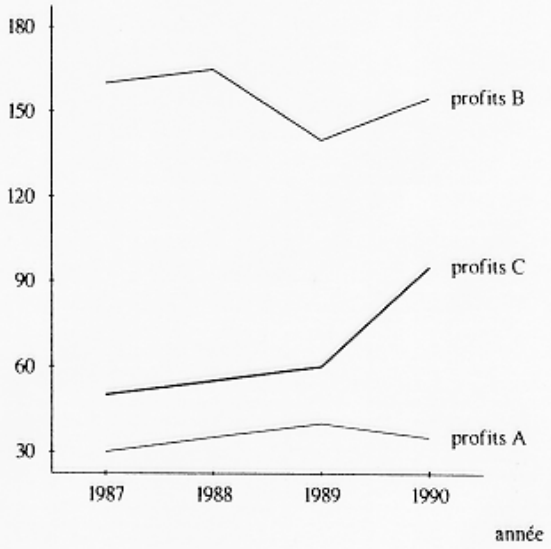
La figure 2.5 montre la sortie du système: le tableau vise à satisfaire l'intention de présentation tandis que les courbes superposées et le texte vise à satisfaire simultanément les intentions de comparaison et d'évolution.

Le texte de cet exemple est le seul type de phrase qui puisse être généré par Postgraphe: l'énumération des tendances de tous les segments de chacune des courbes. Même s'il correspond à une intention du rédacteur, ce texte est répétitif et ne fait pas ressortir une information pertinente.

```
data(% noms des variables
     [année,compagnie,profits],
     % types des variables
     [annee/[symbolique],
      etiquette,
      dollar/[pluriel(profit)]],
     % candidats pour les clés
     [année,compagnie],
     % non-candidats pour les clés
     [profits],
     % intentions du rédacteur
     [% section 1
      [presentation(année),
       presentation(compagnie),
       presentation(profits)],
      % section 2
      [comparaison([profits],[compagnie]),
       evolution(profits,année)]],
     % les données brutes
     [[1987,'A',30],
      [1988,'A',35],
      [1989,'A',40],
      [1990,'A',35],
      [1987,'B',160],
      [1988,'B',165],
      [1989,'B',140],
      [1990,'B',155],
      [1987,'C',50],
      [1988,'C',55],
      [1989,'C',60],
      [1990,'C',95]]).
```

FIG. 2.4 - Exemple d'entrée du système Postgraphe

année / compagnie	1987 profits	1988 profits	1989 profits	1990 profits
A	30	35	40	35
B	160	165	140	155
C	50	55	60	95



De 1987 à 1989 les profits de la compagnie A ont augmenté de 30 $ à 40 $. Jusqu'en 1990 ils ont diminué de 40 $ à 35 $.

De 1987 à 1988 les profits de B ont augmenté de 160 $ à 165 $. Pendant 1 année ils ont diminué de 25 $. Jusqu'en 1990 ils ont augmenté de 140 $ à 155 $.

De 1987 à 1990 les profits de C ont augmenté de 50 $ à 95 $.

FIG. 2.5 - Exemple de sortie du système Postgraphe

2.3.5 Assistant Postgraphe pour Excel

Deux étudiants au baccalauréat, Guy St-Denis et Francis Fauteux ont intégré le système d'intentions de Postgraphe à l'assistant graphique de Microsoft Excel [FAU97]. Ainsi, lorsque l'usager veut un graphique pour représenter des données, il n'a pas à choisir parmi une multitude de type de graphiques; il choisit simplement l'intention et les variables appropriées (voir figure 2.7), et le graphique sera sélectionné selon le même processus que Postgraphe.

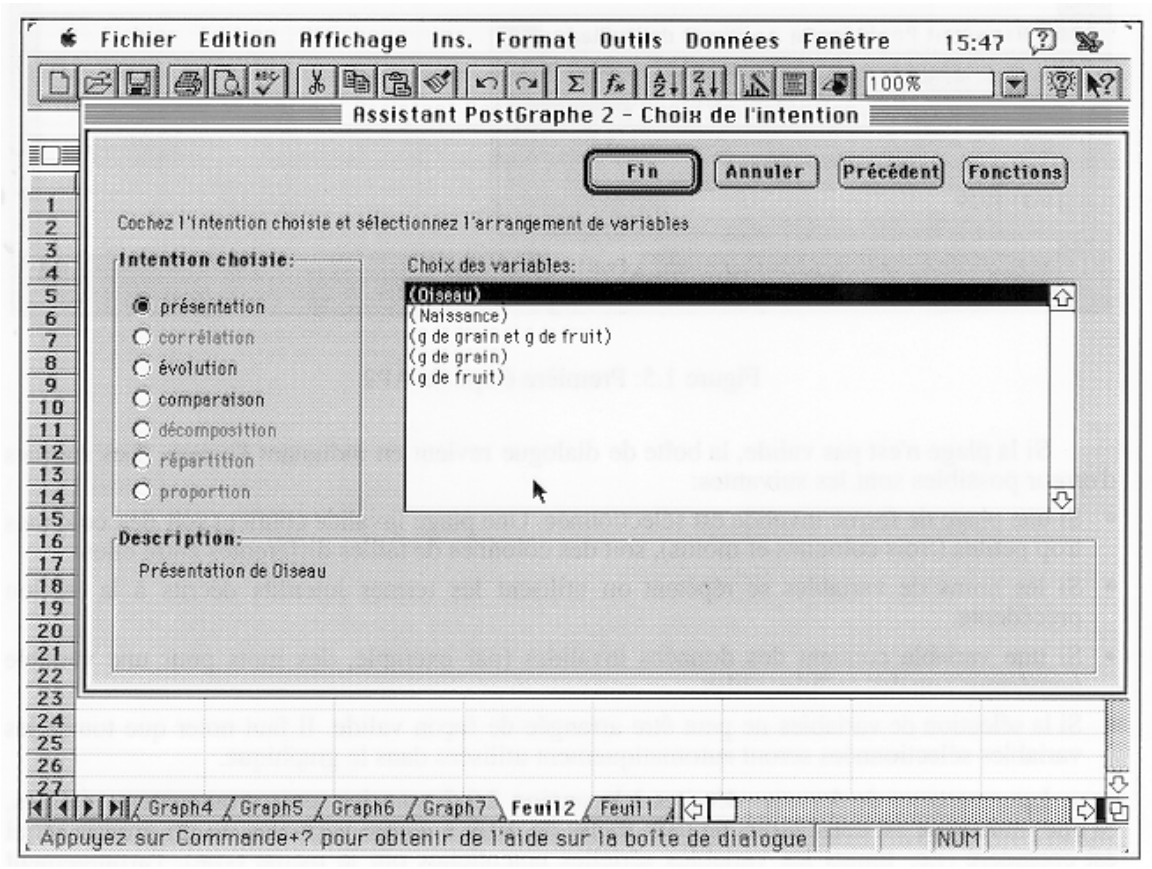


FIG. 2.6 - Assistant Postgraphe

Les groupes sériels

Dans son rapport sur l'assistant Postgraphe publié en août 1997 [FAU97], Francis Fauteux introduit la notion de groupes sériels. Nous définissons cette notion ici, car elle nous a été fort utile pour notre étude corpus ainsi que pour l'implantation.

Une série représente une relation, c'est-à-dire un ensemble de tuples complets permettant de porter sur un graphique un ensemble de points. On appelle groupe sériel un groupe de variables propres à chaque série.

Ce concept s'exprime mieux par un exemple. Supposons un ensemble de tuples sous la forme [X,Y1,Y2] indiquant qu'il y a deux valeurs de Y pour un même X, par exemple:

X représente un pays.
Y1 représente le poids moyen des femmes de ce pays.
Y2 représente le poids moyen des hommes de ce pays.

Ces mêmes tuples peuvent être exprimés sous la forme [X, Y, Z] ou X représente le pays, Y représente le poids moyen et Z représente le sexe. On dit alors que Z, le sexe, représente un groupe sériel. Sur un graphique en colonnes, chaque sexe est représenté par une **série** de colonnes.

Lorsque les tuples sont formés de plus de 2 variables, il y a donc toujours présence de groupes sériels.

2.4 ANA et FRANA

Dans la version originale de Postgraphe [FAS96], la composante de réalisation textuelle utilisée est une version modifiée du générateur **PréTexte** [GAG93] qui est basé sur la théorie de la grammaire systémique. Le choix de PréTexte était justifié d'abord parce qu'il est entièrement écrit en Prolog et aussi parce qu'il est spécialisé en traitement linguistique des expressions temporelles. Toutefois, il y a un inconvénient important: l'interface entre Postgraphe est PréTexte décrite à la section 6.6.3 de la thèse de Massimo Fasciano est beaucoup trop lourde et s'intègre assez difficilement. Un seul schéma de texte a été développé avec cette interface et son format est plutôt rigide: l'énumération des évolutions pour CHAQUE segment de CHAQUE courbe.

Nous avons donc décidé de ne plus utiliser Prétexte pour la génération de texte avec Postgraphe, nous devions trouver une solution moins lourde qui assurerait une meilleure intégration. Nous avons cherché parmi ce qui existait déjà comme génération de texte à partir de données statistiques. Nous avons retenu le modèle utilisé dans les systèmes ANA et FRANA.

Karen Kukich a développé le système ANA en 1983 [KUK83] dans le cadre de sa thèse de doctorat. Ce système génère des rapports boursiers automatiquement, en langue anglaise, à partir des bases de données quotidiennes des cotes de l'indice Dow Jones. Le processus de génération a été développé en 4 modules:

- Générateur de faits
- Générateur de messages
- Organisateur du discours
- Générateur de texte

Le générateur de faits est écrit en langage C et les trois autres modules sont écrits en OPS-5, un langage à base de faits et de règles similaire à Prolog. La sélection du contenu informatif est traité par le générateur de faits, qui s'occupe de sélectionner les statistiques pertinentes et par le générateur de messages qui sélectionne le type de messages à transmettre. L'organisateur du discours détermine l'ordre d'énonciation des différents messages en provenance du générateur de messages.

Enfin, le générateur de texte s'occupe à lui seul de la partie COMMENT LE DIRE?. Il effectue tout le traitement linguistique: les choix grammaticaux, syntaxiques, lexicaux, morphologiques, la ponctuation, etc.

C'est en 1985 que Chantal Contant a développé FRANA [CON85] pour son projet de maîtrise. Suite à une étude de corpus de rapports boursiers en français, elle a développé un module linguistique francophone, en langage OPS-5, qu'elle a substitué au dernier module de ANA.

> *"Le module linguistique (générateur de texte) 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 pour produire un texte élégant grâce à diverses contraintes."* [CON85]

C'est sur ce principe que nous avons construit SelTex, en utilisant les deux mêmes types d'entrée syntagmatique: les syntagmes nominaux pour les sujets et les syntagmes verbaux et/ou prépositionnels pour les prédicats.

Comme point de départ pour l'implantation, nous avons récupéré une adaptation en Prolog de FRANA qui a été écrite par Sylvie Giroux en 1987 dans le cadre du projet de session du cours IFT6010 avec Michel Boyer.

2.5 Autres systèmes

Danc cette section, nous décrivons brièvement d'autres systèmes intéressants de génération de texte qui produisent des rapports statistiques mais qui ont eu peu d'influence sur nos travaux.

LFS

Le système LFS [IOR92] produit des rapports statistiques bilingues sur le niveau d'emploi, le niveau de chômage, la population active, etc.

Dans ce système, l'emphase a été mise sur les points suivants:

- La réalisation grammaticale et les fonctions lexicales.
- Le modèle de planification dont une partie importante consiste à identifier les messages qui peuvent être regroupés dans des structures qui donneront lieu à une seule phrase. Par exemple, deux messages avec un thème identique peuvent être liés par une conjonction.
- Le bilinguisme: LFS peut utiliser des structures syntaxiques différentes pour le même message: le système génère par exemple '*Employment remained virtually unchanged'* en anglais et '*L'emploi a peu varié'* en français.

LFS se concentre principalement sur l'aspect COMMENT LE DIRE de la génération de texte. B. Lavoie a étudié ce système dans son projet de maîtrise [LAV94] où il note que les plans textuels sont trop rigides et qu'on ne tient pas compte des intentions du rédacteur.

RAREAS

Le système RAREAS [GOL86] génère des prévisions météorologiques sous forme de texte à partir de données provenant d'un format particulier.

Dans le cadre de ce projet, les efforts les plus importants concernent l'aspect linguistique (Comment le dire?) comme c'est souvent le cas dans les travaux en génération de texte. Le rapport produit suit un schéma prédéterminé mais le système possède des connaissances géographiques et diverses statistiques météorologiques qui permettent de varier le contenu, par exemple pour détecter des conditions dangereuses et générer un avertissement.

RAREAS produit des prévisions météorologiques pour la marine où on insiste principalement sur la vitesse et la direction des vents. Mais la flexibilité du système permet aisément l'adaptation pour l'agriculture, par exemple, ou pour les bulletins météorologiques publics. De plus, l'approche utilisée peut facilement être adaptée pour produire des textes multilingues.

COMMENT

Le système *Comment* (il s'agit du mot anglais pour *commentaire*) a été présenté par Aurélie Bridault [BRI95] dans le cadre d'un stage réalisé au sein du service RDT (Recherche Développement et Technologies avancées) pour la société Informatique CDC (Caisse des Dépôts et Consignations) à Bagneux, France.

Ce système prototype génère des documents d'analyse financière qui comportent des tableaux numériques illustrées par des graphiques et explicitées par des commentaires. Le module de génération s'appuie sur les techniques de traitement du langage naturel.

Les objectifs visés dans le cadre de ce projet rejoignent les nôtres: accroître la pertinence des commentaires générés, diversifier les contenus sémantiques et multiplier les formes d'expression d'un même contenu pour éliminer la redondance. Pour la détermination du contenu informatif (Quoi Dire?), le système utilise des règles statistiques pour faire une sélection parmi 9 unités de commentaire regroupées dans 3 types d'analyse: les comparaisons aux moyennes, les analyses horizontales et les analyses verticales.

Les résultats obtenus sont encourageants mais l'application est conçue pour produire un rapport très spécifique sur l'état financier des communes. Les données brutes traitées par le système correspondent toujours aux données financières par poste de fonctionnement et d'investissement.

Notre système SelTex utilise des techniques similaires pour la sélection du contenu mais il est plus générique car les règles utilisées peuvent s'appliquer à une plus grande variété de types de données statistiques.

2.6 Nos objectifs

Améliorer la génération de texte dans Postgraphe

Postgraphe possède une variété intéressante de schémas graphiques (21) mais, par manque de temps, seulement 2 schémas textuels ont été conçus. La figure 2.5 montre le seul type de texte implanté: l'énumération des segments d'évolution pour un graphique comportant une ou plusieurs courbes.

Nous avons enrichi la gamme de schémas textuels en développant SelTex**,** une application qui s'intègre à Postgraphe en choisissant des modèles de texte pertinents pour chaque type d'intention et de graphique. Pour ce faire, nous avons remplacé l'interface originale avec Prétexte par une nouvelle basée sur le modèle de FRANA.

Établir des critères de sélection pour déterminer le "QUOI DIRE"

Avant de développer SelTex, nous avons déterminé comment choisir les informations pertinentes et intéressantes à placer dans un rapport statistique, en faisant le lien avec les caractéristiques principales énumérées à la section 2.3.4. Pour atteindre cet objectif, nous avons réalisé une étude de corpus qui fait l'objet du prochain chapitre.

Chapitre 3 - Le corpus

Pourquoi un corpus? Avant d'implanter un générateur de texte, il est intéressant d'observer comment les professionnels écrivent des légendes ou des textes qui accompagnent un graphique. Nous souhaitons que le texte généré soit le plus naturel possible, en nous basant sur de bons exemples. Plus la taille du corpus est importante, plus les leçons à tirer deviendront significatives et mieux on sera en mesure de modéliser adéquatement le processus de génération.

Nous avons extrait de différentes publications (revues, livres, journaux, thèses, rapports) 411 extraits de textes associés à des graphiques statistiques.

3.1 Méthodologie

3.1.1 Le sous-domaine

Les travaux antérieurs en génération de texte ont démontré l'intérêt de se restreindre à un sous-domaine du langage afin d'obtenir des résultats intéressants.

Plus le sous-domaine est restreint, mieux on réussit à générer des rapports automatiques qui se rapprochent des rapports manuels. Lorsque le sous-domaine devient plus vaste, le volume d'expressions fréquentes diminue: on est alors contraint à utiliser un vocabulaire plus général et le style du texte risque d'être plus morne. Nous croyons toutefois qu'il est possible de générer un texte de qualité avec notre sous-domaine même si le résultat généré diffère sensiblement des rapports manuels.

Aussi, plus le sous-domaine est vaste, plus la tâche de construire un corpus représentatif du sous-domaine devient ardue.

3.1.2 Sources

Nous avons cherché des publications qui contenaient une bonne quantité de graphiques statistiques avec du texte qui les accompagne.

Statistique Canada publie le magazine Tendances Sociales Canadiennes (TSC) à tous les 3 mois depuis 1987. Cette publication s'est avérée idéale pour construire notre corpus puisqu'on y retrouve une multitude de graphiques statistiques et de textes sur des sujets très diversifiés.

Jusqu'en 1994, la légende qui accompagne les graphiques du magazine est simplement un titre générique dans la plupart des cas, comme dans la figure 3.1. Mais depuis le dernier numéro de 1994 (TSC no. 35), les rédacteurs du magazine semblent avoir mis en pratique les conseils de Zelazny puisque la légende textuelle qui accompagne les graphiques contient souvent une phrase à message avec une intention précise comme dans la figure 3.2. (À noter cependant que ce n'est pas toujours le cas, comme le témoigne la figure 3.1 qui a été publiée en 1997).

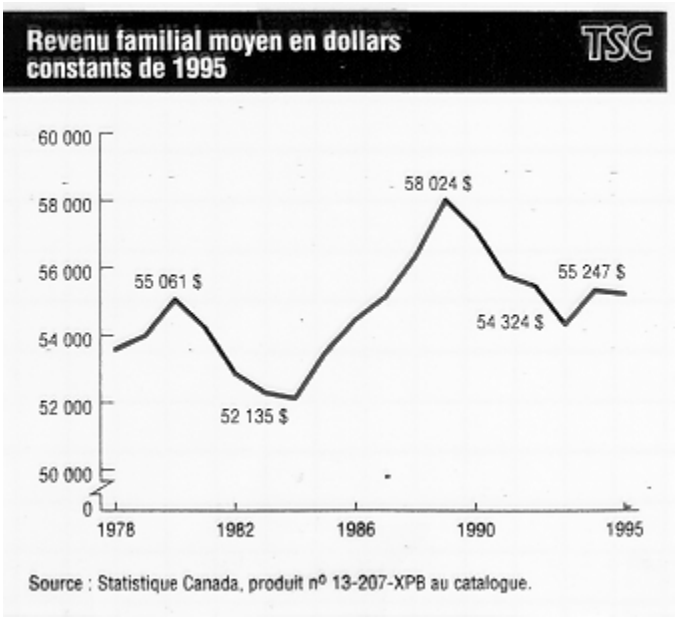


FIG. 3.1 - Graphique avec titre générique provenant de TSC [ts45]

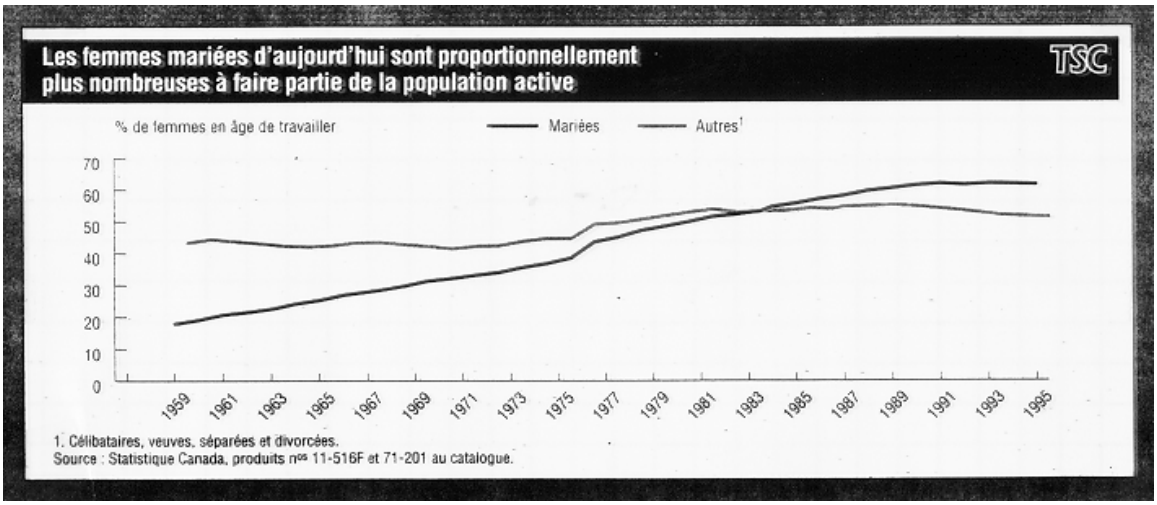


FIG. 3.2 - Graphique avec phrase à message provenant de TSC [ts46]

56% des extraits de notre corpus proviennent du magazine Tendances Sociales Canadiennes. Parmi les autres sources importantes, citons les statistiques du CRSNG

[sng], diverses publications des fonds d'investissement Altamira [alt1, alt2, alt3, alt4, alt5, alt6, alt7] ainsi que le livre de Zelazny [ZEL89]. Le tableau 3.1 énumère les différentes sources de notre corpus et sont présentées de façon plus détaillée à l'**annexe A**.

Référence	Qté
Tendances Sociales Canadiennes, Statistique Canada.	233
Statistiques du CRSNG.	66
Zelazny G. Dites-le avec des graphiques.	52
Fonds d'investissement Altamira.	45
Livres de Blackjack (pour exemples de corrélation).	6
La Presse.	4
Bélanger, Micheline. Génération intégrée de textes et de graphiques.	3
Fasciano, Massimo. Génération intégrée de textes et de graphiques statistiques.	2
Total	411

TAB. 3.1 - Sources du corpus .

Uniformité du corpus

Est-ce que le corpus est représentatif? Le domaine des rapports statistiques est trop vaste pour que nous puissions estimer une marge d'erreur sur nos résultats obtenus. Nous avons fait le maximum d'efforts possibles pour le rendre le plus représentatif possible. Si nous avons sélectionné plus de la moitié des extraits du corpus dans le magazine TSC, c'est que le mandat de cette publication est de publier des statistiques sur tous les sujets sociaux qui peuvent intéresser les gens. Nous avons complété le corpus avec des sources qui se spécialisent dans des thèmes économiques, politiques et scientifiques car ces sujets sont peu présents dans le magazine TSC.

Pour que le corpus soit le moins biaisé possible, nous avons la plupart du temps retenu toutes les légendes textuelles de nos sources sans choisir les plus intéressantes sauf deux exceptions: nous retenions seulement les textes associés à un des 5 graphiques de base et nous avons rejeté une grande quantité de titres génériques ne comportant aucun message comme celui de la figure 3.1; ceux-ci sont malheureusement très souvent présents dans les rapport statistiques!

3.1.3 Informations recueillies pour chaque entrée

Pour chaque extrait de texte de notre corpus, nous indiquons le type de graphique, l'intention du rédacteur, les types de données ainsi que des codes de classification qui ont pour but de classer les extraits de texte dans différentes catégories. Une phrase de texte correspond à une entrée du corpus: nous sélectionnons parfois plusieurs phrases accompagnant le même graphique. Le tableau 3.2 illustre quelques entrées du corpus. Le corpus est présenté en entier à l'**annexe B**.

NO	GRAPH.	SER	TEXTE	RÉF.	INT.	X	Y	CODE1	CODE2
1	Barres		Les aurifères sont restées faibles, non seulement parce que le prix du métal jaune a chuté à un bas de quatre ans, mais aussi à cause du scandale Bre-X.	alt1-7	COMP	%	N	BAS1	RAIS
2	Colonnes	X	Le fonds a affiché un rendement de 7,7% au cours de la pire période de cinq ans (d'août 1976 à juillet 1981).	alt3-7	COMP	T	%	BAS1	SPEC1
3	Barres		Le secteur des aurifères a affiché le pire rendement de l'indice TSE 300, ayant perdu 26,3% depuis le début de l'année.	alt2-7	COMP	N	%	BAS1	
4	Colonnes	X	Les libéraux sont les moins insatisfaits.	presse	COMP	N	%	BAS1	
5	Barres	X	Le taux d'abandon scolaire est le moins élevé chez les personnes qui ont occupé un emploi pendant un nombre modéré d'heures.	ts35-21	COMP	%	N	BAS1	
6	Tarte		La société A détient la plus petite part de marché de sa profession.	zel-45	COMP	N	%	BAS1	
7	Barres		C'est dans la région C que la productivité est la moins bonne	zel-72	COMP	Q	N	BAS1	
8	Barres		Groboucan se situe au dernier rang pour la rentabilité des actifs en 1989	zel-78	COMP	%	N	BAS1	
9	Colonnes	X	Peu de jeunes femmes célibataires mettent leur enfant en adoption	ts32-3	COMP	T	%	BAS2	
10	Colonnes	X	Les travailleurs à temps partiel sont proportionnellement moins nombreux à bénéficier des avantages sociaux	ts43-18	COMP	N	%	BAS2	

TAB. 3.2 - Quelques entrées du corpus

a) Le type de graphique

Pour notre corpus, nous avons retenu seulement des extraits de texte qui accompagnaient un des 5 types de graphiques suivants: tarte, courbe, barres, colonnes ou points. Il s'agit des 5 graphiques de base du modèle de Zelazny (voir section 2.1) qui sont aussi ceux générés par Postgraphe (section 2.3).

Le tableau 3.3 illustre la répartition des types de graphique dans notre corpus. Les colonnes reviennent le plus fréquemment : elles sont utilisées pour tous les types d'intention du rédacteur. Les points sont plus rares : ils sont surtout utilisés dans les rapports scientifiques, peu présents dans notre corpus.

Graphique	Fréquence
Colonnes	35%
Courbe	31%
Barres	22%
Tarte	9%
Points	3%
Total	100%

TAB. 3.3 - Fréquence des types de graphiques dans le corpus

b) Les types de données

Pour chaque entrée de notre corpus, nous avons indiqué le type de données en X et en Y. Le seul type de graphique n'étant pas représenté par 2 axes est la tarte: nous avons alors appelé X la donnée qui identifie chaque secteur de tarte et Y la donnée qui en indique la taille.

Au départ, nous avions classé les données dans 3 catégories soit nominale, ordinale et quantitative mais nous avons ensuite décidé de définir quelques catégories plus spécifiques afin d'obtenir des résultats d'analyse plus concluants. Nous avons distingué

les types de données qui apparaissent le plus souvent mais non de façon aussi détaillée que le système de types utilisé dans Postgraphe.

Les 5 premières lignes du tableau 3.4 indiquent les types de données spécifiques les plus fréquemment rencontrés ainsi que le nombre d'occurrences. Les autres ont été simplement classées comme nominales, ordinales ou quantitatives.

Type de données	Description	Occurrences en X	Occurrences en Y	% global
%	Pourcentage	58	166	28 %
T	Données temporelles (surtout années)	181	1	23 %
$	Argent	16	85	13 %
A	Âge (êtres humains)	28	11	5 %
G	Données géographiques (pays, provinces, villes)	17	13	4 %
N	Données nominales	64	53	15 %
Q	Données quantitatives	30	69	12 %
O	Données ordinales	7	3	1 %
Total		411	411	100%

TAB. 3.4 - Fréquence des types de données du corpus

c) Les groupes sériels

Nous avons indiqué par un 'X' dans notre corpus les graphiques où il y a présence d'une troisième donnée, qu'on appelle les groupes sériels (voir section 2.3). Visuellement, il y a des groupes sériels lorsqu'on a des courbes superposées, des multi-colonnes, des multi-barres, des multi-points ou 2 tartes juxtaposées.

54% des extraits de texte du corpus sont associés à des graphiques **avec** groupes sériels.

Mais l'extrait de texte ne tient pas nécessairement compte des groupes sériels: à la figure 3.3, les deux sexes représentent un groupe sériel, mais l'extrait de texte n'y réfère pas.

d) Référence

Pour chaque entrée du corpus, nous avons indiqué la source et la page. Les sources sont discutées à la section 3.1.2.

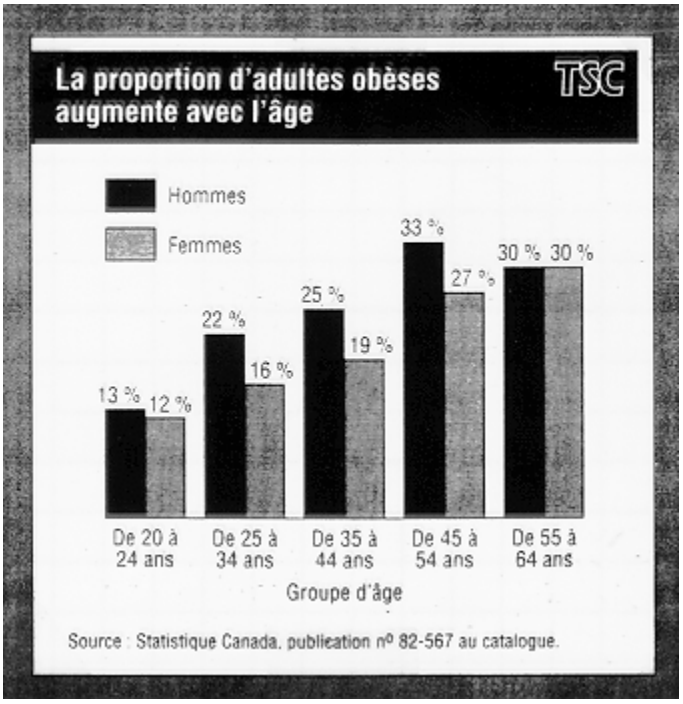


FIG. 3.3 - Graphique avec groupes sériels [ts40]

e) L'intention du rédacteur

Les intentions du rédacteur de notre corpus correspondent aux 5 messages de base de la thèse de Massimo Fasciano décrites à la section 2.3: présentation, évolution, comparaison, distribution et corrélation. L'intention indiquée dans notre corpus est celle qui correspond le mieux au message qu'on retrouve uniquement dans l'extrait de texte; pas dans le graphique ni dans l'ensemble du texte. Il faut garder à l'esprit qu'un des buts de notre recherche est de pouvoir générer du texte associé à une intention.

Pour déterminer l'intention qui caractérise le mieux chaque extrait de texte, nous nous sommes inspiré des critères exprimés dans [ZEL89] et [FAS96] (voir chapitre 2) Dans certains cas, le choix de l'intention est un peu subjectif, par exemple dans le titre du graphique de la figure 3.4.

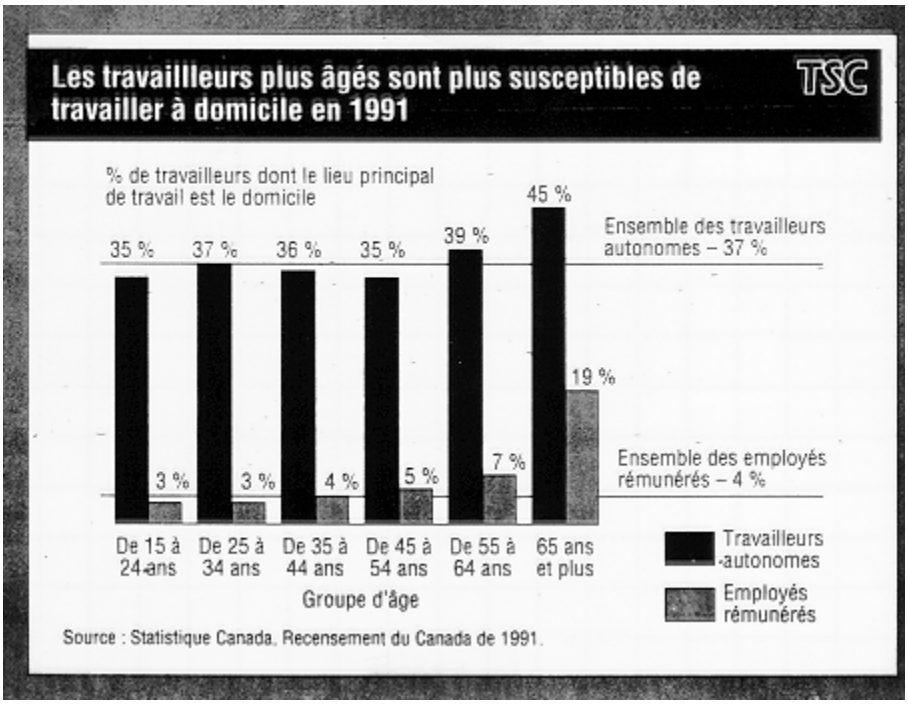


FIG. 3.4 - Choix subjectif de l'intention

La tournure de la phrase nous a incité à classer cet extrait de texte comme CORRÉLATION puisqu'elle évoque un lien entre 2 variables: l'âge et le taux de personnes travaillant à domicile. Si on observe le graphique, on remarque que l'extrait de texte commente la colonne la plus élevée du graphique, ce qui est habituellement considéré comme une COMPARAISON. Aussi, puisque chaque colonne représente à une catégorie d'âge, on a affaire à une DISTRIBUTION.

Les cas similaires sont heureusement rares: la plupart du temps, l'intention associée à l'extrait de texte est sans équivoque, ce qui rend la classification plus facile.

Il peut arriver qu'on retrouve clairement une combinaison de deux intentions dans une même phrase, comme dans le titre de la figure 3.5.

On retrouve une ÉVOLUTION de 1951 à 1991 ainsi qu'une COMPARAISON en mentionnant spécifiquement les personnes âgées.

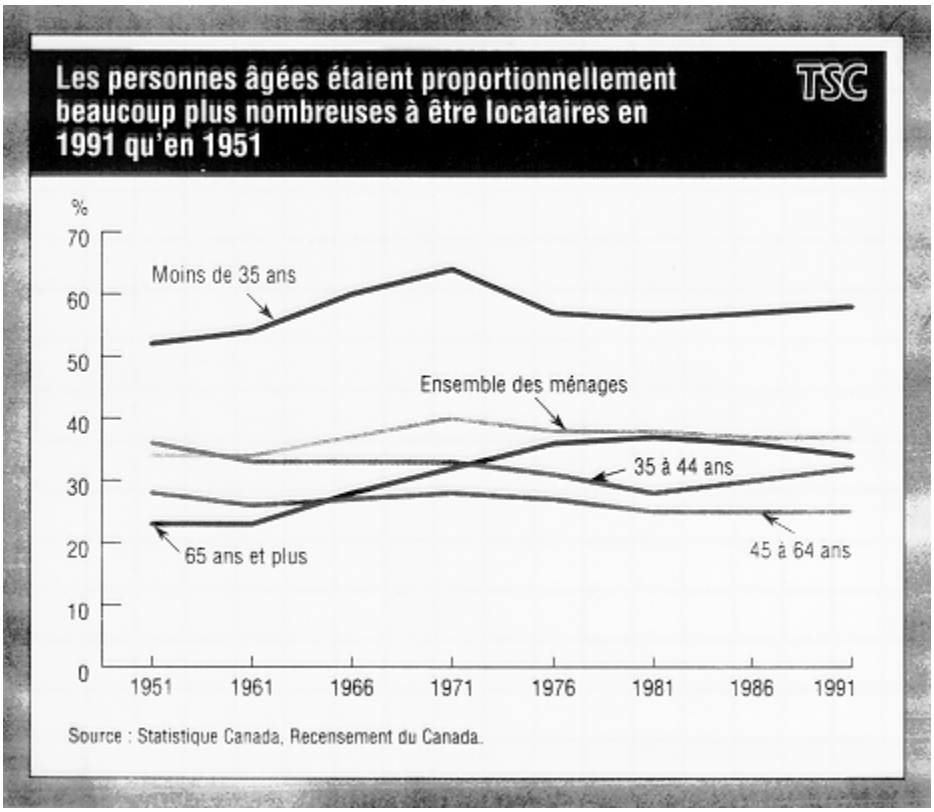


FIG. 3.5 - Combinaison de COMPARAISON et d'ÉVOLUTION

Le tableau 3.5 indique le pourcentage d'extraits de texte de notre corpus associé à chaque intention. On retrouve une combinaison de deux intentions dans 7% des phrases du corpus, ce qui explique le total de 107%. On constate que les intentions de corrélation et de distribution sont beaucoup moins fréquentes.

INTENTION	Fréquence
Comparaison	38 %
Présentation	32 %

Évolution	26 %
Corrélation	8 %
Distribution	3 %
Total	107 %

TAB. 3.5 - Fréquence des intentions dans le corpus

f) Codes de classification

Pour chaque extrait de texte de notre corpus, nous attribuons un ou deux codes de classification qui indiquent la sorte d'information divulguée et/ou la façon dont cette information est exprimée. Par la suite, nous tenterons d'analyser les points communs qui caractérisent les extraits de texte associées au même code. Ces points communs pourront éventuellement servir de critère de décision lors de l'implantation du générateur de texte.

Nous avons révisé cette classification à plusieurs reprises lors de notre analyse du corpus pour la structurer. Cette classification ne devait être ni trop générale car elle serait difficile à exploiter (par exemple une dizaine de codes pour tout le corpus), ni trop spécifique car il serait impossible d'en dégager des règles de formation.

La description des codes de classification que nous attribuons aux extraits de texte répond à la question suivante: *Ce que dit cette phrase, de quelle façon peut-on le voir sur le graphique?*

Chaque code de classification représente une façon d'exprimer un message par du texte. Parmi les 55 codes, plusieurs sont associés à des messages de même nature, mais qui diffèrent par la façon dont ils sont associés à un ou plusieurs éléments du graphique.

Un exemple: Les codes HAUT1 et HAUT4 identifient des données plus élevées. Pour un graphique en colonnes, le code HAUT1 donne la valeur en Y de la colonne la plus élevée tandis que le code HAUT4 identifie le groupe sériel ou les valeurs en Y sont toujours plus élevées que celles des autres groupes sériels.

Le tableau 3.6 donne la liste complète de nos 55 codes de classification. Ils sont présentés ici simplement en ordre alphabétique. Nous les analysons en détail dans les deux prochaines sections: la section 3.2 décrit les codes les plus fréquents pour chaque intention et la section 3.3 les regroupe parmi 7 thèmes principaux.

Codes	Description	Qté
BAS1	La colonne la plus basse, barre la plus courte ou secteur de tarte le plus petit.	8
BAS2	Le groupe sériel le plus bas	3
CARAC	Caractéristique commune d'un groupe plusieurs données	4
CONC	Conclusion générale à tirer.	25
CORR1	Il y a corrélation	7
CORR2	Il y a non-corrélation	6
CORR3	Caractéristiques plus précises d'une corrélation	3
CS1	Comparaison de l'intensité des fluctuations.	4
CS2	A un certain Y, une courbe a rejoint telle autre courbe.	4
CS3	Tendance(s) générales(s) de la (les) courbe(s) qui se démarque(nt) des autres	10
DEB1	La première valeur de Y pour X temporel.	3
DEBFIN	Comparaison du premier et du dernier Y pour un X temporel	6
DIREC	Lien direct entre 2 variables quantitatives	12
ECART	Évolution de l'écart entre les courbes	5
ENUM1	Énumérer les évolutions pour chaque groupe sériel	7
EQUI1	Équilibre entre les différentes données.	3
EQUI2	Équilibre entre 2 données spécifiques.	1
EQUI3	Constance du rapport entre les groupes sériels	1
EQUI4	Fluctuations identiques de courbes superposées	1
EXCEP	Exception dans la constance de l'évolution	2

TAB. 3.6 - Codes de classification (début)

Codes	Description	Qté
FIN1	La dernière valeur de Y pour X temporel.	10
FIN2	Comparaison des dernières valeurs en Y pour X temporel (groupes sériels)	5
FIN3	Groupe sériel dont la dernière valeur en Y est la plus élevée	4
GS1	Comparaison entre 2 groupes sériels.	7
GS2	Colonne ou barre ou le ratio entre les groupes sériels se démarque des autres	3
GS3	Deux (ou +) colonnes ou barres ou le ratio entre les groupes sériels se démarque des autres	1

Codes	Description	Qté
HAUT1	Colonne la plus haute, barre la plus longue ou secteur de tarte le plus gros.	49
HAUT2	Deux (ou +) colonnes les plus hautes, barres les plus longues ou secteurs les plus gros.	7
HAUT3	Sommet de la courbe.	1
HAUT4	Groupe sériel le plus élevé	4
HAUT5	Deux groupes sériels les plus élevés	2
INEQ1	Inégalité entre les données	3
INT	Présentation de l'intention	11
INTER	Indiquer un intervalle de données en X où les données en Y sont plus élevées	4
INVER	Lien inverse entre 2 variables quantitatives	5
MOY1	Y moyen pour courbes ou colonnes, X moyen pour barres.	3
MOY2	Augmentation ou diminution moyenne par unité de temps.	6
OUT	Information qui réfère à des données qu'on ne retrouve pas dans le graphique	21
RAIS	Raison explicative	5
RENV	Point de renversement de la tendance	2
SEUIL	Deux ou plusieurs colonnes, barres ou secteurs qui dépassent un seuil spécifié.	3
SPEC1	Valeur d'une colonne, barre ou secteur qui nous intéresse plus particulièrement.	18
SPEC2	Comparaison de la donnée qui nous intéresse plus particulièrement avec les autres.	3
SPEC3	Présentation de la donnée qui nous intéresse plus particulièrement	2
SPEC4	Tendance de la courbe qui nous intéresse le plus.	2
TAIL	Évaluation de la taille d'une valeur quantitative	2
TEND1	Tendance générale: hausse, baisse ou stabilité.	21
TEND2	Tendance en mentionnant les valeurs de début et de fin et/ou le ratio entre ces 2 valeurs.	11
TEND3	Tendance dans un sous-intervalle de l'axe temporel.	13
TEND4	X temporel où on atteint ou on dépasse un certain seuil en Y	1
TEND5	Description de l'intensité des fluctuations	4
TITRE1	Titre générique mentionnant la description d'une seule donnée	23
TITRE2	Titre générique incluant la description de 2 données ou plus	61
TITRE3	Présentation des données: mention du nombre de valeurs nominales.	4
TOT1	Totaux de toutes les colonnes, barres ou secteurs de tarte.	6
Total		442

TAB. 3.6 - Codes de classification (suite)

3.2 Analyse et description des données

3.2.1 Analyse par type d'intention

Pour chaque type d'intention, nous présentons un tableau des codes de classification les plus fréquents. Notre objectif est d'adapter Postgraphe afin qu'il puisse générer ces types d'extraits de texte de façon automatique à l'aide de schémas de texte appropriés. Nous décrivons quelques exemples de codes caractéristiques à chaque intention avec quelques exemples. Le numéro (#) qui accompagne chaque exemple permet de se référer à l'annexe B pour consulter les informations complètes du corpus.

a) Présentation

Le tableau 3.7 indique les codes de classification associés à des intentions de présentation.

Code	Description	Qté	%
TITRE2	Titre générique incluant la description de 2 données ou plus	61	50%
TITRE1	Titre générique mentionnant la description d'une seule donnée	23	19%
CONC	Conclusion générale à tirer.	11	9%
INT	Présentation de l'intention	9	7%
TITRE3	Présentation des données: mention du nombre de valeurs nominales.	4	3%
TOT1	Mention des totaux de toutes les colonnes, barres ou secteurs de tarte.	3	2%
OUT	Information qui réfère à des données qu'on ne retrouve pas dans le graphique	3	2%
SPEC1	Mention de la valeur d'une colonne, barre ou secteur qui nous intéresse plus particulièrement.	2	2%
MOY1	Mention du Y moyen pour courbes ou colonnes, X moyen pour barres.	2	2%
FIN1	Mention de la dernière valeur de Y pour X temporel.	2	2%
DEB1	Mention de la première valeur de Y pour X temporel.	2	2%
SPEC3	Présentation de la donnée qui nous intéresse plus particulièrement	1	1%
Total		123	100%

TAB. 3.7 - Codes de classification associés à une intention de présentation

Code TITRE2: Le type de texte le plus fréquemment associé à une intention de présentation décrit le type des données en X et Y ainsi que celui des groupes sériels, le cas échéant. Dans une majorité des cas, l'axe horizontal est temporel et on mentionne l'année de début et de fin.

#370: Salaire annuel moyen réel, selon le sexe, 1920-1990.
#379: Immigrants en pourcentage de la population totale, selon la province, 1901-1991.

Code TITRE1: Similaire au précédent, mais le texte ne décrit qu'une seule donnée. Ce code est fréquemment associé à un graphique en tarte. La donnée non mentionnée dans l'extrait est le plus souvent un pourcentage ou ensemble de valeurs nominales. Dans les 2 exemples suivants, l'année est une constante (et non une variable du graphique).

#324: Budget des programmes de bourses, 1994-1995.
#329: Personnes âgées de 15 ans et plus qui ont déménagé, 1989.

Code TITRE3: Ce titre générique spécifie que le graphique illustre un "TOP 10" ou autre quantité de valeurs nominales. Dans notre corpus, il s'agit de données de type géographique pour 3 des 4 occurrences.

#402: Les dix pays de l'OCDE ayant les plus forts pourcentages de jeunes adultes inscrits à l'université, 1988.
#403: Cinq principaux lieux de naissance des résidents non permanents, 1991.

Codes MOY1 et TOT1: Ces codes mentionnent une donnée statistique, total ou moyenne, qui peut être visible ou non sur le graphique. Dans le corpus, il s'agit de dollars dans la majorité des cas. Ce type de message est généré automatiquement dans les travaux de Micheline Bélanger [BEL90].

#225: Le prix moyen des automobiles est de 24224$
#407: Au total, 421,1 millions de $ ou 97,1 % des dépenses du CRSNG en subventions et bourses pour 1993-1994 ont été versés aux universités canadiennes.

b) Évolution

Le tableau 3.8 indique les codes de classification associés à des intentions d'évolution.

Code	Description	Qté	%
TEND1	Tendance générale: hausse, baisse ou stabilité.	20	19%
TEND3	Tendance dans un sous-intervalle de l'axe temporel.	13	12%
TEND2	Tendance en mentionnant les valeurs de début et de fin et/ou le ratio entre ces 2 valeurs.	11	10%
CS3	Tendance(s) générales(s) de la (les) courbe(s) qui se démarque(nt) des autres	7	7%
OUT	Information qui réfère à des données qu'on ne retrouve pas dans le graphique	7	7%
CONC	Conclusion générale à tirer.	5	5%
ENUM1	Énumérer les évolutions pour chaque groupe sériel	5	5%
MOY2	Augmentation ou diminution moyenne par unité de temps.	5	5%
TEND5	Description de l'intensité des fluctuations	4	4%
CS2	A un certain Y, une courbe a rejoint telle autre courbe.	3	3%
DEBFIN	Comparaison du premier et du dernier Y pour un X temporel	3	3%
FIN1	La dernière valeur de Y pour X temporel.	3	3%
	14 codes avec 1 ou 2 occurrences	21	21%
Total		107	100%

TAB. 3.8 - Codes de classification associés à une intention d'évolution

Code TEND1: On décrit la tendance générale de la variation d'une donnée. Ce texte est typiquement associé à un graphique en courbe ou en colonnes. L'axe horizontal est généralement temporel mais le type de données en Y est très variable: pourcentage, dollars ou autres données quantitatives.

#277: La construction d'appartements a diminué depuis 1971.
#283: Hausse constante du nombre de cas de sida diagnostiqués annuellement chez des adultes au Canada depuis 1979.
#287: Les nouvelles mères sont plus nombreuses à commencer leur carrière avant la naissance de leur premier enfant.

Code TEND2: Ce type de texte est plus spécifique que TEND1 mais on le retrouve à peu près dans les mêmes conditions. On observe qu'il est moins pertinent de mentionner la valeur de début dans le cas d'une stabilité ou dans le cas de fortes fluctuations.

#298: Le taux de financement régresse depuis quelques années; d'un maximum de 50% atteint en 1984, il est passé au niveau actuel de 40%.
#299: Le nombre de bourses postdoctorales s'accroît constamment depuis 1983-84, passant de 156 à 353 actuellement, une augmentation de 126%.
#300: La proportion de locataires vivant seuls était quatre fois plus élevée en 1991 qu'en 1951.

Code TEND3: Lorsque la tendance n'est pas constante pour toutes les données du graphique, on mentionne souvent la tendance depuis la dernière fois où elle a été renversée jusqu'à la donnée la plus récente.

#308: Les revenus d'emploi des jeunes hommes sont en baisse depuis la fin des années 70.

On peut aussi faire l'énumération de tendances dans plusieurs segments. Les 3 phrases suivantes se suivent dans un même texte.

#205: Depuis 1961, deux tendances bien nettes se sont dessinées.
#296: De 1961 à 1975, le taux d'homicides n'a pas cessé d'augmenter, passant de 1,3 pour 100000 habitants à un sommet de 3,0, ce qui représente une augmentation de 131%.
#295: De 1975 à 1994, en dépit des fluctuations annuelles, le taux d'homicides a graduellement diminué, passant de 3,0 pour 100 000 habitants à 2,0.

c) Comparaison

Le tableau 3.9 indique les codes de classification associés à des intentions de comparaison.

Code HAUT1: 25% des extraits de texte de notre corpus associés à une intention de type comparaison mentionnent la valeur la plus élevée. Dans une proportion beaucoup moins fréquente (5%), on mentionne parfois la valeur la moins élevée (code BAS1).

#159: La Suède se classe au premier rang, au plan des capacités de lecture et d'écriture.

#144: L'Université de la Colombie-Britannique a reçu la proportion la plus importante, soit 40,5 millions de $ ou 9,6% du total du financement accordé aux universités canadiennes.

Code	Description	Qté	%
HAUT1	Colonne la plus haute, barre la plus longue ou secteur de tarte le plus gros.	36	25%
SPEC1	Valeur d'une colonne, barre ou secteur qui nous intéresse plus particulièrement.	15	10%
BAS1	La colonne la plus basse, barre la plus courte ou secteur de tarte le plus petit.	8	5%
HAUT2	Deux (ou +) colonnes les plus hautes, barres les plus longues ou secteurs de tartes les plus gros.	7	5%
GS1	Comparaison entre 2 groupes sériels.	7	5%
CONC	Conclusion générale à tirer.	5	3%
OUT	Information qui réfère à des données qu'on ne retrouve pas dans le graphique	5	3%
FIN2	Comparaison des dernières valeurs en Y pour X temporel (groupes sériels)	5	3%
FIN1	La dernière valeur de Y pour X temporel.	4	3%
CARAC	Caractéristique commune d'un groupe plusieurs données	4	3%
HAUT4	Groupe sériel le plus élevé	4	3%
RAIS	Raison explicative	4	3%
INEQ1	Inégalité entre les données	3	2%
ECART	Évolution de l'écart entre les courbes	3	2%
CS3	Tendance(s) générales(s) de la (les) courbe(s) qui se démarque(nt) des autres	3	2%
GS2	Colonne ou barre ou le ratio entre les groupes sériels se démarque des autres	3	2%
FIN3	Groupe sériel dont la dernière valeur en Y est la plus élevée	3	2%
BAS2	Le groupe sériel le plus bas	3	2%
DEBFIN	Comparaison du premier et du dernier Y pour un X temporel	3	2%
SPEC2	Comparaison de la donnée qui nous intéresse plus particulièrement avec les autres.	3	2%
	13 codes avec 1 ou 2 occurrences	18	12%
Total		146	100%

TAB. 3.9 - Codes de classification associés à une intention de comparaison

Code HAUT2: Ce type de texte peut être plus pertinent que le précédent selon l'écart entre les valeurs. Par exemple, si l'écart est relativement petit entre les 3 premières valeurs mais qu'il devient plus important entre la 3ème et la 4ème valeur, on énumérera les trois valeurs les plus élevées.

#190: Les adultes de la Colombie-Britannique, de la Nouvelle-Écosse et de l'Alberta sont les plus nombreux à faire des déplacements d'intérêt faunique.

#188: Le financement de la R et D universitaire en sciences naturelles et en génie provient de trois sources principales: le gouvernement fédéral, les gouvernements provinciaux et les universités elles-mêmes.

Code HAUT4: Ce type de texte nécessite évidemment la présence d'un groupe sériel, par exemple des multi-colonnes ou des courbes superposées.

#196: Les immigrants d'autres régions que les États-Unis ou l'Europe sont proportionnellement plus nombreux à vivre avec des parents.

Code BAS2: Dans une proportion beaucoup moins fréquente, on mentionne parfois le groupe sériel le moins élevé (code BAS2).

#10: Les travailleurs à temps partiel sont proportionnellement moins nombreux à bénéficier des avantages sociaux.

d) Corrélation

Le tableau 3.10 indique les codes de classification associés à des intentions de corrélation.

Code	Description	Qté	%
DIREC	Lien direct entre 2 variables quantitatives	12	32%
CORR1	Il y a corrélation	7	19%
CORR2	Il y a non-corrélation	6	16%
INVER	Lien inverse entre 2 variables quantitatives	5	14%
CORR3	Caractéristiques plus précises d'une corrélation	3	8%
CONC	Conclusion générale à tirer.	2	5%
EQUI4	Fluctuations identiques de courbes superposées	1	3%
RAIS	Raison explicative	1	3%
Total		37	100%

TAB. 3.10 - Codes de classification associés à une intention de corrélation

Codes DIREC et INVER: 46% des extraits de texte associés à une intention de corrélation mentionnent que 2 données sont reliées entre elles de façon directe ou inverse. Le graphique associé peut être un graphique à points ou à colonnes. Le plus souvent, le lien se situe entre les données en X et en Y.

#88: La proportion d'adultes obèses augmente avec l'âge.

#219: À mesure que la période de placement s'allonge, la volatilité des placements s'amenuise.

On peut également faire mention d'une dépendance entre les différents éléments d'un groupe sériel dans le cas de courbes superposées ou de multi-colonnes.

#84: Les adultes dans les ménages à faible revenu sont en moins bonne santé.
#109: Depuis 1986, les voyages des Canadiens aux États-Unis ont suivi les fluctuations du dollar canadien.

Codes CORR1 et CORR2: dans ce type de texte, on mentionne simplement si un lien existe ou non.

#44: L'opinion des femmes quant à l'importance d'occuper un emploi rémunéré pour être heureuses varie selon l'âge
#51: Les tarifs plus élevés de certaines marques de carburant ne sont pas liés à une supériorité de leurs performances

Code CORR3: on peut donner des caractéristiques plus précises d'une courbe de régression par exemple en donnant la pente ou en qualifiant la forme de la courbe. Bien que ce type de texte est rare dans notre corpus, il est probablement plus fréquent dans des publications scientifiques.

#54: Note the bowed, parabolic, nature suggested for the regression function.
#55: So on average we expect the running count to rise linearly with number of decks already played, such that the slope is +4 per deck.

e) Distribution

Le tableau 3.11 indique les codes de classification associés à des intentions de distribution.

Code	Description	Qté	%

HAUT1	Colonne la plus haute, barre la plus longue ou secteur de tarte le plus gros.	13	65%
INTER	Indiquer un intervalle de données en X où les données en Y sont plus élevées	3	15%
CONC	Conclusion générale à tirer.	2	10%
EQUI1	Équilibre entre les différentes données.	1	5%
SEUIL1	Deux ou plusieurs colonnes, barres ou secteurs qui dépassent un seuil spécifié.	1	5%
Total		20	100%

TAB. 3.11 - Codes de classification associés à une intention de distribution

Code HAUT1: 65% des extraits de texte associés à une intention de distribution mentionnent la valeur la plus élevée pour un graphique ou chaque barre, colonne ou pointe de tarte représente un intervalle. Pour la plupart, il s'agit d'intervalles d'âge ou de revenus.

#175: La pratique d'un sport est la plus répandue chez les jeunes hommes.
#176: Les gains des mères de 35 à 54 ans travaillant à temps plein ont connu la progression la plus rapide de 1980 à 1990.
#179: Les membres des ménages appartenant à la catégorie des revenus les plus élevés sont proportionnellement plus nombreux à se déclarer en excellent ou en très bonne santé.

Code INTER: Ces extraits de texte sont similaires à ceux du code HAUT1 mais ils diffèrent par la façon dont ils se réfèrent au graphique: on englobe plusieurs colonnes ou barres ou encore une portion de l'aire sous la courbe.

#215: La plupart des adultes âgés de moins de 25 ans vivent au domicile familial.
#216: Il y a davantage de diplômés dans les tranches de salaires les plus élevées.

f) Combinaison de plusieurs intentions

Parmi les 411 phrases de notre corpus, nous en avons relevé 35 où il y a combinaison de deux intentions, soit environ 9%. Pour 31 de ces phrases, la comparaison est combinée à l'évolution; les 4 autres combinent la présentation et l'évolution. Nous croyons que d'autres types de combinaisons peuvent exister, mais nous nous limitons ici aux 2 types rencontrés dans notre corpus.

Combinaison Comparaison/Évolution

Cette combinaison d'intention est exprimée de deux façons différentes: une comparaison d'évolutions ou une évolution de la comparaison. Le premier cas est plus fréquent.

Comparaison d'évolutions: on compare les tendances des groupes sériels. Il s'agit le plus souvent de courbes superposées, mais on retrouve aussi quelques cas de multi-colonnes ou multi-barres.

#93: Comme vous pouvez le constater, les placements effectués dans le cadre d'un RÉER fructifient beaucoup plus vite que les autres.
#98: Quelques conclusions préliminaires se dégagent: - les libéraux ont perdu du terrain durant la campagne; -les réformistes ont fait des gains; - les autres partis ont reçu le jour des élections à peu près le même niveau d'appui qu'au moment de son déclenchement.
#136: Les Canadiens consacrent moins de temps à regarder la télévision et plus de temps à d'autres activités de loisirs.

Évolution de la comparaison: dans ce type de message, on commente la variation de l'écart entre 2 groupes sériels, en général des courbes superposées.

#59: En 1995, les ventes de temps d'antenne ont correspondu aux dépenses d'exploitation de la télévision privée, et ce, pour la première fois depuis la fin des années 80.
#61: D'ici 2030, il devrait y avoir excédent des décès sur les naissances.
#275: Bien que les taux d'homicide commis à l'aide de carabines ou de fusils de chasse aient progressivement diminué depuis 1974, ils ont, par le passé, représenté la majorité des homicides commis à l'aide d'une arme à feu.

Combinaison Présentation/Évolution

Ces extraits de texte décrivent principalement une tendance générale (évolution) mais on y ajoute un aspect de présentation qui peut être quantitatif ou qualitatif. Le premier exemple ci-dessous inclut un message quantitatif: la mention du total. Le second

exemple inclut un aspect qualitatif qui évalue la taille des données. À la section 3.3, nous reviendrons sur le concept des messages quantitatifs et qualitatifs.

#407: Le budget des subventions s'élevait à 400 millions de $ en 1994-1995, une légère baisse par rapport à l'année précédente.
#277: Les propriétaires-occupants d'unités condominiales sont encore rares, mais leur nombre s'accroît.

3.2.2 Analyse par type de graphique

Même si notre projet ne s'attarde pas au choix du graphique, il est intéressant de vérifier, avec le tableau 3.12, si les associations entre intentions et graphiques suggérées par Zelazny (ref. chap 2) se reflètent dans notre corpus. De façon majoritaire, on constate les mêmes liens: les points montrent les corrélations, les barres et la tarte expriment les comparaisons (appelées décomposition et position dans le modèle de Zelazny) tandis que les colonnes et les courbes représentent les évolutions et les distributions.

Parmi les exceptions, la principale est le choix des colonnes pour 30% des intentions de comparaisons. Rappelons que l'intention de présentation n'est pas incluse dans les travaux de Zelazny. On constate qu'elle peut être liée à tous les types de graphique.

Graphique	COMP	CORR	DIST	EVOL	PRES	TOTAL
Barres	57%	7%	4%	8%	25%	100%
Colonnes	30%	11%	10%	18%	31%	100%
Courbe	23%	1%	1%	47%	29%	100%
Points	7%	87%			7%	100%
Tarte	47%			13%	39%	100%
TOTAL	34%	8%	4%	25%	29%	100%

TAB. 3.12 - Intention vs type de graphique dans notre corpus

Nous ne présentons pas une description détaillée des principaux codes de classification pour chaque type de graphique comme nous l'avons fait pour les intentions, car l'analyse n'apporte pas d'éléments nouveaux. Dans notre corpus, il n'y a pas de forte cohésion entre le texte et le type de graphique; le choix du texte est plus fortement relié au type de données. L'interaction entre graphique et texte est discutée plus en détail à la section 3.3.2. Le tableau 3.13 permet quand même d'observer comment se répartissent les 10 codes de classification les plus fréquents, selon le type de graphique.

Code	Barres	Colonnes	Courbe	Points	Tarte	Total
TITRE2	7	23	28	1	2	61
HAUT1	18	25			6	49
CONC	6	11	5		3	25
TITRE1	5	7	3		8	23
TEND1	2	5	14			21
OUT	4	5	6		6	21
SPEC1	8	7			3	18
TEND3		2	11			13
DIREC	1	10		1		12
INT	2	4	5			11
TEND2		6	5			11

TAB. 3.13 - Codes de classification les plus fréquents selon le type de graphique.

3.3 Évaluation globale du corpus

Les codes de classification du tableau 3.6 ont été définis en gardant à l'esprit qu'ils auront un lien direct avec les schémas textuels pour l'implantation de SelTex. Mais de façon globale, que dit-on dans tous ces extraits de texte? L'analyse du corpus nous a permis de distinguer **sept** thèmes principaux dans la nature des messages de nos 411 extraits de texte.

3.3.1 Les 7 thèmes du QUOI DIRE

Les 7 thèmes sont énumérés dans le tableau 3.14 en ordre décroissant de fréquence d'apparition dans le corpus. Ces pourcentages ont été obtenus en associant les 55 codes de classification du tableau 3.6 au thème le plus représentatif.

Thème	Fréquence
Descriptif	30%
Quantitatif	22%
Domination	19%
Déductif	15%
Discriminant	9%
Qualitatif	3%
Justificatif	1%
Total	100%

TAB. 3.14 - Les 7 thèmes des messages du corpus

Dans les paragraphes qui suivent, nous donnons une courte description de chacun des 7 thèmes avec quelques exemples et le lien avec les intentions. A la fin de la section, le tableau 3.15 résume les liens entre thèmes et intentions.

Les messages descriptifs: Ce type de message décrit ce que le graphique représente dans son ensemble ou identifie son aspect visuel principal. On explique par un titre ou une légende ce que représentent les données en X et Y (codes TITRE1 et TITRE2). On

décrit la tendance générale d'une courbe ou l'évolution de l'écart entre 2 courbes (codes TEND1 et ECART). On identifie l'intention du graphique (code INT).

Certains messages descriptifs indiquent que le graphique ne montre qu'un sous-ensemble des données (code TITRE3): le texte "*Les 10 pays de l'OCDE ayant les plus forts pourcentages de jeunes adultes inscrits à l'université, 1988*" informe le lecteur que le graphique ne représente pas tous les pays membres de l'OCDE.

Les messages descriptifs sont associés principalement à des intentions de **présentation** (73%) et **d'évolution** (22%). Des travaux antérieurs (ANA/FRANA) ont démontré qu'il était possible d'obtenir des résultats intéressants pour générer du texte décrivant une évolution.

Les messages quantitatifs: Les chiffres bruts sont les éléments essentiels de ce type de message. On sélectionne des données qui intéresseront le lecteur. Quel est le prix moyen des automobiles ? A combien s'élevait le budget des subventions de l'année précédente? A quel rang se situe notre compagnie par rapport à l'ensemble de l'industrie? A combien se situe le taux d'escompte présentement? Les messages quantitatifs visent à satisfaire les curiosités statistiques du lecteur. On peut également mentionner que tel item est plus élevé que tel autre sans spécifier les chiffres.

Ce type de message peut être associé aux intentions de **comparaison** (46%), **d'évolution** (30%) ou de **présentation** (23%). Souvent, on mentionne la valeur d'une donnée qui intéresse plus particulièrement l'usager (code SPEC1) comme dans l'extrait "*Le revenu annuel moyen d'une famille à Vancouver était de 59 700$ en 1993*" qui accompagne un graphique où chaque barre est représentée par une ville. Si Vancouver était au premier rang ou au dernier rang, on pourrait parler d'un message de domination (thème suivant) mais ce n'est pas le cas: l'article au complet s'intéresse spécifiquement à la ville de Vancouver.

Une bonne proportion des messages quantitatifs de notre corpus réfère à des données qui n'apparaissent pas sur le graphique (code OUT). Par exemple, on a une tarte qui illustre la décomposition du budget pour 1997 et le texte qui l'accompagne compare ces chiffres à ceux de l'année précédente.

Les messages de domination: Ces messages expriment les valeurs les plus élevées ou les plus basses (codes HAUT* et BAS*). On suppose que le lecteur sera intéressé à savoir qui remporte la palme du succès ou celle de la médiocrité, que ce soit au sens réel ou imagé. Qui est le plus riche? Quelle compagnie a obtenu le pire déficit? Quelle région a été le plus touchée par la tempête de verglas? Quels sont les logiciels les plus utilisés? Quel fonds d'investissement a obtenu le pire rendement? Quelle est la principale cause de mortalité chez les jeunes? Dans quelle catégorie d'âge retrouve-t-on le plus de fumeurs? Les messages de domination répondent à ce genre de question.

Ce type de message est utilisé lorsque l'écart est significatif dans le cas d'une donnée unique ou relativement constant dans le cas d'un groupe sériel. L'analyse du corpus révèle qu'on nomme parfois 2 ou 3 données dominantes lorsque celles-ci se détachent du reste du groupe.

Les messages de domination sont surtout associés à une intention de **comparaison** (76%) ou de **distribution** (20%). Pour une intention de distribution, la domination est manifestée par un intervalle plutôt que par des données ponctuelles (voir exemples de la section 3.2.1). Avec les modificateurs de décomposition ou de proportion, les données sont exprimées en pourcentage.

Le message peut parfois indiquer si les données constituent une énumération complète. On ne pourrait affirmer que *"Le secteur des aurifères a affiché le pire rendement de l'indice TSE 300"* si le graphique montre seulement les 10 secteurs ayant obtenu le meilleur rendement. On peut par contre affirmer que *"Au Canada, les adultes*

de Terre-Neuve sont les moins sportifs" si toutes les provinces canadiennes sont incluses dans le graphique.

Les messages déductifs: Dans ce type de message, on déduit quelque chose à partir de l'allure du graphique ou de la valeur des données. On tire une conclusion (code CONC). On fait remarquer une caractéristique commune entre certaines données (code CARAC). On identifie une corrélation (code CORR1). On constate un équilibre ou une constance dans les données (codes EQUI*).

Ces messages utilisent parfois une information additionnelle pour tirer une conclusion comme dans l'extrait *"Les provinces de l'ouest du Canada avaient le taux d'emploi le plus élevé chez les adolescents en 1993"*, où des connaissances géographiques plus complètes sont nécessaires: il faut savoir que la Colombie-Britannique, l'Alberta et la Saskatchewan sont situées à l'ouest du Canada alors que cette connaissance n'est pas exprimée par les données elles-mêmes.

Dans notre corpus, les messages déductifs sont répartis dans les 5 intentions de base mais ceux qu'il est possible de générer automatiquement sont associés à des intentions de **corrélation** (49%) et de **comparaison** (24%). La plupart des messages déductifs associés aux 3 autres intentions sont des conclusions générales (code CONC).

Les messages discriminants: Dans ce type de message, on identifie un fait saillant qui se démarque de l'ensemble des données. On identifie une irrégularité. On mentionne le point de renversement d'une courbe (code RENV), celui où 2 courbes se coupent (code CS2) ou bien une exception dans la constance de l'évolution (code EXCEP). On distingue une donnée où le ratio entre groupes sériels diverge du reste des données (code GS2).

Les messages discriminants sont liés à une intention d'**évolution** (82%) ou de **comparaison** (18%).

Les messages qualitatifs: Dans ce type de message, on qualifie les données avec des adjectifs tel *rare, faible, multiple, forte, élevée, fréquente,* etc. (code TAIL); on donne la forme de la courbe de régression (code CORR3). On qualifie l'intensité des fluctuations (code TEND5).

Ce type des messages a un aspect subjectif car il dépend du jugement de l'auteur; une même donnée peut être qualifiée de différentes façons, selon le contexte.

Les messages qualitatifs sont le plus souvent associés à une intention de **comparaison** (62%) mais aussi à **l'évolution** (23%) ou à la **corrélation** (15%).

Les messages justificatifs: Dans ce type de message, on cherche à identifier les causes. Pourquoi telle colonne est la plus élevée? Pourquoi la courbe s'est renversée en telle année? Pourquoi le dollar a chuté? Pourquoi les intentions de vote d'un parti sont plus élevées qu'au dernier sondage? Les messages justificatifs répondent à ce genre de question pour justifier la valeur ou la tendance des données.

L'échantillon de ce type de message dans notre corpus est trop faible pour pouvoir l'associer à une intention spécifique. Les phrases de notre corpus étaient tirées de légendes ou de courts textes accompagnant un graphique. Les messages justificatifs apparaissent souvent dans des textes plus longs.

THEME	COMP	CORR	EVOL	PRES	DIST	Total
Descriptif	4%		22%	73%		100%
Quantitatif	46%		30%	23%		100%
Domination	76%		4%		20%	100%
Déductif	24%	49%	7%	16%	4%	100%
Discriminant	18%		82%			100%
Qualitatif	15%	23%	62%			100%
Justificatif[1]	80%	20%				100%

[1] Échantillon non significatif

TAB. 3.15 - Lien entre thèmes et intentions

3.3.2 Interaction graphique et texte

Complétion vs redondance

On pourrait croire que pour obtenir une bonne complétion entre texte et graphique, le texte doit rapporter des informations que le graphique ne montre pas de façon évidente. C'est souvent le contraire: de façon générale, le texte sert à renforcer ce qui ressort déjà du graphique. Il y a redondance seulement lorsque le texte répète *toute* l'information de façon exhaustive: ce phénomène s'explique bien par le fait que 29% des extraits de texte associés à une intention de comparaison servent à mentionner la donnée la plus élevée, comme pour confirmer au lecteur: "Oui, ce que vous voyez ressortir de ce graphique, c'est bien ça qui est important."

Cohésion

Pour obtenir une forte intégration entre le texte et le graphique, il est surtout nécessaire de se référer au type de données. La cohésion entre le graphique et le texte ne dépend pas principalement du type de graphique (barres, tarte, courbe, colonnes ou points); elle dépend beaucoup plus souvent du type de **données** utilisé sur le graphique en X et en Y.

Par exemple, dans l'extrait "*Il y a davantage de diplômés dans les tranches de salaires les plus élevés*", on voit sur le graphique que les données sont représentées en tranches de salaires mais celles-ci peuvent être représentées graphiquement par des barres, colonnes, régions sous la courbe ou même secteurs de tarte. Ainsi, chaque type de données a son propre lexique qui assure la cohésion: les mots *tendances* et *évolution* vont référer à un axe de données temporelles, peu importe que le graphique soit une courbe ou en colonnes.

Dans notre corpus, les coréférences aux éléments visuels du graphique sont plutôt rares mais nous croyons que ce phénomène est spécifique au sous-domaine des rapports statistiques. Dans le sous-domaine des manuels d'instructions, les coréférences aux éléments visuels du graphique sont plus importantes.

3.3.3 Faisabilité pour la génération automatique

Cette étude de corpus nous a permis de noter certaines observations quant à la façon de sélectionner l'information appropriée pour la génération de texte; on obtient une idée générale sur les types de messages qu'il sera possible ou impossible à générer de façon automatique.

Heureusement, plusieurs types de messages sont faciles à générer en utilisant des fonctions statistiques: déduire des corrélations (déductif), décrire l'allure générale d'une courbe (descriptif) ou identifier les données les plus élevées (domination). Également, on peut, la plupart du temps, générer un message quantitatif en calculant une moyenne ou une somme pour un ensemble de données ou mentionner les valeurs de début et de fin. Ceci n'élimine pas pour autant les difficultés habituelles à l'étape du "comment le dire".

Dans plusieurs cas, il faut demander à l'usager des informations supplémentaires pour pouvoir générer du texte pertinent. En plus de donner ses intentions, l'usager doit identifier les données pour lesquelles il a un intérêt plus important, comme dans l'exemple de Vancouver à la section précédente. Aussi, le système doit savoir qu'un ensemble de données nominales forme ou non une énumération complète pour pouvoir dire qu'une donnée est au "dernier rang", ou qu'il s'agit des "10 principaux pays".

Pour pouvoir générer des messages descriptifs explicites, l'usager doit s'assurer de préciser suffisamment de détails dans l'identification des données. Par exemple, pour générer le titre "*Taux de personnes accusées de conduite avec facultés affaiblies, selon la province, 1978 à 1994*", l'usager devra fournir la description complète "personnes

accusées de conduite avec facultés affaiblies" associées aux données brutes appropriées. Autrement dit, pour ce qui relève de la description, le système ne pourra pas en dire plus que ce qu'on lui dit.

Pour générer des messages qualitatifs pertinents, le système doit connaître un vocabulaire riche et juste, en fonction du type de données utilisé. Par exemple, une donnée jugée élevée peut être qualifiée de *coûteuse, répandue, fréquente* ou *populaire* selon le type de données représenté. D'autre part, une bonne connaissance du contexte peut aussi être nécessaire pour savoir si une donnée est forte, normale ou faible: par exemple un taux de 5% peut-être *bas* si on parle de l'imposition du revenu mais d'*élevé* si on parle du taux d'inflation.

Pour certains types de messages discriminants, le principal problème n'est pas de savoir comment s'y prendre pour les générer mais de juger si l'information est pertinente ou non. Par exemple, est-il toujours approprié de mentionner le point de rencontre de 2 courbes? Ou de mentionner le renversement de la tendance? Établir des critères pour juger de la pertinence d'une telle information peut s'avérer un problème complexe.

Les messages qui tirent des conclusions générales sont plus compliqués à générer de façon automatique, comme l'extrait "*Ce graphique montre pourquoi il vaut mieux cotiser à un RÉER plutôt qu'à un régime non enregistré*." ou encore "*Les ménages canadiens se sont rapidement adaptés aux nouvelles technologies d'application domestique*." Même si nous ne réalisons pas ce genre de message "intelligent" dans le cadre de notre projet, nous croyons qu'il serait intéressant d'en étudier la faisabilité dans de futurs travaux.

En général, les messages justificatifs sont ceux que le système ne pourra pas générer automatiquement. Pour être cynique, on pourrait dire que ce type de message est "trop intelligent" pour le système mais c'est surtout parce qu'ils contiennent des informations que le système ne connaît pas et qu'il ne peut déduire à partir de ce qu'il connaît. Par

exemple, il est impossible de déduire que "*Le prix de l'or a chuté à cause du scandale Bre-X*" en analysant les données brutes sur l'évolution du prix de l'or.

Un des principaux problèmes de faisabilité est relié à l'étendue du sous-domaine, tel que discuté à la section 3.1.1. Dans Postgraphe, on traite les rapports statistiques de façon générale sans a priori sur les données de l'usager. Si, par exemple, le système est utilisé par Statistique Canada, il serait raisonnable de fournir au système les règles et connaissances nécessaires pour déduire que la Colombie-Britannique et l'Alberta sont les provinces les plus à l'Ouest. En demeurant au niveau général, il est irréaliste de s'attendre à ce que le système paraisse trop intelligent! Il devient alors nécessaire que le fichier d'entrée du système contienne une quantité raisonnable de paramètres pour guider le générateur. Évidemment, on tentera d'obtenir le maximum d'information de l'usager sans trop lui alourdir la tâche.

Chapitre 4 - Techniques de sélection

La génération de texte dans Postgraphe ne se limite qu'à un seul schéma de texte pour une intention d'évolution: il énumère une à une les hausses et les baisses de chaque segment de chaque courbe (voir figure 2.5), ce qui n'est pas toujours pertinent. Pour les intentions autres que l'évolution, Postgraphe ne génère aucun texte. Les résultats de l'analyse du corpus du chapitre précédent nous ont guidé pour choisir les types de texte à générer pour chaque intention.

Dans ce chapitre, nous décrivons nos choix pour l'implantation de SelTex. Pour chacune des 5 intentions de base, nous indiquons les codes de classification du tableau 3.6 que nous avons sélectionnés pour l'implantation ainsi que le thème associé le plus fréquemment à chaque code et nous motivons nos choix par des exemples. En plus des 5 intentions de base, nous avons consacré une section à la combinaison de la comparaison et de l'évolution.

Les phrases en italiques dans ce chapitre ont été générées par SelTex, sauf celles qui sont indiquées avec # et un nombre qui réfère à un exemple de notre corpus.

4.1 Comparaison - sans variable temporelle

Le tableau 4.1 indique les codes de classification sélectionnés pour l'implantation de la comparaison.

De façon générale, les messages de comparaison indiquent les données les plus hautes, parfois les plus basses et celles qui intéressent plus particulièrement l'usager.

Code	Thème	Qté	Description	Exemple
BAS1	Domination	8	La colonne la plus basse, barre la plus courte ou secteur de tarte le plus petit.	Le secteur des aurifères a affiché le pire rendement de l'indice TSE 300, ayant perdu 26,3% depuis le début de l'année.
BAS2	Domination	3	Le groupe sériel le plus bas	Les travailleurs à temps partiel sont proportionnellement moins nombreux à bénéficier des avantages sociaux
GS1	Quantitatif	7	Comparaison entre 2 groupes sériels.	Les femmes sont deux fois plus susceptibles que les hommes de lire des livres durant leurs heures de loisirs
HAUT1	Domination	49	Colonne la plus haute, barre la plus longue ou secteur de tarte le plus gros.	L'Université de la Colombie-Britannique a reçu la proportion la plus importante, soit 40,5 millions de $ ou 9,6% du total du financement accordé aux universités canadiennes.
HAUT2	Domination	7	Deux (ou +) colonnes les plus hautes, barres les plus longues ou secteurs de tartes les plus gros.	Les adultes de la Colombie-Britannique, de la Nouvelle-Ecosse et de l'Alberta sont les plus nombreux à faire des déplacements d'intérêt faunique.
HAUT4	Domination	4	Groupe sériel le plus élevé	Les Autochtones sont proportionnellement plus nombreux à vivre avec des parents
SPEC1	Quantitatif	18	Valeur d'une colonne, barre ou secteur qui nous intéresse plus particulièrement.	Le revenu annuel moyen d'une famille à Vancouver était de 59700$ en 1993
EQUI1	Déductif	3	Équilibre entre les différentes données.	L'opinion des hommes quant à l'importance d'occuper un emploi rémunéré pour être heureux est constante d'un groupe d'âge et d'un niveau de scolarité à l'autre

TAB. 4.1 - Comparaison: Codes choisis pour l'implantation

Pour bien saisir le genre de message pertinent à sélectionner, nous reprenons l'exemple de la fin du chapitre 1, soit le nombre de ménages possédant un ordinateur à domicile, pour chaque province du Canada. La colonne A du tableau 4.2 représente les données réelles de 1996 tandis que les autres colonnes représentent des données fictives pour alimenter la discussion qui suit.

	A	B	C	D
Terre-Neuve	21.5	21.5	25.5	25.5
Île-du-Prince-Édouard	22.4	22.4	22.4	22.4
Nouvelle-Écosse	26.1	26.1	26.1	26.1
Nouveau-Brunswick	21.4	21.4	17.0	27.0

Québec	24.0	24.0	24.0	24.0
Ontario	35.8	35.8	25.8	25.8
Manitoba	25.7	25.7	25.7	25.7
Saskatchewan	28.3	28.3	28.3	28.3
Alberta	37.9	48.2	27.9	27.9
Colombie-Britannique	37.7	37.7	27.7	27.7

TAB. 4.2 - Exemples de données pour la comparaison

Que peut-on dire d'intéressant sur les données de la première colonne? On constate que les 3 données les plus hautes se distinguent des autres, l'écart le plus considérable se situant entre la troisième province (l'Ontario) et la quatrième province (la Saskatchewan).

Nous avons vu au chapitre 3 que le code HAUT1 est le plus fréquemment utilisé pour une intention de comparaison, c'est à dire la mention de la donnée la plus élevée. Si on utilise ce code pour les données du tableau 4.2, on pourrait générer *"C'est l'Alberta qui a le taux le plus élevé de ménages possédant un ordinateur"*. Ce n'est pas inexact, mais il est inapproprié de mettre seulement l'Alberta en évidence puisque 2 autres provinces suivent de très près l'Alberta alors que les 7 autres viennent loin derrière. Dans ce cas, il est plus convenable d'utiliser le code HAUT2 et SelTex génère: *"L'Alberta, la Colombie-Britannique et l'Ontario ont un taux plus élevé de ménages possédant un ordinateur."*

Observons maintenant les données de la colonne B: tous les taux sont identiques à ceux de la colonne A sauf un, celui de l'Alberta qui est de 48.2%. Dans ce cas, est-il plus pertinent de choisir le code HAUT1 ou HAUT2? La réponse nous paraît claire: l'information la plus représentative est que l'Alberta se démarque des autres provinces et nous utiliserons alors le code HAUT1.

Données qui intéressent plus particulièrement l'usager (code SPEC1).

Supposons que les données statistiques du tableau 4.2 (les données réelles de la colonne A) soient publiées dans un quotidien montréalais. Que voudra savoir le lecteur? Il sera intéressé à connaître où se situe le Québec par rapport aux autres provinces.

Puisque le système ne peut deviner les données qui intéressent plus particulièrement l'usager, nous avons modifé quelque peu le format du fichier d'entrée de Postgraphe afin que l'usager puisse les spécifier, et SelTex en tient compte lors de la sélection du texte à générer.

Ainsi, lorsque l'usager demande une comparaison entre les provinces et spécifie qu'il s'intéresse particulièrement au Québec, on obtient:

L'Alberta, la Colombie-Britannique et l'Ontario ont un taux le plus élevé de ménages possédant un ordinateur, tandis que le Québec se situe au septième rang avec 24.0%.

Donnée la plus basse

L'étude de notre corpus démontre que la mention de la ou les données les plus basses sont environ 5 fois moins fréquentes que la mention de données les plus élevées. C'est dire qu'on s'intéresse beaucoup plus aux gagnants qu'aux perdants! Nous remarquons que la donnée la plus basse sera mise en évidence surtout lorsqu'elle est anormalement basse, ou bien lorsqu'il n'y a pas de données élevées qui se démarquent. Dans notre exemple, c'est le cas de la colonne C: aucune province n'a un taux significativement plus élevée tandis que le Nouveau-Brunswick est suffisamment loin derrière les autres pour que SelTex génère:

Avec seulement 17%, le Nouveau-Brunswick a le taux le moins élevé de ménages possédant un ordinateur.

Finalement, il arrive également que l'écart entre toutes les données soit faible, y compris entre la donnée la plus haute et la plus basse, comme c'est le cas à la colonne D de notre exemple. Dans ce cas, SelTex mentionne simplement qu'il y a équilibre:

Le taux de ménages possédant un ordinateur est relativement constant d'une province à l'autre.

Suite à nos observations, nous avons donc établi les règles suivantes, de façon empirique, pour obtenir les messages à générer :

- Mentionner la donnée la plus élevée si sa valeur est au moins 10% plus grande que la seconde et que le nombre de données > 2 (code HAUT1).
- Sinon, mentionner les 2 données les plus élevées si la 2ème valeur est au moins 15% plus grande que la troisième et nombre de données > 4 (code HAUT2).
- Sinon, mentionner les 3 données les plus élevées si la 3ème valeur est au moins 20% plus grande que la quatrième et nombre de données > 6 (code HAUT2).

On s'arrête à 3, c'est le maximum observé dans notre corpus.

- Mentionner la donnée la plus BASSE si sa valeur est au moins 30% plus basse que la seconde plus basse et que le nombre de données > 2 (code BAS1).
- Mentionner l'équilibre entre les données si le ratio entre la valeur la plus élevée et la valeur la plus basse est inférieure à 1.2 et que le nombre de données > 3. (code EQUI1)

SelTex s'assure également d'exclure les données telles que "Autres", "divers," "total" que l'on retrouve souvent dans les graphiques statistiques de notre corpus. Il serait en effet erroné de les comparer avec les données individuelles.

On aura également une option de défaut pour la comparaison: si aucun type de message n'est satisfait par les contraintes, on donnera quand même la donnée la plus élevée.

Mentionnons que ces seuils en pourcentage ne constituent pas "une règle de vérité": nous les avons déterminés de façon approximative en nous basant sur les exemples de notre corpus. De plus, les seuils ne sont pas invariants par translation. Nous suggérons l'implantation d'une notion d'échelle dans une future version de SelTex.

4.2 Évolution (sans groupe sériel)

Le tableau 4.3 indique les codes de classification sélectionnés pour l'implantation de l'évolution.

Code	Thème	Qté	Description	Exemple
DEBFIN	Quantitatif	6	Comparaison du premier et du dernier Y pour un X temporel	En 1992-1993, 22% des étudiants inscrits à temps plein au baccalauréat poursuivaient des études en sciences naturelles et en génie, comparativement à 24,9% en 1984-1985.
FIN1	Quantitatif	10	La derniere valeur de Y pour X temporel.	Il y a eu 108 affaires d'agression sexuelle déclarées pour 100000 habitants en 1994.
TEND1	Descriptif	21	Tendance générale: hausse, baisse ou stabilité.	Hausse du pourcentage des aînés qui touchent des prestations de retraite du RPC
TEND2	Quantitatif	11	Tendance en mentionnant les valeurs de début et de fin et/ou le ratio entre ces 2 valeurs.	Le financement provenant des universités a également connu une baisse, passant de 30% en 1986 à 24% en 1993.
TEND3	Discriminant	13	Tendance dans un sous-intervalle de l'axe temporel.	Les revenus d'emploi des jeunes hommes sont en baisse depuis la fin des années 70

TAB. 4.3 - Évolution: Codes choisis pour l'implantation

Le tableau 4.4 indique l'évolution de différents indicateurs sociaux provenant du magazine Tendances Sociales Canadiennes, été 1997. Notre code de classification TEND1 sert à mettre en évidence la tendance générale. De façon évidente, on pourra parler de hausse ou de croissance lorsque chaque donnée est supérieure à la précédente ou

d'une baisse lorsque chaque donnée est inférieure à la précédente. La première ligne du tableau 4.4 donne le nombre de doctorats décernés au Canada de 1989 à 1995. Puisque la quantité augmente d'année en année, SelTex utilise le code TEND2 et génère le texte suivant:

Le nombre de doctorats décernés au Canada augmente constamment depuis 1989, passant de 2573 à 3621 en 1995.

	1989	1990	1991	1992	1993	1994	1995
Nombre de doctorats décernés	2573	2673	2947	3136	3356	3552	3621
Taux de natalité	15,0	15,3	14,3	14,0	13,4	13,2	12,9
Taux de chômage	7,5	8,1	10,4	11,3	11,2	10,4	9,5
Taux d'inflation	5,0	4,8	5,6	1,5	1,8	0,2	2,1

TAB. 4.4 - Exemples de données pour l'évolution

SelTex décrit les tendances générales sans les **qualifier**: comme nous l'avons fait remarquer à la section 3.3, il est possible de qualifier les tendances seulement si le système est restreint à un domaine très précis.

S'il y a un renversement de tendance, le lecteur s'intéressera à la tendance la plus récente, ce qui correspond au dernier segment de la courbe sur le graphique correspondant. Selon le tableau 4.4 on constate que le taux de chômage au Canada a d'abord augmenté de 1989 à 1992 et diminué de 1992 à 1995. Dans ce cas, SelTex utilise le code TEND3 pour générer:

Le taux de chômage au Canada est en baisse depuis 1992.

Si on note plusieurs renversements de tendance, ou une évolution en dents de scie, SelTex utilise le code DEBFIN: on donne les X et Y de début et de fin. Ce code constitue une solution pratique pour un générateur de texte automatique car il permet de présenter l'évolution de façon pertinente sans essayer de qualifier l'évolution. Selon le

tableau 4.4, les données sur le taux d'inflation n'ont pas suivi une tendance constante et SelTex génère:

Le taux d'inflation au Canada se situe à 2.1% en 1995 comparativement à 5.0% en 1989.

Lors de très grandes fluctuations, il devient inopportun de mentionner la valeur de début: par exemple si le taux d'inflation était à 12% en 1990, il devient absurde de comparer 1989 à 1995. Dans ce cas, SelTex utiliserait plutôt le code FIN1 pour ne donner que la valeur de fin, ce qui donne:

Le taux d'inflation au Canada se situe à 2.1% en 1995.

Les codes FIN1 et TEND3 génèrent un texte qui donne au lecteur "les dernières nouvelles": le texte décrit la situation actuelle alors que le graphique montre l'ensemble de l'évolution.

4.3 Combinaison de comparaison et d'évolution

Au chapitre précédent, notre analyse de corpus a révélé que la combinaison de 2 intentions dans une même phrase que l'on retrouve le plus souvent est la comparaison et l'évolution. Nous avons sélectionné les codes du tableau 4.5 pour implanter cette combinaison.

Code	Thème	Qté	Description	Exemple
ECART	Descriptif	5	Évolution de l'écart entre les courbes	L'écart se creuse entre les habitudes de voyage des jeunes adultes et celles des adultes plus âgés
ENUM1	Descriptif	7	Énumérer les évolutions pour chaque groupe sériel	La population a augmenté dans la RMR de St. John's, mais elle a diminué dans le reste de Terre-Neuve

FIN2	Quantitatif	5	Comparaison des dernières valeurs en Y pour X temporel (groupes sériels)	Les immigrants allophones récents établis à Montréal étaient plus nombreux à parler le français que l'anglais à la maison en 1991.
FIN3	Domination	4	Groupe sériel dont la dernière valeur en Y est la plus élevée	Les immigrants allophones récents établis au Québec sont proportionnellement plus nombreux que ceux des autres provinces à avoir fait un transfert linguistique.

TAB. 4.5 - Combinaison comparaison/évolution: codes choisis pour l'implantation

Il est pertinent d'utiliser ces codes dans un des cas suivants:

- l'usager demande une intention de comparaison et il y a une variable temporelle.
- l'usager demande une intention d'évolution et il y a des groupes sériels.

Reprenons notre exemple favori en nous transportant jusqu'en l'an 2012. Le tableau 4.6 reprend les données réelles de 1996 du pourcentage de ménages possédant un ordinateur montre avec une évolution fictive par province sur une période de 16 ans. Lorsqu'on veut plutôt insister sur la comparaison, le graphique donnera une vue d'ensemble, alors que le texte insiste généralement sur la situation la plus récente (codes FIN2 et FIN3).

	1996	2000	2004	2008	2012
Terre-Neuve	21.5	30.1	57.4	71.8	97.1
Île-du-Prince-Édouard	22.4	27.4	52.2	64.0	91.2
Nouvelle-Écosse	26.1	35.6	55.1	75.1	95.4
Nouveau-Brunswick	21.4	27.5	57.2	74.3	92.6
Québec	24.0	37.1	58.4	77.7	96.4
Ontario	35.8	46.1	63.2	80.1	98.0
Manitoba	25.7	53.2	65.1	73.3	95.0
Saskatchewan	28.3	62.1	63.0	70.6	94.3
Alberta	37.9	50.1	64.2	72.1	95.1
Colombie-Britannique	37.7	45.1	66.8	76.7	93.8

TAB. 4.6 - Exemples de données pour la combinaison d'intentions

Le code FIN3 est le plus "facile" à utiliser pour SelTex car il peut être utilisé en tout temps. Le choix du message à transmettre est sélectionné de la même façon qu'à la section précédente (simple comparaison), sauf que la comparaison est effectuée sur les

données à droite de l'axe temporel, soit l'année 2012 (c'est encore l'idée de mettre en évidence les dernières nouvelles).

En 2012, le pourcentage de ménages possédant un ordinateur à domicile est constant d'une province à l'autre, la moyenne se situant à 94.9%.

Lorsqu'il y a seulement 2 groupes sériels (par exemple hommes/femmes ou francophones/anglophones), SelTex utilise le code FIN2 pour faire une comparaison explicite entre les 2 groupes, toujours selon les données les plus récentes, à droite de l'axe temporel.

Évolution de la comparaison

SelTex peut utiliser le code ECART lorsqu'il y a seulement 2 groupes sériels OU que l'usager s'intéresse à DEUX données spécifiques. Ce code est utilisé lorsque l'évolution dénote un rapprochement ou un éloignement (évolution de la comparaison). Situons-nous maintenant en l'an 2008 avec les données du tableau 4.6; l'usager demande une évolution et spécifie qu'il s'intéresse plus particulièrement au Québec et à l'Ontario. On remarque que l'écart entre ces 2 provinces est de 11.8% en 1996 et diminue constamment par la suite. SelTex génère: *Le pourcentage de ménages possédant un ordinateur au Québec se rapproche de celui de l'Ontario.*

Comparaison de l'évolution

Le code ENUM1 est utilisé par SelTex pour énumérer les tendances de chaque groupe sériel, ce qui donne une comparaison de l'évolution. Si les données du tableau 4.6 sont soumises à SelTex sans spécification particulière, ce code ne sera pas utilisé: il serait en effet absurde de commenter l'évolution de chacune des 10 provinces. SelTex peut faire une comparaison de l'évolution pour un maximum de 4 groupes sériels OU pour des groupes sériels d'un intérêt particulier pour l'usager. Si on se situe en l'an 2008

et que l'usager spécifie un intérêt particulier pour Terre-neuve et l'Île-du-Prince-Edouard, SelTex génère:

Le taux de ménages possédant un ordinateur à domicile augmente plus rapidement à Terre-Neuve qu'à l'Île-du-Prince-Edouard.

4.4 Corrélation

Le tableau 4.7 indique les codes de classification sélectionnés pour l'implantation de la corrélation.

Code	Thème	Qté	Description	Exemple
CORR1	Déductif	7	Il y a correlation	Le prix de la majorité des automobiles varie en fonction du poids.
CORR2	Déductif	6	Il y a non-correlation	L'importance des augmentations au mérite ne dépend pas de l'ancienneté
DIREC	Déductif	12	Lien direct entre 2 variables quantitatives	La pratique sportive augmente avec le revenu de ménage
INVER	Déductif	5	Lien inverse entre 2 variables quantitatives	L'état de santé se détériore avec l'âge.

TAB. 4.7 - Corrélation: codes choisis pour l'implantation

Lorsque l'usager en spécifie l'intention, SelTex peut vérifier la corrélation entre deux variables quantitatives, qu'on nomme ici X et Y.

Pour déterminer s'il y a corrélation significative, on utilise le coefficient de corrélation linéaire déterminé par la formule de la figure 4.1 tirée de [BAI84].

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

où n représente le nombre de couples (x_i, y_i).

FIG. 4.1 - Calcul du coefficient de corrélation linéaire

Le coefficient *r,* qui se situe entre -1 et 1, mesure l'intensité de la liaison linéaire entre deux variables observées : plus il s'éloigne de 0, plus la corrélation est forte. La loi de Student détermine si la corrélation est significative selon 3 paramètres:

- le coefficient de corrélation
- le nombre de couples (X,Y)
- la marge d'erreur accordée

Par exemple, pour 10 couples de données (x,y) et une marge d'erreur de 5%, le coefficient (en valeur absolue) doit être supérieur à 0,549 ; pour 50 couples (x,y) il doit être supérieur à 0.235. La table des valeurs critiques utilisées par SelTex est présentée à l'**annexe C**.

Si la corrélation n'est pas significative, SelTex génère la phrase:

"Il n'y a pas de relation entre <description de X> *et* <description de Y>*"* ou encore

"<description de X> *ne dépend pas de* <description de Y>".

Si la corrélation est significative, elle est directe si le coefficient est positif et inverse si le coefficient est négatif. SelTex génère alors une phrase avec une expression telle "*augmente avec*", "*diminue avec*" ou "*sont plus susceptibles de…*" selon le cas.

Dans quels cas mentionnera-t-on simplement l'existence d'une corrélation sans spécifier si elle est directe ou inverse avec une expression comme "varie selon", "dépend de" ou "est lié" ? Pour certains exemples tirés du corpus, il s'agit simplement d'un choix stylistique: dans *"L'alphabétisme est étroitement lié au niveau de scolarité en 1994 (#43)"*, l'expression *"augmente avec"* aurait été tout aussi convenable.

Mais cet autre exemple est intéressant:

"L'opinion des femmes quant à l'importance d'occuper un emploi rémunéré pour être heureuses varie selon l'âge. (#44)"

Ici, il y aurait ambiguïté sémantique si on écrivait *"augmente avec l'âge"*: on risquerait de comprendre que les femmes changent d'opinion lorsqu'elles vieillissent alors que le message veut plutôt indiquer les différences entre les générations.

Puisque c'est la seule exception que nous avons observée dans notre corpus, SelTex génère un message de code CORR1 seulement lorsque la donnée X ou Y est de type âge.

4.5 Distribution

Le code INTER est le seul code de l'intention de distribution qui a plusieurs occurrences dans notre corpus et c'est également le seul qui est implanté dans SelTex (tableau 4.8).

En fait c'est l'équivalent du code HAUT1 pour la comparaison. La plupart des distributions du corpus sont pour des catégories d'âge ou de salaire. Ainsi, lorsqu'une des données est constituée d'intervalles et que l'usager demande une répartition, SelTex détermine l'intervalle dominant.

Code	Thème	Qté	Description	Exemple
INTER	Domination	4	Indiquer un intervalle de données en X où les données en Y sont plus élevées	La plupart de nos ventes aux Etats-Unis sont comprises entre 30 et 50 dollars.

TAB. 4.8 - Distribution: code choisi pour l'implantation

4.6 Présentation

Le tableau 4.9 indique les codes de classification sélectionnés pour l'implantation de la présentation. SelTex crée les titres génériques à partir de la description des données fournies par l'usager dans le fichier d'entrée. Lorsqu'il y a une donnée temporelle, la valeur de début et de fin est mentionnée, comme par exemple: "*Taux d'inflation au Canada de 1989 à 1995*".

Le code TITRE3 ne sera utilisé par SelTex que si l'usager spécifie dans le fichier d'entrée que les valeurs d'une certaine donnée ne constituent pas une énumération complète.

Code	Thème	Qté	Description	Exemple
INT	Descriptif	11	Présentation de l'intention	Le graphique ci-dessous illustre la corrélation entre le marché canadien et certains marchés étrangers.
TITRE1	Descriptif	23	Titre générique mentionnant la description d'une seule donnée	Raisons pour lesquelles les femmes agressées n'ont pas signalé les actes de violence à la police, 1993
TITRE2	Descriptif	61	Titre générique incluant la description de 2 données ou plus	Dépenses des subventions de R et D coopérative, de 1984-1985 à 1993-1994
TITRE3	Descriptif	4	Présentation des données: mention du nombre de valeurs nominales.	Les 10 principaux pays de naissance des nouveaux immigrants, Québec et reste du Canada, 1991

TAB. 4.9 - Présentation: codes choisis pour l'implantation

4.7 Couverture de SelTex

En résumé, nous avons implanté des schémas de texte pour 26 des 55 codes de classification de notre corpus (soit 47% des codes). Ces codes couvrent 71% des extraits de texte de notre corpus: nous avons surtout sélectionné les codes les plus fréquents mais aussi ceux pour lesquelles les règles de décision sont les plus claires.

Dans ce chapitre, nous avons principalement expliqué ce que fait SelTex sans entrer dans les détails techniques de son fonctionnement. Dans le chapitre suivant, nous décrivons la façon dont chacune des étapes du processus de génération a été implantée ainsi qu'une évaluation des résultats obtenus.

Chapitre 5 - Implantation de SelTex

SelTex est implanté entièrement en Prolog, tout comme Postgraphe. Nous avons examiné différents systèmes Prolog et notre choix s'est arrêté sur SICStus Prolog [SIC97], pour différentes raisons techniques, dont sa portabilité sur Unix et Windows.

Prolog est choisi régulièrement depuis plusieurs années dans le domaine du traitement de la langue naturelle et de la génération de texte. Outre la documentation de SICStus [SIC97], nos principales sources de référence ont été [GAL91] pour les nombreux exemples d'utilisation du Prolog dans le traitement de la langue naturelle, [DER96] pour la définition des standards du langage ainsi que [STE94] pour les techniques de programmation avancées.

La première section de ce chapitre décrit les principales modifications apportées à Postgraphe. La section 5.2 décrit le coeur de SelTex en donnant un exemple de schéma textuel, et les fonctions associées pour chaque étape de l'exécution. À la section 5.3, nous évaluons les résultats obtenus.

5.1 - Modifications à Postgraphe

Fichier d'entrée

Dans notre introduction à Postgraphe au chapitre 2, nous avons donné un exemple du fichier d'entrée (figure 2.5). Nous avons fait certains ajouts au format du fichier d'entrée, qui sont montrés à la figure 5.1.

Après l'énumération des variables, une ligne additionnelle peut donner une description plus explicite des variables qui seront utilisées dans le texte. La figure 5.1 montre que la description plus explicite de la variable "ordinateur" est "ménages ayant un ordinateur". La variable "province" conserve la même description.

Après l'énumération des types des variables, une ligne additionnelle donne une description du type des variables pour les cas où on veut que le texte exprime le type de données d'une façon différente que celle définie dans le système de types. Dans notre exemple de la figure 5.1, on préfère parler de "taux" plutôt que de "pourcentage".

A la ligne suivante, on énumère les données qui intéressent plus particulièrement l'usager ("Québec" dans l'exemple). S'il y en a aucune, les crochets vides [] sont requis.

```
data(
        % 1-noms des variables
     [province,ordinateur],
        % 2-description des variables
     [province,'ménages ayant un ordinateur'],
        % 3-types des variables
     [province,pourcentage],
        % 4-description des types des variables
     [province,taux],
        % 5-donnees qui interessent plus particulierement l'usager.
     ['Québec'],
        % 6-variables pouvant faire partie des clés relationnelles
     [mois,province],
       %  7-variables  pouvant  ne  pas  faire  partie  des  clés
relationnelles
     [ordinateur],
        % 8-intentions du rédacteur
     [[comparaison([ordinateur],[province])]],
        % 9-les données brutes sous forme de tuples
     [['Terre-Neuve',21.5],
      ['Île-du-Prince-Édouard',22.4],
      ['Nouvelle-Écosse',26.1],
      ['Nouveau-Brunswick',21.4],
      ['Québec',24.0],
      ['Ontario',35.8],
      ['Manitoba',25.7],
      ['Saskatchewan',28.3],
      ['Alberta',37.9],
      ['Colombie-Britannique',37.7]]).
```

FIG. 5.1 - Exemple de fichier d'entrée pour Postgraphe après modifications

Fonctions linguistiques

Comme nous l'avons mentionné au chapitre 2, la version originale de Postgraphe utilisait une interface avec le générateur de Prétexte. Nous avons retiré cette interface et l'avons remplacé par une adaptation en Prolog de FRANA. Nous avons cependant récupéré quelques fonctions linguistiques provenant de Prétexte, par exemple la fonction d'élision : dans FRANA, les élisions sont codées directement dans des phrases prédéfinies, ce qui ne convenait pas aux besoins de SelTex.

Nous avons également développé quelques fonctions linguistiques comme l'ajout d'un article défini ou indéfini et la correspondance entre les adjectifs numéraux cardinaux et ordinaux.

Corrélation

Nous avons implanté le calcul du coefficient de corrélation linéaire correspondant à la figure 4.1 qui nécessitait des fonctions de covariance et de somme de carrés. Une autre fonction détermine si le coefficient est significatif selon le nombre de couples. Le code est présenté à l'annexe C.

Adaptation à Sicstus

La première version de Postgraphe fonctionnait avec SWI-Prolog. Puisque nous avons opté pour Sicstus, il y a eu quelques adaptations à faire pour assurer la compatibilité, par exemple il a fallu redéfénir les fonctions **`gensym`**, **`concat`** et

reverse. Également, le code de Sylvie Giroux sur l'adaptation de FRANA a nécessité quelques ajustements.

5.2 - Les schémas textuels

SelTex comporte 15 nouveaux schémas textuels qui sont associés aux codes de classification du tableau 3.6.

Nous allons suivre les étapes pour l'exemple du taux de ménages possédant un ordinateur à domicile par province, pour le fichier d'entrée de la figure 5.1. La sortie du système donne le graphique en barres de la figure 1.1 avec le texte suivant:

L'Alberta, la Colombie-Britannique et l'Ontario ont un taux plus élevé de ménages ayant un ordinateur, alors que le Québec se situe au septième rang avec 24%.

L'usager a spécifié l'intention suivante dans le fichier d'entrée:

```
[comparaison([modem],[province])]]
```

Avant d'entrer dans le module SelTex, Postgraphe vérifie si l'intention peut être satisfaite avec un schéma graphique: elle l'est effectivement, par un graphique en barres. Le programme passe ensuite à l'étape texte. La phase de groupement d'intentions n'a aucun travail à effectuer puisqu'il y a une seule intention à traiter. Le système regarde si SelTex possède un ou plusieurs schémas textuels qui peuvent satisfaire l'intention désirée; dans notre exemple, l'intention est satisfaite par le schéma **position1 :**

```
c_txt(comparaison([Y],[X]),position1,comp_txt,[X,Y],100,[]).
```

Cette entrée indique qu'il y a un schéma nommé **position1** appartenant à la classe **comp_txt** qui peut satisfaire une intention de comparaison de la variable Y par rapport à X pour des tuples à 2 variables [X,Y] et que le seuil de qualité est de 100.

Une fois le schéma obtenu, la première étape pour SelTex est de trouver le prédicat d'exécution **texte** correspondant au schéma textuel **position1** (figure 5.2). Cette fonction s'assure d'abord que X est une clé de Y (si ce n'est pas le cas, il y a d'autres schémas qui pourront convenir). Dans notre exemple, la variable "ordinateur" est une clé de la variable "province". On génère les tuples, et on trie les **données** (**prepare_texte**) avant la prochaine étape de l'exécution du schéma textuel.

```
texte(position1,comp_txt,[X,Y],Proc) :-
      cle([X],L),member(Y,L),
      gentuples([X,Y],Tup),
      prepare_texte(comp_txt(X,Y,Tup),Res),
      texte_procedure(position1(X,Y,Res),Proc).
```

FIG. 5.2 - Fonction texte pour le schéma textuel **position1**

La prochaine étape pour SelTex est de trouver le prédicat d'exécution **texte_réalise** (figure 5.3) correspondant au schéma textuel **position1**. Le prédicat se charge de trouver le message à générer. Celui-ci peut-être associé à 5 différents codes de classification, selon les contraintes: HAUT1, HAUT2, BAS1, EQUI1 et SPEC1.

La première section représente le cas où on énumère les 3 variables les plus élevées: selon les conditions décrites à la section 4.1: la valeur la plus élevée doit être moins de 10% supérieure à la seconde (Val1/Val2 < 1.1), la seconde valeur la plus élevée doit être moins de 10% supérieure à la troisième (Val2/Val3 < 1.1), mais la troisième valeur la plus élevée doit être plus de 10% supérieure à la quatrième (Val3/Val4 > 1.1). L'instruction "Len > 2" s'assure que le nombre total tuples est au moins de 7 (car Len représente le représente le reste de la liste à partir de la 5ème valeur).

```
/* HAUT2 */
/* 3 premiers */
texte_realise(position1(_X,Y,Liste)) :-
        Liste=[[Top1,Val1],[Top2,Val2],[Top3,Val3],[_Top4,Val4]|Reste],
        length(Reste,Len),
        Len > 2,
        Val1/Val2 < 1.1,
        Val2/Val3 < 1.1,
        Val3/Val4 > 1.1,
        var_desc(Y,DescY),
        unite_desc(Y,UnitY),
        assert(message(c1_comp, c2_3prem, nil, s1_3prem, _,
                                  [top1:Top1,   top2:Top2,   top3:Top3,
            unity:UnitY, descy:DescY])),
        verifie_specifique(3,Y,Liste).
/* SPEC1 */
verifie_specifique(N,Y,Liste):-
        dataget([interessants],Specs),
        Specs=[Spec|_ResteSpecs],
        pos([Spec,ValSpec],Liste,Pos),
        RangSpec is Pos + 1,
        RangSpec > N,
        var_desc(Y,DescY),
        assert(message(c1_comp, c2_spec, nil, s1_spec, _,
            [spec:Spec,rangspec:RangSpec,valspec:ValSpec,descy:DescY]))
            .
verifie_specifique(_,_,_).
```

FIG. 5.3 - Fonction **`texte_réalise`** pour le schéma textuel **`position1`**

Si toutes les conditions sont respectées, un message est généré (appel à assert(message)) avec les entrées syntagmatiques correspondantes et les variables requises pour le texte qui sont pour notre exemple:

Top1: “Alberta”

Top2: “Ontario”

Top3: “Colombie-Britannique”

UnitY: “taux”

DescY: “ménages ayant un ordinateur”

Il est possible qu’un second message soit généré si l’usager a indiqué des données d’un intérêt particulier, comme c’est le cas dans notre exemple: l’usager a mentionné le

Québec. La fonction **verifie_specifique** (figure 5.4) est appelée et celle-ci générera un message avec comme paramètre le rang du Québec, à condition que celui-ci n'ait pas déjà été mentionné dans les valeurs les plus élevées. Pour notre exemple, il s'agit du 7ème rang.

```
verifie_specifique(N,Y,Liste):-
        dataget([interessants],Specs),
        Specs=[Spec|_ResteSpecs],
        pos([Spec,ValSpec],Liste,Pos),
        RangSpec is Pos + 1,
        RangSpec > N,
        var_desc(Y,DescY),
        assert(message(c1_comp, c2_spec, nil, s1_spec, _,

[spec:Spec,rangspec:RangSpec,valspec:ValSpec,descy:DescY])).
verifie_specifique(_,_,_).
```

FIG. 5.4 - Fonction vérifiant l'intérêt particulier de l'usager

La dernière étape est de générer le texte comme tel, c'est alors que SelTex appelle l'adaptation Prolog du module FRANA.

Ampleur du code

La version originale de Postgraphe comportait environ 4700 lignes de code et le générateur de texte PréTexte environ 10000 lignes de code ce qui donnait un total de 14700 lignes pour le système intégré.

Dans notre nouvelle version, Postgraphe comporte environ 4300 lignes de code (combinaison d'ajouts et de retrait de fonctions non utilisées), l'adaptation de FRANA a environ 3200 lignes et SelTex compte environ 1100 lignes, ce qui donne un total de 8600 lignes pour tout le système.

5.3 - Évaluation des résultats

Nous avons présenté au chapitre 4 une douzaine d'exemples de phrases produites par notre système SelTex. L'implantation tardive dans le cadre du projet ne nous permet que de courts commentaires pour évaluer les résultats.

Les phrases générées sont dans un français simple et acceptable. Plusieurs phrases générées par SelTex sont de qualité égale à celles du corpus. Toutefois, le vocabulaire utilisé est plus restreint dans le texte généré et on ne retrouve pas une variété de formes syntaxiques aussi intéressante que dans notre corpus. Dans certains cas, SelTex parvient à transmettre le même message que dans la phrase du corpus, mais le texte généré paraissent un peu moins intelligent. Voici un exemple :

Phrase du corpus

Les jeunes adultes étaient plus susceptibles d'avoir l'impression d'être trop qualifiés pour leur emploi que les adultes plus âgés, 1994.

Phrase générée par SelTex

L'impression d'être trop qualifié pour son emploi diminue avec l'âge.

Exemple hors corpus

L'exemple que nous avons présenté en introduction (figure 1.1) peut être comparé au rapport réel de Statistique Canada (qui ne fait pas partie du corpus). Ce dernier contient des informations additionnelles qui n'apparaissent pas sur le graphique associé.

Texte généré par SelTex :

L'Alberta, la Colombie-Britannique et l'Ontario ont un taux plus élevé de ménages ayant un ordinateur alors que le Québec se situe au septième rang avec 24.0%.

Texte tiré de Statistique Canada (hors corpus) :

Le nombre de ménages ayant un ordinateur s'établit à 3,6 millions, ou 31,6%, ce qui représente une hausse par rapport à 28,8 % en 1995 et à 10,3% il y a dix ans. Plus du tiers des ménages en Ontario, en Alberta et en Colombie-Britannique ont un ordinateur.

Comparaison avec la version originale de Postgraphe

SelTex se compare favorablement à la version originale de Postgraphe. Comme les exemples de la section 4.2 le démontrent, SelTex identifie une statistique pertinente pour une intention d'évolution alors que la version originale de Postgraphe énumère tous les segments d'une courbe (figure 2.5). De plus, SelTex peut générer un message associé aux 5 intentions de base, alors que la version originale de Postgraphe générait du texte uniquement pour l'évolution.

Chapitre 6 - Conclusion

6.1 Résumé

Notre projet s'est intéressé à la génération intégrée de texte et de graphique, une discipline de l'Intelligence Artificielle. Le système Postgraphe, implanté dans le cadre de la thèse de doctorat de Massimo Fasciano [FAS96] constituait le point de départ de notre projet. La partie graphique de Postgraphe donnait des résultats satisfaisants mais la partie texte était peu approfondie avec un seul schéma textuel. L'interface choisie, le système Prétexte, ne s'intégrait pas bien au système et l'information contenue dans le texte n'était pas très pertinente.

L'objectif principal de notre projet consistait donc à déterminer comment sélectionner l'information (QUOI DIRE?) pour un texte qui accompagne un graphique statistique, afin de pouvoir construire des schémas de texte pertinents.

Pour atteindre cet objectif, nous avons réalisé une étude de corpus (chapitre 3) afin de déterminer comment choisir les informations pertinentes et intéressantes à présenter dans un rapport statistique. Nous avons recueilli 411 extraits de texte associés à des graphiques statistiques auxquels nous avons attribué des codes de classification qui indiquent la façon dont l'information est exprimée. Nous avons analysé les codes de classification les plus fréquents pour chaque intention et nous avons distingué sept thèmes principaux dans la nature des messages de nos extraits de texte: les messages descriptifs, quantitatifs, de domination, déductifs, discriminants, qualitatifs et justificatifs. Nous avons également éclairci certains aspects généraux sur l'intégration texte et graphique au niveau de la cohésion, la complétion et la redondance.

En nous basant sur les observations de notre corpus, nous avons développé des techniques de sélection pour générer des extraits de texte associés aux codes de classification les plus fréquents qui couvrent 71% des extraits de notre corpus. Pour implanter ces techniques, nous avons développé SelTex un module PROLOG que nous avons intégré à Postgraphe. La technique d'implantation utilisée est basée sur le modèle utilisé dans FRANA.

6.2 Travaux futurs

Nous croyons que notre étude de corpus réalisée ainsi que l'ajout de SelTex à Postgraphe constituent un pas important pour la génération intégrée de graphiques et de texte à partir de données statistiques. Nous décrivons ici quelques lacunes de notre projet ainsi que quelques suggestions pour des travaux futurs.

Lacunes de notre projet

Notre étude de corpus étant concentrée exclusivement aux légendes qui accompagnent les graphiques, nous avons délaissé le texte continu. Il serait important d'examiner les similitudes et différences dans la nature des messages d'un texte continu. Nous croyons que nos codes de classification ne couvrent pas l'ensemble des messages qu'on peut retrouver dans un texte continu.

Notre projet se concentrait exclusivement sur la partie QUOI DIRE de la génération de texte car nous voulions principalement établir des techniques de sélection de l'information. Cette démarche est valable sur le plan théorique mais au moment de l'implantation, il faut choisir des méthodes pour organiser le texte. En se concentrant exclusivement sur les légendes textuelles, nous avons "passé sous le tapis" les étapes importantes du COMMENT LE DIRE. Plusieurs travaux antérieurs ont déjà développé

des techniques intéressantes pour l'organisation du texte que nous n'avons pas mis à profit dans le cadre de notre implantation.

COMMENT LE DIRE

Lorsque nous affirmons avoir implanté des schémas de texte qui couvrent 71% des extraits de notre corpus, il faut réaliser qu'il s'agit du **type de message** uniquement. Dans les textes générés par SelTex, on ne retrouve pas une variété intéressante de formes syntaxiques et de façons différentes d'exprimer un même type de message, comme c'est le cas dans notre corpus. Par exemple, les 49 extraits de texte du corpus auxquels on a assigné le code HAUT1 sont presque tous exprimés différemment (sauf quelques répétitions) alors que SelTex ne produit que 3 légères variantes. Nous avons tenté de choisir un vocabulaire et une syntaxe la plus générale possible pour que le texte demeure acceptable pour la plupart des types de statistiques, mais il reste que le résultat n'est pas toujours approprié. Cet aspect nécessite beaucoup d'amélioration; une partie de la solution consisterait selon nous à fournir aux systèmes des connaissances linguistiques plus solides reliées à plusieurs sous-domaines.

Interface plus conviviale

Le format du fichier d'entrée du système est devenu très lourd, surtout avec les ajouts que nous y avons apportés. Une interface plus intuitive et plus agréable à utiliser nous semble indispensable pour motiver la poursuite de son développement, autant pour intéresser davantage les gens lors d'une démonstration que pour stimuler ceux et celles qui voudront y apporter des améliorations.

Avec une meilleure interface, nous suggérons également que le système puisse fonctionner en plusieurs passes, avec une interaction intermédiaire de la part de l'usager. Dans un premier temps, le système ferait une lecture des données brutes et les options de l'interface n'offrirait que les choix appropriés: on n'offre pas l'évolution s'il n'y a pas de

données temporelles, on n'offre pas la corrélation s'il n'y a pas au moins 2 données quantitatives. Dans un deuxième temps, le système montrerait à l'usager les types de message choisis par défaut mais il serait possible de sélectionner d'autres types de message également acceptables (par exemple en montrant les codes de classification et leur description).

Enrichir les techniques de sélection

Les techniques de sélection décrites au chapitre 4 ne constituent qu'un premier pas pour la partie "intelligente" de la sélection de l'information dans notre système. Nous avons établi de façon empirique les règles qui nous semblaient les plus intuitives pour s'orienter vers tel type de message correspondant à tel code de classification. Certaines de ces règles pourront être enrichies par une approche scientifique, comme pour mesurer et qualifier une corrélation ou une évolution. Les règles de comparaison peuvent être améliorées car les seuils utilisés ne sont pas invariants par translation. D'autres règles devront être raffinées par essais et erreurs en se basant sur de nombreux exemples dans une plus grande variété de domaines.

6.3 Contribution à la recherche

Les systèmes multimédias intégrant graphique et texte constituent un domaine de recherche actif depuis plusieurs années mais à notre connaissance, notre étude de corpus est la première de cette ampleur. Les résultats de cette étude, nos 55 codes de classification qui décrivent le lien entre le texte et le graphique ainsi que l'identification des 7 thèmes principaux pourront être utiles à tous les chercheurs qui développent des systèmes où la sélection de l'information est une étape importante. Nos techniques de sélection proposées serviront d'inspiration de départ pour aider les développeurs à construire des systèmes plus intelligents!

Bibliographie

[BAI84] G. Baillargeon. *Techniques statistiques avec applications en informatique, techniques administratives et sciences humaines.* Les Éditions SMG, 1984.

[BEL90] M. Bélanger. *Génération intégrée de textes et de graphiques.* Mémoire de maîtrise, DIRO, Université de Montréal, 1990.

[BER83] J. Bertin. *Semiology of Graphics.* The University of Wisconsin Press. Translated by William J. Berg, 1983.

[BRI95] A. Bridault. *Génération automatique de commentaires en langage naturel.* Université Pierre & Marie Curie Paris VI, Mémoire de stage pour Informatique CDC (Caisse des Dépôts et Consignations), 1995.

[CON85] C. Contant. *Génération automatique de texte: application au sous-langage boursier français.* Mémoire de maîtrise, Département de linguistique et de philologie, Université de Montréal, 1985.

[COR98] M. Corio et G. Lapalme. *Integrated generation of graphics and text: a corpus study.* Workshop on Content Visualization and Intermedia Representations (CVIR '98), Montréal, pages 63-68, août 1998.

[DER96] P. Deransart, A. Ed-Dbali et L. Cervoni. *Prolog: The Standard.* Springer, 1996.

[FAS96] M. Fasciano. *Génération intégrée de textes et de graphiques statistiques.* Thèse de doctorat, Publication #1039, DIRO, Université de Montréal, 1996.

[FAU97] F. Fauteux et G. Lapalme. *Implantation d'un assistant graphique amélioré pour Microsoft Excel: l'assistant Postgraphe 2.* Rapport interne, DIRO, Université de Montréal, 1997.

[GAL91] A. Gal, G. Lapalme, P. Saint-Dizier et H. Somers. *PROLOG for Natural Language Processing.* John Wiley & Sons, 1991.

[GAG93] M. Gagnon. *Expression de la localisation temporelle dans un générateur de texte*. Thèse de doctorat, DIRO, Université de Montréal, 1993.

[GOL86] E. Goldberg, R. Kittredge et A. Polguère. *Synthesizing weather forecasts from formatted data.* Proceedings of the 11th International Conference on Computational Linguistics, COLING '86 (Bonn), pages 563-565, août 1986.

[IOR92] L. Iordanskaja, M. Kim, R. Kittredge, B. Lavoie, et A. Polguère. *Generation of extented bilingual statistical reports.* Proceedings of the 15th International Conference on Computational Linguistics (COLING-92), pages 1019-1023, août 1992.

[KUK83] K. Kukich. *Knowledge-Based Report Generation: A Knowledge-Engineering Approach to Natural Language Report Generation.* Ph.D. Thesis, University of Pittsburgh, 1983.

[LAV94] B. Lavoie. *Étude sur la planification de texte: Application au sous-domaine de la population active*. Mémoire de maîtrise, Université de Montréal, 1994.

[MAC86] J. D. Mackinlay. *Automatic Design of Graphical Presentations.* Thèse de doctorat, Computer Science Department, Stanford University, 1986.

[SIC97] *SICStus Prolog 3.* http://www.sics.se/sicstus.html

[STE94] L. Sterling et E. Shapiro. *The Art of Prolog*. MIT Press, 1994.

[ZEL89] G. Zelazny. *Dites-le avec des graphiques.* InterÉditions, 1989.

Annexe A - Sources du corpus

Ce tableau énumère les sources de notre corpus pour les 411 extraits de texte, ainsi que le nombre d'extraits de texte recueillis pour chaque source.

Code	Référence	Nbre
alt1	Rapport sur les fonds Altamira, 1er trimestre 1997.	3
alt2	Rapport sur les fonds Altamira, 2e trimestre 1997.	7
alt3	Les Fonds Altamira. Trousse du prospectus simplifié. Altamira, octobre 1997.	9
alt4	Régime enregistré d'épargne-retraite. Altamira, septembre 1996.	14
alt5	Guide de l'investisseur. Altamira, décembre 1996.	10
alt6	Placements Altamira, volume 13 numéro 2, août 1997	1
alt7	Placements Altamira, volume 13 numéro 3, novembre 1997	1
bel	Bélanger, Micheline. *Génération intégrée de textes et de graphiques*. Voir [BEL90].	3
fas	Fasciano, Massimo. *Génération intégrée de textes et de graphiques statistiques*. Voir [FAS96].	2
gri	Griffin, Peter. *The Theory of Blackjack*. Huntington Press, 1996.	2
presse	Journal La Presse, samedi le 7 juin 1997, cahier B, page 4.	4
sng	Statistiques en sciences naturelles et en génie 1994. Publié en 1995 par le CRSNG. No. de cat. NS1-9/1995.	66
ts24	Tendances sociales canadiennes no. 24, Statistique Canada, Printemps 1992.	3
ts25	Tendances sociales canadiennes no. 25, Statistique Canada, Été 1992.	5
ts26	Tendances sociales canadiennes no. 26, Statistique Canada, Automne 1992.	2
ts30	Tendances sociales canadiennes no. 30, Statistique Canada, Automne 1993.	2
ts31	Tendances sociales canadiennes no. 31, Statistique Canada, Hiver 1993.	1
ts32	Tendances sociales canadiennes no. 32, Statistique Canada, Printemps 1994.	5
ts33	Tendances sociales canadiennes no. 33, Statistique Canada, Été 1994.	5
ts34	Tendances sociales canadiennes no. 34, Statistique Canada, Automne 1994.	4
ts35	Tendances sociales canadiennes no. 35, Statistique Canada, Hiver 1994.	9
ts36	Tendances sociales canadiennes no. 36, Statistique Canada, Printemps 1995.	15
ts37	Tendances sociales canadiennes no. 37, Statistique Canada, Été 1995.	15
ts38	Tendances sociales canadiennes no. 38, Statistique Canada, Automne 1995.	13
ts39	Tendances sociales canadiennes no. 39, Statistique Canada, Hiver 1995.	11
ts40	Tendances sociales canadiennes no. 40, Statistique Canada, Printemps 1996.	18
ts41	Tendances sociales canadiennes no. 41, Statistique Canada, Été 1996.	16
ts42	Tendances sociales canadiennes no. 42, Statistique Canada, Automne 1996.	21
ts43	Tendances sociales canadiennes no. 43, Statistique Canada, Hiver 1996.	33
ts44	Tendances sociales canadiennes no. 44, Statistique Canada, Printemps 1997.	14
ts45	Tendances sociales canadiennes no. 45, Statistique Canada, Été 1997.	14
ts46	Tendances sociales canadiennes no. 46, Statistique Canada, Automne 1997.	17
ts47	Tendances sociales canadiennes no. 47, Statistique Canada, Hiver 1997.	10
van	Vancura, Olaf, Ph. D. & Fuchs, Ken. *Knock-Out Blackjack*. Isochoric Publishing, Huntington Press, 1996.	4
zel	Zelazny G. *Dites-le avec des graphiques*. Voir [ZEL89].	52
Total		411

Annexe B - Données du corpus

Cette annexe présente en entier les 411 extraits de texte du corpus et toutes les informations associées décrites à la section 3.1.3.

Les données sont triées selon les codes de classification du tableau 3.6.

La référence pour chaque extrait de texte contient le code de l'annexe A ainsi que le numéro de page.

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
1	Barres		Les aurifères sont restées faibles, non seulement parce que le prix du métal jaune a chuté à un bas de quatre ans, mais aussi à cause du scandale Bre-X.	alt1-7	COMP	%	N	BAS1	RAIS
2	Colonnes	X	Le fonds a affiché un rendement de 7,7% au cours de la pire période de cinq ans (d'août 1976 à juillet 1981).	alt3-7	COMP	T	%	BAS1	SPEC1
3	Barres		Le secteur des aurifères a affiché le pire rendement de l'indice TSE 300, ayant perdu 26,3% depuis le début de l'année.	alt2-7	COMP	N	%	BAS1	
4	Colonnes	X	Les libéraux sont les moins insatisfaits.	presse	COMP	N	%	BAS1	
5	Barres	X	Le taux d'abandon scolaire est le moins élevé chez les personnes qui ont occupé un emploi pendant un nombre modéré d'heures.	ts35-21	COMP	%	N	BAS1	
6	Tarte		La société A détient la plus petite part de marché de sa profession.	zel-45	COMP	N	%	BAS1	
7	Barres		C'est dans la région C que la productivité est la moins bonne	zel-72	COMP	Q	N	BAS1	
8	Barres		Groboucan se situe au dernier rang pour la rentabilité des actifs en 1989	zel-78	COMP	%	N	BAS1	
9	Colonnes	X	Peu de jeunes femmes célibataires mettent leur enfant en adoption	ts32-3	COMP	T	%	BAS2	
10	Colonnes	X	Les travailleurs à temps partiel sont proportionnellement moins nombreux à bénéficier des avantages sociaux	ts43-18	COMP	N	%	BAS2	
11	Colonnes	X	Les personnes âgées ne sont pas de grands utilisateurs de la technologie courante	ts46-28	COMP	N	%	BAS2	
12	Colonnes		Les tâches ménagères incombent en grande partie aux femmes dans les familles où les deux conjoints travaillent à plein temps, 1990	ts31-13	COMP	N	%	CARAC	
13	Colonnes		Les provinces de l'ouest du Canada avaient le taux d'emploi le plus élevé chez les adolescents en 1993	ts35-20	COMP	G	%	CARAC	
14	Barres		Toutes proportions gardées, les habitants des provinces de l'Ouest sont plus susceptibles de ne pas être affiliés à une religion	ts44-31	COMP	%	G	CARAC	
15	Barres		Les ressources en sciences et en technologie étaient les ressources les moins susceptibles de répondre aux besoins de la classe	ts47-29	COMP	%	N	CARAC	
16	Colonnes	X	Le fait d'avoir un conjoint et des enfants n'a pas empêché les hommes et les femmes d'accepter des responsabilités en matière d'aide informelle	ts47-5	COMP	N	%	CONC	
17	Barres	X	L'activité principale des fournisseurs de soins ne les empêchait pas de fournir de l'aide et des soins informels	ts47-5	COMP	%	N	CONC	
18	Tarte		Les charges quotidiennes d'un agent lui laissent peu de temps pour prospecter de nouveaux clients	zel-100	COMP	N	%	CONC	
19	Tarte		Les quatre grandes agences se partagent entre elles une multitude d'activités mineures	zel-101	COMP	N	%	CONC	
20	Barres	X	La nouvelle organisation du travail a réduit les heures supplémentaires à tous les niveaux	zel-109	COMP	Q	N	CONC	
21	Colonnes	X	Le graphique ci-dessous illustre la façon dont le risque diminue au fil du temps.	alt5-5	CORR	T	%	CONC	
22	Colonnes	X	Par conséquent, plus longtemps vous conservez un portefeuille d'actions de sociétés canadiennes à faible capitalisation boursière, plus grandes sont vos chances de faire de l'argent.	alt5-5	CORR	T	%	CONC	
23	Barres	X	Nous vivons effectivement dans une société vieillissante	ts35-6	DIST	Q	A	CONC	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
24	Colonnes		Un grand nombre d'hommes prennent leur retraite avant l'âge de retraite habituel de 65 ans	ts42-11	DIST	A	%	CONC	
25	Colonnes		Si votre revenu de retraite est aujourd'hui de 10 000$, dans 30 ans, il vous faudra 32 400$ pour tenir tête à l'inflation	alt7-6	EVOL	T	$	CONC	
26	Courbe	X	Les ménages canadiens se sont rapidement adaptés aux nouvelles technologies d'application domestique.	ts38-3	EVOL	T	%	CONC	
27	Barres	X	La population a vieiili considérablement dans la RMR de Terre-Neuve	ts47-22	EVOL	%	A	CONC	
28	Colonnes		La nouvelles équipe de direction a réussi à rétablir une situation bénéficiaire	zel-117	EVOL	T	Q	CONC	
29	Courbe		La prévision de croissance paraît irréaliste à la lumière des résultats des 7 dernières années	zel-55	EVOL	T	$	CONC	
30	Tarte		Nous sommes d'avis que ce graphique constitue un argument de poids qui devrait vous inciter à inclure dans votre portefeuille au moins une part de titres étrangers afin de diversifier vos avoirs.	alt3-32	PRES	G	%	CONC	
31	Courbe	X	Un placement de 90000$ fait plus que doubler son rendement s'il est investi dans un RÉER.	alt4	PRES	T	$	CONC	
32	Courbe	X	Ce graphique montre pourquoi il vaut mieux cotiser à un RÉER plutôt qu'à un régime non enregistré.	alt4	PRES	T	$	CONC	
33	Courbe		Les achats périodiques par sommes fixes font jouer les marchés en votre faveur.	alt4	PRES	T	$	CONC	
34	Barres		Jouer avec le marché peut nuire aux rendements.	alt4	PRES	$	O	CONC	
35	Barres		Le graphique démontre pourquoi il faut conserver ses placements à long terme.	alt4	PRES	$	O	CONC	
36	Colonnes	X	Les programmes de télévision autochtones étaient souvent suivis par les Métis qui ne parlaient pas une langue autochtone	ts43-25	PRES	A	%	CONC	
37	Colonnes	X	Les enfants épousent habituellement la religion de leurs parents quand les deux partenaires partagent la même confession	ts44-29	PRES	N	%	CONC	
38	Colonnes		Beaucoup d'enfants qui avaient moins de 2 ans en 1994 ont été mis au monde par une mère ayant eu des comportements à risque durant la grossesse	ts44-6	PRES	N	%	CONC	
39	Colonnes	X	La plupart des gens croient que l'homme et la femme devraient tous deux contribuer au revenu du ménage	ts46-18	PRES	O	%	CONC	
40	Colonnes	X	Bon nombre de personnes croient qu'avoir un emploi est la meilleure façon d'être indépendante	ts46-18	PRES	O	%	CONC	
41	Points		Le prix de la majorité des automobiles varie en fonction du poids.	bel-26	CORR	Q	$	CORR1	
42	Points		Il semblerait que l'augmentation des profits soit liée à l'augmentation des investissements en r/d.	fas-103	CORR	$	$	CORR1	
43	Barres	X	L'alphabétisme est étroitement lié au niveau de scolarité en 1994	ts43-34	CORR	%	O	CORR1	
44	Colonnes	X	L'opinion des femmes quant à l'importance d'occuper un emploi rémunéré pour être heureuses varie selon l'âge	ts46-16	CORR	A	%	CORR1	
45	Points		Il y a une relation entre baisse des prix et accroissement des ventes	zel-141	CORR	Q	$	CORR1	
46	Points	X	Dans l'usine B, les salariés les mieux formés bénéficient de rémunérations plus élevées	zel-142	CORR	Q	$	CORR1	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
47	Barres	X	Il y a un rapport entre la rentabilité et les rémunérations	zel-72	CORR	Q	N	CORR1	
48	Points		Il n'y a pas de relation entre remise et volume des ventes	zel-140	CORR	Q	Q	CORR2	
49	Points		Il n'y a pas de rapport entre les remises et le montant des ventes	zel-63	CORR	Q	$	CORR2	
50	Barres	X	Il n'y a pas de relation entre les remises et le montant des ventes	zel-65	CORR	Q	N	CORR2	
51	Points		Les tarifs plus élevés de certaines marques de carburant ne sont pas liés à une supériorité de leurs performances	zel-71	CORR	Q	$	CORR2	
52	Points		L'importance des augmentations au mérite ne dépend pas de l'ancienneté	zel-71	CORR	Q	Q	CORR2	
53	Barres	X	Il n'y a pas de relation entre part de marché et rentabilité des actifs en 1989	zel-80	CORR	%	N	CORR2	
54	Points		Note the bowed, parabolic, nature suggested for the regression function.	gri-213	CORR	%	%	CORR3	
55	Points		So on average we expect the running count to rise linearly with number of decks already played, such that the slope is +4 per deck.	van-67	CORR	Q	Q	CORR3	
56	Points		Note that the expectation changes in a roughly linear fashion with the running count.	van-75	CORR	Q	%	CORR3	
57	Courbe	X	C'est sur le taux d'emploi des adolescents que les fluctuations économiques influent le plus.	ts35-20	COMP+EVOL	T	%	CS1	
58	Courbe	X	Depuis 1985, le chiffre d'affaires de Groboucan a vivement progressé alors que les bénéfices ont fluctué	zel-82	EVOL+COMP	T	Q	CS1	
59	Courbe	X	En 1995, les ventes de temps d'antenne ont correspondu aux dépenses d'exploitation de la télévision privée, et ce, pour la première fois depuis la fin des années 80	ts44-23	COMP+EVOL	T	$	CS2	
60	Courbe	X	La fécondité des femmes au début de la trentaine dépasse maintenant celle des femmes au début de la vingtaine.	ts39-14	EVOL+COMP	T	Q	CS2	
61	Courbe	X	D'ici 2030, il devrait y avoir excédent des décès sur les naissances.	ts42-5	EVOL+COMP	T	Q	CS2	
62	Courbe	X	Toutefois, en 1991 et 1992, le nombre d'homicides commis à l'aide d'une arme de poing a dépassé le nombre d'homicides commis à l'aide d'une carabine ou d'un fusil de chasse.	ts43-20	EVOL+COMP	T	%	CS2	
63	Courbe	X	L'augmentation la plus forte du taux de mortalité par cancer du poumon s'observe chez les personnes âgées.	ts39-12	COMP	T	Q	CS3	
64	Courbe	X	Les seules familles dont le revenu réel n'a pas varié ou s'est accru sont celles où les deux conjoints travaillent.	ts35-8	COMP+EVOL	T	$	CS3	
65	Colonnes	X	Dans les années 80, la rémunération et les heures de travail ont diminué pour les travailleurs de sexe masculin du quintile de rémunération inférieur	ts46-10	COMP+EVOL	O	%	CS3	
66	Courbe	X	Le travail autonome dans les branches d'activité autres que l'agriculture est demeuré relativement stable depuis 1931.	ts37-27	EVOL	T	%	CS3	
67	Courbe	X	Le taux de mortalité par cancer du poumon est plus élevé aujourd'hui que durant les années 50.	ts39-10	EVOL	T	Q	CS3	
68	Courbe	X	Hausse rapide du nombre de bénéficiaires d'une pension de retraite du RPC	ts40-11	EVOL	T	Q	CS3	
69	Courbe	X	Baisse de la fréquence du faible revenu chez les personnes âgées	ts40-14	EVOL	T	%	CS3	
70	Courbe	X	La rémunération horaire réelle des travailleurs de sexe masculin de moins de 35	ts46-11	EVOL	T	%	CS3	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
			ans a diminué au cours des années 80						
71	Courbe	X	Les personnes âgées étaient proportionnellement beaucoup plus nombreuses à être locataires en 1991 qu'en 1951.	ts36-10	EVOL+COMP	T	%	CS3	
72	Courbe	X	La proportion de jeunes hommes ayant une faible rémunération annuelle a augmenté plus rapidement que la moyenne pour tous les travailleurs de sexe masculin	ts46-11	EVOL+COMP	T	%	CS3	
73	Courbe	X	En 1984-1985, 6400 subventions étaient accordées pour un total de 143,1 millions de $, et la subvention de recherche moyenne s'élevait à 22 600$.	sng-12	PRES	T	$	DEB1	OUT
74	Courbe	X	La subvention stratégique moyenne s'élevait à 59100$ en 1984-1985, année où 546 subventions étaient accordées pour un total de 32,3 millions de $.	sng-15	PRES	T	$	DEB1	OUT
75	Courbe		Le chiffre d'affaires est passé de 1,2 millions de francs en 1984 à 3,4 millions en 1989, malgré le recul provoqué par une grève en 1986.	zel-21	EVOL	T	$	DEBFIN	OUT
76	Courbe	X	La population de la RMR (région métropolitaine de recensement) est passée de moins de 300 000 personnes en 1951 à plus de 900 000 personnes en 1991.	ts37-21	EVOL	T	Q	DEBFIN	SPEC4
77	Colonnes	X	Le secteur des études postsecondaires ou secteur universitaire compte pour 1123 millions de $ en R et D en sciences naturelles et en génie, comparativement à 622 millions de $ en 1984.	sng-30	COMP	T	$	DEBFIN	
78	Courbe	X	En 1992-1993, 22% des étudiants inscrits à temps plein au baccalauréat poursuivaient des études en sciences naturelles et en génie, comparativement à 24,9% en 1984-1985.	sng-33	COMP	T	%	DEBFIN	
79	Courbe	X	De même, les inscriptions au niveau de la maîtrise s'élevaient à 28,8% en 1992-1993, par rapport à 30% en 1984-1985.	sng-33	COMP	T	%	DEBFIN	
80	Courbe	X	La contribution du gouvernement fédéral, qui avait atteint 49% en 1984 a chuté à 45% en 1993.	sng-32	EVOL	T	%	DEBFIN	
81	Points		En effet, plus le poids est élevé, plus le prix l'est aussi.	bel-26	CORR	Q	$	DIREC	
82	Barres	X	La pratique sportive augmente avec le revenu de ménage	ts36-21	CORR	%	$	DIREC	
83	Colonnes	X	Les adultes à faible niveau d'instruction sont en moins bonne santé.	ts37-18	CORR	A	%	DIREC	
84	Colonnes	X	Les adultes dans les ménages à faible revenu sont en moins bonne santé.	ts37-18	CORR	A	%	DIREC	
85	Colonnes	X	Les ménages à revenu élevé sont plus susceptibles de posséder un ordinateur personnel tout comme les familles avec enfants.	ts38-6	CORR	A	%	DIREC	
86	Colonnes	X	Le fonctionnement d'un ordinateur pose moins de difficultés aux jeunes Canadiens et aux membres des ménages à revenu élevé.	ts38-7	CORR	$	%	DIREC	
87	Colonnes	X	Les travailleurs plus âgés sont plus susceptibles de travailler à domicile en 1991	ts40-20	CORR	A	%	DIREC	
88	Colonnes	X	La proportion d'adultes obèses augmente avec l'âge	ts40-29	CORR	A	%	DIREC	
89	Colonnes		La proportion de jeunes fumeurs augmente avec l'âge	ts43-5	CORR	A	%	DIREC	
90	Colonnes	X	Les Canadiens plus scolarisés sont plus susceptibles de lire pendant leurs loisirs	ts46-24	CORR	O	%	DIREC	
91	Colonnes	X	Les personnes dont le niveau de scolarité est plus faible sont moins susceptibles d'avoir recours à la technologie courante	ts46-29	CORR	N	%	DIREC	
92	Colonnes	X	Un revenu élevé donne lieu à une grande utilisation de la technologie courante	ts46-29	CORR	N	%	DIREC	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
93	Courbe	X	Comme vous pouvez le constater, les placements effectués dans le cadre d'un RÉER fructifient beaucoup plus vite que les autres.	alt4	COMP+EVOL	T	$	ECART	
94	Courbe	X	L'écart entre les taux de rendement annuel moyen des actions canadiennes et des obligations du Canada sur les 38 dernières années est d'environ 2,2% par an.	alt5-4	COMP+EVOL	T	$	ECART	
95	Courbe	X	Les niveaux d'emploi des femmes se rapprochent de ceux des hommes.	ts36-31	COMP+EVOL	T	%	ECART	
96	Courbe	X	L'écart se creuse entre les habitudes de voyage des jeunes adultes et celles des adultes plus âgés	ts45-3	EVOL+COMP	T	Q	ECART	
97	Courbe	X	Le chiffre d'affaires net augmente moins vite que les frais commerciaux	zel-127	EVOL+COMP	T	Q	ECART	
98	Courbe	X	Quelques conclusions préliminaires se dégagent: - les libéraux ont perdu du terrain durant la campagne; -les réformistes ont fait des gains; - les autres partis ont reçu le jour des élections à peu près le même niveau d'appui qu'au moment de son déclench	presse	COMP+EVOL	T	%	ENUM1	
99	Colonnes	X	Parmi nos quatre concurrents, B a aussi progressé, alors que C, A et D ont perdu des parts de marché.	zel-22	COMP+EVOL	T	%	ENUM1	
100	Courbe	X	Le pourcentage du financement contribué par le secteur privé a augmenté, tandis que celui des gouvernements provinciaux est demeuré à peu près le même.	sng-32	EVOL+COMP	T	%	ENUM1	
101	Courbe	X	L'emploi a augmenté dans la région d'Ottawa-Hull, mais la proportion de travailleurs dans les services gouvernementaux a diminué.	ts37-23	EVOL+COMP	T	%	ENUM1	
102	Courbe	X	Entre 1961 et 1992, la valeur totale du travail non rémunéré a augmenté, mais sa valeur par adulte est restée presque la même	ts42-33	EVOL+COMP	T	$	ENUM1	
103	Courbe	X	La population a augmenté dans la RMR de St. John's, mais elle a diminué dans le reste de Terre-Neuve	ts47-20	EVOL+COMP	T	Q	ENUM1	
104	Colonnes		Partout au Canada, une proportion importante d'adultes font régulièrement du sport.	ts36-23	COMP	G	%	EQUI1	
105	Barres		Au mois de septembre, le taux de rotation du personnel a été à peu près identique dans les six divisions.	zel-71	COMP	%	N	EQUI1	
106	Colonnes	X	L'opinion des hommes quant à l'importance d'occuper un emploi rémunéré pour être heureux est constante d'un groupe d'âge et d'un niveau de scolarité à l'autre	ts46-16	DIST	A	%	EQUI1	
107	Tarte		Les Canadiens consacrent à peu près autant de temps au travail rémunéré qu'au travail non rémunéré	ts42-29	COMP	N	%	EQUI2	
108	Barres	X	Le rapport entre investissement domestique et investissement à l'étranger est du même ordre pour toute la profession	zel-111	COMP	Q	N	EQUI3	
109	Courbe	X	Depuis 1986, les voyages des Canadiens aux Etats-Unis ont suivi les fluctuations du dollar canadien	ts45-4	CORR	T	Q	EQUI4	
110	Colonnes	X	Le taux des jeunes accusés de crimes de violence a diminué de moins de 1% en 1994, il s'agissait de la première année depuis 1986 où ce taux n'augmentait pas.	ts43-21	EVOL	T	Q	EXCEP	
111	Colonnes		Depuis 1983, les ventes ont augmenté chaque année sauf une	zel-117	EVOL	T	Q	EXCEP	
112	Courbe	X	Celle-ci atteignait 27200$ en 1994-1995, avec 7300 subventions accordées pour un	sng-12	PRES	T	$	FIN1	OUT

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
			total de 200,5 millions de $.						
113	Courbe	X	En comparaison, 481 subventions totalisant 41,2 millions de $ étaient accordées en 1993-1994, et la subvention moyenne atteignait 82 000$.	sng-15	COMP	T	$	FIN1	
114	Colonnes	X	43% des enfants nés en 1993 étaient le premier enfant de leur mère.	ts39-15	COMP	T	%	FIN1	
115	Courbe	X	Les dépenses du CRSNG s'élevaient à 494,9 millions de $ en 1993,1994, une baisse de 0,9% par rapport aux dépenses de 499 millions de $ de l'année précedente	sng-04	COMP+EVOL	T	$	FIN1	
116	Colonnes		Le ratio se situe maintenant à 0,07	sng-06	EVOL	T	%	FIN1	
117	Courbe	X	Les inscriptions au doctorat ont atteint 43% en 1992-1993.	sng-33	EVOL	T	%	FIN1	
118	Courbe	X	Les versement globaux au titre de la SV, du SRG et de l'AC dépassaient 20 milliards de dollars en 1994	ts40-7	EVOL	T	$	FIN1	
119	Courbe		Le taux de réussite se situe présentement à 32%.	sng-16	PRES	T	%	FIN1	
120	Colonnes		Il y a eu 108 affaires d'agression sexuelle déclarées pour 100000 habitants en 1994.	ts43-20	PRES	T	Q	FIN1	
121	Courbe	X	Si vous cotisez la somme de 250$ par mois, votre RÉER s'élevera à 515 710$ au bout de 30 ans, par rapport à 243 565$ dans un compte non enregistré, ce qui représente une différence de 272 145$ pour un placement global de 90 000$.	alt4	COMP	T	$	FIN2	
122	Colonnes	X	Si vous cotisez le même montant chaque mois et en présumant un taux de rendement fixe (10%), votre régime vaudra 82 250$ de plus si vous avez cotisé pendant 30 ans au début de la saison des RÉER plutôt qu'à la fin.	alt4	COMP	T	$	FIN2	
123	Colonnes	X	Les immigrants allophones récents établis à Montréal étaient plus nombreux à parler le français que l'anglais à la maison en 1991.	ts35-25	COMP	T	%	FIN2	
124	Courbe	X	Les Canadiens dépensent plus en voyages à l'étranger que les étrangers en voyages au Canada	ts45-7	COMP	T	$	FIN2	
125	Courbe	X	Les femmes mariées d'aujourd'hui sont proportionnellement plus nombreuses à faire partie de la population active	ts46-15	COMP	T	%	FIN2	
126	Courbe	X	À court terme, les placements dans des actions canadiennes sont généralement soumis à davantage de fluctuations mais les investisseurs en tirent normalement un meilleur rendement à long terme.	alt5-4	COMP	T	$	FIN3	CS1
127	Courbe	X	Les rendements passés montrent que les actions sont plus performantes à long terme.	alt4	COMP	T	$	FIN3	
128	Courbe	X	Les immigrants allophones récents établis au Québec sont proportionnellement plus nombreux que ceux des autres provinces à avoir fait un transfert linguistique.	ts35-24	COMP	T	%	FIN3	
129	Courbe	X	En 1994, le coût des prestations de retraite du RPC atteignait 9,8 milliards de dollars.	ts40-12	EVOL	T	$	FIN3	
130	Barres	X	En 1993-1994, les étudiants étrangers étaient proportionnellement plus nombreux que les étudiants canadiens à étudier en sciences	ts41-22	COMP	%	N	GS1	
131	Barres	X	Toutes proportions gardées, les femmes qui vivent avec des parents sont généralement plus nombreuses à contribuer au soutien du ménage que les	ts42-27	COMP	%	A	GS1	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
			hommes						
132	Barres	X	Les Nord-Américains de 16 à 65 ans étaient proportionnellement plus nombreux que les Allemands à posséder des capacités de lecture de niveau 1 ou de niveau 5 en 1994	ts43-30	COMP	%	G	GS1	
133	Colonnes	X	Les femmes sont deux fois plus susceptibles que les hommes de lire des livres durant leurs heures de loisirs	ts46-25	COMP	N	%	GS1	
134	Colonnes	X	Les résidents des régions urbaines sont de plus grands utilisateurs de la technologie courante que les résidents des régions rurales	ts46-30	COMP	N	%	GS1	
135	Points	X	In a single-deck game, the player has an advantage over 35% of the time, dropping to about 20% in an eight-deck game.	van-112	COMP	%	%	GS1	
136	Barres	X	Les Canadiens consacrent moins de temps à regarder la télévision et plus de temps à d'autres activités de loisirs	ts44-25	COMP+EVOL	Q	N	GS1	
137	Colonnes	X	Les adultes ayant une incapacité étaient, toutes proportions gardées, plus de deux fois plus nombreux que les autres adultes à avoir moins de 9 années de scolarité.	ts38-11	COMP	O	%	GS2	
138	Barres	X	L'instauration d'un intéressement en 1989 n'a pas entraîné d'augmentation de la production de l'usine 3	zel-108	COMP	Q	N	GS2	
139	Barres	X	C'est parmi les jeunes gens et les jeunes femmes que les heures d'écoute ont le plus nettement diminué	ts44-25	COMP+EVOL	Q	A	GS2	
140	Barres		Propulsé par l'essor des actions bancaires, le secteur des services financiers est demeuré en tête du marché pour l'année, ayant affiché un gain de 27,8% au 30 juin.	alt2-7	COMP	N	%	HAUT1	RAIS
141	Colonnes	X	Le fonds a affiché un rendement de 16,4% au cours de la meilleure période de cinq ans (d'août 1981 à juillet 1986).	alt3-7	COMP	T	%	HAUT1	SPEC1
142	Barres		Au cours de cette période, le secteur des services financiers a pris la tête du marché avec une progression de 7,61%.	alt1-7	COMP	%	N	HAUT1	
143	Colonnes	X	Les électeurs du Bloc et du Reform se montrent beaucoup plus critiques.	presse	COMP	N	%	HAUT1	
144	Tarte		L'Université de la Colombie-Britannique a reçu la proportion la plus importante, soit 40,5 millions de $ ou 9,6% du total du financement accordé aux universités canadiennes.	sng-09	COMP	N	$	HAUT1	
145	Tarte		En 1993, le gouvernement fédéral a versé 45,4% (510 millions de $) du financement dans ce secteur.	sng-31	COMP	T	$	HAUT1	
146	Colonnes		Plus du tiers des travailleurs à temps partiel en 1993 avaient ce régime de travail parce qu'ils n'avaient rien trouvé d'autre	ts36-31	COMP	N	%	HAUT1	
147	Barres		20% des agressions sexuelles commises par un inconnu surviennent dans la rue.	ts36-6	COMP	%	N	HAUT1	
148	Barres		En 1991, 24% de la population active d'Ottawa-Hull travaillait dans les services gouvernementaux.	ts37-23	COMP	%	N	HAUT1	
149	Barres	X	Les travailleurs autonomes qui étaient des employeurs avaient le revenu d'emploi moyen le plus élevé en 1990.	ts37-29	COMP	$	N	HAUT1	
150	Colonnes	X	Les adultes de Terre-Neuve chassent ou pêchent le plus.	ts37-34	COMP	G	%	HAUT1	
151	Colonnes		Les jeunes adultes ont plus tendance à vivre au domicile familial si les parents sont	ts38-17	COMP	N	%	HAUT1	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
			encore ensemble.						
152	Barres		En 1991, 44% des immigrants résidant dans la RMR de Vancouver provenaient de pays asiatiques.	ts38-28	COMP	G	%	HAUT1	
153	Barres		Le traitement de textes, l'application informatique la plus répandue.	ts38-7	COMP	%	N	HAUT1	
154	Tarte		La plupart des citoyens d'origine chinoise sont des immigrants.	ts39-24	COMP	N	%	HAUT1	
155	Barres		Les inscriptions dans le programme de commerce, de gestion et d'administration des affaires sont les plus courantes	ts40-24	COMP	%	N	HAUT1	
156	Barres		La lecture de documents est la méthode la plus commune d'apprentissage.	ts40-25	COMP	%	N	HAUT1	
157	Barres	X	Concilier le travail et les responsabilités familiales est le plus gros obstacle pour les apprenants adultes	ts40-25	COMP	%	N	HAUT1	
158	Colonnes	X	Les jeunes sont,toutes proportions gardées, les plus nombreux à avoir subi une blessure l'année précédente	ts40-29	COMP	A	%	HAUT1	
159	Barres	X	La Suède se classe au premier rang, au plan des capacités de lecture et d'écriture	ts41-36	COMP	%	G	HAUT1	
160	Barres		Les hommes ayant des relations homosexuelles forment les trois quarts des cas de sida chez les adultes	ts41-6	COMP	%	N	HAUT1	
161	Colonnes		Les jeune retraités sont proportionnellement plus nombreux à occuper de nouveau un travail rémunéré	ts42-15	COMP	A	%	HAUT1	
162	Colonnes		Le quart des retraités retournent au travail pour des raisons financières	ts42-15	COMP	N	%	HAUT1	
163	Barres	X	Vivre avec des parents se rencontre le plus fréquemment chez les femmes âgées	ts42-24	COMP	%	A	HAUT1	
164	Colonnes	X	Les Métis qui parlaient une langue autochtone étaient proportionnellement plus nombreux dans les régions rurales	ts43-24	COMP	N	%	HAUT1	
165	Colonnes		Les Métis qui participaient aux activités traditionnelles étaient proportionnellement plus nombreux dans les régions rurales	ts43-25	COMP	N	%	HAUT1	
166	Barres		La majorité des enfants élevés dans une famille reconstituée vivaient dans une famille reconstituée complexe en 1994	ts44-10	COMP	%	N	HAUT1	
167	Tarte		Plus d'un tiers du temps d'écoute des Canadiens va à des dramatiques et à des comédies étrangères	ts44-24	COMP	N	%	HAUT1	
168	Tarte		En 1994, la plupart des enfants de 12 ans vivaient dans une famille biparentale avec leurs deux parents biologiques	ts44-3	COMP	N	%	HAUT1	
169	Colonnes	X	Les diplômés universitaires étaient proportionnellement plus nombreux à avoir cessé de fumer, 1994-1995	ts45-21	COMP	N	%	HAUT1	
170	Colonnes	X	Les fumeurs ayant terminé ou non leurs études primaires étaient proportionnellement plus nombreux à avoir diminué leur consommation au cours de l'année précédente, 1994-1995	ts45-23	COMP	N	%	HAUT1	
171	Colonnes	X	Les interruptions de travail rémunéré sont plus courantes chez les jeunes femmes, 1990 à 1994	ts46-3	COMP	A	%	HAUT1	
172	Colonnes	X	Les diplômés occupant des emplois de bureau, de vente ou de services étaient les plus susceptibles d'avoir l'impression d'être trop qualifiés	ts47-14	COMP	N	%	HAUT1	
173	Barres		Les parents participaient plus fréquemment aux collectes de fonds	ts47-27	COMP	%	N	HAUT1	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
174	Tarte		La majeure partie des financements est consacrée à la fabrication	zel-72	COMP	N	%	HAUT1	
175	Barres	X	La pratique d'un sport est la plus répandue chez les jeunes hommes.	ts36-20	DIST	%	A	HAUT1	
176	Barres	X	Les gains des mères de 35 à 54 ans travaillant à temps plein ont connu la progression la plus rapide de 1980 à 1990	ts36-26	DIST	$	A	HAUT1	
177	Barres	X	Les consommateurs actuels appartenant à la tranche de revenu inférieure et âgés de 15 à 24 ans sont les plus susceptibles de faire face à des problèmes d'alcool.	ts38-22	DIST	%	A	HAUT1	
178	Colonnes		Les personnes âgées de 25 à 34 ans sont plus susceptibles de conduire avec des facultés affaiblies comme le sont également les personnes appartenant à la tranche de revenu supérieure.	ts38-23	DIST	A	%	HAUT1	
179	Colonnes	X	Les membres des ménages appartenant à la catégorie des revenus les plus élevés sont proportionnellement plus nombreux à se déclarer en excellent ou en très bonne santé	ts40-28	DIST	O	%	HAUT1	
180	Colonnes	X	En 1993, 20% des fils de pères à revenu élevé bénéficiaient aussi d'une bonne rémunération	ts42-36	DIST	O	%	HAUT1	
181	Colonnes	X	Les personnes du groupe des 55 à 64 ans sont les plus susceptibles de voyager à l'étranger, 1995	ts45-3	DIST	A	Q	HAUT1	
182	Colonnes	X	Les enfants de familles appartenant au groupe de statut socioéconomique supérieur étaient les plus susceptibles d'être parmi les premiers de classe	ts47-28	DIST	O	%	HAUT1	
183	Colonnes	X	Les femmes de 45 à 64 ans étaient plus susceptibles de fournir de l'aide aux personnes ayant un problème de santé de longue durée	ts47-3	DIST	A	%	HAUT1	
184	Colonnes		La majorité des livraisons sont effectuées en 5 ou 6 jours	zel-137	DIST	T	Q	HAUT1	
185	Colonnes		La majorité des salariés gagnent entre 10000 et 15000 francs	zel-71	DIST	$	Q	HAUT1	
186	Colonnes		L'an dernier, la rotation des effectifs a surtout concerné le groupe des 30-35 ans	zel-72	DIST	A	Q	HAUT1	
187	Colonnes	X	En 1989, la plupart des ventes de Groboucan ont porté sur les modèles bon marché de bandersnatchs frumieux	zel-84	DIST	$	Q	HAUT1	
188	Tarte		Le financement de la R et D universitaire en sciences naturelles et en génie provient de trois sources principales: le gouvernement fédéral, les gouvernements provinciaux et les universités elles-mêmes.	sng-31	COMP	N	$	HAUT2	
189	Colonnes	X	La préservation des espèces en déclin ou menacées est très importante pour les résidents adultes de l'Alberta et de la Colombie-Britannique.	ts37-31	COMP	G	%	HAUT2	
190	Colonnes		Les adultes de la Colombie-Britannique, de la Nouvelle-Ecosse et de l'Alberta sont les plus nombreux à faire des déplacements d'intérêt faunique.	ts37-32	COMP	G	%	HAUT2	
191	Barres	X	Les enseignants, administrateurs et directeurs du secteur culturel avaient le revenu moyen le plus élevé parmi les travailleurs de ce secteur en 1993	ts41-26	COMP	$	N	HAUT2	
192	Barres		La santé et le choix personnel sont les principaux motifs de retraite chez les hommes	ts42-11	COMP	%	N	HAUT2	
193	Barres	X	D'ici 2016, l'Ontario et la Colombie-Britannique s'attendent à avoir une plus grande proportion de la population des personnes ayant une incapacité de travail	ts42-2	COMP	%	G	HAUT2	
194	Barres		Au mois d'août, deux usines ont produit largement plus que les six autres	zel-72	COMP	Q	N	HAUT2	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
195	Colonnes	X	Les Autochtones sont proportionnellement plus nombreux à vivre avec des parents	ts42-25	COMP	N	%	HAUT4	
196	Colonnes	X	Les immigrants d'autres régions que les Etats-Unis ou l'Europe sont proportionnellement plus nombreux à vivre avec des parents	ts42-25	COMP	N	%	HAUT4	
197	Barres	X	Les enfants élevés dans une famille dirigée par une mère seule étaient plus susceptibles que les autres d'éprouver des problèmes	ts44-8	COMP	%	N	HAUT4	
198	Colonnes	X	Les femmes ayant terminé ou non leurs études primaires étaient proportionnellement plus nombreuses à avoir essayé de cesser de fumer au cours de l'année précédente, 1994-1995	ts45-23	COMP	N	%	HAUT4	
199	Colonnes	X	En 1991, les propriétaires d'unités condominiales étaient le plus souvent des gens seuls ou des couples âgés sans enfants	ts41-30	COMP	A	%	HAUT5	
200	Colonnes	X	Les travailleurs qui font des semaines comprimées et les travailleurs sur demande sont proportionnellement les plus nombreux à déclarer un stress lié au manque de temps élevé	ts43-17	COMP	N	%	HAUT5	
201	Barres		Les résultats commerciaux sont inégaux	zel-106	COMP	Q	N	INEQ1	
202	Barres		Le taux de rotation des effectifs varie selon le département	zel-106	COMP	Q	N	INEQ1	
203	Barres	X	La part respective des produits varie selon les régions	zel-110	COMP	Q	N	INEQ1	
204	Courbe	X	De 1984 à 1993, le financement de la R et D universitaire a évolué considérablement.	sng-32	EVOL	T	%	INT	
205	Courbe		Depuis 1961, deux tendances bien nettes se sont dessinées.	ts43-20	EVOL	T	Q	INT	
206	Barres	X	Pondération du marché canadien comparativement à l'indice boursier mondial.	alt3-36	PRES	%	N	INT	
207	Colonnes	X	Cette analyse indique le rendement le plus élevé, le rendement moyen et le rendement le moins élevé d'un fonds sur diverses périodes.	alt3-7	PRES	T	%	INT	
208	Colonnes		Croissance semestrielle de 10 000$ investis à partie de juin 1991.	alt3-8	PRES	T	$	INT	
209	Courbe		Comme l'illustre ce graphique, si vous devez investir 100$ par mois, le coût des titres auxquels vous souscrivez va varier chaque mois.	alt4	PRES	T	$	INT	
210	Colonnes	X	Comme l'illustre ce graphique, la différence sur plusieurs années peut être considérable.	alt4	PRES	T	$	INT	
211	Courbe	X	Rendement de diverses catégories de placements: évolution d'un placement de 1$ de décembre 1957 à septembre 1996	alt5-4	PRES	T	$	INT	
212	Barres		Le graphique ci-dessous illustre la corrélation entre le marché canadien et certains marchés étrangers.	alt6	PRES	Q	Q	INT	
213	Colonnes	X	La figure 3.19 montre l'évolution des profits de la compagnie Xyz inc. et de ses investissements en recherche et développement entre 1985 et 1994	fas-68	PRES	T	$	INT	
214	Courbe	X	Le graphique ci-contre montre l'évolution des intentions de vote durant la campagne.	presse	PRES	T	%	INT	
215	Colonnes		La plupart des adultes âgés de moins de 25 ans vivent au domicile familial.	ts38-17	DIST	A	%	INTER	
216	Colonnes	X	Il y a davantage de diplômés dans les tranches de salaires les plus élevées	zel-139	DIST	$	Q	INTER	
217	Courbe		La plupart de nos ventes aux Etats-Unis sont comprises entre 30 et 50 dollars.	zel-59	DIST	$	Q	INTER	
218	Courbe		La migration nette a été la plus élevée entre 1951 et 1961	ts35-27	EVOL	T	Q	INTER	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
219	Colonnes	X	À mesure que la période de placement s'allonge, la volatilité des placements s'amenuise.	alt5-5	CORR	T	%	INVER	
220	Colonnes	X	L'état de santé se détériore avec l'âge.	ts37-17	CORR	A	%	INVER	
221	Colonnes		Une proportion plus faible de femmes âgées étaient occupées	ts42-12	CORR	A	%	INVER	
222	Barres	X	Les capacités de lecture des adultes d'âge plus avancé étaient inférieures à celles des adultes plus jeunes en 1994	ts43-32	CORR	%	A	INVER	
223	Colonnes	X	Les jeunes adultes étaient plus susceptibles d'avoir l'impression d'être trop qualifiés pour leur emploi que les adultes plus âgés, 1994	ts47-17	CORR	A	%	INVER	
224	Barres		Moins de 20% des adultes ayant une incapacité auraient besoin d'aménagement en milieu de travail.	ts38-13	COMP	%	N	MOY1	
225	Barres		Le prix moyen des automobiles est de 24224$	bel-27	PRES	$	N	MOY1	
226	Barres		En 1994, 1 enfant de moins de 12 ans sur 4 vivait dans une famille à faible revenu	ts44-3	PRES	%	G	MOY1	
227	Courbe	X	Le taux de croissance annuelle moyen pour cette période était de 6,9% en dollars courants et de 3,5% en dollars constants.	sng-29	COMP+EVOL	T	$	MOY2	
228	Courbe	X	Depuis 1984-85, les dépenses du CRSNG ont augmenté en moyenne chaque année de 5,3% en dollars courants ou de 2,2 % en dollars constants de 1984.	sng-04	EVOL	T	$	MOY2	
229	Colonnes	X	Le taux de croissance annuelle moyenne se situait à 6,8% en dollars courants et à 3,7% en dollars constants.	sng-30	EVOL	T	$	MOY2	
230	Courbe		De 1978 à 1993, le taux de personnes accusées de conduite avec facultés affaiblies a enregistré une baisse annuelle moyenne de 4%.	ts43-21	EVOL	T	Q	MOY2	
231	Colonnes	X	Au cours de cette période, ce taux a augmenté en moyenne chaque année de 11%.	ts43-21	EVOL	T	Q	MOY2	
232	Colonnes	X	De 1986 à 1994, le taux des jeunes accusés de crimes contre les biens a diminué en moyenne de 2% par année.	ts43-21	EVOL	T	Q	MOY2	
233	Barres		Le secteur des papiers et produits forestiers est le seul secteur des ressources à avoir surpassé le TSE 300, sous l'impulsion de la hausse des prix des pâtes et du papier journal.	alt2-7	COMP	N	%	OUT	RAIS
234	Barres		Les métaux et minéraux, le secteur des ressources le moins performant de l'indice l'année dernière, sont repartis en force sous l'effet de la montée des prix des métaux communs.	alt1-7	COMP	%	N	OUT	
235	Tarte		Les 437 millions de $ versés par le CRSNG comptent pour 85,7% du financement du gouvernement fédéral de la R et D universitaire en sciences naturelles et en génie.	sng-31	COMP	T	$	OUT	
236	Tarte		Depuis 1984-1985, le budget des subventions a augmenté de 61% en dollars courants ou de 23% en dollars constants.	sng-11	EVOL	N	$	OUT	
237	Tarte		Depuis 1984-1985, le budget des Programmes de bourses s'est accru de 46% en dollars courants et de 12% en dollars constants.	sng-22	EVOL	N	$	OUT	
238	Colonnes	X	Fléchissement de l'écoute de la télévision chez les jeunes	ts40-40	EVOL	A	Q	OUT	
239	Colonnes		La grève de 1988 a temporairement ralenti la progression des ventes	zel-116	EVOL	T	Q	OUT	
240	Barres		Les prix du cuivre et du zinc ont augmenté du fait des bonnes données	alt2-7	PRES	N	%	OUT	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
			fondamentales liées à la demande et au bas niveau des stocks, ce qui n'a pas empêché le secteur des métaux et minéraux d'accuser une perte de 1,6%.						
241	Colonnes	X	Si vous cotisez en tout début d'année, votre régime fructifiera à l'abri de l'impôt pendant 14 mois supplémentaires.	alt4	PRES	T	$	OUT	
242	Courbe	X	La quasi-totalité des personnes âgées reçoivent des prestations de la SV.	ts40-7	PRES	T	$	OUT	
243	Colonnes		En 1994, la police a signalé 31690 affaires d'agression sexuelle, ce qui représente 10% de tous les crimes de violence commis pendant l'année.	ts43-20	PRES	T	Q	OUT	
244	Barres		Deux divisions ont enregistré des pertes à la suite de l'annulation de contrats public	zel-107	COMP	Q	N	RAIS	
245	Points		They illustrate that the poor correlations are due more to the large deviations from the regression functions than to any peculiar non-linear nature of these curves.	gri-227	CORR	%	%	RAIS	
246	Colonnes		La tendance à la baisse depuis 1984 s'est inversée en 1990.	sng-06	EVOL	T	%	RENV	
247	Colonnes		Après avoir augmenté rapidement au début des années 80, le nombre de bourses d'études supérieures s'est stabilisé aux alentours de 2500 bourses par année au cours des 10 dernières années.	sng-25	EVOL	T	Q	RENV	
248	Colonnes		Quatre domaines se sont partagés plus de 50% (226 millions de $) du financement du CRSNG en 1993-94: l'agriculture, les pêches et la foresterie; la production industrielle et la technologie; et la recherche générale en …	sng-10	COMP	N	$	SEUIL1	
249	Tarte		Des 57 universités canadiennes qui reçoivent un appui, les dix premières universités se sont partagées 60,5% des fonds en 1993-1994, une légère augmentation par rapport à 57,3% en 1984-1985.	sng-09	COMP+EVOL	G	$	SEUIL1	
250	Colonnes		75% de nos salariés gagnent plus de 150 000 francs	zel-59	DIST	$	Q	SEUIL1	
251	Colonnes	X	Un investisseur détenant des actions de sociétés canadiennes à faible capitalisation boursière a obtenu des rendements allant de -24% à +56% sur une période d'un an.	alt5-5	COMP	T	%	SPEC1	DEB1
252	Colonnes	X	Si ces mêmes actions avaient été détenues pendant vingt ans, leur taux de rendement se serait situé entre 12 et 16%.	alt5-5	COMP	T	%	SPEC1	FIN1
253	Colonnes	X	En 1993, 28% des femmes âgées seules ne touchaient pas de prestations du RPC et du RRQ	ts40-14	COMP	$	%	SPEC1	HAUT1
254	Barres		L'indice du TSE 300 a augmenté de 10,5% au deuxième trimestre, annulant la légère perte enregistrée au premier trimestre et affichant un gain de 9,6% pour le premier semestre.	alt2-7	EVOL	N	%	SPEC1	OUT
255	Barres		Les actions de produits industriels, auxquelles le TSE 300 attribue la plus forte pondération après les services financiers, ont également surpassé l'indice avec un rendement de 15% pour le semestre.	alt2-7	COMP	N	%	SPEC1	
256	Barres	X	Le quart des médecins et dentistes sont des femmes	ts36-32	COMP	%	N	SPEC1	
257	Barres	X	Le revenu annuel moyen d'une famille à Vancouver était de 59700$ en 1993	ts38-29	COMP	$	G	SPEC1	
258	Tarte		La conception du produit représente moins de 10% du prix du revient total	zel-100	COMP	N	Q	SPEC1	
259	Barres	X	Notre société se situe au-dessus de la moyenne dans les deux catégories commerciales	zel-108	COMP	Q	N	SPEC1	
260	Barres		Par rapport à notre quatre principaux concurrents, nous sommes au 1er rang pour	zel-21	COMP	%	N	SPEC1	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
			le taux de rendement du capital investi, avec 14%.						
261	Barres		La marge bénéficiaire du client se situe au quatrième rang	zel-50	COMP	%	N	SPEC1	
262	Tarte		Le chef des ventes ne passe que 15% de son temps sur le terrain	zel-71	COMP	N	%	SPEC1	
263	Tarte		En 1989, Groboucan détient la seconde part de marché de son secteur d'activité	zel-76	COMP	N	%	SPEC1	
264	Colonnes	X	Notre part de marché a augmenté de 4 points, passant de 11% en 1979 à 15% aujourd'hui.	zel-22	COMP+EVOL	T	%	SPEC1	
265	Colonnes	X	Le rendement moyen du fonds détenu pour n'importe quelle période de cinq ans est de 11,8%.	alt3-7	PRES	T	%	SPEC1	
266	Barres		L'inuktitut est la langue maternelle de plus de 7 personnes sur 10 au Nunavut	ts44-21	PRES	%	N	SPEC1	
267	Barres		Le graphique ci-dessous montre que bon nombre des marchés étrangers ont surpassé le marché boursier canadien sur la période de dix ans terminée le 30 septembre 1996.	alt4	COMP	%	G	SPEC2	
268	Barres		De nombreux marchés étrangers offrent des rendements plus élevés.	alt4	COMP	%	G	SPEC2	
269	Tarte		Le secteur privé compte pour un pourcentage important (10,7%) comparativement à l'industrie américaine, dont la contribution à la R et D post-secondaire est un peu plus de la moitié des montants investis au Canada.	sng-31	COMP	T	$	SPEC2	
270	Barres	X	D'ici 2016, 20% des Canadiens appartiendront à une minorité visible	ts41-3	EVOL	%	N	SPEC3	
271	Tarte		Le graphique ci-dessous présente la capitalisation boursière totale au Canada sous forme de pourcentage de la capitalisation boursière à l'échelle mondiale.	alt3-32	PRES	G	%	SPEC3	
272	Courbe	X	Le nombre de voyages d'affaires aux Etats-Unis augmente constamment	ts45-4	EVOL	T	Q	SPEC4	
273	Courbe	X	Le taux de chômage a diminué dans les régions métropolitaines, mais il varie davantage	ts42-36	EVOL+COMP	A	%	TEND1	CS1
274	Colonnes	X	La proportion de ménages locataires a augmenté dans toutes les provinces, sauf au Québec et au Nouveau-Brunswick	ts36-9	EVOL+COMP	G	%	TEND1	GS3
275	Courbe	X	Bien que les taux d'homicide commis à l'aide de carabines ou de fusils de chasse aient progressivement diminué depuis 1974, ils ont, par le passé, représenté la majorité des homicides commis à l'aide d'une arme à feu.	ts43-20	COMP+EVOL	T	%	TEND1	OUT
276	Colonnes		Les ventes continuent à progresser malgré le recul dû à la grève de 1988	zel-53	EVOL	T	$	TEND1	OUT
277	Courbe		Les propriétaires-occupants d'unités condominiales sont encore rares, mais leur nombre s'accroît	ts41-29	EVOL+PRES	T	%	TEND1	TAIL
278	Courbe		Croissance démographique faible, mais continue, au cours des prochaines décennies	ts42-3	EVOL+PRES	T	Q	TEND1	TAIL
279	Courbe		Le ratio DBRD/PIB correspondant pour ce secteur s'est maintenu relativement stable entre 0,14% et 0,16% au cours de la décennie.	sng-30	EVOL	T	%	TEND1	
280	Courbe		Le revenu moyen des ménages s'est stabilisé	ts35-13	EVOL	T	$	TEND1	
281	Courbe		La construction d'appartements a diminué depuis 1971	ts36-11	EVOL	T	Q	TEND1	
282	Barres	X	Les adultes de 15 ans et plus sont maintenant moins enclins à consommer de l'alcool.	ts38-21	EVOL	%	T	TEND1	
283	Courbe	X	La fécondité a chuté dans les trois plus grandes provinces canadiennes	ts39-16	EVOL	T	Q	TEND1	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
284	Courbe		Hausse du pourcentage des aînés qui touchent des prestations de retraite du RPC	ts40-11	EVOL	T	%	TEND1	
285	Courbe	X	Hausse de la prestation de retraite mensuelle moyenne	ts40-12	EVOL	T	$	TEND1	
286	Courbe		Nombre croissant d'élèves étrangers au Canada	ts41-20	EVOL	T	Q	TEND1	
287	Courbe		Hausse constante du nombre de cas de sida diagnostiqués annuellement chez des adultes au Canada depuis 1979	ts41-5	EVOL	T	Q	TEND1	
288	Courbe		L'année 1994 a marqué la onzième année consécutive où le taux de personnes accusées de conduite avec facultés affaiblies a diminué	ts43-21	EVOL	T	Q	TEND1	
289	Barres		La population a augmenté rapidement dans le territoire du Nunavut	ts44-19	EVOL	%	N	TEND1	
290	Courbe		Nombre d'ordinateurs personnelles à la hausse	ts44-36	EVOL	T	%	TEND1	
291	Colonnes		Les nouvelles mères sont plus nombreuses à commencer leur carrière avant la naissance de leur premier enfant	ts46-5	EVOL	T	%	TEND1	
292	Colonnes	X	Les raisons familiales sont moins couramment invoquées par les femmes	ts46-5	EVOL	T	%	TEND1	
293	Colonnes		Le bénéfice par action de notre société diminue	zel-72	EVOL	T	$	TEND1	
294	Courbe	X	Au cours de la période de 1983 à 1992, les dépenses brutes en R et D en sciences naturelles et en génie se sont accrues de 82% (ou de 37% après le rajustement pour tenir compte de l'inflation), et sont passées de 4,3 milliards de $ en 1983 à 7,9 milliard	sng-29	EVOL+COMP	T	$	TEND2	ENUM1
295	Courbe	X	Le financement provenant des universités a également connu une baisse, passant de 30% en 1986 à 24% en 1993.	sng-32	EVOL	T	%	TEND2	TEND3
296	Courbe		De 1975 à 1994, en dépit des fluctuations annuelles, le taux d'homicides a graduellement diminué, passant de 3,0 pour 100 000 habitants à 2,0. Il s'agit d'une diminution de 33%.	ts43-20	EVOL	T	Q	TEND2	TEND3
297	Courbe		De 1961 à 1975, le taux d'homicides n'a pas cessé d'augmenter, passant de 1,3 pour 100000 habitants à un sommet de 3,0, ce qui représente une augmentation de 131%.	ts43-20	EVOL	T	Q	TEND2	TEND3
298	Colonnes		Alors que le taux de ces affaires a enregistré une augmentation annuelle moyenne de 10% entre 1983 et 1994, il a diminué de 10% entre 1993 et 1994.	ts43-20	EVOL	T	Q	TEND2	TEND3
299	Courbe		Le taux de financement régresse depuis quelques années; d'un maximum de 50% atteint en 1984, il est passé au niveau actuel de 40%.	sng-14	EVOL	T	%	TEND2	
300	Colonnes		Le nombre de bourses postdoctorales s'accroît constamment depuis 1983-84, passant de 156 à 353 actuellement, une augmentation de 126%.	sng-26	EVOL	T	Q	TEND2	
301	Colonnes	X	La proportion de locataires vivant seuls était quatre fois plus élevée en 1991 qu'en 1951.	ts36-10	EVOL	T	%	TEND2	
302	Colonnes	X	Le taux des jeunes accusés dans des affaires relatives à des crimes de violence a plus que doublé.	ts43-21	EVOL	T	Q	TEND2	
303	Colonnes	X	Selon les projections, la prévalence de la démence va tripler au cours des 35 prochaines années	ts45-25	EVOL	T	Q	TEND2	
304	Colonnes		Les ventes de la société ont été multiplié par 6 depuis 1983	zel-116	EVOL	T	Q	TEND2	
305	Courbe	X	La migration internationale nette a connu des variations, mais une hausse soutenue	ts42-5	EVOL	T	Q	TEND3	TEND5

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
			depuis le milieu des années 80						
306	Courbe		Les taux d'intérêt ont diminué depuis la fin des années 80	ts36-11	EVOL	T	Q	TEND3	
307	Courbe		Le taux d'inoccupation des appartements a augmenté depuis le milieu des années 80	ts36-11	EVOL	T	%	TEND3	
308	Courbe	X	Les revenus d'emploi des jeunes hommes sont en baisse depuis la fin des années 70	ts42-19	EVOL	T	Q	TEND3	
309	Courbe	X	Le pourcentage des déclarants ayant fait un don à un organisme de bienfaisance a diminué depuis 1991	ts43-10	EVOL	T	Q	TEND3	
310	Courbe	X	La moyenne des dons provenant de particuliers a peu changé au cours des dernières années	ts43-10	EVOL	T	$	TEND3	
311	Colonnes	X	Le taux des jeunes accusés de crimes contre les biens en 1994 a diminué de 9% par rapport à l'année précédente, ce qui représente une baisse annuelle pour la troisième année consécutive.	ts43-21	EVOL	T	Q	TEND3	
312	Courbe	X	Les jeunes gens de 15 à 19 ans étaient plus nombreux à fumer en 1994 et en 1995 qu'en 1990	ts43-4	EVOL	T	%	TEND3	
313	Courbe		On prévoit une augmentation des ventes pendant les dix prochaines années	zel-71	EVOL	T	Q	TEND3	
314	Courbe		Les personnes âgées représenteraient plus de 20% de la population du Canada en l'an 2031	ts40-8	EVOL	T	%	TEND4	
315	Courbe	X	Les dépenses du CRSNG en subventions d'appareillage ont fluctué au cours des dix dernières années, chutant à 20,3 millions de $ en 1984-1985 pour plafonner à 45,2 millions de $ en 1992-1993.	sng-21	EVOL	T	$	TEND5	HAUT3
316	Courbe		Notre activité a été relativement stable	zel-96	EVOL	T	Q	TEND5	
317	Courbe		Notre activité a été relativement instable	zel-96	EVOL	T	Q	TEND5	
318	Tarte		Budget du CRSNG pour 1994-1995	sng-03	PRES	$	%	TITRE1	
319	Colonnes		Cette figure illustre les dépenses du CRSNG exprimées en pourcentage du produit intérieur brut (PIB).	sng-06	PRES	T	%	TITRE1	
320	Colonnes		Dépenses du CRSNG en pourcentage du PIB	sng-06	PRES	T	%	TITRE1	
321	Tarte		Appui aux universités canadiennes, 1993-1994	sng-09	PRES	G	$	TITRE1	
322	Colonnes		Dépenses du CRSNG en subventions et bourses, 1993-1994	sng-10	PRES	N	$	TITRE1	
323	Tarte		Budget des subventions, 1994-1995	sng-11	PRES	N	$	TITRE1	
324	Tarte		Budget des programmes de bourses, 1994-1995	sng-22	PRES	N	$	TITRE1	
325	Tarte		Financement de la R et D dans les universités canadiennes - SNG, 1993	sng-31	PRES	T	$	TITRE1	
326	Tarte	X	Polluants atmosphériques selon la source, 1985	ts24-25	PRES	N	%	TITRE1	
327	Tarte		Bénéficiaires d'aide sociale, mars 1990	ts24-9	PRES	N	%	TITRE1	
328	Colonnes	X	Nature des activités physiques hebdomadaires, 1988	ts25-19	PRES	N	%	TITRE1	
329	Colonnes		Personnes âgées de 15 ans et plus qui ont déménagé, 1989	ts25-33	PRES	G	%	TITRE1	
330	Colonnes	X	Usage des médicaments chez les Canadiens, 1989	ts33-26	PRES	N	%	TITRE1	
331	Barres		Proportion de certains ménages qui possédaient ou louaient un véhicule, 1992	ts34-29	PRES	%	N	TITRE1	
332	Barres		Raisons pour lesquelles les femmes agressées n'ont pas signalé les actes de	ts34-6	PRES	%	N	TITRE1	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
			violence à la police, 1993						
333	Barres	X	Les sports les plus pratiqués au Canada	ts36-20	PRES	Q	N	TITRE1	
334	Colonnes		Proportion des naissances pour les femmes âgées de 30 à 34 ans dont c'était le premier enfant.	ts39-15	PRES	T	%	TITRE1	
335	Barres	X	Pourcentage des personnes âgées dans la population active occupée qui travaillaient à domicile en 1991	ts40-19	PRES	%	G	TITRE1	
336	Courbe		Proportion d'enfants vivant dans des familles à faible revenu	ts42-18	PRES	T	%	TITRE1	
337	Courbe	X	Familles à faible revenu avec des enfants qui dépendent davantage des transferts	ts42-19	PRES	T	%	TITRE1	
338	Barres	X	Population du Canada, 1995 et 2016	ts42-7	PRES	Q	A	TITRE1	
339	Tarte		Panier de 1992 utilisé dans l'IPC	ts45-14	PRES	N	%	TITRE1	
340	Courbe		Revenu familial moyen en dollars constants de 1995	ts45-32	PRES	T	$	TITRE1	
341	Barres		Rendement des sous-indices du TSE 300 pour les six mois terminés le 30 juin 1997	alt2-7	PRES	N	%	TITRE2	
342	Colonnes	X	Analyse de toutes les périodes de placement de 1, 3 et 5 ans, du 1er juillet 1991 au 30 juin 1997.	alt3-8	PRES	T	%	TITRE2	
343	Courbe	X	Le graphique ci-contre montre le rendement relatif des diverses catégories d'actif sur une période de 38 ans.	alt5-4	PRES	T	$	TITRE2	
344	Colonnes	X	Chaque barre représente la fourchette de rendement d'une catégorie de placements particulière sur une période donnée.	alt5-5	PRES	T	%	TITRE2	
345	Courbe	X	Dépenses du CRSNG, de 1984-1985 à 1993-1994	sng-04	PRES	T	$	TITRE2	
346	Colonnes		Cette figure illustre la recherche et la formation appuyées par le CRSNG par principal domaine d'application	sng-10	PRES	$	N	TITRE2	
347	Courbe	X	Subvention de recherche moyenne, de 1984-1984 à 1994-1995	sng-12	PRES	T	$	TITRE2	
348	Colonnes	X	Résultats du concours de subventions de recherche, de 1984 à 1994	sng-14	PRES	T	$	TITRE2	
349	Colonnes	X	Cette figure illustre le montant total des demandes et des subventions accordées dans le cadre des subventions de recherche, par année de concours.	sng-14	PRES	T	$	TITRE2	
350	Courbe	X	Subvention stratégique moyenne, de 1984-1985 à 1993-1994	sng-15	PRES	T	$	TITRE2	
351	Colonnes	X	Résultats du concours de subventions stratégiques, de 1984 à 1994	sng-17	PRES	T	$	TITRE2	
352	Colonnes	X	Cette figure illustre les montants totalisant les demandes, les montants accordés en subventions et les taux de réussite correspondants	sng-17	PRES	T	$	TITRE2	
353	Colonnes	X	Dépenses des subventions de R et D coopérative, de 1984-1985 à 1993-1994	sng-18	PRES	T	$	TITRE2	
354	Colonnes	X	Dépenses du programme de professeurs-chercheurs industriels, de 1984-1985 à 1993-1994	sng-19	PRES	T	$	TITRE2	
355	Courbe	X	Dépenses d'appareillage, de 1984-1985 à 1993-1994	sng-20	PRES	T	$	TITRE2	
356	Colonnes		Bourses d'études supérieures, de 1984-85 à 1993-1994	sng-25	PRES	T	Q	TITRE2	
357	Colonnes		Bourses postdoctorales de 1983-1984 à 1992-1993	sng-26	PRES	T	Q	TITRE2	
358	Courbe	X	Dépenses brutes en R et D - SNG, de 1983 à 1992	sng-29	PRES	T	$	TITRE2	
359	Colonnes	X	DBRD (SNG) par secteur d'exécution, études supérieures, de 1984 à 1993	sng-30	PRES	T	$	TITRE2	
360	Courbe	X	Financement de la R et D dans les universités canadiennes - SNG, de 1984 à 1993	sng-32	PRES	T	%	TITRE2	
361	Courbe	X	Inscriptions à temps plein (SNG), de 1982-1983 à 1992-1993	sng-33	PRES	T	%	TITRE2	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
362	Courbe	X	Ce graphique illustre le pourcentage des inscriptions à temps plein en sciences naturelles et en génie au niveau du baccalauréat, de la maîtrise et du doctorat, par rapport au total des inscriptions pour chaque cycle.	sng-33	PRES	T	%	TITRE2	
363	Courbe		Bénéficiaires d'aide sociale, mars 1970 à mars 1990	ts24-9	PRES	T	Q	TITRE2	
364	Courbe	X	Revenu réel moyen des enfants et des personnes âgées, 1971-1989	ts25-13	PRES	T	$	TITRE2	
365	Barres		Pourcentage de déclarants demandant une déduction pour dons de charité, selon la tranche de revenu, 1990	ts25-24	PRES	%	$	TITRE2	
366	Courbe		Ensemble des dépenses publiques associées aux tribunaux, 1971-1990	ts25-6	PRES	T	$	TITRE2	
367	Colonnes	X	Distribution de la population de la ville de Montréal, selon le groupe d'âge, 1971 et 1991	ts26	PRES	A	%	TITRE2	
368	Tarte		Causes entendues devant les tribunaux de la jeunesse, selon le genre d'infraction 1990-91	ts26-5	PRES	N	%	TITRE2	
369	Tarte		Composition ethnique de la population du Canada, 1991	ts30-20	PRES	N	%	TITRE2	
370	Colonnes	X	Salaire annuel moyen réel, selon le sexe, 1920-1990	ts32-17	PRES	T	$	TITRE2	
371	Colonnes		Progression du salaire annuel moyen réel 1920-1990	ts32-18	PRES	T	%	TITRE2	
372	Colonnes	X	Rapports de masculinité des personnes seules, selon la province, 1991	ts32-27	PRES	G	Q	TITRE2	
373	Colonnes		Proportion d'enfants nés de femmes non mariées, selon la province, 1991	ts33-17	PRES	G	%	TITRE2	
374	Courbe	X	Age moyen au premier mariage, 1921-1992.	ts33-4	PRES	T	A	TITRE2	
375	Courbe		Proportion de mariages où au moins un des époux avait déjà été marié, 1921-1988.	ts33-5	PRES	T	%	TITRE2	
376	Colonnes	X	Proportion de couples vivant en union libre, selon la province, 1981 et 1991.	ts33-9	PRES	G	%	TITRE2	
377	Courbe	X	Crimes liés aux véhicules à moteur, Canada et Etats-Unis, 1980 à 1992	ts34-23	PRES	T	Q	TITRE2	
378	Colonnes		Vol de véhicules à moteur, selon la province, 1992	ts34-24	PRES	G	Q	TITRE2	
379	Courbe	X	Immigrants en pourcentage de la population totale, selon la province, 1901-1991	ts37-10	PRES	T	%	TITRE2	
380	Barres		Répartition des adultes parmi les minorités visibles, 1991	ts37-4	PRES	%	N	TITRE2	
381	Colonnes		Taux d'activité comparatifs des membres des minorités visibles,1991	ts37-5	PRES	N	%	TITRE2	
382	Barres	X	Proportion des femmes inscrites à temps plein, pour l'ensemble des niveaux, 1972-73 et 1992-93.	ts39-19	PRES	%	N	TITRE2	
383	Courbe	X	Rapport entre les femmes et les hommes au premier cycle, 1992-1993.	ts39-20	PRES	A	Q	TITRE2	
384	Courbe	X	Immigration annuelle en provenance de Hong Kong et de la République populaire de Chine, en proportion de l'immigration totale.	ts39-23	PRES	T	%	TITRE2	
385	Courbe	X	Incidence du cancer de la prostate et taux de mortalité par ce cancer de 1969 à 1995.	ts39-4	PRES	T	Q	TITRE2	
386	Courbe	X	Taux de mortalité par âge, toutes causes confondues, Canada, 1993	ts41-13	PRES	A	Q	TITRE2	
387	Courbe	X	Taux de mortalité par âge, toutes causes extérieures confondues, Canada, 1993	ts41-13	PRES	A	Q	TITRE2	
388	Courbe		Rapport hommes-femmes dans les taux comparatifs de mortalité, Canada, de 1950 à 1993	ts41-14	PRES	T	Q	TITRE2	
389	Courbe	X	Taux comparatifs de mortalité pour les principales causes de décès, Canada, de 1950 à 1993	ts41-17	PRES	T	Q	TITRE2	
390	Colonnes	X	Concentration des élèves étrangers dans les établissements d'enseignement de	ts41-21	PRES	G	%	TITRE2	

NO	GRAPHIQUE	SER	TEXTE	RÉF.	INTENTION	X	Y	CODE1	CODE2
			l'Ontario en 1993-1994						
391	Barres	X	Concentration des hommes et des femmes dans les diverses professions du secteur culturel en 1993	ts41-24	PRES	%	N	TITRE2	
392	Barres		Taux des cas de sida dans certains pays, 1994	ts41-8	PRES	Q	G	TITRE2	
393	Courbe	X	Homicides avec une arme à feu en proportion de tous les homicides par type d'arme à feu, Canada, 1974 à 1994	ts43-20	PRES	T	%	TITRE2	
394	Courbe		Taux d'homicides, Canada, 1961 à 1994	ts43-20	PRES	T	Q	TITRE2	
395	Colonnes		Affaires reliées aux agressions sexuelles, Canada, 1983 à 1994	ts43-20	PRES	T	Q	TITRE2	
396	Courbe		Taux de personnes accusées de conduite avec facultés affaiblies, Canada, 1978 à 1994	ts43-21	PRES	T	Q	TITRE2	
397	Colonnes	X	Taux des jeunes accusés selon le type de crime, Canada, 1986 à 1994	ts43-21	PRES	T	Q	TITRE2	
398	Barres	X	Population totale des régions métropolitaines de recensement, 1995 et 2000	ts43-7	PRES	Q	G	TITRE2	
399	Courbe		Population de Halifax, 1851 à 1961	ts45-10	PRES	T	Q	TITRE2	
400	Courbe		Taux de variation annuel de l'IPC en pourcentage, 1915 à 1996	ts45-16	PRES	T	%	TITRE2	
401	Points		This figure shows the expectation (%) as a function of K-O running count.	van-75	PRES	Q	%	TITRE2	
402	Barres	X	Les dix pays de l'OCDE ayant les plus forts pourcentages de jeunes adultes inscrits à l'université, 1988	ts30-9	PRES	%	G	TITRE3	
403	Colonnes	X	Cinq principaux lieux de naissance des résidents non permanents, 1991	ts32-14	PRES	G	%	TITRE3	
404	Barres	X	Les 10 principaux pays de naissance des nouveaux immigrants, Québec et reste du Canada, 1991	ts37-11	PRES	%	G	TITRE3	
405	Barres	X	Pourcentage des employés dans les 10 principales industries de Halifax, comparativement à celui dans l'ensemble des RMR	ts45-11	PRES	%	N	TITRE3	
406	Tarte		En 1994-95, le budget du CRSNG s'élevait à 493,4 millions de $ comparativement à 494,9 millions de $ en 1993-94.	sng-03	COMP	N	$	TOT1	OUT
407	Tarte		Le budget des subventions s'élevait à 400 millions de $ en 1994-1995, une légère baisse par rapport à l'année précédente.	sng-11	PRES+EVOL	N	$	TOT1	OUT
408	Tarte		Le budget des programmes de bourses s'élevait à 75,1 millions de $ en 1994-1995, une baisse comparativement aux 78,1 millions de $ l'année précédente.	sng-22	PRES+EVOL	N	$	TOT1	OUT
409	Tarte		Au total, 421,1 millions de $ ou 97,1 % des dépenses du CRSNG en subventions et bourses pour 1993-1994 ont été versés aux universités canadiennes.	sng-09	PRES	N	$	TOT1	
410	Colonnes	X	Au cours des dix dernières années, le CRSNG et l'industrie ont versé respectivement 138 millions de $ et 176 millions de $ dans la recherche coopérative.	sng-18	PRES	T	$	TOT1	
411	Colonnes	X	Depuis 1984-1985, les dépenses du CRSNG et de l'industrie ont atteint respectivement 89 millions de $ et 55 millions de $.	sng-19	PRES	T	$	TOT1	

Annexe C - Corrélation linéaire

Fonctions pour déterminer le coefficient de corrélation linéaire selon une liste couples (x,y).

```
correl(ListeCouples, CoefCorr):-
        moyennes(ListeCouples, Xmoy, Ymoy),
        covariance(ListeCouples, Xmoy, Ymoy, Covariance),
        somme_carres_diff(ListeCouples, Xmoy, Ymoy, Xsomme, Ysomme),
        ProduitEcartTypes is sqrt(Xsomme) * sqrt(Ysomme),
        CoefCorr is Covariance / ProduitEcartTypes.

moyennes(ListeCouples, Xmoy, Ymoy) :-
        somme(ListeCouples, Xsomme, Ysomme),
        length(ListeCouples, Len),
        Xmoy is Xsomme / Len,
        Ymoy is Ysomme / Len.

somme([[X,Y]], X, Y).
somme([[X,Y] | ResteListeCouples], Xsomme, Ysomme):-
   somme(ResteListeCouples, Xtemp, Ytemp),
   Xsomme is X + Xtemp,
   Ysomme is Y + Ytemp.

covariance([[X,Y]], Xmoy, Ymoy, Covariance):-
   Covariance is (X - Xmoy) * (Y- Ymoy).
covariance([[X,Y] | ResteListeCouples], Xmoy, Ymoy, Covariance):-
   covariance(ResteListeCouples, Xmoy, Ymoy, Ctemp),
   Covariance is Ctemp + ((X - Xmoy) * (Y- Ymoy)).

somme_carres_diff([[X,Y]], Xmoy, Ymoy, exp(X-Xmoy,2), exp(Y-Ymoy,2)).
somme_carres_diff([[X,Y]  |  ResteListeCouples],  Xmoy,  Ymoy,  Xsomme,
Ysomme):-
   somme_carres_diff(ResteListeCouples, Xmoy, Ymoy, Xtemp, Ytemp),
   Xsomme is Xtemp + exp(X-Xmoy,2),
   Ysomme is Ytemp + exp(Y-Ymoy,2).
```

Seuil de signification du coefficient de corrélation linéaire r_c pour un échantillon de *n* couples selon [BAI84].

n	r_c
3	0,908
4	0,900
5	0,805
7	0,669
10	0,549
12	0,497
15	0,441
17	0,412
20	0,378
22	0,360
25	0,337
27	0,323
30	0,306
32	0,296
35	0,283
40	0,264
50	0,235
60	0,214
80	0,185
90	0,174
100	0,165