MARTHE SIMARD

LA CORRECTION AUTOMATIQUE DE LA PONCTUATION

PERSPECTIVES ET LIMITES

Thèse
présentée
à la Faculté des études supérieures
de l'Université Laval
pour l'obtention
du grade de Philosophiae Doctor (Ph.D.)

Département des langues, de linguistique et de traduction
FACULTÉ DES LETTRES
UNIVERSITÉ LAVAL
QUÉBEC

SEPTEMBRE 2000

Sommaire

Cette recherche veut identifier les conditions d'automatisation de la correction des erreurs de ponctuation. Nous avons appliqué une méthode empirique pour effectuer une analyse comparée de 150 textes professionnels publiés dans les éditions électroniques de grands journaux et de magazines d'expression française ainsi que de 75 textes non publiés, tirés aléatoirement parmi une population de 16 084 productions de rédacteurs occasionnels. Nous avons ensuite élaboré une grille de calibrage permettant de décrire un texte en utilisant une banque d'indices simples. Nous introduisons le concept nouveau d'indice de maîtrise. Nous combinons ces indices de maîtrise à un relevé d'indices de faiblesse pour bâtir un modèle théorique autorisant le calibrage automatique des correcteurs grammaticaux. Un correcteur ainsi calibré en fonction du profil linguistique de son utilisateur pourrait prédire, entre autres, le nombre, la nature et la distribution probables des erreurs de ponctuation de ses textes au moyen des formules de calculs statistiques robustes que nous proposons.

Marthe Simard,
candidate

Jacques Ladouceur,
directeur de recherche

À la mémoire de mon père, Joseph-Léonce Simard

Résumé

Cette recherche veut identifier les conditions d'automatisation de la correction de la ponctuation. En appliquant une méthode empirique à de petits corpus, nous avons comparé un échantillon statistiquement représentatif tiré aléatoirement d'une population de 16 084 textes écrits par des rédacteurs occasionnels à un corpus de 150 textes non littéraires publiés par des rédacteurs professionnels. Cette comparaison s'intéressait à la ponctuation tout autant qu'aux aspects proprement linguistiques du texte.

Sur le plan de la correction automatique de texte, nous démontrons que des indices linguistiques, aussi bien de maîtrise que de faiblesse, peuvent aider à décrire n'importe quel texte. Sur le plan de la correction automatique de la ponctuation, nous arguons qu'il existe en fait seulement deux catégories d'erreurs : celles qui ont un impact sur la définition automatique des frontières de phrase et les autres. Les premières rendent la segmentation du texte hasardeuse et se retrouvent exclusivement dans les deux tiers des productions écrites par des rédacteurs occasionnels. Les secondes, spécialement les erreurs d'omission de virgules, se retrouvent à tous les niveaux d'expertise linguistique. Nous établissons finalement qu'il existe une relation linéaire positive simple entre l'occurrence de certains indices linguistiques simples et le nombre et la nature des erreurs de ponctuation d'un texte.

Notre recherche introduit donc la notion de calibrage comme outil de correction automatique de texte, y compris des erreurs de ponctuation. Le calibrage consiste à dresser le profil linguistique d'un texte au moyen de la détection d'indices de calibrage simples. Notre matrice de calibrage permet ainsi d'identifier automatiquement le niveau d'expertise linguistique d'un texte en reconnaissant, non seulement ceux qu'une correction automatique ne pourra pas

aider, mais également les textes professionnels de haut calibre. En outre, un correcteur grammatical appliquant nos formules de prédiction statistiques sera en mesure de repérer les textes qu'il ne pourra pas segmenter sans se tromper tout autant que de prédire le nombre, la nature et la distribution des erreurs de ponctuation probables dans les productions écrites non littéraires d'expression francaise.

Marthe Simard,
candidate

Jacques Ladouceur,
directeur de recherche

Avertissement

Les conclusions de la présente étude n'engagent en rien le Secteur de l'évaluation et de la sanction des études collégiales de la Direction de l'enseignement collégial du ministère de l'Éducation du Québec, qui a rendu disponibles les textes des candidats à l'épreuve de français de mars 1998.

Dans cet ouvrage, le masculin est employé pour le féminin pour des raisons purement grammaticales.

Remerciements

Cette recherche n'aurait pas été possible sans la contribution des personnes et organismes suivants: le Fonds des chercheurs et de l'aide à la recherche (F.C.A.R.) et le Conseil de recherches en sciences humaines du Canada (C.R.S.H), qui ont financé notre recherche; le ministère de l'Éducation du Québec, notamment M. Robert Poulin, coordonnateur de l'évaluation et de la sanction des études, qui a mis à notre disposition, sous réserve des dispositions d'une entente de confidentialité, les textes des candidats à l'épreuve de français de mars 1998, et le statisticien Jean-Denis Moffet, qui a tiré et documenté l'échantillon constituant la partie essentielle de notre corpus; le linguiste John Chandioux, qui a accepté de discuter de son expérience en correction automatique de la langue; l'entreprise Machina Sapiens, qui nous a gracieusement prêté son produit, *Le Correcteur 101.*

Tout au long de cet exercice, le travail de notre directeur de recherche, le professeur Jacques Ladouceur, a été extraordinaire. C'est avec la plus vive reconnaissance que nous rendons ici hommage à ses remarquables qualités aussi bien humaines que professionnelles. Puissent de nombreux autres étudiants profiter de son savoir et de sa générosité!

Nous ne saurions passer non plus sous silence la contribution toute personnelle de Jeff MacHan dont la présence constante, la vive intelligence et le sens de l'humour ont accompagné et nourri la chercheuse tout autant que l'épouse. Qu'il en soit affectueusement remercié!

Préambule

Mon père n'était pas un homme instruit. Né en 1905 et orphelin cinq ans plus tard, il avait grandi chez sa grand-mère, qui gouvernait d'une main de fer une maisonnée sans éducation ni culture. Il en avait toujours souffert et avait travaillé fort pour apprendre par lui-même tout ce que l'école n'avait pu lui enseigner. Il avait récolté le fruit de ses efforts, non seulement dans une vie réussie d'homme d'affaires, mais également en se liant d'une amitié indéfectible avec le docteur Savary, un médecin et ami sincère qui avait agi comme guide affectif et intellectuel tout au long de sa quête vers le savoir.

L'un des rêves de mon père avait été d'écrire. Il avait quelque peu atteint ce but en publiant dans le journal local, sous un pseudonyme, des chroniques fougueuses commentant les aléas de la vie politique canadienne des années quarante et cinquante. Sa plume, riche d'images et pleine d'ironie, démontrait un talent certain pour l'écriture pamphlétaire.

Pourtant, si papa avait réussi à bien contrôler la grammaire française — ce qui en dit assez sur son acharnement à apprendre, lui qui n'avait fait qu'une deuxième année —, il n'avait jamais su ponctuer. Combien de fois est-il venu me voir, dans ma chambre de jeune fille, pour me demander de corriger l'un de ses textes! Combien de fois m'a t-il demandé de lui expliquer comment je faisais pour mettre les points et les virgules "à la bonne place"! Mais, hélas, je ne savais pas pourquoi ni comment je ponctuais ainsi que je le faisais. Malgré mes efforts et les siens, il n'a jamais pu ponctuer correctement. Lui qui avait vaincu des embûches invraisemblables pour échapper à la misère, à l'ignorance et à la médiocrité, ce sont de petits signes anodins qui eurent finalement le dessus sur lui.

À la mort de mon père, en février 1985, ma mère m'a donné les grammaires et les nombreux guides de ponctuation qu'il avait achetés au fil des années, dans la guerre courageuse qu'il livrait aux virgules et aux points. Je me suis juré alors que, si jamais je faisais des études avancées, je les consacrerais à la ponctuation.

Table des matières

Table des illustrations

Liste des tableaux

Liste des figures

Introduction

La langue écrite a évolué vers un emploi de plus en plus fréquent de la ponctuation. Dans les textes précédant l'invention de l'imprimerie, en effet, l'usage des signes de ponctuation est pratiquement inexistant (Demanuelli, 1977: 39; Drillon, 1991: 24, 26). Bien que la virgule et le point aient été connus dès l'époque latine, des manuscrits du V^e^ et du VI^e^ siècle n'en présentent aucun. En fait, ils ne présentent même pas de blancs (Drillon, *loc. cit.*). C'est à la fin du Moyen-Âge que l'habitude de séparer des unités phrastiques par l'usage de certains signes s'est étendu: les blancs d'abord, puis la barre oblique, les deux-points et un signe ressemblant aux deux points, disparu aujourd'hui (Demanuelli, *loc. cit.*).

Il fallut l'invention de l'imprimerie pour permettre la diffusion de nouvelles habitudes de ponctuation. En 1471, à Paris, Jean Heynlin effectue une première liste des ponctèmes alors usités: la *virgula* [,], le *colon* [˙], le *periodus* [;], le *comma* [point moyen avec virgule suscrite], le *punctus interrogativus* [?], la *parenthesis* [( )]; également la barre oblique simple [/], pour une division simple, et la barre oblique double [//], pour la césure (Tournier dans Drillon, 1991: 26). Au XVI^e^ siècle, l'imprimeur Tory introduit la majuscule pour marquer le commencement de la phrase et l'apostrophe, pour séparer un article d'un nom. À la fin du XVI^e^, les imprimeurs ont remplacé les copistes et leur volonté fait loi dans le traitement typographique du texte[1].

Néanmoins, une étude quantitative de la ponctuation des manuscrits autographes de la correspondance de Racine fait ressortir, outre *l'instabilité foncière*

---

[1] On était obligé de passer par [les typographes] , comme par les écrivains publics , ce qu'ils étaient en réalité. *Tout se décidait donc dans les secret des officines (...). Aucun de leurs registres internes n'étaient connu des usagers. Comme nul ne pouvait passer ces murs sans être de leur coterie , et que l'on ne savait pas ce qui se passait de l'autre côté , on décida de s'aligner sur leurs manies et leurs fantaisies. Tout était bon qui venait d'eux. Ils s'assurèrent ainsi petit à petit de la gestion des signes, des usages , de la ponctuation et même du style des auteurs.* (Catach , 1989 : 152)

de la ponctuation (Barko, 1977: 98), une *tendance très marquée à l'absence de ponctuation* (*loc. cit.*). En fait, précise Barko, seuls le point et la virgule apparaissaient chez Racine de façon régulière, les autres signes ne se voyant employés que très rarement. Barko souligne d'ailleurs, dans son ouvrage, la surponctuation des versions modernes des textes de Racine par rapport aux versions originales. Le corpus de Barko, avec ses 3 234 points et ses 2 327 virgules, met également en évidence un usage du point plus étendu que celui de la virgule, dont l'emploi se révèle par ailleurs «particulièrement instable» (*ibid.*: 102).

Des articles généralement écrits par des correcteurs d'épreuves (Catach, 1989: 152) recommandent, à partir du XVIII^e^ siècle, l'emploi systématique de la ponctuation. Nicholas Beauzée, le premier, fait état du besoin de réglementer l'usage de la ponctuation (Barko, *op. cit.*: 68). Cette volonté de dicter des instructions (Catach, *op. cit.*: 234) permettant de régulariser un usage encore instable se prolonge chez les encyclopédistes du XVIII^e^ et trouve son apogée avec les travaux de Pierre Larousse au XIX^e^ siècle. Tout au long de ce processus, la ponctuation est régie et appliquée, souvent à l'encontre de la ponctuation de l'auteur[2], par des correcteurs payés par l'éditeur ou l'imprimeur. Seuls quelques auteurs – George Sand, s'opposant publiquement à Larousse; Valery Larbaud, dans sa *Lettre aux imprimeurs*; André Gide, contre Roger Martin du Gard – semblent vouloir réclamer le droit au maintien de l'intégrité totale de leurs textes (Barko, 1977: 66; Vaarlot, 1977: 20 -21; Drillon, 1991: 32).

Un texte sans ponctuation est considéré comme illisible aujourd'hui, comme en fait foi la démonstration de Jones (1996c : 6, 7). Les textes de l'époque romantique du XIX^e^ siècle étudiés par Claude Gruaz (1980) montrent déjà un usage plus étendu de la virgule que du point. En 1958, Brun et Doppagne, dans *La Ponctuation et l'Art d'écrire*, déclarent la virgule le signe le plus utilisé. Les traités de Jacques Damourette (1930), de Henri Sensine (1939), de Brun et Doppagne ([1958]), et plus récemment, de Jean-Pierre Colignon (1988), d'Albert Doppagne (1989) et de Jacques Drillon

---

2 *La ponctuation*, écrivait un certain Chapoulaud, imprimeur à Limoges, en 1865, *est une des parties les plus difficiles de la Grammaire. Seul l'imprimeur instruit et expérimenté est conséquent dans sa manière de ponctuer, et sur ce point, l'auteur doit s'en rapporter à lui...* (Catach, 1989 : 197).

(1991) soulignent aujourd'hui l'importance de la ponctuation, notamment de la virgule, dans la clarté d'un texte.

Les rédacteurs de la fin du vingtième siècle se distinguent fondamentalement de ceux dont les textes ont traditionnellement servi à décrire les règles de la ponctuation : ils disposent du micro-ordinateur, un outil puissant d'écriture et d'analyse se retrouvant aujourd'hui, grâce à la popularité d'Internet, dans un nombre grandissant de foyers. Avec chaque ordinateur, vient souvent un texteur, et avec chaque texteur, un correcteur orthographique et grammatical.

Le phénomène de la correction automatique de textes n'est pas nouveau. Dès les débuts de la micro-informatique, il s'est trouvé d'intrépides développeurs pour proposer des logiciels réputés pouvoir repérer et corriger les fautes des utilisateurs. En 1982, Chandioux (1996) finissait l'élaboration d'un nouveau langage de programmation, *GramR* ®, qui allait lui permettre de développer rapidement, entre autres, un correcteur orthographique et grammatical fonctionnel, *GramR - Le détecteur*. En 1993, Lesage *et al.* (1993) listaient 46 outils d'aide à l'écriture d'expression française, dont 39 se targuant de détecter les erreurs orthographiques et grammaticales. Pourtant, en 1999, aucun de ces logiciels n'est encore disponible. À leur place, existe une liste très réduite, dans laquelle on retrouve deux logiciels québécois, *Le Correcteur 101* et *Antidote*. Même *Word* a préféré, dans sa version 1998 pour Macintosh, un autre correcteur que *Hugo*, développé et mis en marché par Logidisque en 1987, que Microsoft avait intégré à son populaire texteur au début des années 90.

Dans leur *Enquête sur l'état d'utilisation des outils informatisés d'aide à la rédaction dans les organisations*, Lesage *et al.* (1993) font ressortir l'insatisfaction des utilisateurs face à la performance des correcteurs. Cette performance est marquée par les erreurs de détection et les bruits, c'est-à-dire les signalements de cas non pertinents à l'analyse corrective (Valcheshini, 1995). Il ne faut donc pas s'étonner du haut taux d'abandon de ces outils par les utilisateurs, même quand les correcteurs sont intégrés aux systèmes de traitement de texte (Lesage *et al.*, *op cit.*).

Parmi tous les aspects identifiés par Dale (1996) dans le traitement automatique d'un texte comportant des erreurs (*copy-editing*), nous en trouvons un particulièrement significatif : le développement de mécanismes d'analyse robuste pour permettre la reconnaissance et le traitement de textes à large couverture, c'est-à-dire de textes du type de ceux que les correcteurs grammaticaux sont appelés à analyser.

Ted Briscoe (1996b), exposant la problématique de la robustesse des mécanismes d'analyse, détermine trois types de problèmes : la segmentation du texte en unités analysables (*chunking*), le choix de l'analyse sémantique et syntaxique correcte parmi toutes les analyses proposées par le parseur (*disambiguation*) et le traitement des unités sortant du champ de couverture de l'analyseur (*undergeneration*).

La segmentation automatique du texte pose en elle-même une difficulté particulière. Jones (1996c : 144) souligne le besoin de trouver une solution à ce problème en raison de l'étendue des corpus modernes comportant souvent plusieurs millions de mots. Il mentionne ainsi l'expérience de deux linguistes qui ont réussi une telle segmentation en se basant sur les signes de ponctuation.

Par ailleurs, l'attribution d'une seule catégorie grammaticale aux mots d'un texte (*tagging*) dans un lemmatiseur statistique pose un problème plus important pour le français que pour l'anglais (Chanod et Tapanainen, 1994). En effet, à cause de la complexité morphologique du français et du processus d'étiquetage des options d'analyse, Chanod et Tapanainen (1994 : 2) estiment que le choix des combinaisons d'étiquettes d'analyses possibles pour un seul mot du vocabulaire français dépasse, dans certains cas, les 6 500 :

> *The French lexicon was not originally designed for a (statistical) tagger, and the number of different tag combinations is quite high. The size of the tagset is only 88; but because a Word is typically associated with a sequence of tags, the number of different combinations is higher, 353 possible sequences of single French Words. If we also consider Words joined with an article or a clitic pronoun, the number of different combinations is much higher, namely 6525.*

Quant à la difficulté pour les parseurs de reconnaître des unités tombant en dehors de leur champ de couverture (*undergeneration*), Chanod (1993) l'explique

par ce que nous pourrions appeler leur caractère « a rtificiel » En effet, discutant de l'approche répandue chez les linguistes informaticiens de développer des parseurs reconnaissant seulement le « n oyau » des phrases, il rappelle qu'au contraire, *l'analyse automatique robuste, surtout en phase initiale, est avant tout confrontée à des particularismes* (Chanod, 1993 : 3), réalisés surtout dans ce qu'il appelle des structures « périphériques ». En fait, ajoute-t-il, *l'insertion d'éléments inessentiels constitue l'une des principales difficultés de l'analyse robuste, en raison de la multiplicité des formes des éléments insérés (adverbes au sens large (Gross, 1986), incises, groupes nominaux errants, vocatifs, parenthèses, ponctuation, etc.) et de leur mobilité* [...] (Chanod, 1993 : 4). De telles structures se marquent graphiquement par des signes de ponctuation, le plus souvent par des virgules.

Or selon l'étude statistique de textes étudiants (Ministère de l'Enseignement supérieur et de la Science, 1993), nous pouvons nous attendre à ce qu'une très forte majorité des textes que les correcteurs grammaticaux sont appelées à réviser comportent des erreurs de ponctuation. Nous savons aussi que la ponctuation constitue un repère naturel dans le découpage automatique d'un texte en unités analysables et que les développeurs d'outils automatiques de correction de texte sont donc dépendants de la ponctuation des textes qu'ils analysent (Chandioux, 1996). Ce que cette recherche aura démontré, entre autres éléments, c'est jusqu'à quel point le pouvoir de diagnostic des grammaires informatisées est négativement influencé par les erreurs de ponctuation.

Cet ouvrage comprend sept chapitres. Le premier, « Ob jectifs et Problématique », fixe l'étendue de la présente étude et fait état des questions essentielles posées par la performance en ponctuation des groupes de rédacteurs observés: des rédacteurs occasionnels, c'est-à-dire n'ayant pas d'expérience professionnelle de l'écriture, et des rédacteurs dits professionnels ou experts, dont les textes identifiés à leur nom ont été publiés dans des livres, des journaux ou des magazines. Le deuxième, « Éta t de la question », fait une revue de la littérature du domaine de même que de la performance, en matière de diagnostic et de correction des erreurs de ponctuation, des correcteurs grammaticaux les plus populaires. Le troisième chapitre, « M éthodologie », décrit notre cheminement

méthodologique pendant les phases exploratoire et expérimentale de notre recherche. Le quatrième chapitre présente notre grille de calibrage. Les cinquième et sixième chapitres montrent comment nous en sommes arrivés au choix des indices sur lesquels s'appuient les calculs d'indice de faiblesse linguistique (chapitre 5) et de maîtrise linguistique (chapitre 6). Finalement, le septième chapitre élabore une matrice de calibrage et propose des éléments de méthode pour exploiter la grille de calibrage dans l'estimation de la proportion des erreurs de ponctuation à partir du profil linguistique d'un utilisateur.

Rappelons en terminant que la rédaction de cette thèse nous a imposé au moins deux choix importants. Le premier concerne la structure non traditionnelle de la présente étude. Si les trois premiers chapitres suivent assez le schéma usuel — objectifs et problématique; état de la question; méthodologie —, les chapitres subséquents s'en éloignent quelque peu en effet. Au lieu d'un chapitre *Résultats*, nous exposons nos découvertes sur quatre chapitres distincts, puisque l'alternative aurait créé un chapitre fort maladroit de près de 200 pages! Le deuxième choix touche notre insistance à reprendre plusieurs fois les explications de nos calculs de régression linéaire et celles des diagrammes de dispersion qui les représentent. Introduits d'abord en méthodologie, les commentaires guidant l'interprétation d'une régression linéaire et d'un diagramme de dispersion sont repris ensuite au chapitre 7, au fil de la présentation de notre matrice de calibrage, et également répétés dans des notes en bas de page. Conscients que les linguistes ne sont pas nécessairement tous à l'aise avec certaines formes de calculs statistiques, nous avons cru bon d'errer ici dans le sens de la prudence plutôt que de risquer que le lecteur se méprenne sur le sens à donner à des représentations statistiques somme toutes assez difficiles à comprendre pour un non-initié.

Notre recherche propose une solution alternative à la détection tout azimut des erreurs pratiquée traditionnellement par les correcteurs grammaticaux, détection exigeant une analyse syntaxique robuste. Nous défendons en effet la thèse selon laquelle il est possible de calibrer la correction automatique de textes selon le profil linguistique de l'utilisateur, profil que nous exprimons par les mathématiques au moyen de formules statistiques robustes. Notre matrice de calibrage permet ainsi à

un module de correction d'identifier automatiquement le niveau d'expertise linguistique d'un texte en reconnaissant, non seulement les textes qu'une correction automatique ne pourrait pas aider, mais également les textes professionnels de haut calibre qui n'ont pas besoin d'un tel support. En outre, un correcteur grammatical appliquant nos formules de prédiction statistiques à partir de la détection automatique de nos indices de calibrage pourra repérer les textes qu'il ne pourra pas segmenter sans se tromper tout autant que prédire le nombre, la nature et la distribution probables des erreurs de ponctuation de n'importe quel texte qui lui sera soumis en format électronique.

Un
Objectifs et problématique

1.1 Objectifs de la recherche

Notre recherche a pour objectif principal de déterminer les conditions de correction automatique des erreurs de ponctuation produites dans des textes non littéraires d'expression française, notamment les erreurs de virgule. Nous tenterons de répondre aux deux questions suivantes :

* Comment les erreurs de ponctuation peuvent-elles être détectées automatiquement?
* Une fois détectées (en supposant l'exercice possible), comment les erreurs de ponctuation peuvent-elles être corrigées automatiquement?

Nous poursuivrons plusieurs objectifs secondaires au cours de cette recherche.

Nous tenterons d'abord d'identifier des indices permettant de mesurer automatiquement la qualité du français écrit d'un rédacteur à partir de la prémisse que la nature et le nombre de ses erreurs de ponctuation varient en fonction de son expertise linguistique .

Nous chercherons ensuite à vérifier s'il existe une relation statistique entre le niveau de contrôle de la ponctuation et le niveau de contrôle linguistique de l'utilisateur tel qu'établi par les indices identifiés.

Dans l'éventualité d'une confirmation du lien statistique entre la ponctuation et les indices de niveau de contrôle linguistique, nous chercherons à déterminer si les indices linguistiques peuvent être utilisés pour détecter certaines catégories d'erreurs de ponctuation.

Finalement, nous proposerons, le cas échéant, des éléments d'algorithme permettant de corriger automatiquement les erreurs détectées.

1.2 Limites de la recherche

Notre objectif principal impose une approche normative du phénomène de la ponctuation. Or les grammaires normatives et les traités de ponctuation proposent souvent des critères contradictoires pour identifier les erreurs de ponctuation (Purnelle, 1998; Catach, 1994; Simard, 1993). L'absence de consensus sur le choix d'une norme définitive en ponctuation française pose donc le problème épineux de la définition de l'erreur. Comment l'ordinateur peut-il en effet détecter une erreur de ponctuation si les humains eux-mêmes n'arrivent pas à décider ce qui constitue en fait une erreur ? Autrement dit, à quelle école normative pourrions-nous demander à un correcteur de ponctuation de souscrire?

Cette question constitue en fait un faux problème du point de vue informatique. En effet, l'ordinateur « se soucie » peu de savoir quelle école normative a raison; il « s'intéresse » plutôt à l'application de la grille d'analyse du concepteur. Pour élaborer un outil de correction automatique des erreurs de ponctuation, il nous faut donc nous arrêter, non pas à la nature d'une grille d'analyse[3], mais plutôt à sa méthode d'application.

C'est dans cet esprit que nous avons choisi de nous inspirer de la grille[4] de Guénette, Lépine et Roy (1995). Cette grille comporte de nombreux avantages dans une recherche comme la nôtre : elle est relativement récente, ce qui signifie qu'elle tient compte de la plus grande couverture possible du discours normatif portant sur la ponctuation française; elle est formulée en catégories hiérarchisées convertibles en langage informatique et surtout, elle est montée à partir d'*un catalogue de fautes* (Guénette, Lépine et Roy, *op. cit.* : VI) repérées dans la pratique d'écriture de

---

[3] En fait, cette affirmation n'est elle-même qu'une demi-vérité, puisque l'expérience enseigne que la nature — ou plutôt le format — d'une grille de détection d'erreurs aura une incidence sur ses chances de portabilité en langage informatique. Par exemple, les traités de ponctuation exprimés dans un discours prescriptif élaboré souvent enjolivé de maintes fioritures stylistiques, doivent être filtrés pour permettre la mise en relief de règles décodables par un ordinateur. Nous pouvons donc nous attendre à ce que certains traités se prêtent mieux que d'autres à une telle conversion.

rédacteurs occasionnels, la clientèle même à qui un correcteur de ponctuation pourrait être destiné. Toutefois, notre choix ne représente pas l'endossement de la norme avancée par Guénette, Lépine et Roy (*op. cit.*) non plus qu'une déclaration universelle de sa validité.

Traditionnellement, les études normatives de la ponctuation se penchaient sur les textes littéraires bien connus. Ces textes sont exclus de la présente recherche parce que la correction automatique de la ponctuation demeure davantage pertinente, du point de vue pratique, pour les rédacteurs occasionnels. C'est pourquoi nous nous intéresserons exclusivement à la ponctuation des textes informatifs et argumentatifs du modèle de ceux qui sont publiés dans les magazines (par exemple *Actualité*), les revues scientifiques, les journaux ou certains recueils spécialisés (par exemple, la série *État du monde* publiée annuellement chez Boréal) parce que leur intention de communication se compare à celle des rédacteurs occasionnels dans leur pratique de l'écriture, c'est-à-dire informer ou défendre une idée.

Par ailleurs, nous exploitons un corpus de 225 textes d'expression française. Bien que nous ayons pris soin de fonder notre étude des productions de rédacteurs occasionnels sur un échantillon statistiquement représentatif de la population d'où il est tiré, nous demeurons conscients que ces textes ne constituent pas une représentation statistique du français de leurs auteurs, non plus que de la langue française en général.

---

[4] Voir Annexe 1 *Grille descriptive de la ponctuation.*

1.3 Problématique

Considérant nos objectifs, la problématique de cette recherche comporte deux aspects principaux : le repérage automatique des erreurs de ponctuation et la correction automatique des erreurs de ponctuation.

1.31 Repérage automatique des erreurs de ponctuation

À quelles erreurs de ponctuation pouvons-nous nous attendre de la part de rédacteurs d'expression française? Guénette, Lépine et Roy (1995) décrivent trois catégories d'erreurs : l'omission de signes, la confusion de signes et l'introduction abusive de signes (signes indus). L'étude de la distribution de ces erreurs de ponctuation dans notre corpus permet d'effectuer deux observations capitales :

* l'omission de signes requis constitue l'erreur de ponctuation la plus fréquente;
* certaines erreurs de ponctuation ont un effet négatif sur la définition automatique des frontières de phrase.

L'impact de ces observations nous amène à prendre la mesure réelle du niveau de difficulté de cette recherche tout en en déterminant les avenues potentielles.

Incidence des erreurs d'omission de signes

Notre corpus présente 1 6 02 erreurs de ponctuation parmi lesquelles 71% (1 139) correspondent à des signes manquants (Fig. 1).

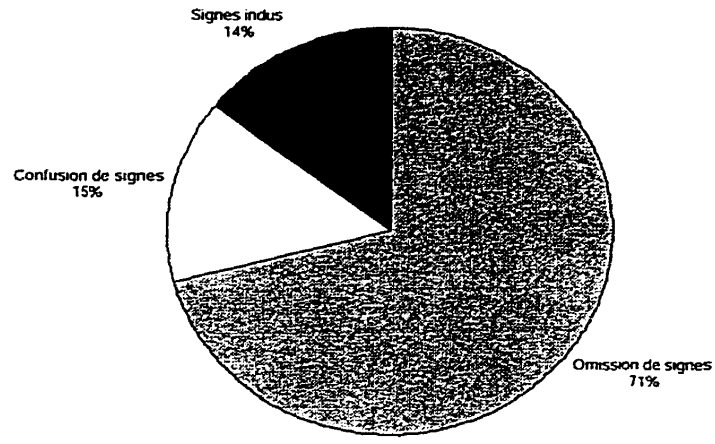


Figure 1

Distribution des erreurs de ponctuation pour l'ensemble du corpus

Parmi les signes omis, c'est la virgule qui représente le signe de ponctuation le plus fréquemment oublié (Tableau 1).

Tableau 1

Place des virgules manquantes dans les erreurs de ponctuation

Corpus	Omission de virgules	Omissions de signes	Erreurs de ponctuation
Moffet	682	916	1339
Experts	168	172	212
Bissonnette	51	51	51
Total	901	1139	1602

Parmi toutes les erreurs d'omission de virgules, c'est l'omission d'une virgule à fonction de délimiteur qui constitue l'erreur de ponctuation la plus fréquente (Fig. 2).

---

5 Voir Chapitre 3. *Méthodologie : Corpus de recherche*

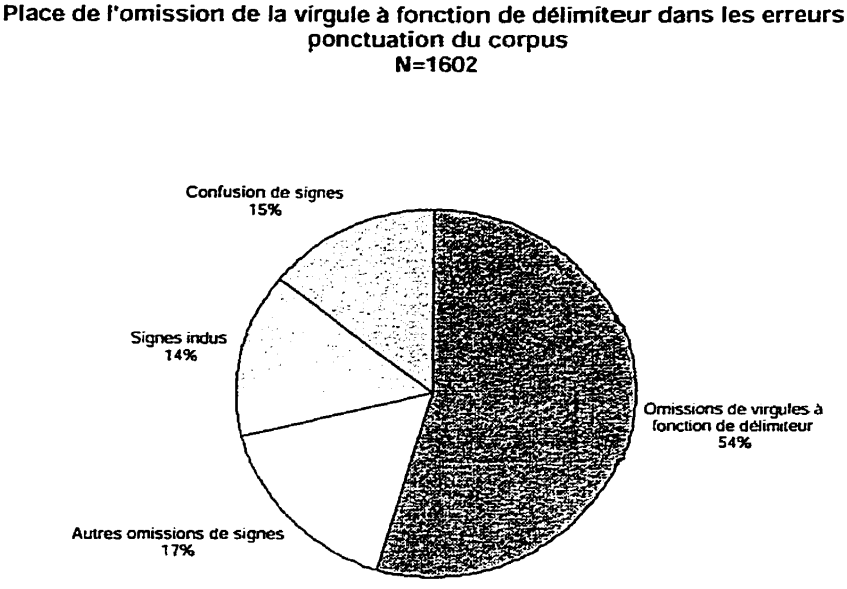


Figure 2

Place des virgules à fonction de délimiteur manquantes dans les erreurs de ponctuation du corpus

La prépondérance des erreurs d'omission de signes pose donc le vrai problème de la détection automatique des erreurs de ponctuation : comment détecter la non-ponctuation? Autrement dit, comment repérer automatiquement les contextes dans lesquels un signe oublié devrait être introduit?

Cette question recadre en fait la problématique réelle du traitement automatique de la ponctuation. En effet, comme la non-ponctuation constitue l'erreur graphique la plus fréquente, il faudrait en fait développer un moyen, non pas pour détecter automatiquement les erreurs de ponctuation, mais pour ponctuer automatiquement un texte.

Voilà qui nous amène tout droit à la discussion de l'utilité de la ponctuation en analyse automatique de texte. Puisque, traditionnellement, la ponctuation a toujours été automatiquement retirée du texte avant son traitement (Jones, 1996c), il faudrait considérer le texte ainsi produit comme le texte à travailler et limiter nos efforts à l'introduction automatique des signes de ponctuation.

En théorie donc, le problème de détection automatique des erreurs de ponctuation serait un faux problème. Il ne servirait à rien en effet de dépister les erreurs de ponctuation (ou erreurs du système graphique [d'après Nunberg, 1990[6]]) d'un texte à partir du moment où nous arriverions à le ponctuer automatiquement : peu importe l'erreur réalisée par l'utilisateur, la version révisée du texte ne présenterait que la ponctuation correcte, si bien qu'à la limite, nous pourrions considérer la ponctuation originale comme non pertinente.

Or Jones (1996c) et Briscoe (1994; 1996b), entre autres chercheurs, ont démontré l'importance de la ponctuation dans la désambiguïsation des unités analysables lors de l'analyse automatique de texte. Jones (*ibid.*) et Briscoe (*ibid.*) soulignent d'ailleurs l'amélioration significative de la performance des grammaires utilisant la ponctuation pour interpréter les segments de texte. Par conséquent, l'absence de nombreux signes requis diminue considérablement le pouvoir de détection d'un analyseur en raison des problèmes d'ambiguïté auxquels il fait maintenant face.

Dans la pratique donc, la prépondérance des erreurs d'omission de signes complexifie de façon significative le processus de détection d'erreurs de ponctuation à cause de l'impact négatif de ces erreurs sur la performance de l'analyseur. Autrement dit, c'est au moment où l'analyseur devrait le plus pouvoir interpréter correctement les contextes ambigus qu'il se trouve le moins en mesure de le faire.

Le rôle des virgules dans la détermination du sens avec les propositions relatives explicative et déterminative constitue un exemple typique de ce phénomène. Considérons par exemple

---

[6] Nunberg (1990 : 17) décrit la ponctuation comme un ensemble de signes graphiques:

> *a set of non-alphanumérical characters that are used to provide information about structural relations among elements of a text, including commas, semicolons, colons, periods, parentheses, quotation marks and so forth. From the point of view of function, however, punctuation must be considered together with a variety of other graphical features of the text, including font- and face-alternations, capitalizations, indentation and spacing, all of which can be used to the same sorts of purposes. From here on, I will talk about all these graphical devices as instances of* text-category indicators *of written language.*

[1] Il a critiqué ses amis, qui l'avaient trahi, avant de quitter la soirée.

[2] Il a critiqué ses amis, qui l'avaient trahi avant de quitter la soirée.

Comment savoir, en cas d'absence du membre droit de la paire de virgules explicatives, si la relative s'achève ou non après « trahi »? Qui a quitté la soirée : celui qui a critiqué ses amis ou les amis qui l'avaient trahi? C'est la présence ou l'absence de la virgule qui permet d'interpréter le sens.

Par ailleurs, une paire de virgules à fonction de délimiteur rend possible, on s'en souviendra, la distinction entre les emplois explicatif et déterminatif des propositions relatives. Soit

[3] La blonde, qui discutait avec mon frère, m'a salué.

[4] La blonde qui discutait avec mon frère m'a salué.

[3] et [4] traduisent deux renseignements bien différents. Dans le premier cas, une blonde a salué le rédacteur pendant qu'elle discutait avec le frère de celui-ci; le sens correspond à une structure "La blonde m'a salué et elle discutait avec mon frère". La relative explicative dans [3] constitue une *structure périphérique* (Chanod, 1993) syntaxiquement évacuable. Dans le deuxième cas, c'est la blonde qui discutait avec le frère du rédacteur qui l'a salué, et non pas d'autres blondes. La relative déterminative dans [4] fait partie intégrante d'un SN se réécrivant SN —> SN + P et ne peut être évacuée.

Nous n'avons pas de raison de douter du sens favorisé par l'auteur dans les exemples que nous venons d'apporter. Cependant, qu'arrive-t-il si nous commençons à douter que la ponctuation soit correcte? Manque-t-il une virgule à [2]? Les virgules de [3] sont-elles de trop? Manque-t-il la paire de virgules à [4]? Il n'est pas possible de répondre à ces questions à moins de comprendre le sens de la phrase et son contexte, ce qu'il n'est pas encore possible de faire aujourd'hui avec un analyseur.

La catégorisation grammaticale des structures dans les suites [1] et [2] devient vulnérable à l'erreur si elle dépend de la ponctuation réalisée en raison de l'ambiguïté du contexte. Cet étiquetage devient carrément inexact si l'analyseur considère [3] et [4] comme équivalentes parce que la ponctuation, retirée dans [3]

pour faciliter l'analyse, rend maintenant les deux suites en tout point semblables. Dans tous les cas, la détection de ces erreurs de ponctuation devient pratiquement impossible à réussir automatiquement en raison de l'écueil de l'interprétation du sens.

L'une des solutions possibles consisterait bien sûr à ignorer, malgré leur fréquence, les erreurs d'omission de signes. Nous pourrions par exemple concentrer nos efforts sur la détection des autres erreurs de ponctuation, celles qui touchent les signes apparaissant effectivement dans le texte. Ces erreurs, qui constituent 29% de l'ensemble des erreurs de ponctuation du corpus, se divisent en proportions à peu près égales (Fig. 2) entre la présence abusive d'un signe de ponctuation et la confusion d'un signe avec un autre.

Dans le cas de la ponctuation abusive cependant, le problème de la vulnérabilité de l'analyseur à l'ambiguïté du contexte se pose toujours. En effet, pour déterminer l'erreur, il faudrait réussir une analyse en comparant les suites avec et sans le signe suspect et nous demander laquelle correspond à l'intention de l'utilisateur et respecte la syntaxe française. Comme nous venons de le démontrer avec nos exemples, seul le sens permettrait un diagnostic juste dans certains contextes.

Quant au problème de la détection des erreurs de confusion de signes, il pose une difficulté apparemment incontournable, que nous examinerons dans un instant.

La détection des erreurs de ponctuation constitue donc un problème très difficile, non seulement à cause de la fréquence élevée de ces erreurs dans les textes de rédacteurs occasionnels, mais surtout en raison de leur impact négatif sur le processus de l'analyse syntaxique automatique lui-même.

Impact des erreurs de ponctuation sur la segmentation automatique du texte

Pour déterminer automatiquement les limites d'une phrase, il est commun de la définir par des signes graphiques aisément repérables par un ordinateur : généralement la majuscule à l'initiale de phrase et le point à la finale. Bien que cette

convention introduise en soi plusieurs problèmes complexes comme, entre autres, la désambiguïsation des divers emplois des majuscules et du point[7], il reste qu'elle est réalisée actuellement par les correcteurs grammaticaux en s'appuyant sur la ponctuation, assumée correcte, déjà présente dans le texte.

Cependant, plusieurs erreurs de ponctuation modifient de façon significative la syntaxe de la phrase. Nous en avons identifié sept :

* l'omission du point;
* l'introduction d'un point abusif (assertif ou expressif);
* l'emploi d'une virgule au lieu du point;
* l'emploi d'un point au lieu du deux-points;
* l'emploi d'un point au lieu d'une virgule;
* l'emploi d'un point au lieu d'un point-virgule;
* l'emploi d'un signe autre que la virgule au lieu du point.

De telles erreurs ont une incidence sur la définition automatique des frontières de phrase, et donc sur le processus essentiel de la segmentation du texte en unités analysables (ce que Briscoe (1994) appelle *chunking*).

Il importe donc de distinguer deux types d'erreurs de ponctuation : les erreurs avec impact sur le traitement automatique de la langue (que nous désignerons à partir de maintenant par « type I ») et toutes les autres (que nous appellerons à présent « type II »). Les premières devraient en effet être théoriquement corrigées pour permettre la correction des secondes, puisque la segmentation d'un texte en unités analysables, généralement les phrases, devrait idéalement s'effectuer à partir de suites constituant réellement des phrases. Il s'ensuit donc que le problème des signes de ponctuation requis mais absents, particulièrement des virgules à fonction de délimiteur, (l'une des catégories d'erreurs de type II), s'il reste le plus fréquent (Fig. 2), demeure secondaire par rapport au problème plus sérieux des erreurs de ponctuation avec impact sur la définition automatique des frontières de phrase (type I).

---

7 D'autres chercheurs se sont penchés sur ces questions particulières. Par exemple Dister (1998) propose un transducteur pour désambiguïser le point et permettre le repérage automatique de la fin des phrases françaises et Palmer et Hearst (1997), un algorithme pour désambiguïser automatiquement les frontières de phrase à partir des majuscules et du point.

Nous espérions que les erreurs de type I soient exceptionnelles. Il est vrai que, dans le bassin des erreurs de ponctuation relevées dans l'ensemble de notre corpus, seulement 12% d'entre elles tombent dans cette catégorie. Malheureusement, ces erreurs se retrouvent dans près des deux tiers (63%; 47 textes sur 75) des productions non professionnelles, ce qui étend considérablement l'impact du problème.

Par ailleurs, l'examen de la distribution de ces erreurs traduites en format logique[8] (Fig. 3) montre que 42% des erreurs de type I produites par les rédacteurs occasionnels impliquent la confusion entre la virgule et le point (*virgule au lieu du point*, 22%; *point au lieu de virgule*, 20%).

[8] Rappelons qu'une donnée en format logique s'exprime par l'opposition 0 (non), 1 (oui) par contraste avec le format nombre usuel. Nous exprimons le nombre d'erreurs de ponctuation de type I en format logique parce que ce format reflète mieux leur distribution dans notre corpus : une erreur en format logique équivaut en effet à un texte, puisque toutes les occurrences d'une même erreur sont comptées une fois seulement pour chaque texte. Les fréquences sont par conséquent calculées à partir de l'ensemble du corpus.

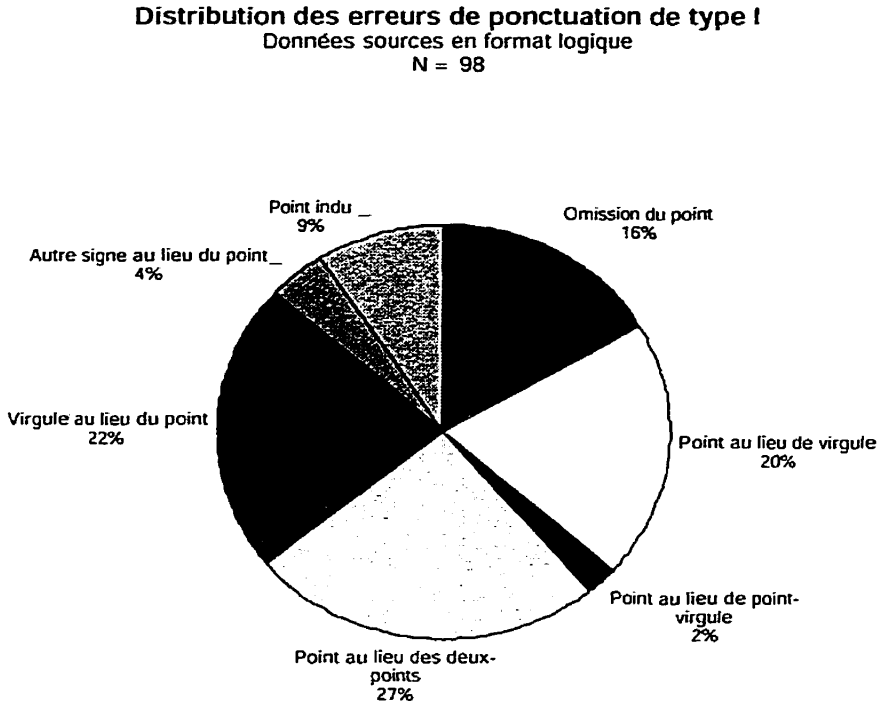


Figure 3

Distribution des erreurs de ponctuation avec incidence sur la définition automatique des frontières de phrase: données en format logique

La fréquence élevée de textes présentant des erreurs de ponctuation de type I nous oblige à remettre en question les méthodes mêmes de segmentation de texte, puisque les chances sont bonnes que, dans une production donnée, au moins un signe manquant, abusif ou mal placé — et souvent plusieurs — auront un impact négatif sur la segmentation du texte.

La difficulté consiste ici à évaluer la validité de la ponctuation présente, ce qui pose en soi une difficulté encore plus importante. En effet, puisque la segmentation valide d'un texte présentant des erreurs de type I n'est pas possible sans préalablement identifier les signes qui ne sont pas fiables, il faudrait avoir détecté ces signes et les avoir corrigés avant de procéder, non seulement au repérage des autres types d'erreurs de ponctuation, mais également au repérage des erreurs du texte en général.

Autrement dit, l'analyse automatique d'un texte pour fins de correction devrait idéalement procéder à deux exercices de segmentation : un premier, considéré comme préliminaire, pour permettre le dépistage et la correction des erreurs de type I et un deuxième, pour assurer la détection et la correction des autres erreurs du texte, erreurs de ponctuation comprises.

Par ailleurs, l'examen de la distribution des erreurs de type I fait ressortir que, si elles sont très probables dans un texte de rédacteur occasionnel, elles ne sont pas nombreuses proportionnellement à l'ensemble des autres erreurs de ponctuation du texte. Cela signifie donc que, en plus de devoir identifier les erreurs de type I avant de segmenter un texte, il faudrait encore, pour pouvoir les dépister, avoir été capable de les distinguer parmi toutes les autres erreurs de ponctuation de l'utilisateur, ce que par ailleurs nous ne pouvons faire sans segmenter le texte.

Nous pourrions toujours arguer que, vu la proportion réduite des erreurs de type I présentes dans un texte donné, il ne sert à rien de s'arrêter à ce problème. Cependant, comme il nous est encore impossible de prédire si un texte donné comporte ou non ce type de problème et combien d'erreurs de type I il peut comporter exactement quand il en comporte, nous ne pouvons réellement prédire non plus jusqu'à quel point la segmentation automatique du texte effectuée sans tenir compte de ces occurrences peut être fiable, et subséquemment jusqu'à quel point les détections réalisées et les corrections suggérées en se basant sur cette segmentation sont utilisables.

Existe-t-il une différence significative entre la ponctuation des rédacteurs occasionnels et celle des rédacteurs professionnels? Oui, si nous en croyons l'examen de notre corpus.

Notre corpus révèle en effet que les erreurs de ponctuation se retrouvent de façon significativement plus importante chez les rédacteurs occasionnels que chez les rédacteurs professionnels (Fig. 4).

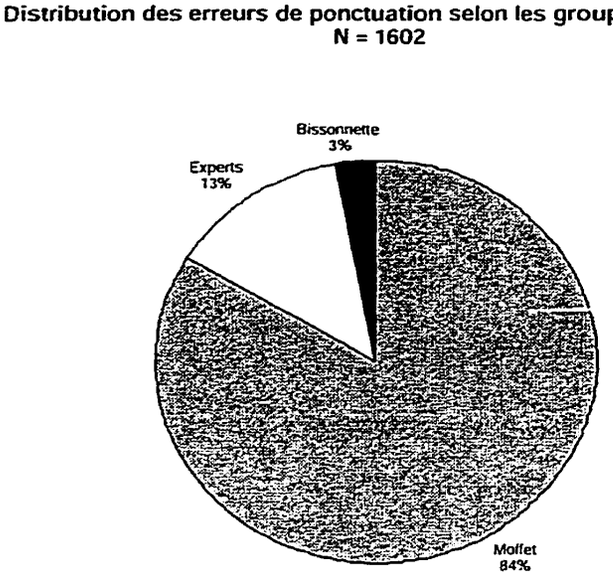


Figure 4

Distribution des erreurs de ponctuation selon les groupes de textes du corpus

Bien que des erreurs de ponctuation aient été relevées dans tous les groupes de textes, plus de 8 erreurs de ponctuation sur dix l'ont été dans un texte écrit par un rédacteur occasionnel.

Nous avons vu que, dans notre corpus, ce sont les erreurs d'omission de signes qui sont les plus fréquentes. Or les résultats des erreurs d'omission de signes, traduits en fréquences relatives (Tableau 2), indiquent une progression significative dans la proportion des erreurs d'omission par rapport aux autres catégories d'erreurs de ponctuation selon les rédacteurs du corpus.

Tableau 2

Proportion des erreurs d'omission de virgules et d'omission de signes dans l'ensemble des erreurs de ponctuation

Corpus[9]	Omission de virgules	Omissions de signes
Moffet	50%	68%
Experts	74%	81%
Bissonnette	100%	100%

Les rédacteurs occasionnels commettent donc plus souvent d'autres types d'erreurs de ponctuation (confusion de signes, ponctuation abusive) que les rédacteurs professionnels.

Cette progression est également visible dans les erreurs d'omission de virgules. Alors que la moitié des erreurs de ponctuation du corpus Moffet concernent des virgules absentes mais requises par le contexte, près des trois-quarts des erreurs de ponctuation du corpus « Experts » tombent dans cette catégorie alors que ce sont toutes les erreurs de ponctuation relevées dans les textes Bissonnette qui concernent des virgules omises.

Les signes de ponctuation placés par les rédacteurs occasionnels dans leurs textes l'ont été de façon incohérente. Nous pouvons démontrer cet aspect particulier du problème en étudiant les écarts-types des résultats en ponctuation pour les trois groupes de textes.

L'écart-type (représenté, pour une population statistique, par la lettre grecque σ) constitue en effet une mesure de dispersion des résultats par rapport à la moyenne (représentée, pour une population statistique, par la lettre grecque μ). Racine carrée d'une autre mesure de dispersion, la variance, l'écart-type *permet de retrouver les unités de mesure dans lesquelles s'exprimaient les observations originales* (Allaire, 1998 : 8-2). Or plus les écarts sont élevés entre les résultats de chaque individu et la moyenne du groupe, moins cette moyenne est représentative des résultats de

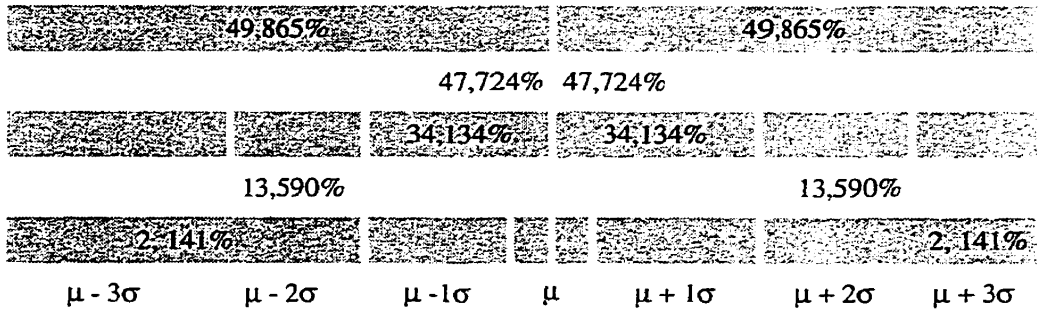
[9] Voir Chapitre 3. *Méthodologie.*

l'ensemble. L'écart-type, qui traduit la moyenne des écarts à la moyenne, donne donc une bonne indication de leur variabilité.

Par ailleurs, c'est également au moyen, entre autres, de l'écart-type que les statisticiens estiment la position exacte occupée par un sujet dans une distribution normale. Ce calcul, appelé cote Z ou *écarts-réduits* (Allaire, 1998 : 10-1) permet de quantifier l'écart entre un résultat donné et la moyenne du groupe exprimée en nombre d'écarts-types (Allaire, *ibid.*). En effet, selon la courbe normale, chaque résultat s'établit dans une zone située par rapport à la moyenne (Allaire, 1998 : 12-1), comme le résume le tableau 3.

Tableau 3

Distribution des résultats d'une population selon la courbe normale



En d'autres termes, 68,27% (34,134% + 34,134%) des sujets d'une population se distribuant selon la courbe normale présenteront des résultats variant entre plus et moins un écart-type de la moyenne; 95,45%, entre plus et moins 2 écarts-types et 99,73%, entre plus et moins 3 écarts-types (Allaire, *ibid.*). C'est pourquoi les scores Z s'expriment au moyen d'un chiffre variant de -3,091 (c'est-à-dire moins 3,091 écart-types) à +3,091 (c'est-à-dire plus 3,091 écarts-types), situant ainsi chaque résultat dans la zone appropriée de la courbe normale.

Les écarts-types des groupes de textes de notre corpus indiquent une différence significative de la performance en ponctuation des rédacteurs occasionnels par rapport à celle des rédacteurs experts (Tableau 4).

Tableau 4

Moyennes et écarts-types pour les résultats en ponctuation des rédacteurs du corpus

Données	Moffet	Experts	Bissonnette
Erreurs de ponctuation	1339	212	51
μ	18	3	1
σ	12	3	1

L'écart-type (σ) des résultats en ponctuation des experts montre peu de dispersion par rapport à la moyenne (μ). Considérant les calculs du tableau 3, les textes des rédacteurs professionnels pourront présenter de 0 à 12 erreurs de ponctuation, alors que les textes du corpus Bissonnette pourront contenir de 0 à 4 erreurs de ponctuation. Les diagrammes des figures 5 et 6 illustrent en effet une dispersion de données assez similaire aux chiffres estimés.

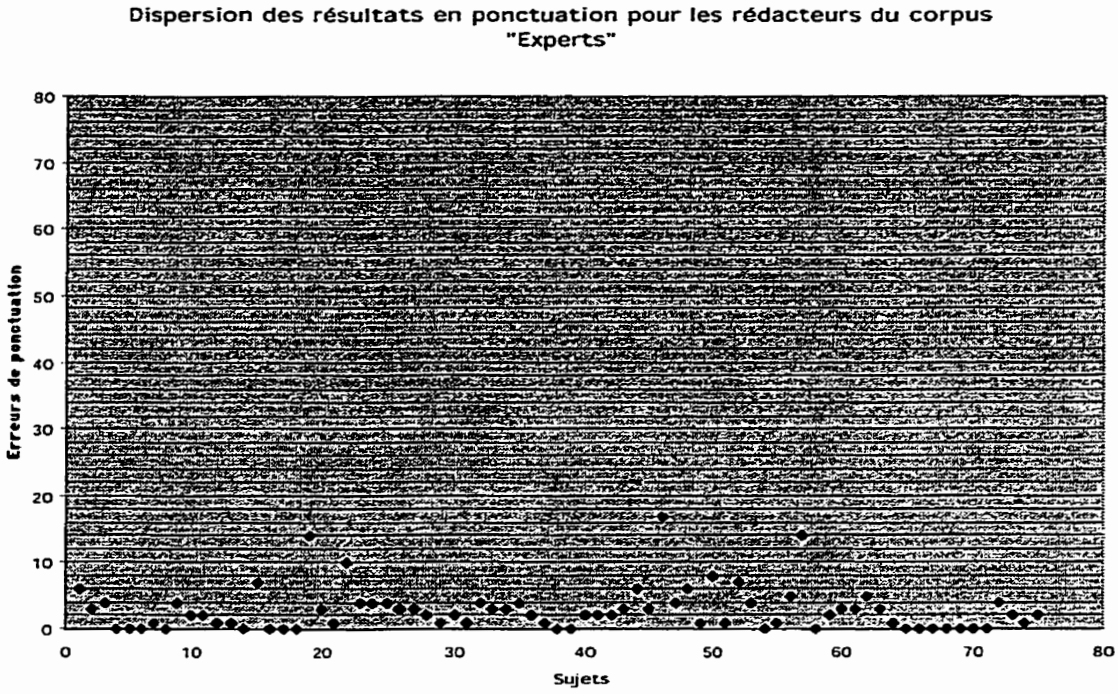


Figure 5

Dispersion des résultats en ponctuation pour les rédacteurs du corpus "Experts"

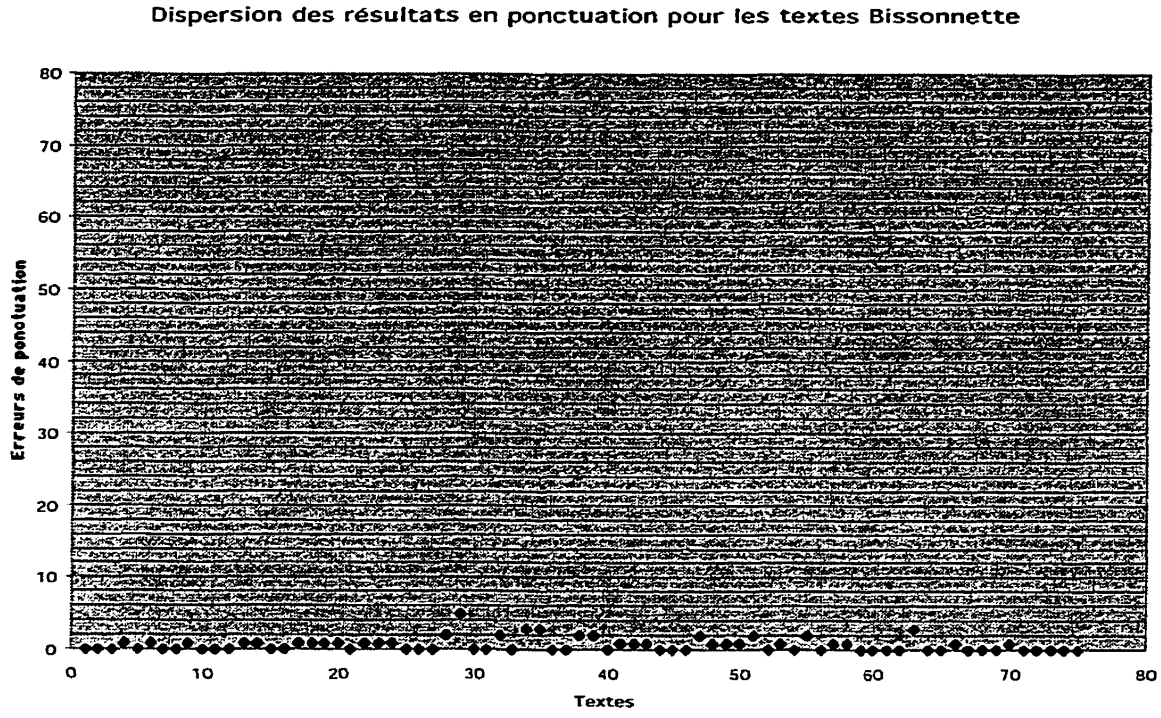


Figure 6

Dispersion des résultats en ponctuation pour les textes Bissonnette

En revanche, l'écart-type des résultats en ponctuation des rédacteurs occasionnels montre une grande dispersion par rapport à la moyenne. En appliquant en effet les calculs du tableau 4 pour les textes Moffet, nous obtenons un éventail de résultats possibles variant de 0 à 48 erreurs de ponctuation. Cet écart autour de la moyenne se révèle encore plus important dans le diagramme de dispersion de la figure 7:

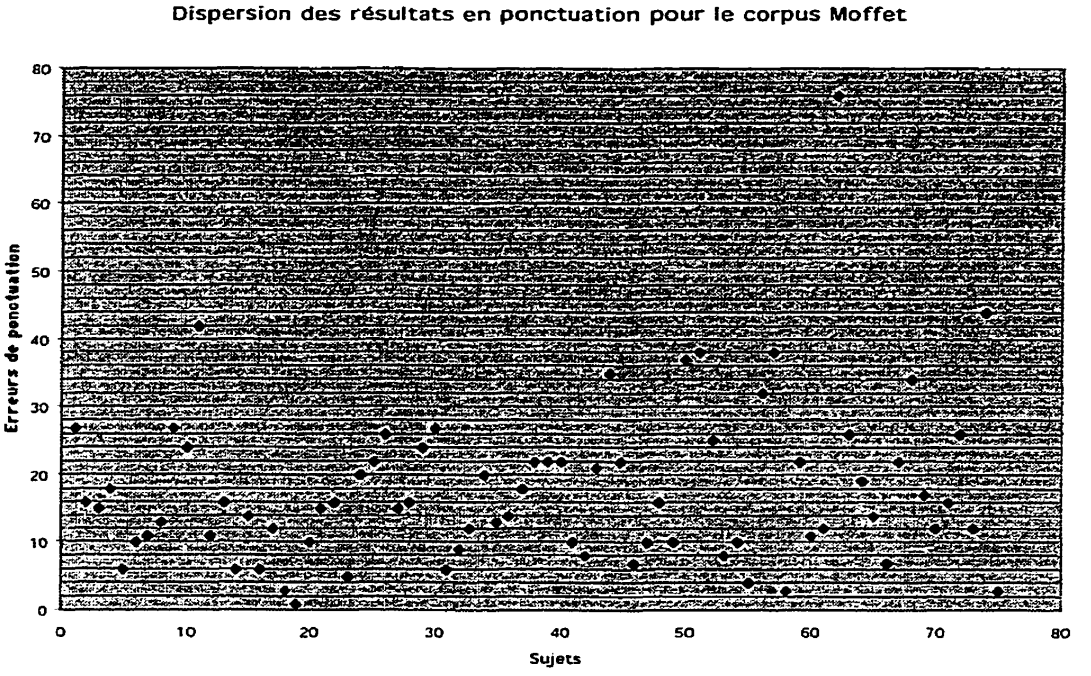


Figure 7

Dispersion des résultats en ponctuation pour le corpus Moffet

La dispersion importante des résultats en ponctuation des rédacteurs occasionnels suggère ainsi l'absence d'un système commun de règles pour placer les signes de ponctuation. Dans l'échantillon Moffet, les signes semblent donc avoir été choisis et placés dans les textes plus ou moins au hasard, ce qui nous amène à conclure que, du moins dans la population des 16 2 00 rédacteurs occasionnels d'où cet échantillon a été tiré, il n'est pas possible de repérer, sur la base d'un système de règles normatives, où une erreur de ponctuation pourra se produire.

Au fond, le repérage automatique de l'erreur de ponctuation est-il même possible? Il semblerait que non. Toutefois, en fermant toutes les portes susceptibles de déboucher sur une solution praticable, nous avons ouvert la proverbiale fenêtre. En effet, nous mentionnions que nous n'avions pas de raison de douter du sens généré par la ponctuation des suites [1], [2], [3] et [4]. Or dans quelles circonstances ne douterions-nous pas de cette ponctuation? Quand nous pouvons raisonnablement supposer que la ponctuation est correcte parce que la production de l'auteur est recadrée dans un contexte extralinguistique d'expertise, tel qu'il arriverait dans le cas de textes publiés sous la supervision d'une équipe de production reconnue comme

celle d'un magazine ou d'un journal. C'est le cas en fait, ainsi que nous venons de le voir, dans les textes professionnels de notre corpus.

Or supposons que quelqu'un nous signale que les suites [1], [2], [3] et [4] peuvent contenir des erreurs, comme il arriverait si ces phrases se retrouvaient, non pas dans des textes de rédacteurs professionnels, mais dans des textes soumis dans le cadre d'une épreuve de français. Soudain l'éclairage change et le doute survient. C'est alors que les problèmes de détection automatique d'erreurs de ponctuation se posent.

Par conséquent, s'il nous était possible d'identifier automatiquement à quel type de rédacteur nous avons affaire, nous pourrions déjà lever le doute sur les textes rédigés par les rédacteurs « professionnels ». Quant aux autres, nous saurions du moins à quoi nous attendre.

1.3.2 Correction automatique de la ponctuation

Les utilisateurs des traitements de texte ont été habitués à une correction automatique en temps réel : les suites jugées incorrectes par le module de révision linguistique sont mises en relief et des suggestions sont apportées à l'utilisateur pour corriger le problème. La décision, et donc, la responsabilité de la correction revient ainsi à l'utilisateur.

Il est également possible de faire une lecture corrective automatique du texte en temps différé. Dans ce cas, le texte à analyser est saisi et soumis au correcteur, qui en fait un diagnostic. Encore une fois, la responsabilité de la correction revient à l'utilisateur.

Nous venons de voir cependant que la correction automatique d'erreurs de ponctuation en temps réel semble problématique, considérant la complexité du problème de détection : 1) repérer l'absence d'un signe requis ou poser un jugement sur la validité d'un signe présent en partant de l'analyse automatique de segments de textes dont les limites peuvent ne pas être fiables et 2) identifier correctement l'interprétation syntaxique appropriée dans un contexte rendu possiblement plus ambigu en raison de l'absence probable de la ponctuation requise. Inutile de préciser

que, dans un tel contexte, la suggestion d'une ponctuation alternative semble bien peu pertinente.

Les exigences posées par les habitudes d'utilisation des correcteurs grammaticaux constituent donc également un élément de notre problématique. Si nous devons en effet suggérer des solutions aux erreurs de ponctuation que nous avons également à repérer en temps réel, il devient carrément impossible de nous baser sur la ponctuation, à la fois pour segmenter le texte et le corriger, même si un diagnostic automatique ne posait en soi aucun problème. Voilà qui nous place devant un mur apparemment insurmontable.

Correction automatique et approche prédictive

Si nous admettons que l'action sur les erreurs déjà produites mène à une impasse, ne pourrions-nous pas trouver, à défaut, un moyen de prédire les contextes où elles risquent de se produire? Une approche prédictive permet en effet de contourner certains problèmes de traitement identifiés.

Nous savons que près des deux tiers des textes de notre échantillon montre des erreurs de ponctuation de type I. Si nous pouvions prédire l'occurrence de ces textes, nous pourrions alors, au lieu de chercher à reconnaître ces erreurs dans le texte même, essayer plutôt d'identifier automatiquement les textes où la segmentation de l'analyseur risque de ne pas être utilisable. Un exercice de correction automatique consisterait alors principalement à diriger l'attention de l'utilisateur sur ce problème.

Par ailleurs, serions-nous en mesure de prédire la fréquence et le type d'erreurs de ponctuation le plus susceptible de se produire dans un texte donné? Si oui, alors notre méthode de correction consisterait à prévenir l'utilisateur des « risques » au moment où les contextes pertinents apparaîtraient. Cette intervention pourrait s'effectuer alors en temps réel, bien que les mesures de prédiction aient pu, elles, se calculer en temps différé, par exemple à partir d'autres textes du même rédacteur.

L'approche prédictive en résolution de problème n'est pas nouvelle. En santé, l'apparition d'une épidémie s'accompagne de l'identification de populations à risque et d'actions préventives au sein de ces populations. À défaut de pouvoir empêcher les tremblements de terre, les sismologues tentent d'identifier les facteurs pouvant permettre de les prédire. Les volcanologues font de même avec les éruptions volcaniques et les météorologues, avec les ouragans et les tornades. Des modèles sont développés et appliqués pour estimer les risques qu'un incident désastreux ne se produise. Tous reconnaissent qu'il s'agit là de projections basées sur des estimations. Pourtant, malgré leurs limites, ces estimations demeurent utiles pour augmenter les chances des humains de survivre aux grands désastres.

Bien que la gravité du problème des erreurs de ponctuation ne constitue bien sûr en rien une « c rise » humaine, elle peut être vue comme telle dans l'univers du texte. Nous pouvons donc nous demander jusqu'à quel point il serait possible de développer et appliquer un modèle mathématique pour nous aider à « p rédire » automatiquement l'état de la ponctuation d'un texte, et peut-être même celui de sa qualité linguistique.

Identification des outils de prédiction

L'approche prédictive soulève en soi toute une nouvelle série de questions.

Ces questions sont d'abord linguistiques. Existent-ils des indices utilisables pour prédire la qualité de la ponctuation d'un rédacteur? Si oui, quels peuvent-ils être? Ont-ils tous le même poids? Sont-ils signifiants séparément ou doivent-ils être considérés dans leur ensemble pour montrer une valeur prédictive?

Les problèmes sont ensuite informatiques. Pourrons-nous repérer automatiquement les indices significatifs? Et si oui, sous quelles conditions et avec quelle mesure de fiabilité? Quelles seront les difficultés propres à cet exercice? Pourrons-nous résoudre ces difficultés en faisant appel aux connaissances actuelles du domaine?

Les problèmes sont également mathématiques. Sera-t-il possible de traduire ces indices en langage mathématique pour qu'un ordinateur puisse les interpréter en apportant des prédictions utiles et exploitables?

Finalement, les questions posées par notre approche prédictive sont d'ordre méthodologique. De quelle façon en effet pourrons-nous mesurer la valeur des prédictions obtenues, si nous en obtenons, et comment les exploiterons-nous pour aider l'utilisateur?

C'est à toutes ces questions que notre recherche tentera de répondre.

Deux
État de la question

La correction automatique de la ponctuation représente un domaine de recherche entièrement nouveau. C'est assez tôt dans notre étude en effet que nous nous sommes rendu compte qu'il n'existait pas beaucoup de littérature scientifique dans ce domaine de recherche. Grâce à des contacts personnels cependant, nous avons quand même réussi à obtenir (et enregistrer) une entrevue avec le développeur d'un correcteur grammatical, le linguiste John Chandioux (1996), mieux connu dans le milieu de la linguistique informatique international pour des logiciels de traduction automatique comme *Météo* et le *Général Tao*. Nous n'avons jamais interrompu notre quête cependant, même s'il devenait de plus en plus clair au fil des mois que la recherche active en correction automatique de textes, principalement financée par l'entreprise privée, tombait sous le sceau du secret et n'était tout simplement pas diffusée.

Cependant, le traitement automatique de la ponctuation a commencé à générer de l'intérêt chez les linguistes informaticiens (Jones, 1996d). Palmer et Hearst (1997: 241) soulignent par exemple la difficulté extrême d'interpréter la ponctuation dans l'analyse automatique de textes :

> *Segmenting a text into sentences is a nontrivial task, however, since in English and many other languages the* [period, the exclamation point and the question mark] *are ambiguous. A period, for example can denote a decimal point, an abbreviation, the end of a sentence, or even an abbreviation at the end of a sentence. Exclamation points and question marks can occur within quotation marks or parentheses as well as at the end of a sentence. Ellipsis, a series of periods (...), can occur both within sentences and at sentence boundaries.*
>
> .....
>
> *The existence of punctuation in grammatical subsentences suggests the possibility of a further decomposition of the sentence boundary problem into types of sentences boundaries, one of which would be " embedded sentence boundary."*

Palmer (1994), Palmer et Hearst (1997) et Dister (1998), entre autres, ont tenté d'identifier un moyen de désambiguïser les points et les indices de frontières de phrase. Palmer et Hearst (*ibid.* : 246) proposent par exemple un système, le *Satz system*, interprétant les contextes entourant les marques de fin de phrase, et non les mots eux-mêmes. Le choix de cette interprétation repose sur une analyse de fréquences intégrée au lexique du système, une approche à qui les auteurs attribuent le haut taux de fiabilité des analyses. Dister (1998 : 437) souligne pour sa part la double ambiguïté du point et de la majuscule et propose un transducteur capable de lever les ambiguïtés entre les points assertif et abréviatif.

Considérant le haut niveau de difficulté représenté par l'interprétation correcte des signes de ponctuation, faut-il donc s'étonner que les chercheurs aient traditionnellement pris la décision d'évacuer la ponctuation au moment de l'analyse automatique (Jones, 1995) :

> *There are no current text based natural language analysis or generation systems that make full use of punctuation, and while there are some that make limited use, like the Editor's Assistant (Dale, 1990), they tend to be the exception rather than the rule. Instead, punctuation is usually stripped out of the text before processing, and is not included in generated text.*

Jones (1996a et b; 1994), dans une expérience de parsing avec une grammaire intégrant des règles de ponctuation, a fait ressortir l'utilité d'une grammaire tenant compte de la ponctuation pour l'analyse automatique de phrases complexes *: For the longer sentences of real language, however, a grammar which makes use of punctuation massively outperforms an otherwise similar grammar that ignores it.* Jones va plus loin en doutant du succès d'une grammaire qui ignorerait la ponctuation: *Indeed, it is difficult to see how any grammar that takes no notice of punctuation could ever become successful at analysing such sentences unless some huge amount of semantic and pragmatic knowledge is used to disambiguate the analysis.* L'analyse automatique d'une phrase devrait donc tenir compte de la ponctuation (Briscoe, 1996a; Jones, 1996a et b; Dale, 1990).

En effet, la ponctuation, et spécialement la virgule, joue un rôle déterminant dans la syntaxe et le sens de la phrase. Ce rôle demeure abondamment documenté aussi bien dans la littérature du domaine (entre autres, Briscoe, 1996a; Jones, 1994; Catach, 1994; Nunberg, 1990; Védénina, 1989, Fonagy, 1980) que dans la littérature

normative (entre autres, Ramat, 1994; Colignon, 1993; Drillon, 1991; Brun et Doppagne, [1958], Damourette, 1939).

Nous nous trouvons ainsi devant un problème apparemment insoluble. D'une part, l'intégration des signes de ponctuation dans l'analyse complexifie l'exercice déjà difficile de l'analyse automatique robuste tel que documenté par, entre autres, Dale (1996), Briscoe (1996b) et Jones (1996c). D'autre part, sans la ponctuation, le résultat d'une telle analyse demeure médiocre. Or ce problème s'appuie quand même sur un présupposé : la ponctuation à interpréter est correcte.

Mais qu'arrive-t-il si la ponctuation du texte est incorrecte?

En novembre 1993, la direction générale de l'Enseignement collégial (DGEC) du ministère de l'Enseignement supérieur et de la Science publiait une analyse des erreurs détectées dans les 20 2 41 copies des candidats à l'université soumis à l'épreuve de français écrit du mois de mars 1993. Cette analyse révèle que 92,75% des copies contenaient des erreurs de virgule.

Pour faire face aux difficultés d'emploi des signes de ponctuation, les rédacteurs disposent de trois types d'outils : les traités consacrés à la ponctuation et autres guides d'écriture ou de correction; les grammaires normatives et les manuels scolaires et les correcteurs grammaticaux, logiciels utilitaires indépendants ou intégrés dans les texteurs comme aide à la production.

Les traités de ponctuation, guides d'écriture et grammaires usuelles sont incapables de proposer des règles d'emploi sans se contredire les uns les autres (Catach, 1994; Simard, 1993; Nunberg, 1990). Quant aux détecteurs orthographiques,

la ponctuation y constitue un "très gros problème" (John Chandioux[10], 23 janvier 1996) :

> *De façon très étonnante, la ponctuation se révèle une source fréquente d'erreurs dans les textes rédigés, même par les professionnels de l'écriture. Ces erreurs sont à la source de nombreuses fausses détections et fausses analyses effectuées par les détecteurs orthographiques, et GramR ne fait pas exception. Les détecteurs sont souvent dépendants de la ponctuation pour réussir la lecture des phrases qu'ils traitent. Quand la ponctuation est erronée, ce qui arrive extrêmement souvent, on se trouve devant un très gros problème.*

La reconnaissance par M. Chandioux de l'influence négative des erreurs de ponctuation sur la performance des correcteurs orthographiques nous ramène au problème complexe de la ponctuation dans le traitement automatique de la langue.

Face à l'absence de littérature scientifique en matière de correction automatique de la ponctuation, devons-nous nous déclarer sans ressources pour autant? Non, puisque les traces de ces recherches se manifestent dans les logiciels d'aide à l'écriture. En effet, au moyen d'exercices de rétroingénierie (en anglais, *reverse engineering*), nous pouvons induire les décisions mises de l'avant dans le traitement automatique des erreurs de ponctuation en observant le « comportement » des correcteurs grammaticaux.

En 1997, nous nous sommes arrêtés au fonctionnement de deux correcteurs choisis en raison de leur renommée et de leur popularité : la grammaire française de *Word7*®[11], (*Hugo*) et *Antidote*®[12]. Les résultats de cette première étude ont été publiés dans la 33° édition de la Revue Informatique et Statistiques dans les Sciences humaines (RISSH) sous le titre « Du traitement automatique de la ponctuation». Cette étude a été enrichie et mise à jour en 1999 avec de nouveaux logiciels : la grammaire de *Word* 98, le *Correcteur 101*®[13] et la version la plus récente d'*Antidote*. Cette

---

10 John Chandioux, président directeur général du Groupe John Chandioux inc., siégeait, au moment de notre entrevue avec lui, au comité scientifique et au comité directeur de l'Institut canadien linguistique de Moncton de même qu'au comité directeur et au comité international de l'observatoire québécois des Industries de la langue. Il est l'auteur de plusieurs systèmes de traduction automatique (*Météo-2*, *Général Tao* et *Lexicum*) déjà implantés dans les entreprises et les gouvernements, du langage de programmation GramR et du détecteur orthographique GramR mis en marché au Canada, en France et en Belgique. John Chandioux a été rencontré par l'auteure le 23 janvier 1996 dans les bureaux de son entreprise à Montréal.

11 *Word* est une marque déposée de Microsoft Corporation.

12 *Antidote* est une marque déposée de Druide informatique.

13 *Le Correcteur 101* est une marque déposée de Machina Sapiens.

seconde étude est intégrée dans le présent chapitre sous le titre « Du traitement automatique de la ponctuation : mise à jour».

Les deux études font ressortir la dépendance des analyseurs syntaxiques face à la ponctuation du texte de même que l'incidence négative des erreurs de ponctuation sur la fiabilité du repérage et de la correction automatiques des erreurs tout court.

2.1 De la correction automatique de la ponctuation[14]

L'usage des virgules constitue, en français, le problème de ponctuation majeur chez les rédacteurs professionnels. D'une part, les grammairiens ne s'entendent pas sur la norme à respecter en matière de ponctuation (Catach, 1994; Simard, 1993), si bien qu'il devient difficile de s'y retrouver, et d'autre part, en conséquence de ce flottement, l'emploi de la virgule semble faire la plupart du temps l'objet d'une décision arbitraire. En outre, la ponctuation, souvent ressentie comme une signature stylistique, s'accompagne souvent de fantaisies personnelles : certains rédacteurs « virguliseront» leurs textes avec générosité alors que d'autres utiliseront les virgules avec parcimonie. Pourtant, entre ces deux écarts, tous s'entendront pour reconnaître l'existence de virgules qu'il aurait été préférable d'éviter tout autant que de virgules qu'il aurait fallu utiliser.

Parmi tous les problèmes d'orthographe compliquant l'écriture du français, celui de la virgule pourrait sembler superflu. Pourtant tel n'est pas le cas. Dans le monde de la publication seulement, où les critères stricts des standards des maisons d'édition se doublent des critères propres aux règles stylistiques, syntaxiques et typographiques du français écrit, le processus de correction d'épreuves consomme beaucoup de temps — et d'argent — et requiert l'expertise d'un grand nombre de professionnels de l'écriture. Souvent, la révision consommant le plus de temps et d'énergie s'applique à la vérification des erreurs de bas de gamme comme les erreurs de ponctuation (Dale, 1990). C'est pourquoi la conception d'un outil présentant des fonctions de correction d'erreurs de ponctuation, même partiellement automatisée, permettrait d'améliorer l'efficacité des rédacteurs d'expression française durant

l'étape de réécriture et contribuerait à diminuer les coûts de production du matériel imprimé. Or le développement réussi d'un tel outil semble bien improbable à l'heure actuelle si nous considérons la piètre performance des correcteurs grammaticaux en face des erreurs de ponctuation.

Nous avons étudié deux correcteurs grammaticaux en raison de leur renommée et de l'étendue de leur utilisation : *Antidote*® développé et mis en marché par la compagnie Druide informatique[15], et la grammaire française de *Word7*®, *Hugo*©, développée par Logidisque inc. et mise en marché par Softissimo.

2.1.1 Correcteurs étudiés

Antidote®

Antidote® représente la toute dernière génération des outils de révision de textes. Ce logiciel propose un ensemble intégré de quatre outils linguistiques (un correcteur grammatical, un dictionnaire en ligne, un conjugueur et une grammaire des difficultés de la langue française) pouvant travailler à partir des texteurs les plus populaires et être consultés en mode interactif. Depuis son lancement en 1996, *Antidote*® s'est mérité plusieurs prix, particulièrement le Mérite du français de l'Office de la langue française du Québec et une place de finaliste, en 1997, au prestigieux concours d'excellence du Canada français, le gala des Octas, organisé par la Fédération de l'informatique du Québec et parrainé par les plus grandes entreprises en technologies de l'information, dont IBM et Microsoft. Le plus récent exploit d'*Antidote*® a consisté à devancer de façon significative Le *Correcteur 101*®, son plus proche rival, et le correcteur de *Word*® lors d'une étude indépendante menée par la grammairienne et auteure Marie-Éva de Villers, qui voulait comparer la performance des trois correcteurs les plus populaires sur le marché en les soumettant à une épreuve spécialement montée à cet effet.

---

[14] L'article est reproduit avec la permission des éditeurs du RISSH, avec cependant quelques adaptations pour le format du présent document.

[15] Pour toute information technique, visiter *www.druide.com*.

Le correcteur d'*Antidote®* permet de régler les paramètres de révision de textes. Ces paramètres touchent surtout les aspects lexicaux et stylistiques de la langue sans faire état de la ponctuation (Fig. 8).

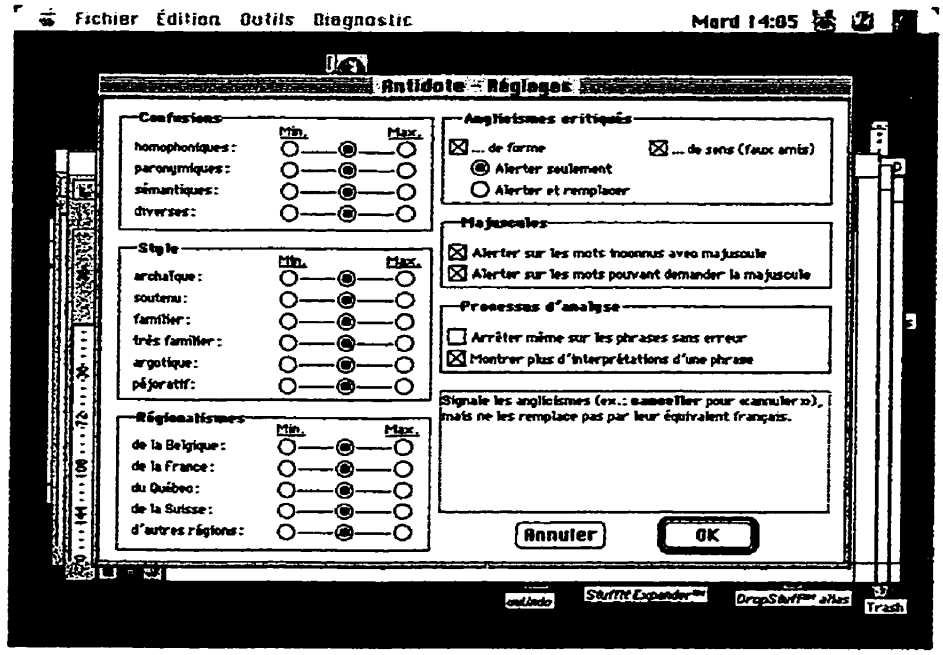


Figure 8

Réglage d'Antidote

Antidote® n'annonce donc pas qu'il traitera les erreurs de ponctuation. Il tente néanmoins de le faire dans les contextes d'écriture soumis et intègre plusieurs remarques sur la ponctuation dans les explications de ses « alertes».

Hugo© dans Word 7®

Après avoir connu un succès commercial sans précédent et s'être également mérité plusieurs prix, la grammaire française *Hugo©* a finalement été intégrée par Microsoft Corporation à son texteur *Word®*. Ce texteur, le plus populaire actuellement dans le monde de la bureautique, assure ainsi à *Hugo©* une place de choix dans la francophonie internationale et lui permet sans doute d'être le correcteur grammatical le plus utilisé.

La grammaire de *Word7®* prévoit de nombreux réglages : d'abord sur les niveaux de langue, ensuite sur le style et finalement sur la typographie (Fig. 9).

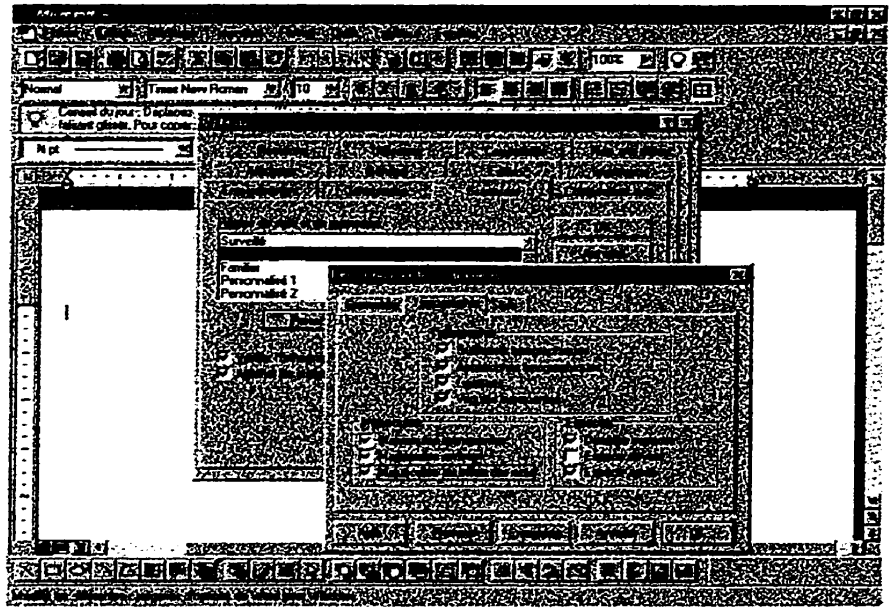


Figure 9

Réglage de la grammaire de Word7

Dans son réglage « Typographie », la grammaire de *Word7®* annonce ainsi qu'elle sera en mesure de détecter et corriger, entre autres, des problèmes de virgules manquantes. Tel ne sera pas le cas cependant dans les contextes d'écriture qui lui seront soumis. Par ailleurs, les mêmes tests effectués en activant le niveau de langue « Surveillé » ont donné les mêmes analyses et les mêmes résultats pour les mêmes contextes d'écriture. Nous pouvons donc penser que l'activation des réglages de niveau de langue dans la configuration « Grammaire » ne semble pas permettre une différentiation notable dans le traitement des textes analysés.

Nous cherchions à placer les deux correcteurs en position de vérifier quelques contextes problématiques d'écriture dans l'espoir de mieux comprendre comment ils traitent les erreurs de ponctuation.

Cependant, il nous a d'abord fallu régler la question toujours débattue en littérature normative de la définition de l'erreur. En appliquant la norme de l'Office de la langue française du Québec, les correcteurs (humains) de l'épreuve annuelle de français du Gouvernement du Québec adressée aux candidats à l'université ont

identifié (Ministère de l'Enseignement supérieur et de la Science, 1993), dans leur échantillon de textes d'étudiants, trois types d'erreurs : des virgules requises par la syntaxe mais absentes (par exemple, une virgule fermante oubliée), des virgules non requises par la syntaxe mais présentes (par exemple, une virgule séparant le sujet de son verbe ou le verbe de son complément d'objet direct) et des virgules substituées à un autre signe (généralement le point). Se fondant sur Catach (1994), Guénette, Lépine et Roy (1995) ont documenté les mêmes catégories d'erreurs dans leur propre corpus :

> *[...] nous avons dépouillé les copies [d'étudiants universitaires] que nous avions à notre disposition, classé les erreurs relevées, en nous interrogeant sur leur nature et leur cause probable. Ce travail a été plus profitable que la consultation de vingt grammaires. Il a entre autres confirmé notre sentiment que les problèmes n'avaient pas pour source unique le manque de connaissances grammaticales, puisque, dans les mêmes copies, on rencontrait, à côté de la faute classique, quantité de verbes et d'adjectifs correctement accordés, de phrases bien construites, etc. S'en est donc trouvée nuancée dans notre esprit la définition du concept de « difficulté », car l'étude des erreurs montrait que les fautes les plus fréquentes — donc les difficultés réelles et quotidiennes pour les étudiants — n'étaient pas nécessairement ce que les grammaires ou les divers ouvrages pédagogiques désignaient comme telles.*

C'est donc à partir de ces trois catégories d'erreurs que nous avons travaillé. Le paragraphe (1) de même que les phrases (2), (3) et (4) représentent des exemples de textes soumis aux correcteurs grammaticaux pour examen.

(1) L'usage des virgules constitue, en français, le problème de ponctuation majeur chez les rédacteurs professionnels. D'une part, les grammairiens ne s'entendent pas sur la norme à respecter en matière de ponctuation (Catach, 1994; Simard, 1993), si bien qu'il devient difficile de s'y retrouver, et d'autre part, en conséquence de ce flottement, l'emploi de la virgule semble faire plus souvent qu'autrement l'objet d'une décision arbitraire. En outre, la présence d'idiolectes d'écriture, où la ponctuation est ressentie comme un signature stylistique, en complique le traitement : certains auront tendance à beaucoup utiliser la virgule alors que d'autres l'utiliseront très peu. Pourtant, entre ces deux écarts, tous s'entendront pour reconnaître l'existence de virgules qu'il aurait été préférable d'éviter tout autant que de virgules qu'il aurait fallu utiliser. C'est pourquoi la conception d'un outil présentant des fonctions de correction des erreurs de virgule, même partiellement automatisée, permettrait d'améliorer sensiblement la rapidité et l'efficacité du rédacteur durant l'étape de réécriture.

(2) Novex, qui a pour mission de diffuser des programmes en haute technologie, a maintenant un portefeuille de 23 placements dans des entreprises.

(3) Je vous donne trois choix : faire ce voyage maintenant, le remettre à plus tard ou accepter de l'argent en échange.

(4) J'ai trois amis : Pierre, Paul et Jacques.

Le paragraphe (1) comporte plusieurs caractéristiques intéressantes. Court et de niveau soutenu, il simplifie d'abord l'analyse des grammaires en leur évitant

d'avoir à interpréter des contextes ambigus en raison d'erreurs d'orthographe ou d'accord. Ensuite, il comporte plusieurs occurrences différentes de la ponctuation française, y compris des parenthèses, des points-virgules et un deux-points, ce qui fournit des contextes diversifiés. Finalement, il présente suffisamment de variations pour permettre l'introduction des trois catégories d'erreurs identifiées.

Par ailleurs, des phrases comme (2), (3) et (4) ont été utilisées à deux fins : soit introduire des emplois aberrants de la virgule, soit modifier l'emploi d'un seul signe — par exemple le deux-points — et voir la réaction des correcteurs. La phrase (4) se distingue de la phrase (3) par l'emploi des noms propres commandant la majuscule.

Cet exercice a permis de faire ressortir des « comportements » particulièrement intéressants de la part des grammaires examinées :

* les correcteurs segmentent les textes à partir de leur ponctuation sans la remettre en question.
* Ils attribuent un pouvoir de segmentation à certains signes seulement.
* Ils ne peuvent généralement pas détecter avec fiabilité les erreurs de ponctuation.

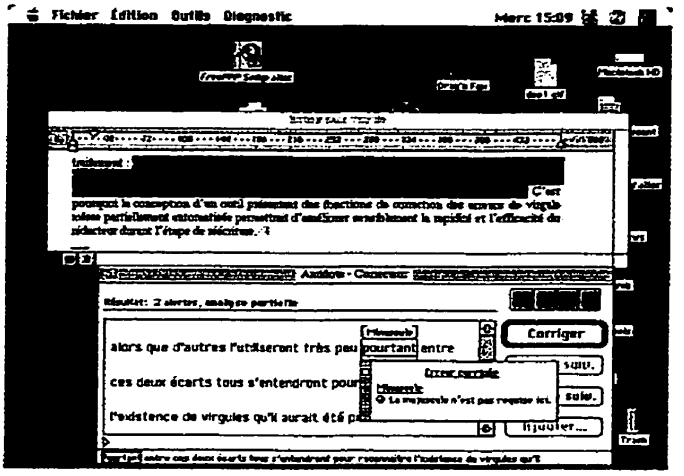
2.1.2 Segmentation des textes à partir de la ponctuation de leur auteur

Les deux correcteurs examinés ont réagi de la même façon face à l'effacement, la substitution ou l'utilisation aberrante d'un signe de ponctuation : ils ont accepté les suites et les ont segmentées en conséquence. Cette segmentation est facilement reconnaissable dans la saisie effectuée par le correcteur au moment de la révision du texte de même que dans le résultat de l'analyse effectuée.

Effacement d'un point

Un point a été effacé dans le paragraphe (1), réunissant la troisième phrase à la quatrième. Pour faciliter le travail des correcteurs, la majuscule qui commençait cette dernière suite a été maintenue comme indice (Fig. 10).

Antidote®



La grammaire de *Word7®*

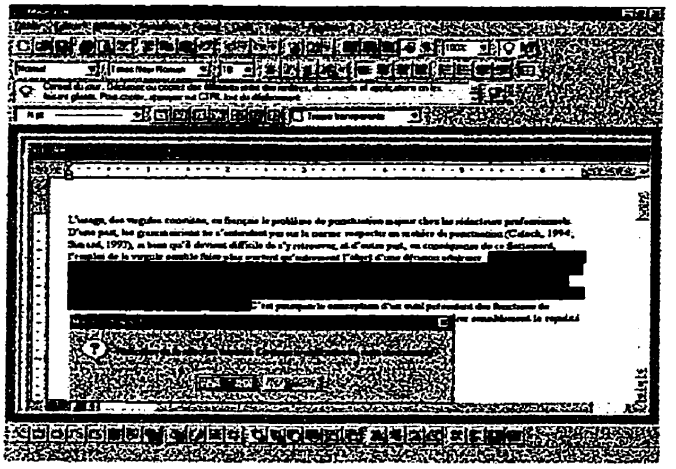


Figure 10

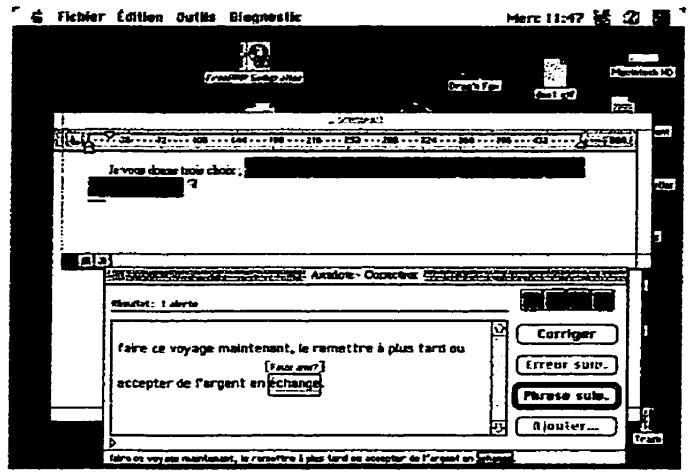
Diagnostic après effacement d'un point

Le point enlevé, *Antidote®* considère la majuscule de « Pourtant » comme une erreur et recommande la minuscule sans envisager la possibilité qu'un point peut avoir été oublié; la grammaire de *Word7®* accepte l'ensemble et ignore la majuscule de « Pourtant » (la reconnaissance des erreurs de majuscule faisait pourtant partie des réglages activés en préférence). Les deux correcteurs segmentent les textes à partir des points, analysant la phrase de sa majuscule initiale jusqu'à son point final.

Substitution d'un deux-points par un point-virgule

La phrase (3) comporte un deux-points que nous remplaçons par un point-virgule. L'analyse est demandée pour l'ensemble de la phrase (Fig. 11).

Antidote®



La grammaire de *Word7®*

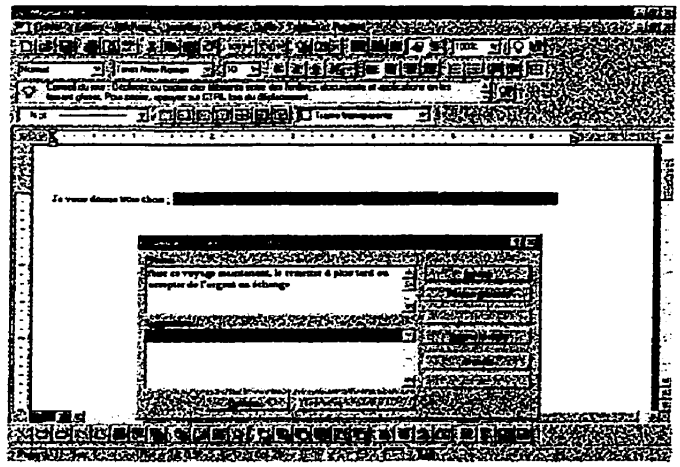


Figure 11

Diagnostic après substitution d'un deux-points par un point-virgule

Une fois le point-virgule placé cependant et malgré la demande de l'utilisateur, la phrase est segmentée en deux parties analysées séparément : de la majuscule au point-virgule d'abord; du point-virgule au point ensuite. Nous verrons plus loin que ce type de segmentation portera des conséquences dans certains contextes comme

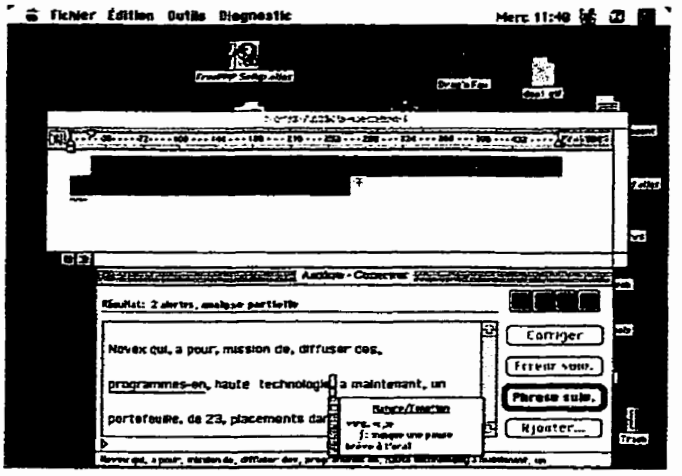
lorsqu'une référence bibliographique de type (Nom d'auteur, date; nom d'auteur, date) est introduite dans le corps du texte.

Introduction d'une virgule placée à tous les deux mots

Chacun de nos correcteurs accepte (2*), où une virgule a pourtant été ajoutée systématiquement à tous les deux mots sans tenir compte des relations entre les suites syntaxiques et leurs connecteurs (Fig. 12).

(2*) Novex qui, a pour, mission de, diffuser des, programmes en, haute technologie, a maintenant, un portefeuille, de 23, placements dans, des entreprises.

Antidote®



La grammaire de *Word7®*

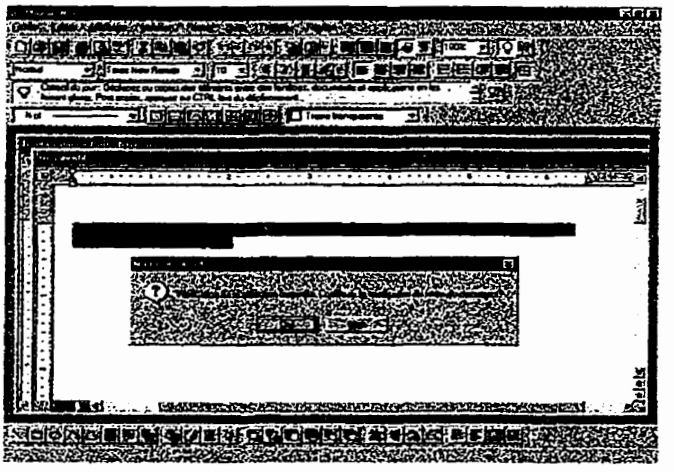


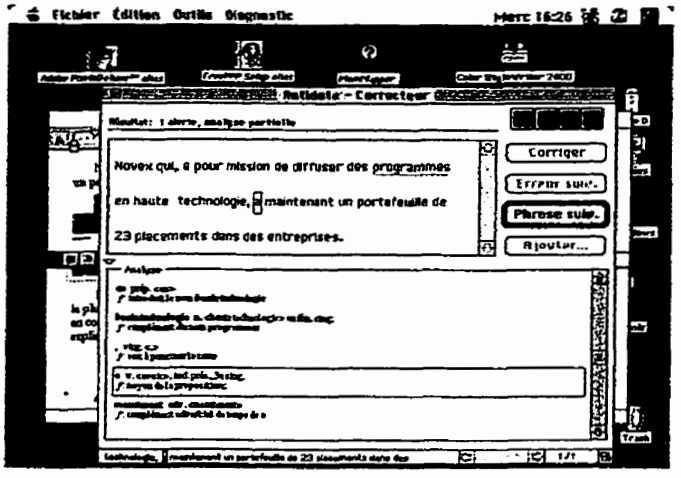
Figure 12

Diagnostic après introduction d'une virgule aberrante

La segmentation diffère cependant avec les correcteurs. La grammaire de *Word7®* segmente la phrase en ignorant les virgules pour considérer l'ensemble comme l'unité à analyser et détermine ensuite que la phrase ne présente pas d'erreurs. *Antidote®*, au contraire, segmente chaque unité à partir des virgules. Cette segmentation apparaît clairement quand une analyse détaillée lui est demandée (Fig. 13) :

Antidote®

Analyse détaillée de (2)



Analyse détaillée de (2*)

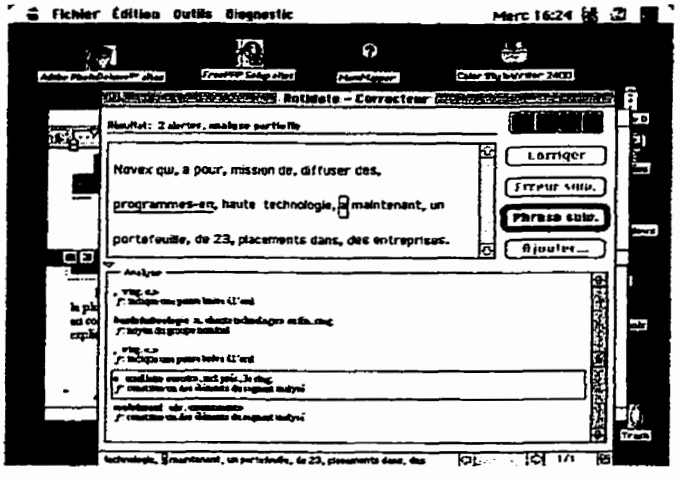


Figure 13

Antidote : analyses comparées

En comparant l'analyse détaillée de (2) à celle de (2*), nous constatons que les liens syntaxiques qu'était capable d'établir *Antidote®* dans la phrase originale (2) pour identifier la fonction des mots sont disparus avec son traitement de (2*). Par exemple, en analysant (2), *Antidote®* reconnaît correctement le rôle joué par le verbe « a » dans les deux propositions : verbe 'avoir', noyau de la proposition; en analysant (2*) par contre, il détermine que le second « a » est employé comme auxiliaire, ce qui est faux dans le contexte soumis, et le décrit simplement comme un des éléments du segment analysé. En analysant (2), *Antidote®* se rend compte que la suite « haute

technologie » est un complément du nom « programme » et que le connecteur « en » a pour fonction d'introduire ce complément; avec (2*), *Antidote*® a perdu sa référence avec sa segmentation autour des virgules : il accepte la virgule précédant « haute technologie », à laquelle il attribue une fonction prosodique — en conformité avec la thèse encore la plus répandue dans la littérature prescriptive[16] —, et s'en tient à une analyse de nature plutôt que de fonction du groupe nominal. De la même manière, « maintenant » est reconnu comme un adverbe avec fonction de complément dans (2) mais pas dans (2*).

La segmentation du texte analysé autour de la ponctuation est désastreuse pour les correcteurs orthographiques. En effet, une étude statistique (Ministère de l'Enseignement supérieur et de la Science, 1993) d'un échantillon de quelque 20 000 copies de candidats à l'université a démontré que près de 93% des textes produits comportaient des erreurs de virgule et près de 30%, des erreurs de point. Parmi toutes les erreurs de ponctuation enregistrées, plus de 80% touchaient l'emploi des virgules. Les chances sont donc très bonnes que les textes analysés régulièrement par une grammaire informatique comportent un grand nombre d'erreurs de ponctuation, surtout dans l'emploi des virgules. Par conséquent, la segmentation autour de la ponctuation du texte, bien qu'elle se justifie du point de vue de l'analyse automatique de la langue, devient un problème dans l'analyse automatique du discours.

---

[16] Il est fascinant de constater que les concepteurs de l'analyseur syntaxique d'*Antidote*® se sont ralliés à la thèse prosodique, héritée du grec, qui ne peut leur être d'aucun secours en traitement automatique de la langue, alors que des linguistes informaticiens comme Briscoe (1996a), Jones (1996b), Nunberg (1996, 1990) et Dale (1991), qui s'intéressent au rôle de la ponctuation dans l'analyse automatique de la langue, l'ont manifestement rejetée.

Attribution sélective du pouvoir de segmentation

Les signes de ponctuation sont généralement classés en un système hiérarchisé aussi bien dans la littérature prescriptive que non prescriptive. Cette hiérarchisation est rendue nécessaire dans le cas possible de l'action de signes utilisés en conjonction avec d'autres en un même point du discours. Catach (1994 : 122) cite ainsi et commente trois « lois » formulées par Tournier (dans Catach, édit., 1980 : 39, 45) :

> *Loi d'exclusion. — Certains ponctuants s'excluent mutuellement (...) Même s'il y a, en un point du discours, plusieurs ponctuances à marquer, un seul ponctuant est réalisé, et une seule fois. (ex. : la virgule est interdite avant une parenthèse).*
>
> *Loi de neutralisation. — Si en un point du discours plusieurs ponctuances doivent être marquées, et ne peuvent normalement l'être que par le même ponctuant, celui-ci n'est réalisé qu'une fois. (ex. : le point final l'emporte en finale sur le point abréviatif).*
>
> *Loi d'absorption. — Il existe des signes qui ne peuvent apparaître l'un à côté de l'autre, bien que comportant des ponctuants et des ponctuances différents : dans de tels cas, un seul ponctuant est réalisé et il se charge alors de sa ponctuance propre et des autres.*
>
> *Cette loi, particulièrement importante, concerne tous les cas où l'une des deux virgules doubles (il s'agit surtout d'elles) disparaît en position de conflit avec un autre signe, ex. :*
>
> *(absorption par la majuscule de début de phrase) « Croyez-moi, il en est toujours ainsi » // (position interne) « Il en est, croyez-moi, toujours ainsi »// (absorption par le point final) « Il en est toujours ainsi, croyez-moi ».*

Dans une telle hiérarchie, la virgule apparaît généralement comme le signe le plus « faible » et le point d'assertion comme le signe le plus « fort », bien qu'entre ces deux pôles, la position des signes puisse varier selon les grammairiens.

Le point-virgule contre le deux-points

Les deux correcteurs étudiés ont analysé nos exemples en ignorant certains signes tout en accordant à d'autres un pouvoir de segmentation de même force que celui du point assertif.

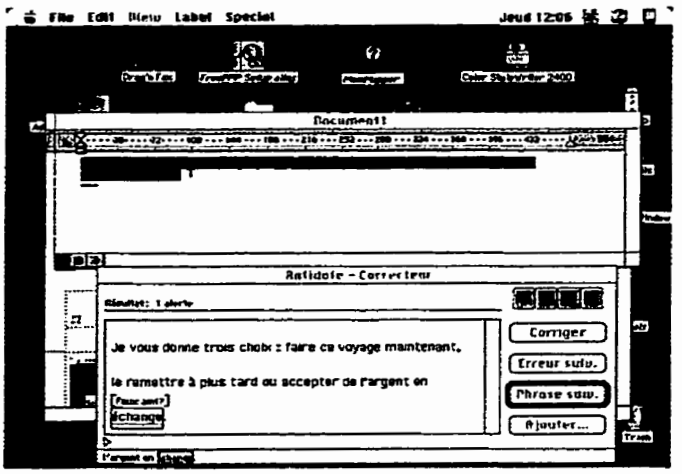
Un bon exemple peut être trouvé dans leur différence de traitement d'une phrase comportant un deux-points ou un point-virgule. La littérature du domaine attribue au deux-points et au point-virgule une « force » relativement équivalente malgré des fonctions discursives très différentes. Par exemple, Védénina (1989 : 134) décrit une fonction de délimiteurs de parties syntaxiques à tous deux mais de phrases elliptiques pour le deux-points; Tournier (1977 : 228-227) voit le deux-points comme

servant à introduire l'insertion de parties de phrases et le point-virgule comme délimiteur d'éléments de la phrase de base. Dans tous les cas, ces deux signes s'excluent mutuellement.

Après que nous avons demandé l'analyse pour la suite (3) avec un deux-points, *Antidote®* procède en segmentant la phrase de la majuscule au point. Cependant, si nous substituons au deux-points un point-virgule, la même phrase est analysée en deux parties indépendantes (Fig. 14).

Antidote®

Suite (3) avec un deux-points



Suite (3) avec un point-virgule

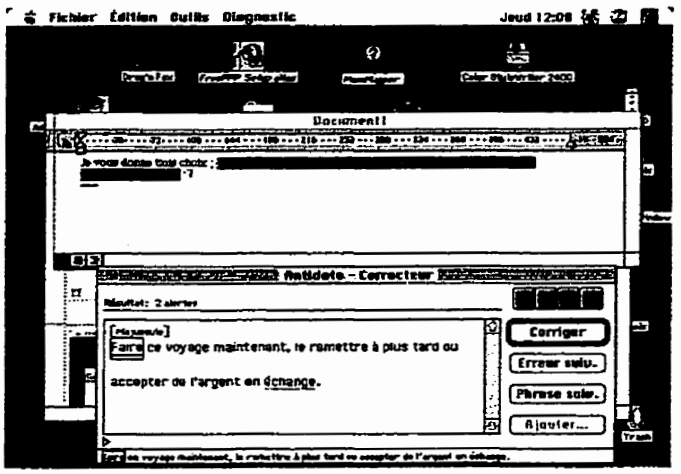


Figure 14

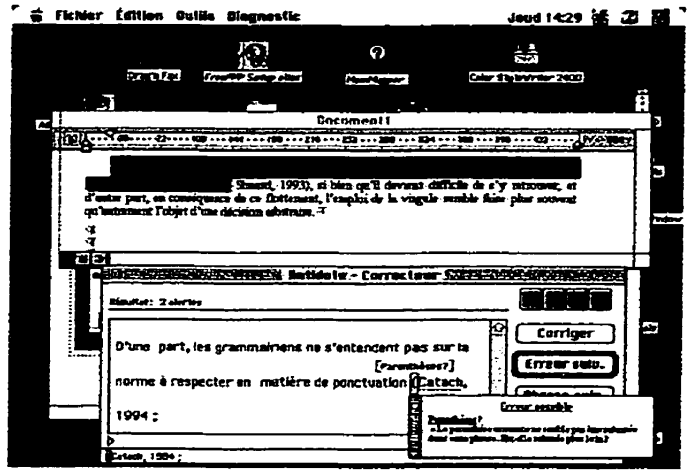
Antidote: segmentation à partir d'autres signes que le point

La grammaire de *Word7®* procède de même.

Les parenthèses et autres signes doubles

Mais voyons ce qui arrive dans l'analyse du paragraphe (1). L'analyse est demandée pour la seconde phrase. Encore une fois, la segmentation s'effectue autour d'un point-virgule. Or ce point-virgule ne fait pas partie de la phrase comme telle. Il sépare plutôt une suite que Chanod (1993 : 4) appelle « périphérique », c'est-à-dire ajoutée au noyau de la phrase et délimitée ici par des parenthèses. L'introduction de ces parenthèses génère un découpage automatique différent pour les correcteurs (Fig. 15)

Antidote®



La grammaire de Word7®

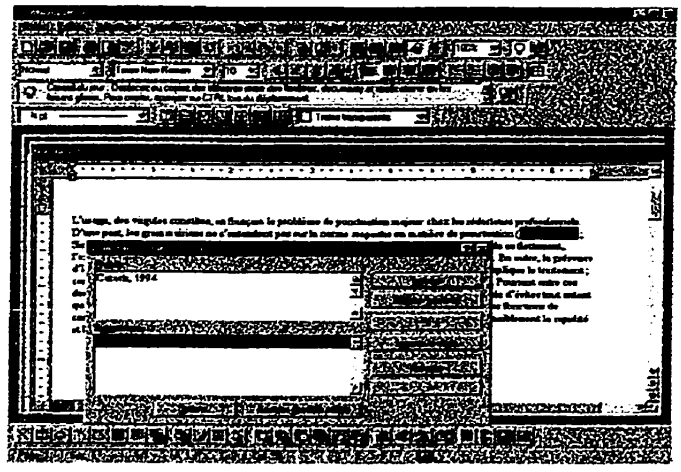


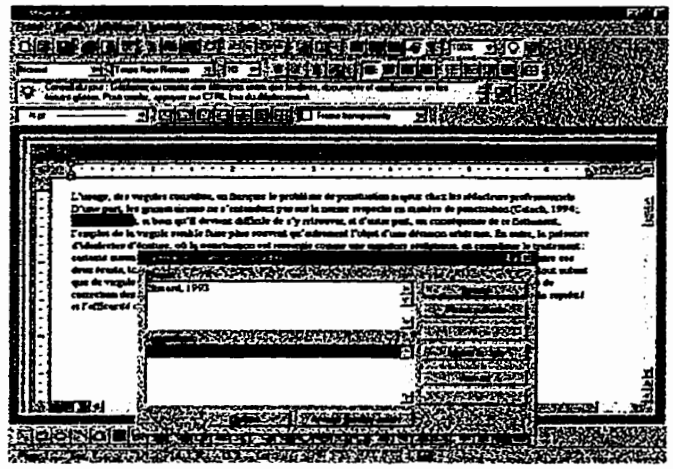
Figure 15

Effet des parenthèses et autres signes doubles sur la segmentation

Antidote® cherche la parenthèse fermante et nous suggère de vérifier si elle se présente plus loin. Si nous demandons l'analyse pour le reste de la phrase, Antidote® ne s'arrête pas à la parenthèse fermante mais segmente du point-virgule jusqu'au point assertif. La grammaire de Word7®, en revanche, découpe la même phrase en quatre segments d'analyse en prenant comme limites les parenthèses selon la séquence suivante : de la majuscule à la parenthèse ouvrante; de la parenthèse ouvrante au point-virgule; du point-virgule à la parenthèse fermante; de la parenthèse fermante au point assertif, considérant chaque segment indépendamment de son voisin. C'est pourquoi la grammaire de Word7® cherche en vain le verbe conjugué dans les segments analysés à partir du point-virgule (Fig. 16).

La grammaire de Word7®

Segmentation du point-virgule à la parenthèse fermante



Segmentation de la parenthèse fermante au point assertif

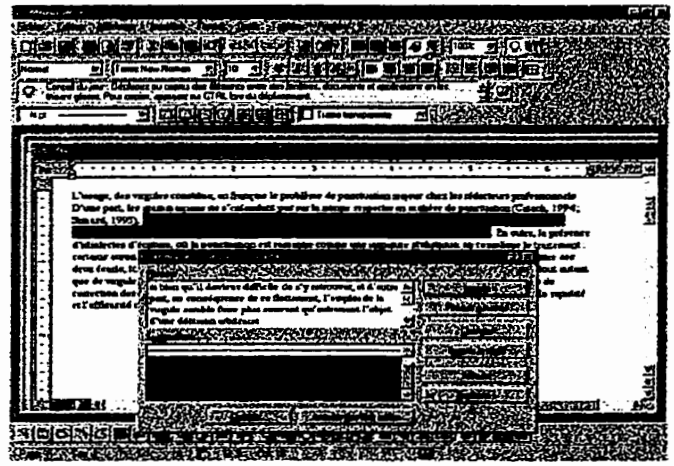


Figure 16

Effet de l'hypersegmentation sur le diagnostic

Pourtant, les délimiteurs (Nunberg, 1990) comme les parenthèses sont considérés comme « forts » dans la littérature du domaine parce qu'ils ne peuvent être absorbés par d'autres : Catach (1994 : 123) souligne que les signes essentiellement doubles (parenthèses, crochets, guillemets) se maintiennent dans tous les cas.

La substitution de signes de même catégorie ne change pas la segmentation effectuée. Ainsi le remplacement des parenthèses par des crochets ou des accolades génère la même analyse à partir du même découpage. De la même façon, le

remplacement du point-virgule par un autre point — point de suspension, point d'interrogation, point d'exclamation — produit également les mêmes résultats.

Segmentation selon les signes de ponctuation : une vue d'ensemble

Le tableau 5 présente une synthèse de la segmentation généralement observée dans nos exemples. Le point signale une segmentation autour du signe de ponctuation; le tiret, une non-segmentation.

Tableau 5

Synthèse de la segmentation dans les exemples analysés

Signes de ponctuation		Antidote®	Grammaire de Word7®
Point assertif		•	•
Autres points		•	•
Point-virgule		•	•
Parenthèses et autres signes doubles		•	—
Virgule	correcte	—	—
	incorrecte	• (le plus souvent)	• (le plus souvent)

Deux éléments sont à noter dans la pratique de segmentation des correcteurs autour de la ponctuation. Premièrement, les grammaires analysent chaque segment de façon indépendante. La principale conséquence de ce mode d'analyse, c'est que, lorsqu'une segmentation est effectuée autour d'un autre signe que le point assertif, les correcteurs sont incapables, en cas de besoin, de compléter leur analyse en faisant appel à des suites figurant en dehors du segment. Deuxièmement, alors que les virgules correctement placées sont généralement ignorées dans l'analyse des correcteurs, les virgules incorrectement placées génèrent le plus souvent une segmentation complète, entraînant tout un lot de fausses analyses et d'erreurs de diagnostic.

2.1.3 Échec des correcteurs en matière de détection d'erreurs de ponctuation

Antidote® et la grammaire de Word7® échouent généralement à détecter et corriger les problèmes de ponctuation.

La grammaire de Word7® a failli, dans presque tous les cas, à reconnaître les erreurs que nous avions placées dans nos exemples. Les virgules manquantes n'ont pas été identifiées comme telles, non plus que les virgules additionnelles, même quand il s'agissait d'aberrations comme dans l'exemple (2*). La grammaire de Word7® ne segmente ni ne corrige autour des virgules : elle semble plutôt les ignorer et accepter la phrase comme si elle n'était pas ponctuée, y laissant des erreurs et en ajoutant même parfois de son propre cru.

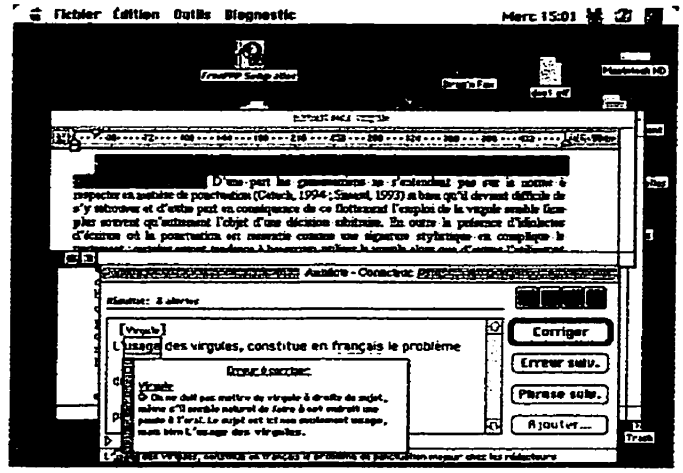
Antidote®, en revanche, réussit à détecter quelques erreurs de ponctuation. C'est pourquoi nous allons plutôt nous attarder à partir de maintenant au travail d'*Antidote*®.

Présence d'une virgule non requise

Une virgule non requise est introduite entre le sujet et son verbe (Fig. 17).

Antidote®
Virgule non requise introduite entre sujet et verbe

Détection réussie



Détection échouée

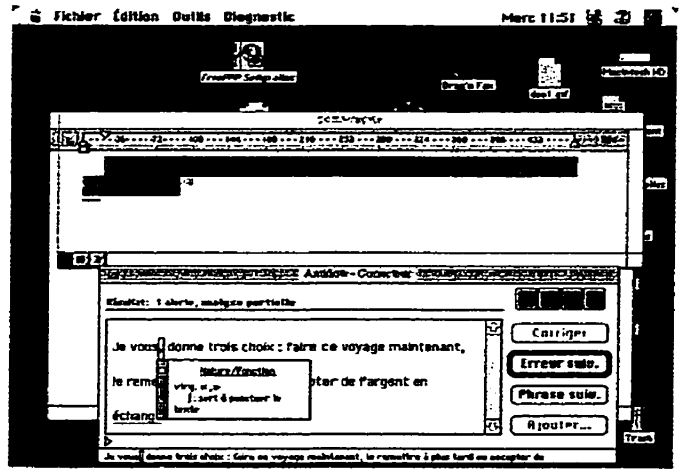


Figure 17

Diagnostic après introduction d'une virgule non requise entre SN et SV

Dans notre premier exemple, nous avons introduit une virgule entre le syntagme nominal sujet «l'usage des virgules » et son verbe « constitue ». *Antidote®* réussit à repérer l'erreur et ne segmente pas après cette virgule. L'explication offerte est juste. Par contre, dans notre deuxième exemple, *Antidote®* ne voit pas l'erreur dans la virgule placée entre le pronom « vous » et le verbe « donne » et déclare que cette virgule « sert à ponctuer le texte ».

Mais voyons ce qui arrive si nous plaçons une virgule entre le syntagme nominal « l'usage » et son complément « des virgules » (Fig. 18).

Antidote®
Détection échouée

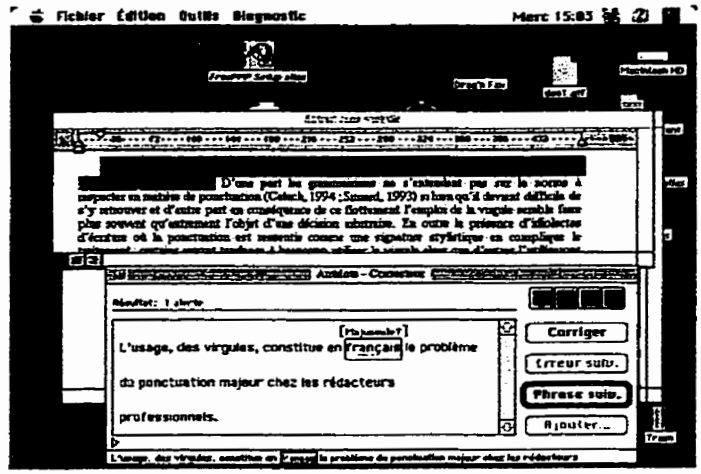


Figure 18

Diagnostic après introduction d'une virgule entre SN et SP dans un SN sujet

Cette fois-ci, *Antidote®* n'est plus en état de reconnaître l'erreur. La segmentation après la première virgule le rend incapable d'établir le lien syntaxique lui permettant de reconnaître le syntagme nominal sujet. Il est entendu alors qu'il n'est plus non plus en mesure de repérer la présence de la virgule entre le sujet et son verbe.

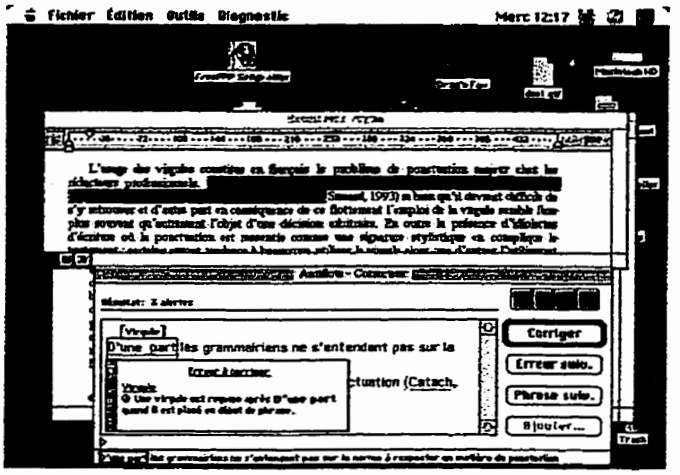
Absence d'une virgule requise

Nous avons effacé deux types de virgules : une virgule requise après un connecteur placé à l'initiale de la phrase (Fig. 19) et l'un des membres d'une paire de virgules.

Le paragraphe (1) fournit plusieurs occurrences de phrases commençant par un connecteur.

Antidote®
Virgule effacée après un connecteur placé en initiale de phrase

« D'une part »
Détection réussie



« Pourtant »
Détection réussie

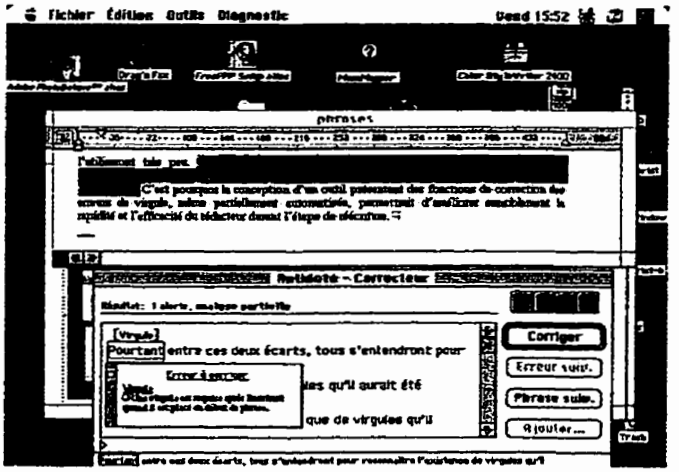


Figure 19

Diagnostic après effacement d'une virgule requise en initiale de phrase

Antidote® repère la virgule manquante et explique correctement le contexte. Cependant, si la virgule requise après « d'autre part » est effacée (Fig. 20), *Antidote®* ne s'en rend pas compte :

Antidote®
Virgule effacée après « d'autre part » placé en médiane de phrase
Détection échouée

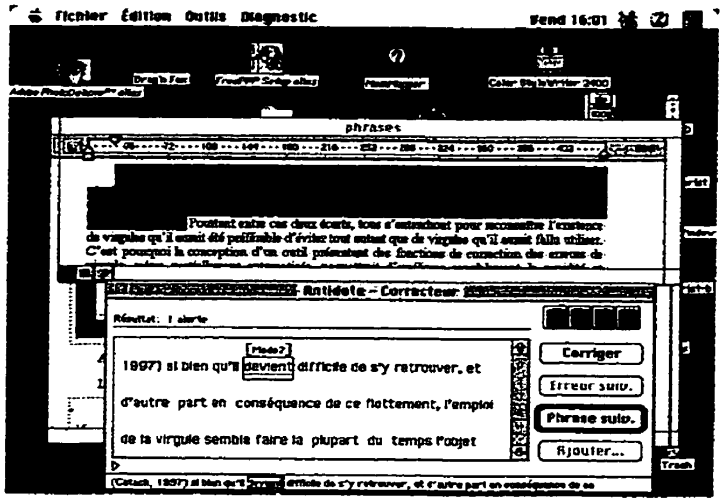


Figure 20

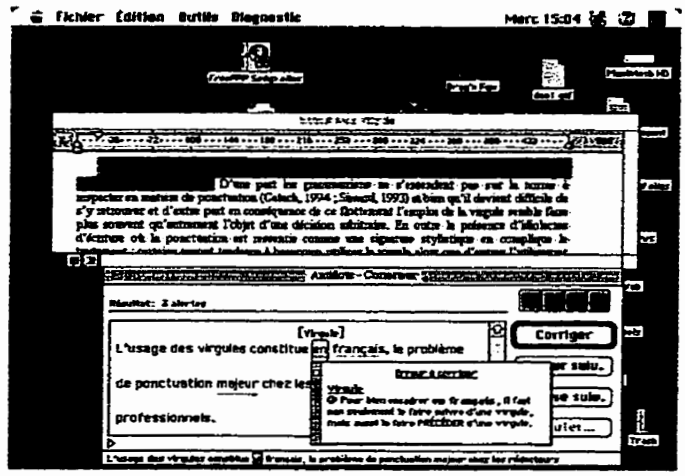
Diagnostic après effacement d'une virgule en médiane de phrase

La suite « d'autre part » ne se trouvant pas en initiale de phrase, *Antidote®* semble avoir perdu son repère : l'erreur n'est pas détectée.

L'un des éléments d'une paire de virgules a été effacé (Fig. 21) dans la première phrase de (1).

Antidote®
Effacement de l'un des membres d'une paire de virgules

Virgule ouvrante effacée
Détection réussie



Virgule fermante effacée
Détection échouée

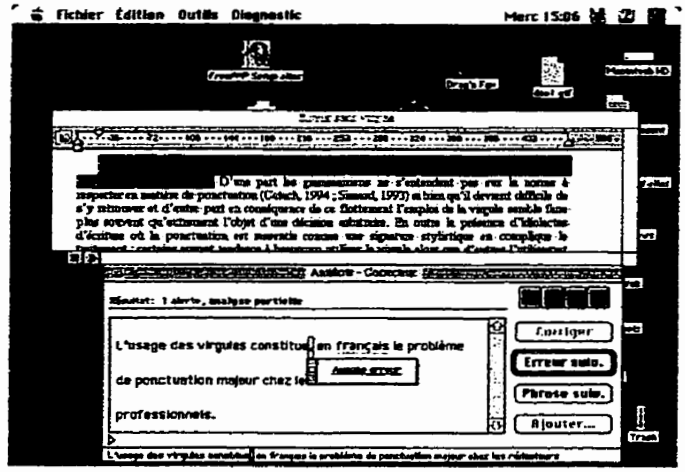


Figure 21

Diagnostic après effacement de l'un des membres d'une paire de virgules

Antidote® réussit à repérer la virgule ouvrante qui manque mais pas la virgule fermante. Cet échec à reconnaître l'absence du deuxième membre d'une paire de virgules a été répété pour chacun de nos exemples. Par ailleurs, si les deux virgules encadrant « en français » sont effacées, *Antidote®* ne réagit pas et accepte la suite telle quelle.

Emploi erroné d'une virgule à la place du point

Le point séparant la dernière phrase du paragraphe (1) de l'avant-dernière est remplacé par une virgule (Fig. 22).

Antidote®
Remplacement du point par une virgule

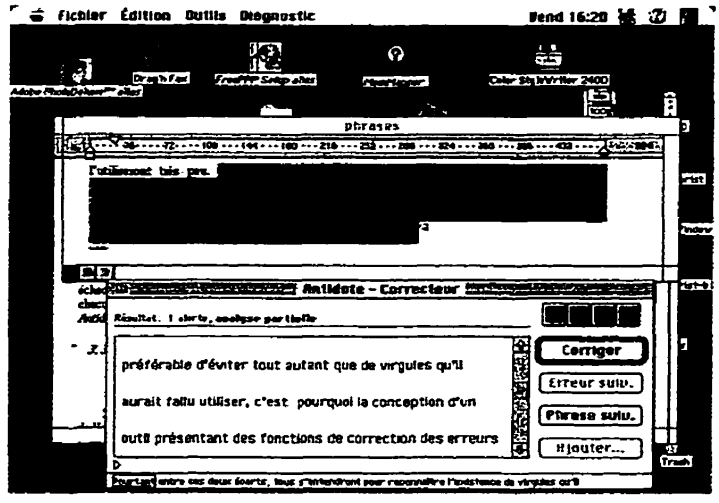


Figure 22

Diagnostic après remplacement du point par une virgule

Antidote® accepte la phrase telle quelle, échouant dans le repérage de l'erreur. Cet échec se répète même si tous les points du paragraphe sont remplacés par des virgules pour constituer une phrase bien improbable de 156 mots. Encore une fois, *Antidote®* accepte la suite telle quelle.

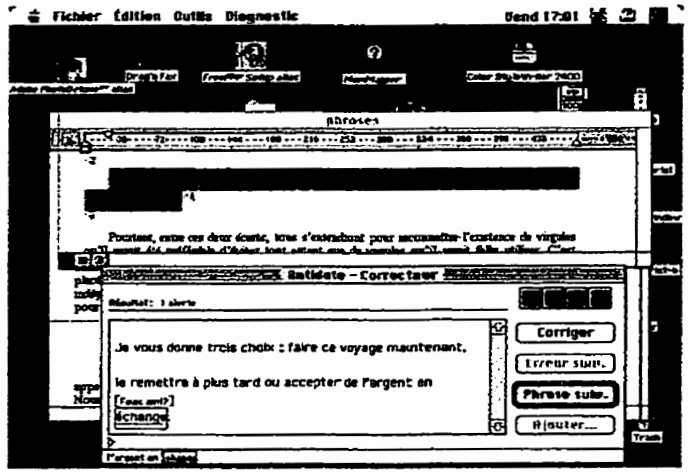
Selon Chandioux (1996), la difficulté la plus importante associée à la détection automatique de l'erreur de rédaction pose le problème de la validité même de la détection : jusqu'à quel point un correcteur reconnaît-il les problèmes réels d'écriture? Autrement dit, comment contrôle-t-il le nombre de fausses détections et de détections manquantes ? Cette difficulté prend toute son importance si nous prenons en considération le traitement de la ponctuation.

Nous avons vu que les grammaires segmentaient le texte pour fins d'analyse sans habituellement remettre en question la ponctuation paraissant dans le texte. Nous avons constaté que les correcteurs segmentaient les textes à réviser avec d'autres signes de ponctuation que le point. Nous avons également vu que chaque

segment était analysé par les correcteurs indépendamment les uns des autres. Nous avons finalement observé que, du moins dans nos exemples, des virgules erronément placées provoquaient le plus souvent une segmentation du texte, et donc une analyse indépendante de ce segment. Il nous reste à nous demander si la capacité d'analyse des correcteurs pour l'ensemble du texte à analyser se trouve entravée du fait d'une erreur de ponctuation. Eh bien, elle l'est, si nous en croyons l'exercice auquel nous avons soumis *Antidote®* et la grammaire de *Word7®* (Fig. 23).

Antidote®

Diagnostic sur suite ponctuée correctement



Diagnostic sur suite incorrectement ponctuée

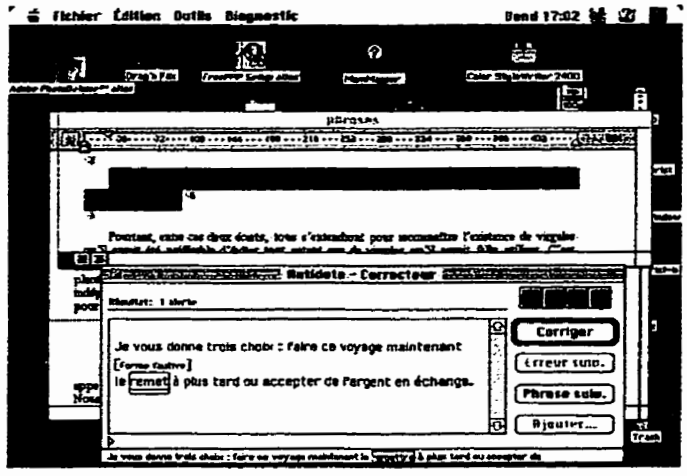


Figure 23

Influence de l'erreur de ponctuation sur le diagnostic

L'effacement de la virgule marquant l'énumération rend *Antidote*® incapable de procéder à son analyse et cette erreur génère une fausse détection. La grammaire de *Word7*® réagit de même une fois l'erreur de ponctuation introduite. En fait, les erreurs de ponctuation, particulièrement de virgules, ont rendu souvent les grammaires étudiées impuissantes à générer un diagnostic fiable. Nous pourrions apporter de nombreux exemples. Qu'il nous suffise de rappeler la différence dans les analyses effectuées par *Antidote*® des phrases (2) et (2*).

Comme nous pouvons le voir à ce rapide tour d'horizon, nous avons imposé à nos correcteurs un exercice difficile. D'abord, l'analyse automatique de phrases ordinaires comportant une combinaison de structures « périphériques » attachées à un « noyau » (Chanod, 1993) constitue en elle-même l'un des obstacles les plus importants de l'analyse automatique robuste (Dale, 1996; Briscoe, 1996b), celle-là même à laquelle sont confrontés les correcteurs grammaticaux. Ensuite, la ponctuation exploitée comme point de repère pour la segmentation du texte par les correcteurs ne constitue pas une donnée fiable dans les textes écrits par des rédacteurs occasionnels. Enfin, la présence d'erreurs de ponctuation empêche souvent un diagnostic juste pour le restant de la phrase analysée. Il ne faut donc pas s'étonner de l'échec des correcteurs en matière de détection et de correction, non seulement d'erreurs de ponctuation, mais d'erreurs grammaticales tout court.

2.2 De la correction automatique de la ponctuation : mise à jour

Cette mise à jour s'intéresse aux correcteurs suivants :

* *Antidote* 98, version 2.0;
* *Cordial*, la grammaire de *Word* 98[17];
* Le *Correcteur 101* Pro, version 4.0.4.

Ce deuxième exercice reprend les phrases que nous avions soumises à *Antidote* et à la grammaire de *Word 7*[18] mais en ajoutant (5) :

---

17 *Word 98* est la version la plus récente du texteur *Word* pour Macintosh.

18 Voir ce chapitre : « À propos de notre exercice ».

(5) *Je vous attend tous et toutes en grand nombre.

En effet, cette phrase, utilisée par un étudiant de premier cycle[19] dans le cadre d'un projet portant sur les correcteurs orthographiques, a attiré notre attention sur un diagnostic inusité de la part du *Correcteur 101*[20]. Intrigués, nous avons étendu ce test à l'ensemble des correcteurs examinés.

Les correcteurs grammaticaux de 1999 sont-ils plus efficaces que leurs prédécesseurs dans le traitement des erreurs de ponctuation? Nous pouvions nous attendre en effet à ce que l'évolution rapide des connaissances en génie linguistique de même que les règles du marché aient une incidence positive sur la performance générale des logiciels que nous avions analysés. Notre seconde étude démontre qu'il n'en est rien.

En fait, les correcteurs étudiés pour notre mise à jour demeurent impuissants à détecter — et *a fortiori* à corriger — les problèmes de ponctuation. Ils continuent de segmenter les textes en s'appuyant sur leur ponctuation originale à partir du point assertif comme marque de l'unité analysable.

2.2.1 Antidote 98, v. 2

Les résultats de notre premier test se sont reproduits avec la version 2 d'*Antidote*, *Antidote* 98.

Maintien du problème de segmentation

Les problèmes de segmentation observés lors de notre premier exercice continuent de se présenter :

* le correcteur appuie sa segmentation sur la ponctuation présente dans le texte sans la remettre en question, ainsi que l'avaient démontré nos exemples de l'effacement d'un point (Fig. 10), de la substitution d'un deux-points par un point-virgule (Fig. 11) et de l'introduction d'une virgule aberrante (Fig. 12);

---

19 Serge Côté, étudiant en linguistique à l'Université Laval, pour le cours « Projet étudiant » (LNG17498). M.Côté rapportait, dans son travail de session, les résultats de son investigation sur la performance grammaticale de quelques correcteurs orthographiques.

20 Voir « Le *Correcteur 101* » plus loin dans ce chapitre.

* l'hypersegmentation[21] à partir du point-virgule (Fig. 14) ou de signes doubles comme les parenthèses ou les crochets (Fig. 15 et 16) continue de se produire.

Performance inchangéee en matière de détection d'erreurs de ponctuation

La performance d'*Antidote* 98 dans la détection des erreurs de ponctuation ne s'est pas non plus améliorée :

* il réussit encore à dépister l'erreur de la virgule introduite entre un sujet et son verbe dans le contexte de (1) mais pas dans celui de (3) (Fig. 17);

* il continue de reconnaître l'erreur dans le cas d'une virgule manquante en initiale de phrase (Fig. 19) mais pas en médiane (Fig. 20);

* il ne reconnaît pas l'emploi erroné d'une virgule à la place du point (Fig. 22).

Diagnostic de la nouvelle suite (5)

L'erreur de conjugaison de la suite (5) se voit corrigée sans problème par *Antidote* 98. Cependant le correcteur introduit une hypercorrection en confondant l'emploi féminin du pronom *toutes* avec l'emploi adverbial de *tout* (Fig. 24).

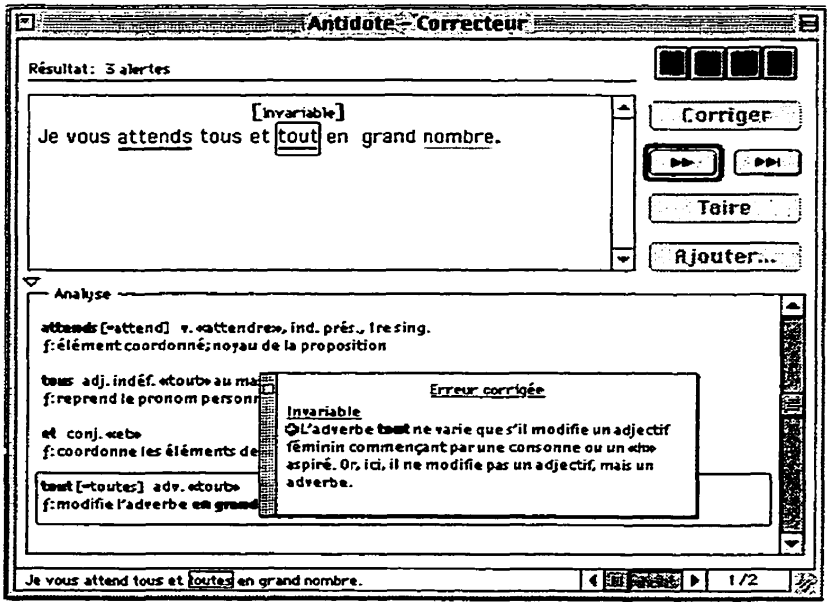


Figure 24

Antidote98: Diagnostic de la phrase (5).

---

21 Nous proposons ce néologisme pour décrire une segmentation effectuée à partir d'un autre signe de ponctuation que le point et ayant une incidence négative sur le diagnostic. Bien que sans le nommer alors, nous avions fait ressortir ce phénomène dans Simard (1997).

La fausse détection d'*Antidote* 98 s'explique par sa lecture de la suite *en grand nombre* dont il ne reconnaît pas la nature de terme complexe. Comme son explication détaillée le montre en effet, il considère que *tout* est un adverbe modifiant un autre adverbe. Cet adverbe, nous apprenons plus loin qu'il s'agit de *en grand*. Fait intéressant, cette erreur d'analyse s'évanouit si *et toutes* disparaît.

Antidote 98 ne modifie pas la ponctuation dans la suite (5).

2.2.2 La grammaire de Word *98*

Le correcteur grammatical intégré à *Word* 98 n'est plus *Hugo* mais *Cordial*, un logiciel fabriqué par une firme française, Synapse Développement[22]. Dans un couriel, Mme A. Marie Chaplain, du Service administratif et commercial de la société, nous présentait *Cordial* comme « un outil global » :

> *« Cordial » est un correcteur orthographique, grammatical, syntaxique de la langue française, mais il est bien plus que cela : un outil global de traitement de la langue (occurrence des mots, analyse sémantique et stylistique, analyse logique de la phrase, comparaison avec 2000 ouvrages de références, intégration du Littré, etc...*[Sic]*). Il s'intègre dans les outils Microsoft en remplaçant*[23] *les fonctions existant dans ces logiciels, et doit être lancé en indépendant pour les autres fonctions.*
>
> .....
>
> *Cordial existe surtout sur PC : actuellement c'est la version 5, et accessoirement sur Mac (pas de mise à jour depuis 2 ans, trop peu de demandes, la version sur mac* [Sic] *est l'équivalent de la version 3 sur PC).*

La version intégrée dans *Word* 98 exploite seulement les modules d'orthographe et de grammaire.

La grammaire de Word 98 : fonctions de paramétrage

Comme la grammaire de *Word* 97, les fonctions de paramétrage de la grammaire de *Word* 98 prévoient la correction de certaines erreurs de virgules (Fig. 25).

---

22 *www.synapse-fr.com.*

23 Souligné dans le texte.

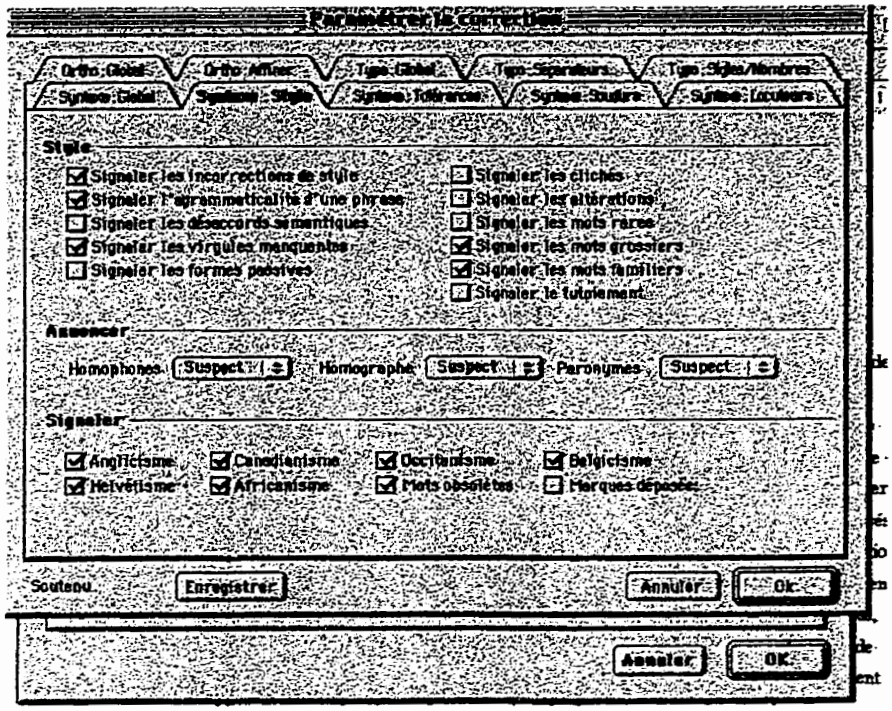


Figure 25

Réglage de la grammaire de Word 98

Traitement de la segmentation

La performance de *Cordial* dans *Word* 98 se compare à celles de *Hugo* dans *Word* 97, mais à une exception près : le problème de l'hypersegmentation ne se présente pas (Fig. 26). En cela, le logiciel se distingue aussi d'*Antidote* 98 :

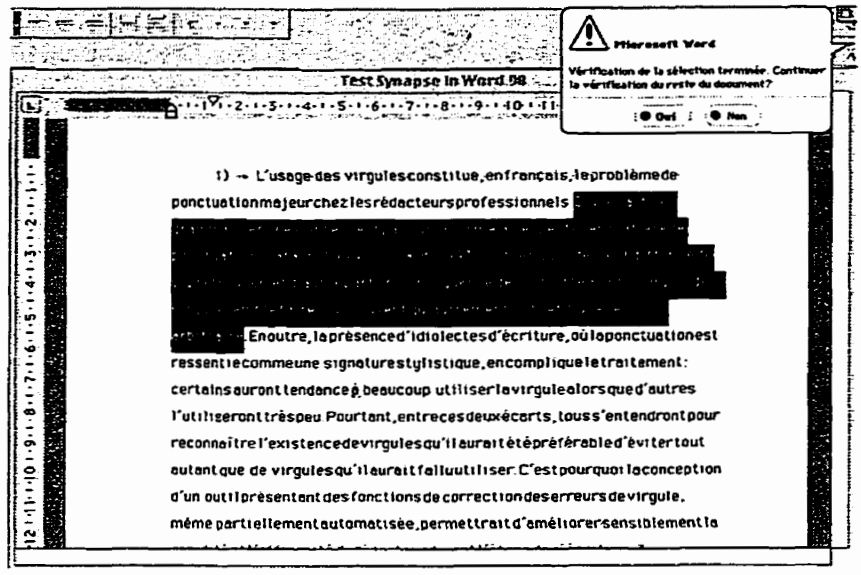


Figure 26

Hypersegmentation absente

La zone d'ombre (Fig. 26) signale l'unité segmentée par *Cordial* pour fins d'analyse. Cette zone s'étend du début de la phrase, marquée par la majuscule, jusqu'au point assertif indiquant la fin de la suite. La grammaire de *Word* 98 ignore avec raison les parenthèses introduisant les références bibliographiques de même que le point-virgule qui les sépare. En conséquence, elle n'introduit pas les fausses détections générées par l'hypersegmentation.

Faible performance en détection d'erreurs de ponctuation

La performance de *Cordial* en détection d'erreurs de ponctuation se révèle cependant aussi pauvre que celle des autres logiciels. Si le correcteur semble pouvoir repérer, comme *Antidote* 98, une virgule oubliée en initiale de phrase, il ne peut pas non plus reconnaître l'erreur en médiane de phrase. En outre, alors qu'*Antidote* 98 réussit à détecter l'absence du membre gauche d'une paire de virgules (Fig. 21), le correcteur de *Word* 98 ne le peut pas (Fig. 27).

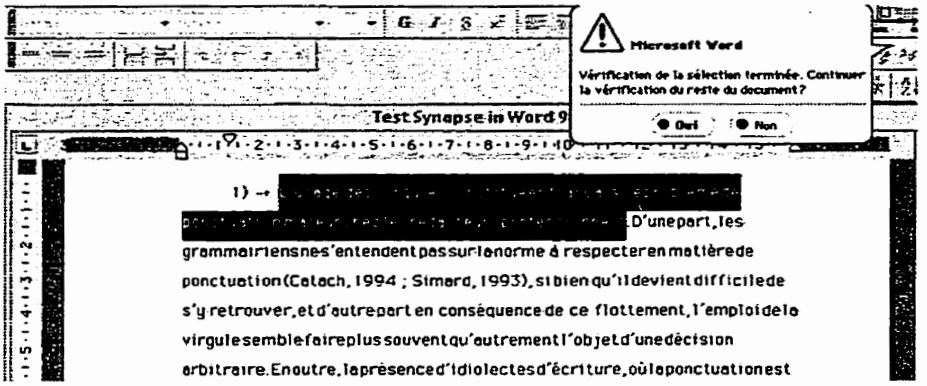


Figure 27

Diagnostic après effacement du membre gauche d'une paire de virgules

En fait, tout se passe comme si la grammaire de *Word* 98 ignorait le plus souvent la ponctuation intra-phrastique. Cependant, si tous les points de (1) sont remplacés par des virgules, le correcteur, bien qu'incapable de reconnaître les points manquants, restera tout de même en mesure de signaler l'occurrence d'une phrase longue (Fig. 28):

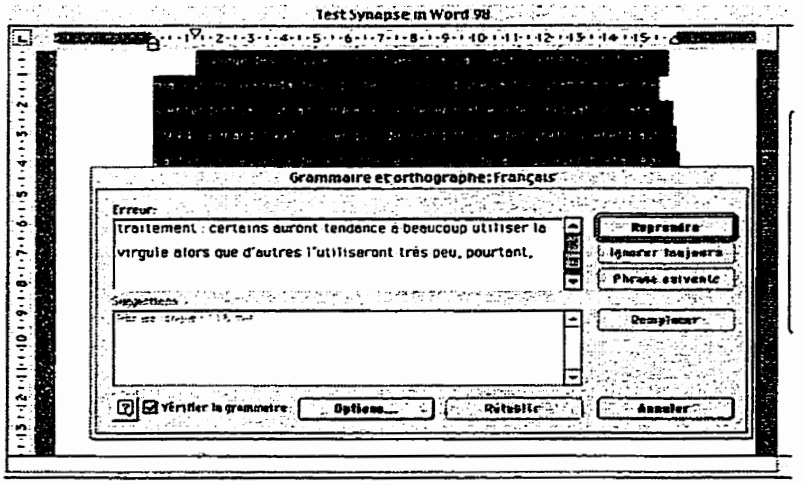


Figure 28

Diagnostic d'une phrase longue

Cette détection peut s'effectuer grâce à un réglage permettant de prédéterminer le nombre de mots acceptable par le correcteur dans une phrase (maximum 128).

Grammaire de *Word* : Diagnostic de la nouvelle suite (5)

Cordial repère avec succès l'erreur de conjugaison de (5) sans introduire d'hypercorrections (Fig. 29).

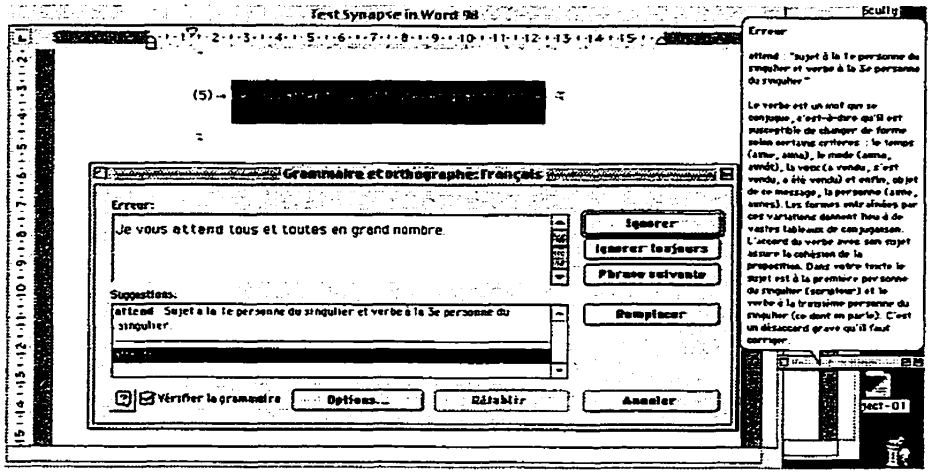


Figure 29

La grammaire de Word 98: diagnostic de la phrase (5)

L'explication du diagnostic est élaborée mais facile à comprendre.

2.2.3 Le Correcteur 101 Pro[24]

Nous avons examiné la version *Pro* 4.0.4. du *Correcteur 101.*

Le *Correcteur 101* : fonctions de paramétrage

Le *Correcteur 101* prévoit la correction de « difficultés typographiques » (Fig. 30).

[24] Le *Correcteur 101* que nous avons étudié est une copie de courtoisie prêtée par *Machina Sapiens* pour une période de 30 jours.

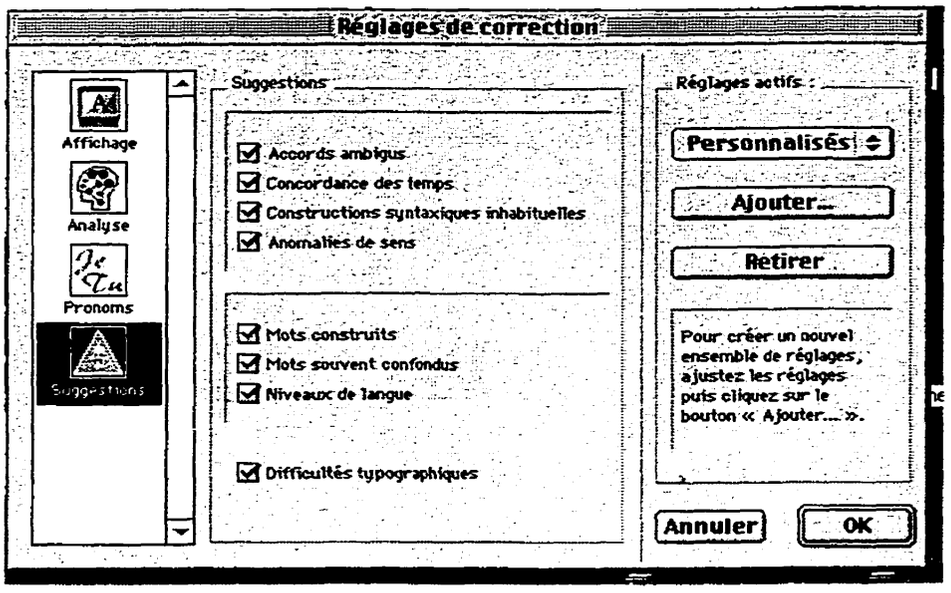


Figure 30

Réglage du Correcteur 101

La documentation accessible par le menu d'aide inclut la ponctuation dans l'ensemble « Difficultés typographiques » (Fig. 31).

Difficultés typographiques

Les difficultés typographiques regroupent les erreurs possibles sur les abréviations, la ponctuation, les dates, etc.

Figure 31

Le Correcteur 101: explication de "Difficultés typographiques"

Par ailleurs, le choix « Préférences » du menu « Édition » propose une option « Typographie ». Cette option ne renvoie cependant qu'à la gestion des espaces précédant ou suivant les signes de ponctuation (Fig. 32).

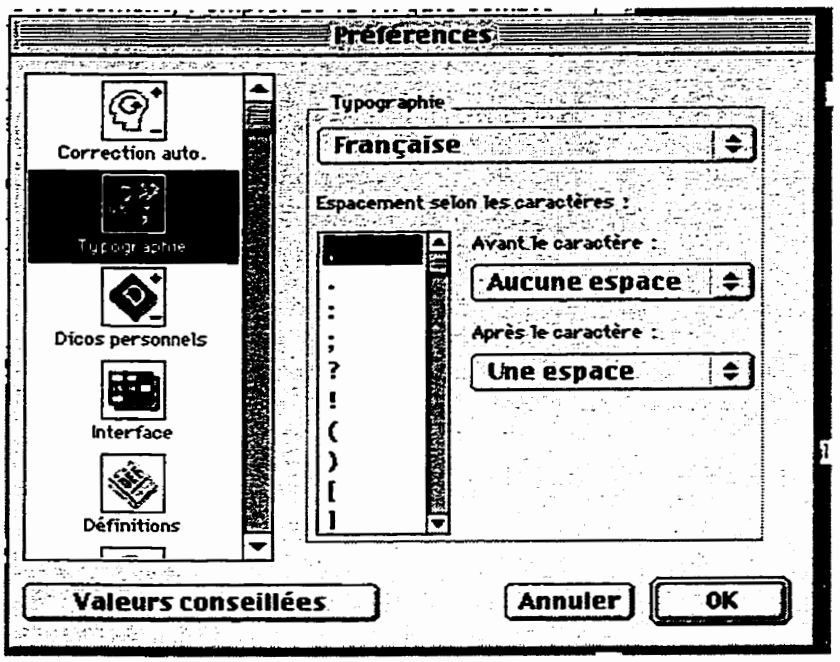


Figure 32

Le Correcteur 101: Préférences pour la révision typographique

Traitement de la segmentation

Comme la grammaire de *Word 98*, le *Correcteur 101* segmente en se fondant sur la ponctuation originale du texte. Dans l'effacement du point précédent *Pourtant* dans (1), le logiciel segmente à partir du point assertif en ignorant l'erreur (Fig. 33).

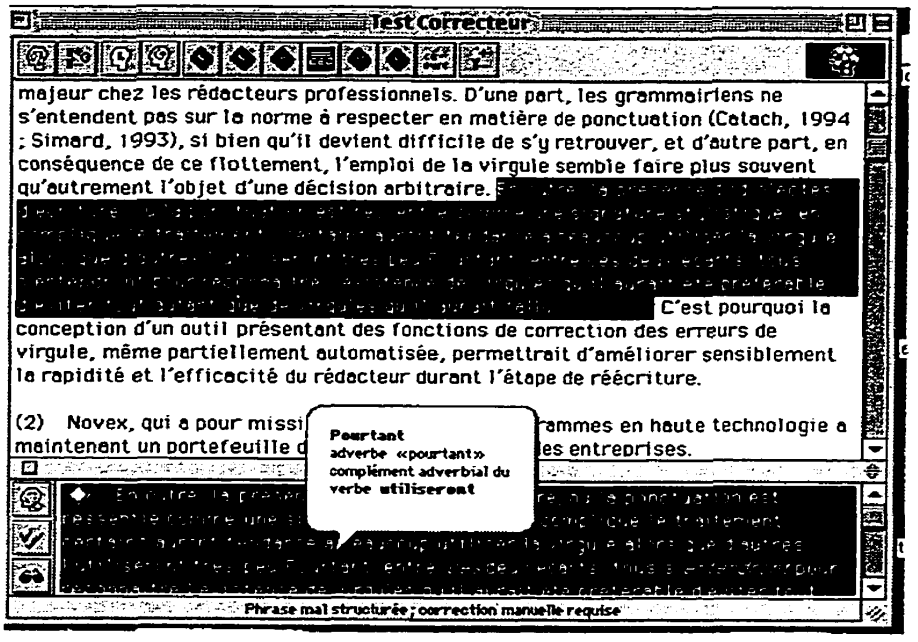


Figure 33

Segmentation à partir de la ponctuation originale

Le signalement d'une phrase mal structurée renvoie à une fausse détection où le *Correcteur 101* déclare ne pas pouvoir identifier le complément d'objet direct de *utiliser* (Fig. 34).

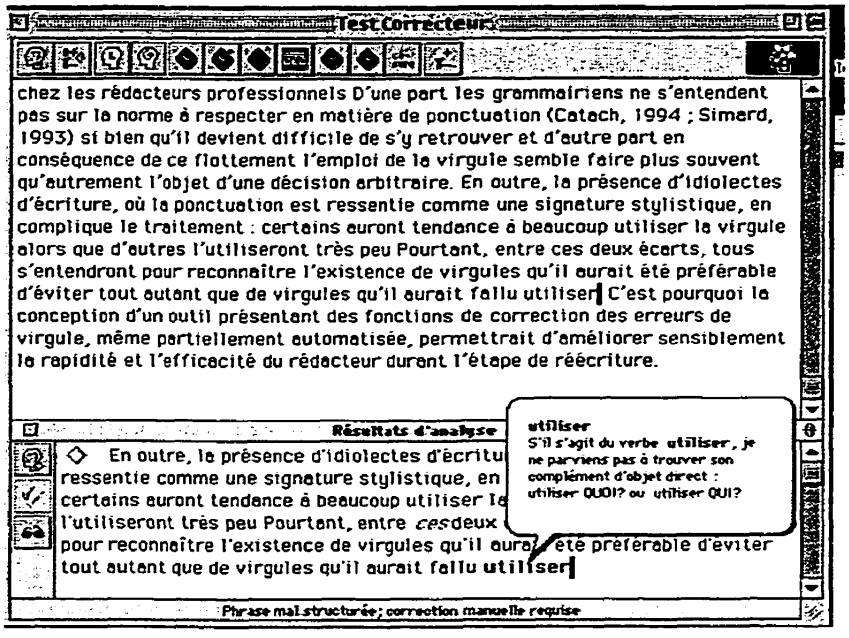


Figure 34

Hypercorrection

La segmentation à partir de la ponctuation originale (Fig. 35) se vérifie également dans le traitement de (3) ...

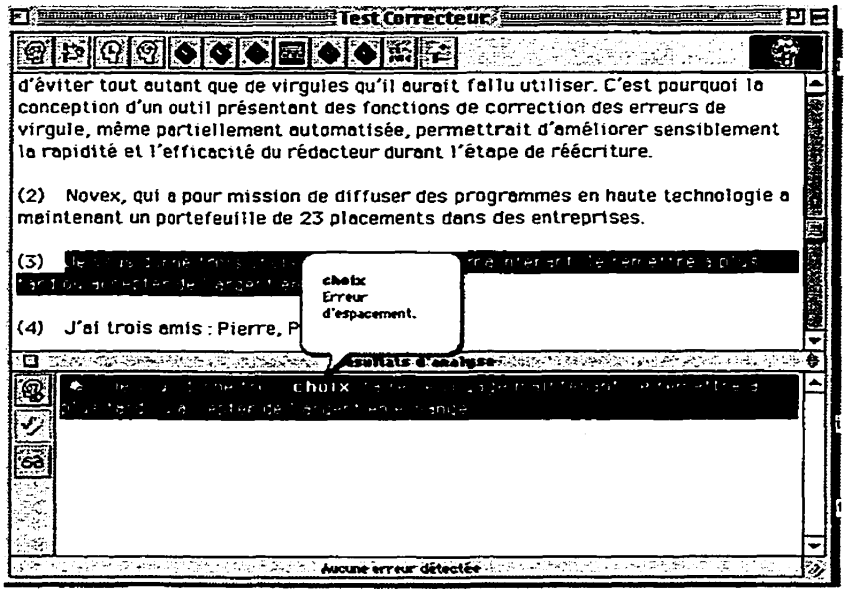


Figure 35

Acceptation de l'erreur comme base de segmentation

... et dans celui de (2*) (Fig. 36).

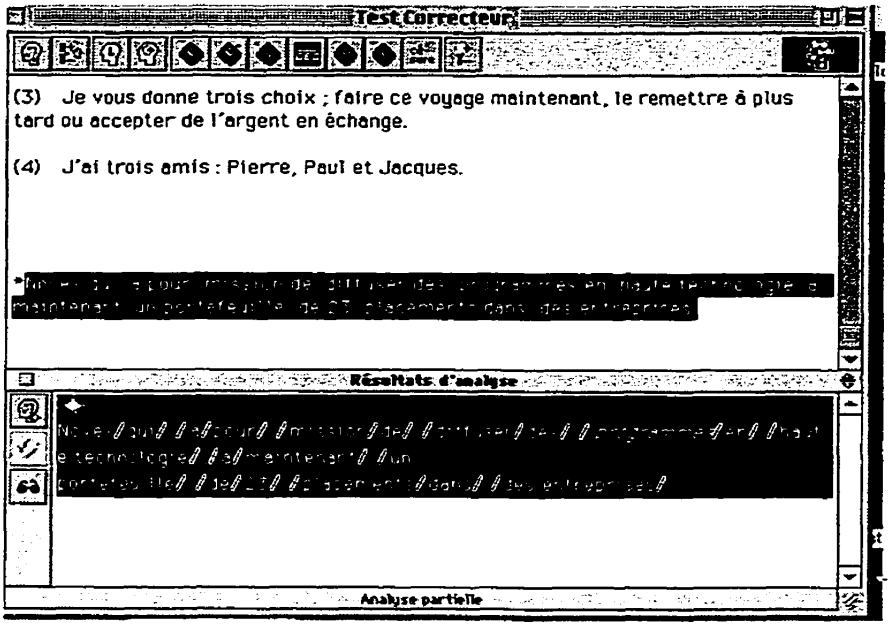
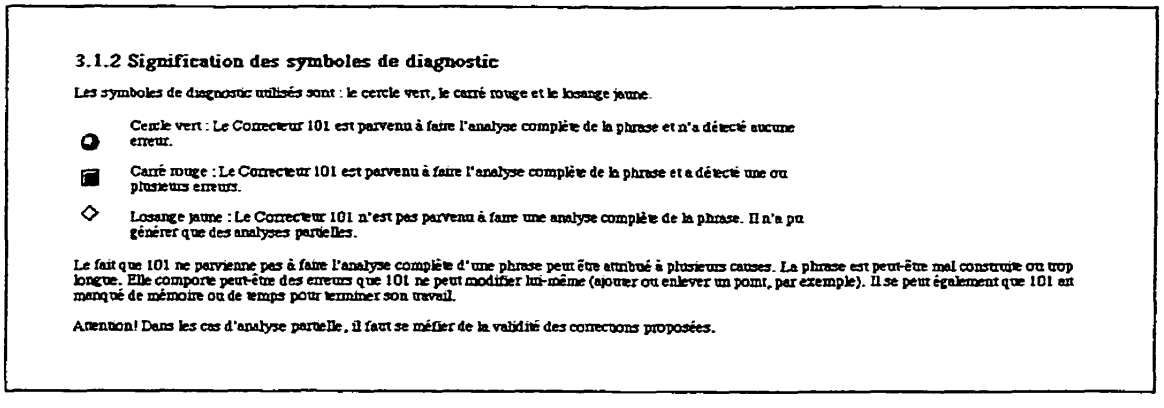


Figure 36

Diagnostic de la suite (2)*

Cependant, le *Correcteur 101* déclare une analyse partielle de (2*), indiquant, selon sa documentation, un problème qu'il ne peut résoudre (Fig. 37).



3.1.2 Signification des symboles de diagnostic

Les symboles de diagnostic utilisés sont : le cercle vert, le carré rouge et le losange jaune.

Cercle vert : Le Correcteur 101 est parvenu à faire l'analyse complète de la phrase et n'a détecté aucune erreur.

Carré rouge : Le Correcteur 101 est parvenu à faire l'analyse complète de la phrase et a détecté une ou plusieurs erreurs.

Losange jaune : Le Correcteur 101 n'est pas parvenu à faire une analyse complète de la phrase. Il n'a pu générer que des analyses partielles.

Le fait que 101 ne parvienne pas à faire l'analyse complète d'une phrase peut être attribué à plusieurs causes. La phrase est peut-être mal construite ou trop longue. Elle comporte peut-être des erreurs que 101 ne peut modifier lui-même (ajouter ou enlever un point, par exemple). Il se peut également que 101 ait manqué de mémoire ou de temps pour terminer son travail.

Attention! Dans les cas d'analyse partielle, il faut se méfier de la validité des corrections proposées.

Figure 37

Le Correcteur 101: Explication de "Analyses partielles"

En revanche, comme la grammaire de *Word 98*, le *Correcteur 101* évite l'hypersegmentation (Fig. 38).

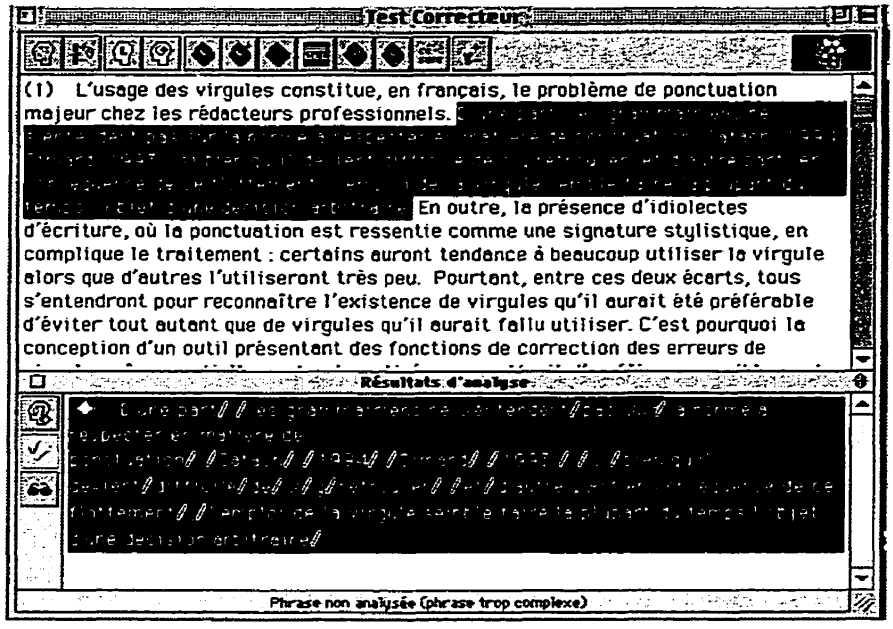


Figure 38

Le Correcteur 101 : exemple de segmentation

La suite est segmentée sans regard pour les parenthèses et le point-virgule. Le *Correcteur 101* considère néanmoins la phrase « trop complexe » pour être analysée.

Contre-performance en détection d'erreurs de ponctuation

Nous avons cherché à déterminer les modifications à apporter à la seconde phrase de (1) pour que le correcteur la juge moins « complexe ». Il s'avère que l'effacement — erroné — des virgules fermantes après « d'une part ... d'autre part » semble être la condition pour simplifier la structure de la phrase à la satisfaction du logiciel (Fig. 39).

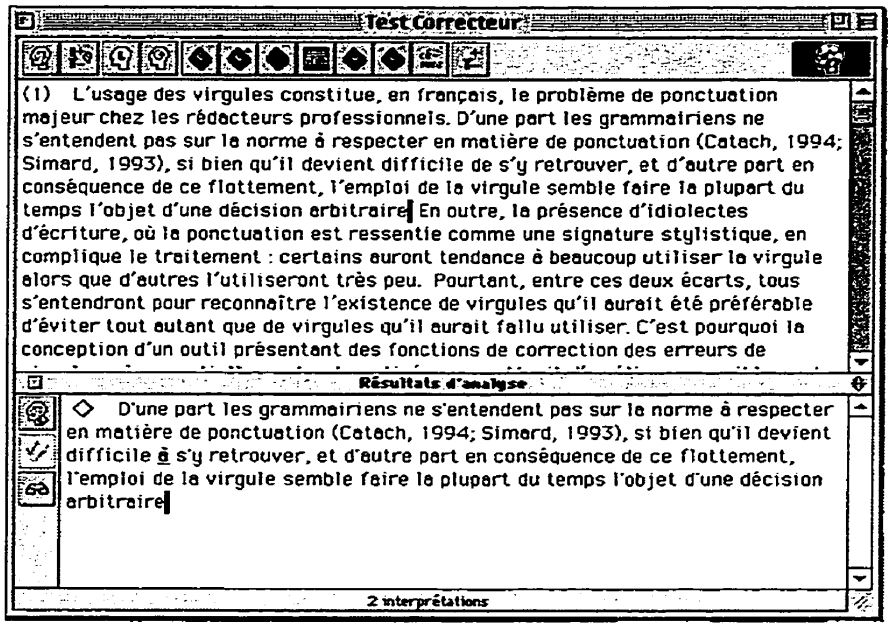


Figure 39

Diagnostic rendu possible après effacement de deux virgules correctes

Notons au passage la zone d'ombre marquant la segmentation à partir du point de même que la suggestion erronée de remplacer « d e » par « à ». Intrigués par ce diagnostic, nous avons carrément effacé toute ponctuation à l'intérieur de la phrase et demandé son avis au *Correcteur 101* (Fig. 40).

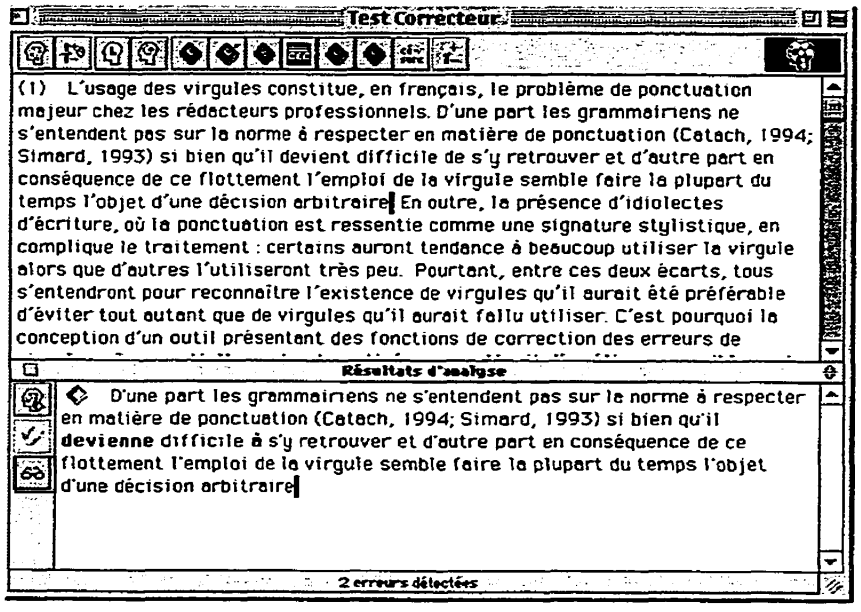


Figure 40

Diagnostic après effacement de toutes les virgules requises

Le *Correcteur 101* continue d'accepter d'analyser la phrase en dépit de l'absence de toutes les virgules requises. Cependant, il introduit une nouvelle hypercorrection, ajoutant un subjonctif non requis par le contexte. Curieusement donc, seule la phrase correctement ponctuée est rejetée par le logiciel.

Cette contre-performance se prolonge avec le diagnostic de la suite (5).

Le *Correcteur 101* : diagnostic de la suite (5)

Le logiciel effectue une lecture tout à fait erronée de la suite (5) (Fig. 41).

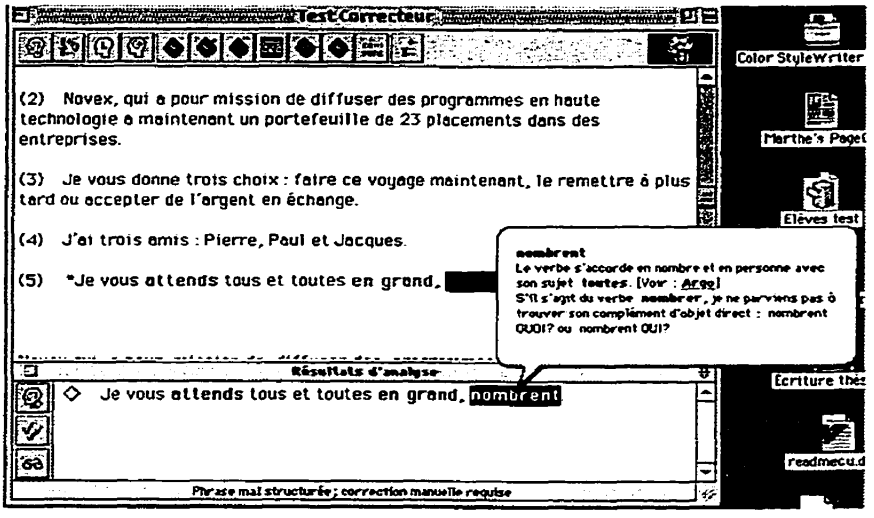


Figure 41

Le Correcteur 101: diagnostic de la suite (5)

Bien que corrigeant l'erreur dans « Je vous attend », le *Correcteur 101* introduit deux erreurs de son crû : une virgule non requise et un barbarisme. Nous pouvons nous demander ce qu'aurait pensé un rédacteur montrant des déficiences importantes en français écrit. Aurait-il pu lui-même se rendre compte de l'aberration suggérée ou le *Correcteur 101* aurait-il ajouté à sa confusion?

Nous venons de démontrer l'incapacité des correcteurs à traiter les problèmes de ponctuation de façon concluante. Bien que certaines erreurs puissent parfois — nous serions tentés de dire « accidentellement » — être corrigées, la plupart demeurent ignorées. Par ailleurs, nous avons également indiqué l'effet négatif des erreurs de ponctuation sur la détection automatique des erreurs grammaticales tout court. Finalement, nous avons également souligné l'importance du traitement de la segmentation dans la performance des correcteurs.

Le tableau 6 présente la synthèse du traitement de la segmentation dans tous les correcteurs examinés. Encore un fois, le point signale une segmentation autour du signe de ponctuation; le tiret, une non-segmentation.

Tableau 6

Nouvelle synthèse de la segmentation dans les exemples analysés

		Antidote®	Grammaire de Word7®	Antidote 98®	Grammaire de Word 98®	Correcteur 101
Point assertif		•	•	•	•	•
Autres points		•	•	•	•	•
Point-virgule		•	•	•	—	—
Parenthèses et autres signes doubles		•	—	•	—	—
Virgule	correcte	—	—	—	—	—
	incorrecte	• (le plus souvent)	— (le plus souvent)	• (le plus souvent)	— (le plus souvent)	— (le plus souvent)

Après notre analyse, nous pouvons en effet retrouver deux tendances : la ponctuation intra-phrastique est prise en compte dans le traitement automatique de la phrase ou elle ne l'est pas. Dans le premier cas, l'hypersegmentation introduit souvent de fausses détections. Dans le deuxième cas, il n'y a pas d'hypersegmentation et le nombre de fausses détections diminue en conséquence. Cependant, les erreurs de ponctuation dans un cas comme dans l'autre sont ignorées. Par conséquent, malgré les efforts des fabricants, les analyseurs syntaxiques des correcteurs grammaticaux échouent dans le traitement correctif des phrases, et particulièrement dans le traitement de leurs erreurs de ponctuation.

Conclusion

Richardson (1994) classe les analyseurs en deux catégories : ceux qui reposent sur des règles (*rule-based*) et ceux qui exploitent une approche statistique (*statistical-based*). Les parseurs à base de règles contiennent des énoncés linguistiques rendant l'ordinateur théoriquement capable de produire des arbres syntaxiques décrivant la structure des phrases d'un texte sans en trahir le sens. Par contre, les parseurs à base statistique tentent de décrire les connaissances linguistiques au moyen de paramètres basés sur les règles de la probabilité. L'approche linguistique demeure cependant la plus commune chez les linguistes informaticiens.

Cependant, nous savons que les parseurs actuels, peu importe leur type, traitent généralement les phrases en ignorant le phénomène de la ponctuation (Jones, 1995) :

> *There are no current text based natural language analysis or generation systems that make full use of punctuation, and while there are some that make limited use, like the Editor's Assistant (Dale, 1990), they tend to be the exception rather than the rule. Instead, punctuation is usually stripped out of the text before processing, and is not included in generated text.*

Cette approche est en fait celle de Chandioux (1996) dans son correcteur *GrammR* (que nous n'avons pas analysé justement pour cette raison). Notre étude de la grammaire de *Word 98* et du *Correcteur 101* nous donne à penser qu'il s'agit là également de la solution préférée par leurs développeurs.

Cependant, dans un exercice comparé, Jones (1996a et b, 1994) a fait ressortir l'avantage du traitement de la ponctuation dans l'analyse automatique de phrases complexes *: For the longer sentences of real language* [...], *a grammar which makes use of punctuation massively outperforms an otherwise similar grammar that ignores it.*

Ce n'est que très récemment (Jones, 1996a) que les linguistes informaticiens se sont penchés sur le problème de l'absence du traitement de la ponctuation par les parseurs. En juin 1996, dans le cadre du 34e congrès[25] de l'*Association for Computational Linguistics*, Bernard Jones[26] organisait une journée d'ateliers, le SIGPARSE 96, qui réunissait pour la première fois les chercheurs les plus impliqués dans ce nouveau domaine.

Bernard Jones (1996a) et Ted Briscoe (1996a; 1994) ont développé chacun un parseur de type linguistique au moyen des *Alvin Natural Language Tools* (Carroll, Briscoe et Grover, 1991). Les résultats de ces parseurs révèlent la part essentielle jouée par la ponctuation dans l'interprétation des éléments constitutifs de la phrase (Jones, 1996a et b, 1994; Briscoe, 1996a). Bien que prometteuses, les succès de Jones et Briscoe sont néanmoins tempérés par deux éléments importants : leurs corpus

---

25 University of California at Santa Cruz, California.

26 Alors au Center for Cognitive Science, University of Edinburgh, Edinburgh, United Kingdom.

étaient manuellement pré-segmentés; la qualité de la ponctuation originale du texte ne constituait pas un élément de validation.

Jones (1996c : 62-64) introduit brièvement la question des erreurs de ponctuation dans l'analyse de ses corpus[27]. Son point de vue vise cependant à valider son parseur par rapport à la « l égalité » des parses obtenues plutôt qu'à estimer l'effet de ces erreurs dans la syntaxe des phrases. Il conclut néanmoins cette partie de l'étude en discutant l'impact de la littérature prescriptive et des erreurs dans l'analyse automatique de la ponctuation :

> *The result from this section of the investigation confirm not only that the most important punctuation marks are the full-stop and comma, and therefore that development of a correct treatment of these will be of the greatest benefit to the field of language analysis, but also that the strictly prescriptive style guides, and also the linguistic treatments of Nunberg (1990), whilst suitable for production of text, are too prescriptive for the analysis of raw text since punctuation patterns that do not occur or are actively disappproved of, do occur in texts. Thus any treatment should have the capacity to assign at least some meaning to the 'incorrect' punctuation patterns, otherwise systems will be of little use. This is more true in the field of punctuation than the related problems in the field of syntax, since there is less regularity in the use of punctuation system, and a great deal of idiosyncratic usage occurs.*

En identifiant ainsi la cause probable de l'inefficacité des correcteurs grammaticaux dans la détection de l'erreur de ponctuation — et souvent de l'erreur grammaticale tout court — ce commentaire de Jones (1996c) fait également ressortir le besoin de déterminer des solutions nouvelles pour traiter automatiquement l'erreur de ponctuation.

---

[27] particulièrement le corpus *Leverhulme* (350 000 mots) qui provenait de textes rédigés par des étudiants du secondaire.

Trois
Méthodologie

La correction automatique de la ponctuation est un domaine de recherche encore largement inexploré. C'est seulement depuis à peine une décennie que des linguistes informaticiens ont commencé à se pencher sérieusement sur le problème posé par la ponctuation dans le traitement automatique des langues naturelles. Les travaux de Nunberg (1990) d'abord, puis de Dale (1990, 1996), de Briscoe (1994, 1996a) de même que, plus récemment, de Jones (1996c) et de Say et Ackman (1997a, b) ont fait ressortir, entre autres, le rôle désambiguïseur de la ponctuation dans la détermination correcte des parses lors d'une analyse automatique de texte. Cependant, dans tous les corpus étudiés, la ponctuation est généralement assumée correcte par les chercheurs, surtout parce que ces corpus sont généralement sélectionnés parmi des textes correctement écrits[28] et traités manuellement pour accélérer l'analyse.

Jusqu'à présent, les travaux ayant porté sur la ponctuation automatique consistent principalement à reconnaître et désambiguïser les signes de ponctuation d'un corpus donné et à les exploiter de façon à améliorer les résultats du parseur. La question de la désambiguïsation des signes, particulièrement du point assertif qu'il faut distinguer du point abréviatif, telle qu'étudiée, par exemple, par Dister (1998), Palmer (1994) et Palmer et Hearst (1997), concerne encore une fois la reconnaissance de signes qu'on assume correctement placés.

Or l'un des problèmes de la correction automatique de la ponctuation consiste justement à tirer de l'information utile pour évacuer les erreurs de ponctuation d'un texte en dépit des complications d'analyse automatique apportées par ces erreurs mêmes. En théorie, un analyseur intégrant un module de correction automatique de

la ponctuation devrait donc pouvoir, non seulement analyser un texte en intégrant la ponctuation malgré les difficultés linguistiques et informatiques que cela pose, mais également mettre en doute l'opportunité des signes de ponctuation apparaissant dans un contexte donné et mettre de l'avant, au besoin, une méthode de correction appropriée. C'est pourquoi l'objectif principal de la présente étude consiste justement à déterminer les conditions pour que le processus de questionnement de la ponctuation d'un texte et sa correction éventuelle puissent s'effectuer automatiquement, même à partir d'un texte mal ponctué.

---

[28] Jones (1996c), dans l'élaboration de sa théorie informatique de la ponctuation, fait l'étude de plusieurs corpus d'expression anglaise, dont un seul réunit des textes d'opinion rédigés par des étudiants du secondaire.

La figure 42 fait la synthèse du déroulement méthodologique de notre recherche et y situe les thèmes discutés dans cette section.

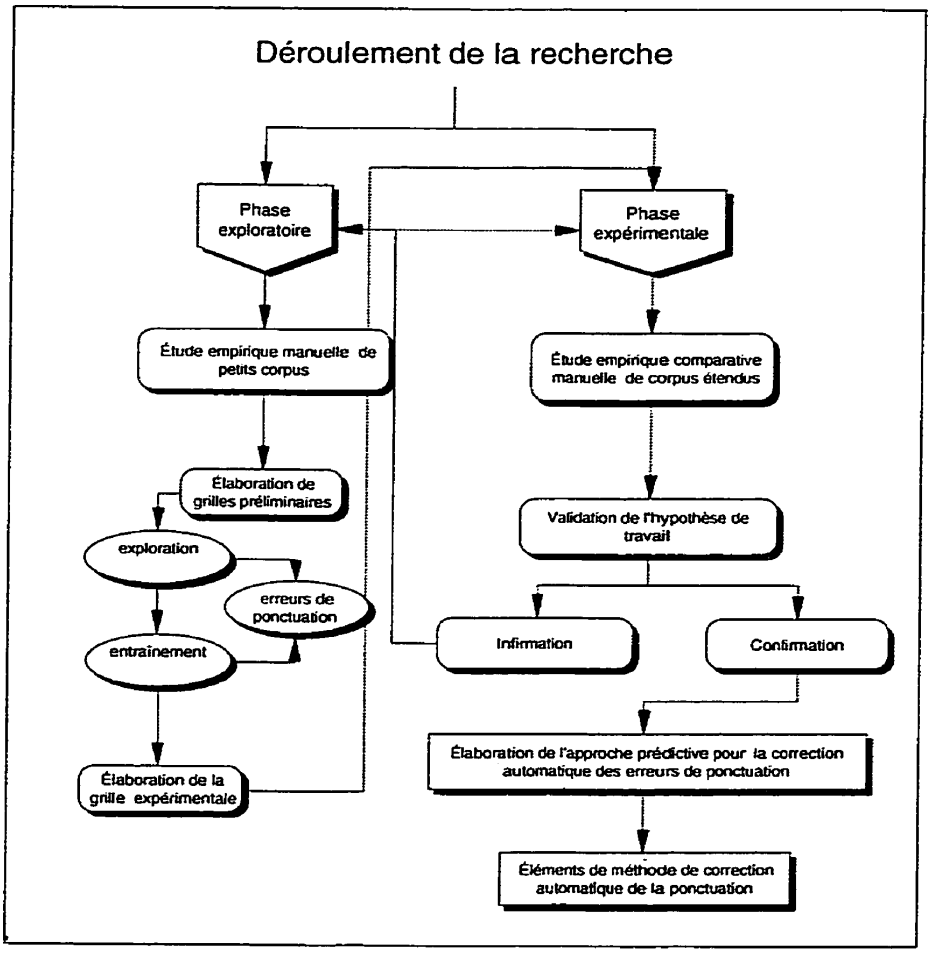


Figure 42

Déroulement méthodologique de la recherche

Notre recherche s'est déroulée en deux phases : une phase exploratoire et une phase expérimentale.

3.1 Phase exploratoire de la recherche

Nous avons trouvé difficile la phase exploratoire de notre recherche en raison surtout de l'incertitude liée à un domaine de recherche que nous n'arrivions pas à documenter directement. C'est pourquoi nous avons avancé avec prudence.

Toutefois, dans sa discussion de la méthodologie en intelligence artificielle, Cohen (1995 : 7) met en garde les chercheurs contre les risques d'éliminer l'étape incertaine de la recherche exploratoire au profit de la « sécurité » des tests et des manipulations :

> *Testing hypotheses has the panache of « real science », whereas exploration seems like fishing and assesment plain dull. In fact, these activities are complementary : one is not more scientific than the other; a research project must involve them all. Indeed, these activities might just as well be considered phases of a research project as individual studies.*

En réponse aux difficultés relevées pendant la première étape de notre recherche, nous avons développé une méthodologie plus adaptée à notre objectif principal.

3.1.1. Étude empirique manuelle de petits corpus

Certains obstacles rencontrés pendant nos premiers mois de recherche nous ont contraints à passer d'une méthodologie basée sur l'étude empirique automatisée de vastes corpus à une étude empirique manuelle de petits corpus.

La figure 43 résume notre cheminement décisionnel en exploitant des figures et des symboles pour illustrer certains aspects intangibles de cette étape. Les crochets indiquent une approche méthodologiquement connue; les cadrans, des étapes nous ayant demandé beaucoup de temps; les points d'interrogation, des aspects ayant posé problème et les boîtes marquées par un X de rature, des voies explorées sans succès ou les critères nous ayant permis d'évaluer la performance de notre logiciel de lisibilité.

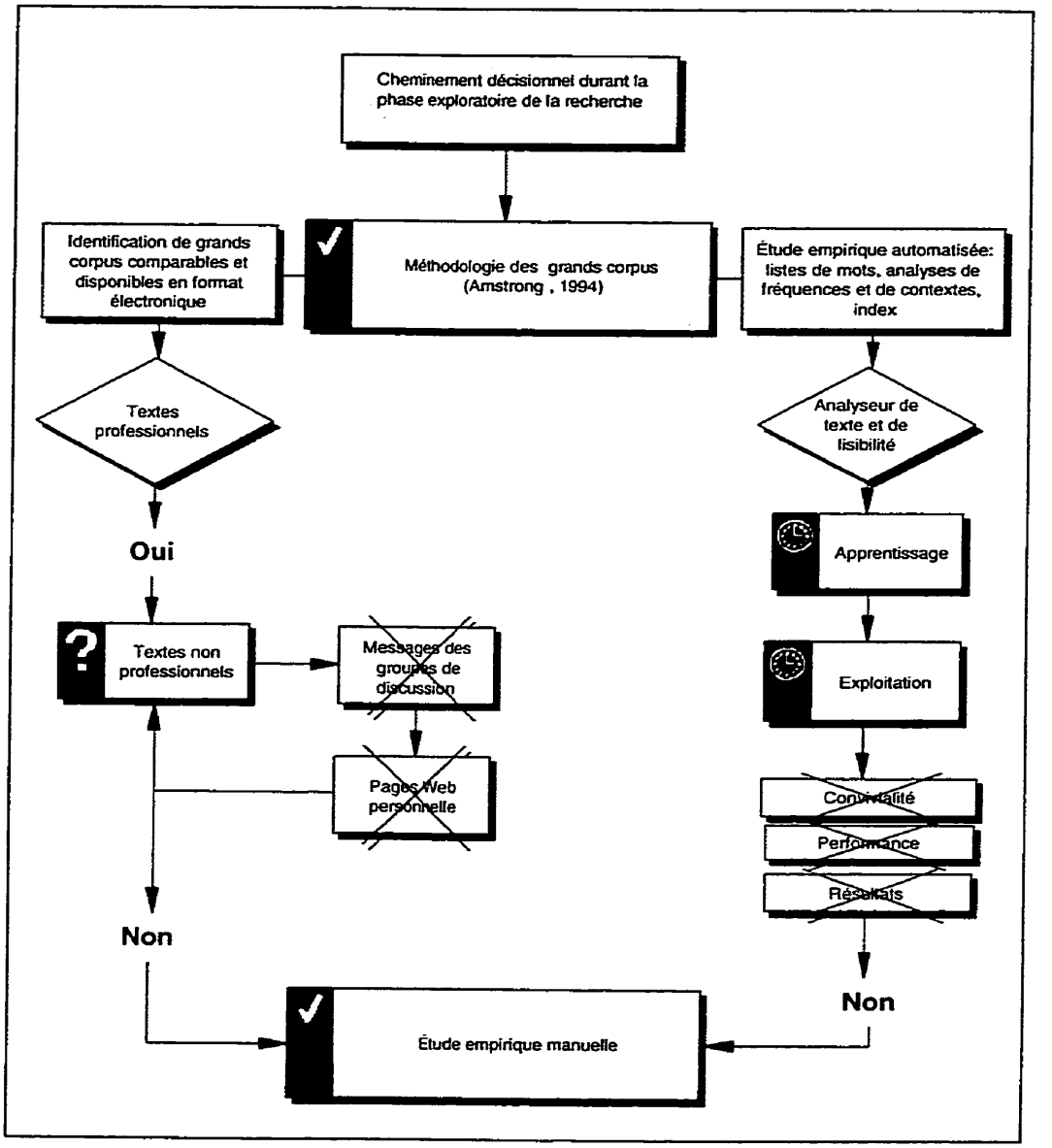


Figure 43

Cheminement décisionnel pendant la phase exploratoire de la recherche

Contre-performance de l'analyseur de lisibilité

Quelques mois après le commencement de notre recherche active, nous avons dû abandonner notre projet de travailler avec le logiciel d'analyse de texte et de lisibilité identifié durant le montage de notre projet d'étude (Simard, 1996b : 27).

Notre projet initial (Simard, 1996a : 12) prévoyait en effet que nous allions nous intéresser exclusivement à la correction d'erreurs de virgule à fonction de délimiteur en tentant ...

> [...] [d']*élaborer une matrice* [...] *listant les contextes et les fréquences les plus souvent associés aux virgules à fonction de délimiteur de même qu'à leurs erreurs d'emploi.*

Cependant, après plusieurs mois de travail, nous avons dû nous rendre à l'évidence que le montage automatique de fréquences et d'index utiles ne serait pas possible, principalement à cause des limites linguistiques et informatiques du logiciel que nous utilisions. Par exemple, l'une de ses fonctions non paramétrables définissait une phrase au moyen du point et d'autres signes de ponctuation forts comme le point-virgule et les points expressifs. Par ailleurs, entre autres limites informatiques, l'analyseur ne permettait pas à l'utilisateur de gérer les multiples fichiers secondaires générés par chacune de ses analyses. Or comme nous avions besoin d'examiner, non pas un seul grand texte, mais une grande quantité de textes de taille plus réduite, il devenait rapidement impossible de se retrouver dans le labyrinthe de fichiers secondaires ainsi créés. De plus, exploitant exclusivement la technologie déjà dépassée du DOS, la convivialité du logiciel et sa gestion désuète de la mémoire de l'ordinateur limitaient considérablement son utilité, en tout cas pour notre étude.

Notre recherche exploratoire, qui aurait déjà été incertaine dans un domaine bien documenté, se compliquait tout à coup du fait que la génération rapide de listes exhaustives de mots et de contextes devenaient à toutes fins pratiques impossible. La mise au rancart de notre analyseur de lisibilité nous empêchait désormais de penser lister les contextes et les fréquences associées aux erreurs de virgule à fonction de délimiteur et, *à fortiori*, à n'importe quel autre type d'erreur de ponctuation.

Absence de vastes corpus pertinents analysables

L'étude linguistique à partir de vastes corpus (*the corpus-based approach*) constitue une approche privilégiée par un nombre grandissant de linguistes (Armstrong, 1994 :vii) :

> *What is it that has brought about this rapid growth of interest in corpus-based NLP? For some, it is simply a rediscovery of empirical and statistical methods popular in the 1950s. Machine translation, for example, was at that time viewed as a 'mere' decoding problem, but computing resources were far from adequate for processing the data according to this model. The technological advances in computer power has certainly favored the reintroduction of this approach, as has the growing availability of large-scale textual resources in machine-readable form.*

C'est justement cette approche que nous avions planifié utiliser. Cependant, à peu près au moment où nous réalisions que le travail d'analyse informatisé serait impossible, nous nous rendions compte aussi que les textes en format électronique rédigés par des auteurs occasionnels n'étaient pas disponibles en nombre suffisant. Autrement dit, si nous pouvions répondre par l'affirmative dans la composition d'un corpus de textes de niveau professionnel, nous devions répondre par la négative au problème de la composition d'un corpus de textes non professionnels (Fig. 43).

Nous n'avions pas soupçonné cette difficulté pendant la planification de notre recherche du fait que Jones (1996c) avait tiré des groupes de discussion (*Newsgroups*), dont les messages sont caractérisés par une écriture immédiate et quotidienne, une partie importante de son large corpus. Cependant, sur examen attentif de ces textes, nous avons réalisé que, généralement courts et souvent rédigés en style télégraphique, ils ne constituaient pas une base comparative valide par rapport aux textes professionnels que nous avions accumulés facilement et en grand nombre.

En effet, les deux groupes de textes différaient trop l'un de l'autre sur le plan de critères non linguistiques. Ils se distinguaient non seulement sur le plan de la taille, mais souvent sur le plan de l'intention de communication. Ils ne se comparaient pas non plus sur le plan de leur contexte, qui favorisait la meilleure qualité possible pour les textes publiés, mais qui ne prédisposait pas à cette qualité pour les autres.

À défaut de recourir aux groupes de discussion, nous nous sommes repliés sur les pages *Web* personnelles. La difficulté, cette fois, a été de trouver des sites comportant des textes assez longs pour constituer une base comparative valide. En outre, plusieurs de ces pages étaient bâties par des tiers, si bien qu'il devenait difficile

d'en identifier l'auteur. Après plusieurs semaines de recherche infructueuse, nous avons finalement abandonné cette avenue.

Bref, non seulement ne pouvions-nous pas méthodologiquement nous guider sur l'expérience antérieure de chercheurs en correction automatique de la ponctuation, mais encore nous ne pouvions appliquer la méthode de travail documentée dans les travaux sur grands corpus. Il nous faudrait nous résoudre à exploiter des textes de petite taille que nous serions capables d'analyser manuellement, aussi bien pendant la phase exploratoire de notre recherche que pendant sa phase expérimentale.

Après avoir identifié des corpus de petite taille analysables manuellement, nous avons élaboré une grille préliminaire en calquant le travail d'un analyseur de texte et de lisibilité : comptage de mots; montage de listes et d'index; repérage de contextes signifiants.

La figure 44 montre notre cheminement après notre réorientation méthodologique. Cette fois-ci, nous pouvons voir que les étapes nous ayant demandé beaucoup de temps sont celles relatives au montage de nos grilles de travail. La boîte référant à notre grille d'erreurs de ponctuation se voit apposer un crochet parce que nous ne l'avons pas créée entièrement: elle a été adaptée d'une grille préexistante, celle de Guénette, Lépine et Roy (1995).

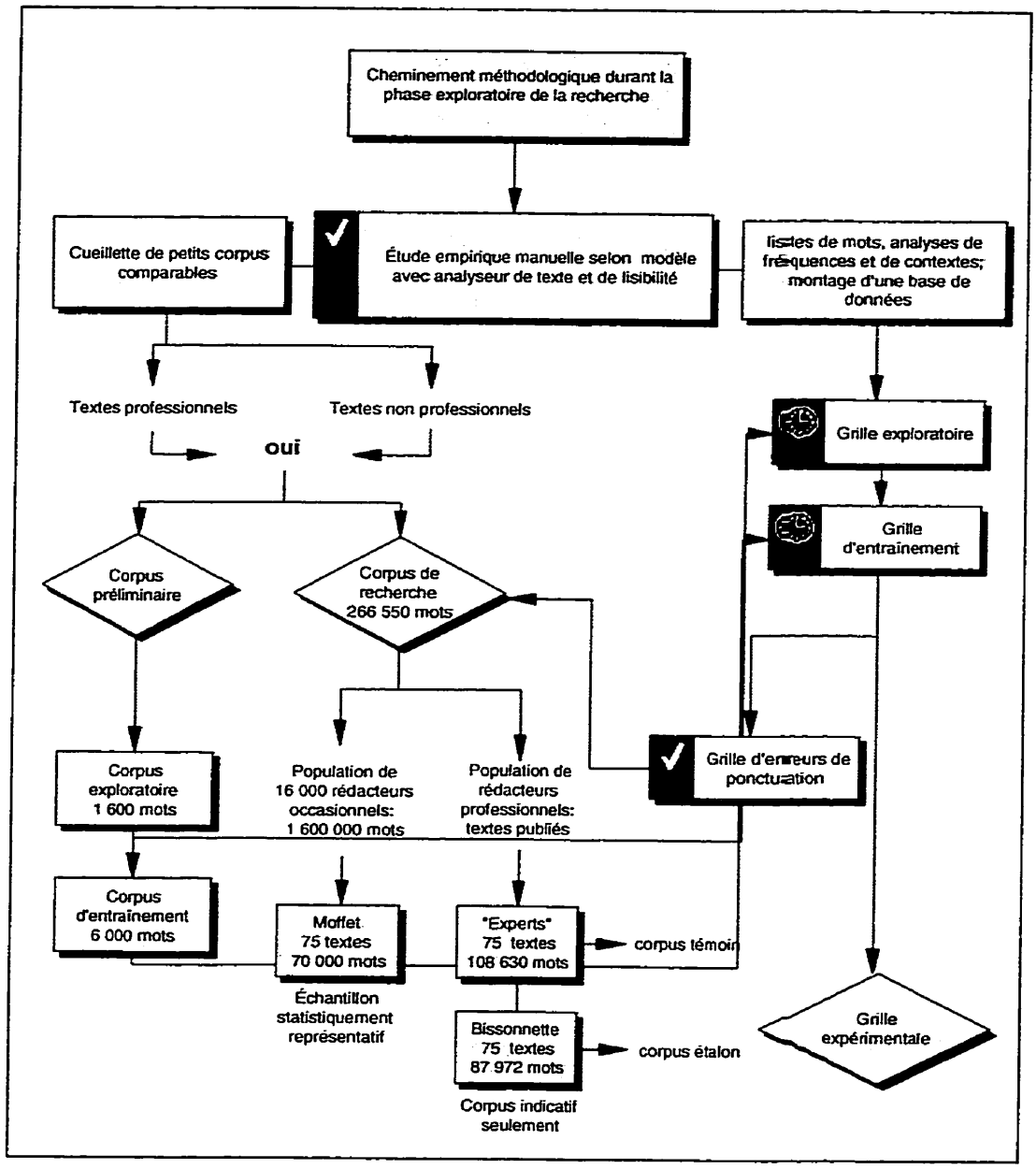


Figure 44

Cheminement de la phase exploratoire après réorientation méthodologique

3.1.2. Cueillette de petits corpus comparables

Nous avons constitué nos corpus en identifiant des textes comparables sur le plan extralinguistique mais distincts linguistiquement selon nos deux populations cible : les rédacteurs contrôlant leur français écrit et les rédacteurs ne le contrôlant pas.

Nous avons monté deux corpus : un corpus préliminaire qui nous permettrait de mettre au point une grille expérimentale et un corpus de recherche avec lequel nous testerions notre hypothèse de travail.

Corpus préliminaire

Le corpus préliminaire réunit en fait deux corpus: un corpus exploratoire et un corpus d'entraînement. Chacun de ces corpus comprend à son tour deux ensembles de textes : les corpus A désignent les textes écrits par des rédacteurs considérés comme professionnels; les corpus B rassemblent les textes de moins bonne qualité linguistique .

Le tableau 7 présente les critères de sélection du corpus préliminaire.

Tableau 7

Critères de sélection du corpus préliminaire

Critères externes (Corpus A et B)		Critères quantitatifs (Corpus A et B)		Critères qualitatifs			
exploratoire	entraînement	exploratoire	entraînement	exploratoire		entraînement	
				A	B	A	B
Format électronique		550 à 999 mots	1000 à 3000 mots	Sans erreur	Avec erreurs	Texte publié	Texte non publié

Corpus exploratoire

Le tableau 8 fait la synthèse de la description externe du corpus exploratoire.

Tableau 8

Description externe du corpus exploratoire

Rubriques	Corpus exploratoire	
	A Français écrit contrôlé	B Français écrit non contrôlé
Auteur	Lise Bissonnette, éditorialiste	Les frères Allot, ébénistes
Titre/s	« Après le rêve »	« Les frères Allot, ébénistes d'art »
Source	Le Devoir 3 septembre 1997	www.freres-allot.com
Nombre total de mots (Grands titres exclus)	932	782
Intention de communication	argumentation	argumentation

Qualitativement, les textes du corpus exploratoire (dont le tableau 9 montre des extraits), se trouvent aux deux pôles de la maîtrise linguistique. Rédigé au Québec dans un français élégant et recherché, le texte de Lise Bissonnette commentait la mort accidentelle de la princesse de Galles. Le texte de l'ébénisterie Allot en revanche, rédigé en France dans une langue maladroite qui cherchait néanmoins à projeter l'élégance et la maîtrise, présentait les mérites de l'entreprise familiale. Le texte Allot est d'autant plus intéressant que les efforts de contrôle de qualité de son auteur y sont visibles aussi bien sur le plan du ton que du style. Il constituait donc, sur le plan extralinguistique, une excellente base comparative.

Tableau 9

Extraits du corpus exploratoire

Lise Bissonnette	Les frères Allot
La princesse Diana était la femme la plus célèbre du monde. Il aurait fallu vivre sur un atoll désert d'un archipel encore inconnu – et encore – pour ignorer son existence. Il était logique, naturel, que les circonstances tragiques, spectaculaires et un brin sulfureuses de sa mort provoquent une réaction dont la démesure est évidente mais qui révèle, tout de même, des antidotes au cynisme des temps.	*Dans un pays de légende au cœur de la Bretagne à Loudéac exactement les meubles ont une histoire. La forêt de Brocéliande, haut lieu de nos traditions, c'est le domaine de Viviane la fée, Lancelot du lac, Merlin l'enchanteur.* *A l'orée de ces bois de Quénécan, de Lanoué, de la Hardounais, de Paimpont et de Loudéac commence l'aventure. Les frères Allot en compagnie de l'exploitant forestier partent à la recherche de leurs trophées : de la bille de bois de chêne, de chataîgner, de merisier, de noyer... A la culée, jugent la qualité des arbres et séjourneront une fois choisi, dans le clos de l'entreprise.*

Corpus d'entraînement

Le tableau 10 décrit le corpus d'entraînement selon des critères externes. Les textes de ce corpus provenaient de deux sources principales : le recueil électronique gravé sur CD-rom de la revue *Cap-aux-diamants* (Les Logiciels de Marque, 1996) et le corpus *Valsheshini*, recueil de textes administratifs et techniques non publiés auxquels nous avons eu accès par permission spéciale du Groupe John Chandioux.

Tableau 10

Critères externes du corpus d'entraînement

Rubriques	Corpus d'entraînement	
	A Texte publié	B Texte non publié
Auteur	Yves Bergeron, ethnologue	— informaticien
Titre/s	« Cuire et conserver les aliments » « Des collections retrouvées » « Une tradition de porteur d'eau »	Texte 07
Source	Cap-aux-diamants no 24, hiver 1991, p. 26-28 no 40, hiver 1995, p. 65 no 13, printemps 1988, p. 49-51	Corpus Valsheshini
Nombre total de mots (Grands titres exclus)	3645 1 324 703 1608	2028
Intention de communication	information	information

Qualitativement, les textes du corpus d'entraînement se distinguaient également par leur qualité linguistique. Le corpus d'entraînement A réunissait trois textes de l'ethnologue Yves Bergeron. Deux de ces textes s'intéressent à différents aspects de la vie quotidienne en Nouvelle-France; le troisième s'arrête à la question de l'inventaire des collections d'objets historiques entreposées dans les musées du Québec. Le corpus d'entraînement B était le septième texte du corpus *Valsheshini*. Rédigé par un informaticien (ou une informaticienne), « Val-07 » documente un logiciel développé spécialement pour accomplir des tâches administratives chez Hydro-Québec.

Les extraits présentés dans le tableau 11 font ressortir les différences entre les textes A et B sur le plan de la force linguistique. Nous n'y corrigeons aucune erreur présente dans le texte original, y compris dans l'extrait du texte A.

Tableau 11

Extraits du corpus d'entraînement

Yves Bergeron, ethnologue	— informaticien
Comment les habitants des villes s'approvisionnaient-ils en eau potable avant l'installation des réseaux d'aqueducs? Bien sûr, ils utilisaient les puits comme à la campagne mais ceux-ci étaient souvent contaminés par la proximité des latrines. On recueillait également l'eau de pluie dans des barils, mais cette eau servait surtout aux travaux ménagers. Aussi s'en remettait-ton le plus souvent aux services des porteurs d'eau. (p.3)	*Le code original fournit par la division contrôle des revenus ne contenait aucun commentaire et aucune documentation sur la structure du code ni sur les variables utilisées. De plus, les fichiers responsables de la sortie à l'écran et à l'imprimante furent complètement réécris pour permettre une modification du code plus facile. Il existe une redondance dans ces fichiers mais cette dernière est nécessaire pour pouvoir modifier le code rapidement et ce sans se demander si quelque-chose fut oublié. Les variables utilisées sont maintenant conforme au principe d'une variable/une signification, ce qui n'était pas le cas auparavant (et je vous jure que c'était mêlant...).. (p. 3)*

Notons que les deux textes du corpus d'entraînement s'éloignaient moins l'un de l'autre sur le plan de la qualité de la langue que ceux du corpus exploratoire. La dernière phrase de cet extrait du texte A contient par exemple une faute (*s'en remettait-t-on*).

Corpus de recherche

Le corpus de recherche réunit 225 textes totalisant un peu plus de 266 000 mots. Il comprend en fait trois groupes différents de textes — les textes Moffet, « Experts » et Bissonnette — mais, encore une fois, deux groupes de rédacteurs: des rédacteurs occasionnels (corpus Moffet) et des rédacteurs professionnels (corpus « Experts » et Bissonnette). Le corpus Moffet est le produit d'un échantillonnage statistique réunissant des textes d'opinion rédigés par des finissants de cégeps dans le cadre des épreuves de français du ministère de l'Éducation du Québec. Le corpus Bissonnette rassemble des éditoriaux signés en 1997 par Lise Bissonnette dans *Le Devoir*. Le corpus « Experts » présente des textes d'opinion parus dans les éditions électroniques de publications d'expression française entre 1990 et 1997. Le tableau 12 présente un sommaire du corpus de recherche.

Tableau 12

Sommaire du corpus de recherche

Corpus	Rédacteurs	Textes	Mots
Moffet	75	75	69 950
Experts	75	75	108 628
Bissonnette	1	75	87 972
Total	**151**	**225**	**266 550**

Textes de rédacteurs occasionnels: corpus Moffet

Le corpus Moffet comprend 75 textes réunissant près de 70 000 mots. Il provient d'un tirage aléatoire systématique parmi 16 084 dissertations critiques rédigées par des candidats à l'université dans le cadre de l'épreuve uniforme québécoise de français, tenue le 13 mai 1998[29]. Lors de cette épreuve, les étudiants sont invités à rédiger un texte d'opinion d'au moins 900 mots. La moyenne des textes colligés est d'environ 1 000 mots par texte, soit une population de textes de près de 2 millions de mots.

La population des candidats à l'université soumis à l'épreuve de mai 1998 provenait entièrement d'établissements collégiaux du Québec (Moffet, 1998 :2-3). Bien que non nécessairement représentative de la francophonie québécoise, canadienne ou internationale, la population de laquelle l'échantillon a été tiré représente néanmoins tous les finissants de céceps admissibles à l'université au moment de l'épreuve. Cette population réunit des sujets que nous pouvons considérer comme « instruits », puisque, au moment de l'épreuve, ils allaient recevoir un diplôme d'études collégiales. En outre, ils souhaitaient poursuivre une carrière professionnelle dont la formation requiert des études supérieures.

Le statisticien du MÉQ, Jean-Denis Moffet, décrit (Moffet, 1998 : 1) comment il a effectué la procédure de sélection de notre corpus:

---

[29] Les règles de cette épreuve de même que le devis et les grilles de correction sont accessibles sur l'Internet à *www.meq.gouv.qc.ca/ens-coll/Eprv_uniforme/mfrancais.htm.*

Pour sélectionner l'échantillon, à l'aide du logiciel Édustat, j'ai procédé de la façon suivante : j'ai déterminé que je désirais un échantillon pour des résultats moyens, puisque je connaissais l'écart type (12,5) de la population pour le critère de la langue, lequel j'ai retenu pour votre étude. J'ai choisi un niveau de confiance de 95 % avec une marge d'erreur de 3 % et un écart type de 13, ce qui a déterminé la taille de l'échantillon : 71 élèves. J'ai ramené ce nombre à 75.

Par la suite, j'ai choisi au hasard les individus devant faire partie de l'échantillon par échantillonnage systématique. J'ai aussi contrôlé le sexe des élèves et le collège d'origine.

L'échantillon constituant le corpus que nous avons appelé *Moffet* est représentatif de la population au chapitre du critère de la langue (Moffet, 1998 : 6; 9-10). Le tableau 13 compare les résultats de la population au chapitre du critère de la langue avec ceux de notre échantillon.

Tableau 13

Comparaison de l'échantillon Moffet avec sa population sur le plan de la langue

Corpus Moffet (Nombre de sujets : 75)		Population (Nombre de sujets : 16 084)	
Moyenne de fautes de syntaxe et ponctuation	Moyenne de fautes en orthographe d'usage et en orthographe grammaticale	Moyenne de fautes de syntaxe et ponctuation	Moyenne de fautes en orthographe d'usage et en orthographe grammaticale
9,4	11,7	9,8	11,6

Cependant, le corpus Moffet n'est pas représentatif de la population sur le plan du vocabulaire : l'échantillon présente une qualité de vocabulaire supérieure (Moffet, 1998 : 6; 9-10) Le tableau 14 met en relief la différence de résultats des sujets de l'échantillon par rapport à ceux des sujets de la population au chapitre du vocabulaire.

Tableau 14

Comparaison de l'échantillon Moffet avec la population sur le plan du vocabulaire

	A (0-9 fautes)	B (10-19 fautes)	C (20-30 fautes)	D (31-45 fautes)	E (46-60 fautes)	F (+ de 60 fautes)
Corpus Moffet Nombre de sujets : 75	77,3%	17,3%	5,3%	0,0%	0,0%	0,0%
Population Nombre de sujets : 16 084	68,8%	25,7%	4,9%	0,5%	0,1%	0,0%

Un élément est à souligner dans le corpus Moffet. On pourrait en effet arguer que si les textes Moffet sont représentatifs de la population d'où ils sont tirés, ils ne sont pas nécessairement représentatifs de la qualité linguistique de leurs auteurs. Nous serions en fait d'accord avec cet énoncé puisque ces textes représentent la meilleure production possible de rédacteurs se trouvant dans le contexte contraignant d'une épreuve officielle associée à un enjeu significatif : leur admission à l'université. Ces rédacteurs étaient donc fortement motivés à produire le meilleur texte possible. Sur le plan extralinguistique donc, leurs textes constituent une très bonne base comparative avec les textes A. L'Annexe 2 décrit les textes de l'échantillon Moffet et les classe selon le numéro d'identification arbitraire que nous leur avons attribué.

Le corpus Moffet se présente sous la forme de photocopies de textes manuscrits (Fig.45). La correction du MÉQ[30] et le résultat du sujet sont visibles sur chacune des copies. Conformément à l'entente de confidentialité qui nous lie au MÉQ, nous n'avons cependant accès à aucun élément d'identification. La page suivante reproduit la première page d'un texte de notre corpus, tel que reçu du MÉQ pour les fins de notre étude.

[30] Le critère de la langue, directement concerné par l'objet de notre recherche, était l'un des trois grands critères servant à juger linguistiquement les candidats. Les deux autres critères avaient trait à la qualité de l'argumentation et à la structure du texte (MÉQ, 1998 : 2).

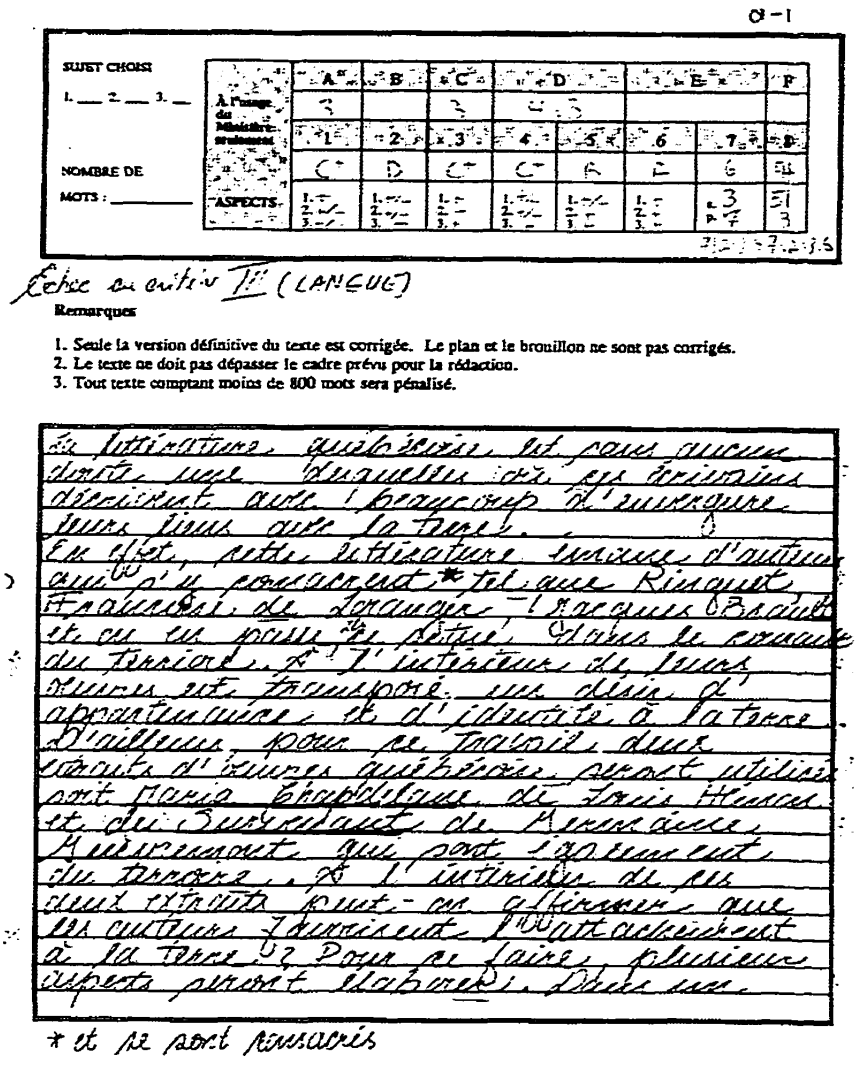


Figure 45

Exemple d'une page d'un texte du corpus Moffet

Textes de rédacteurs professionnels: corpus « Experts »

Le corpus « Experts » rassemble 75 textes d'opinion publiés sous la supervision d'une équipe de production d'un journal ou d'un magazine. En réunissant ces textes, nous avons tâché de diversifier autant que possible les origines professionnelles des auteurs de façon à disposer d'une bonne variété de styles d'écriture. Nous avons également recueilli quelques textes de rédacteurs français pour augmenter encore une fois les chances de diversité linguistique.

Cependant, les textes du corpus « Experts » représentent un ensemble indicatif seulement de la qualité linguistique des rédacteurs professionnels. Ils ne constituent pas un échantillon statistiquement significatif parce que leur tirage n'a pas été effectué selon les règles de la statistique. Il était en effet illusoire de prétendre effectuer un tel tirage en raison de l'importance de la population d'auteurs d'expression française. Il nous est apparu plus réaliste de nous en tenir à un choix de productions publiées en format électronique, disponibles soit sur l'Internet soit sur CD-rom.

Nous nous sommes quand même fixé certains critères pour retenir les textes du corpus « Experts » :

* le texte devait contenir au moins 800 mots, préférablement plus;
* le texte devait avoir fait l'objet d'une publication supervisée par un éditeur (éditorialiste ou directeur de publication);
* le texte devait avoir été rédigé par un seul auteur clairement identifié.

Les textes n'étaient pas lus préalablement, autrement que dans le processus du comptage de mots. Comme notre but était, cette fois-ci, de rechercher la variété, nous n'avons retenu qu'un seul texte par auteur, bien que, dans certaines occasions, plusieurs textes du même auteur aient été disponibles. La page suivante (Fig. 46) montre l'extrait d'un texte du corpus « Experts ».

Des coffres et de constitutions: archives des premières communautés religieuses

Archives

Les remparts de Québec encerclent, en un ruban de pierre de quelques kilomètres, une concentration d'archives documentaires et conventuelles unique au Canada.

par Christine C. Turgeon

Archiviste

Les chercheurs de toutes disciplines trouvent, dans l'intimité de ces dépôts, des collections et des fonds d'archives souvent intacts, encore logés dans les murs de l'institution qui les a produits, et présentant, la plupart du temps, un état remarquable de conservation. Cette pérennité, bien que troublée par plusieurs incendies aux XVIIe et XVIIIe siècles, et par les boulets de canon de la Conquête, fait notre admiration et l'étonnement même des archivistes étrangers qui nous rendent visite.

Les archives des institutions religieuses de Québec ont bénéficié d'une conservation exceptionnelle, grâce à une réelle politique de conservation qui repose, dès le XVIIe siècle, sur des règlements archivistiques précis, mais aussi sur la volonté d'hommes et de femmes d'entreprendre en cette terre d'élection qu'est la Nouvelle-France, leur existence légale, leur mission apostolique et leurs biens temporels.

Il nous a paru intéressant de retracer la genèse de la conservation des archives religieuses au Québec à travers trois de ses institutions les plus anciennes: l'Hôtel-Dieu, les Ursulines et le Séminaire. Plutôt que de décrire le contenu des fonds comme le voudrait la tradition archivistique, nous avons préféré étudier les conditions matérielles de leur conservation et les raisons d'être de ces archives.

Les «amazones du grand Dieu»

Sensibles aux appels lancés en 1635 par le père Paul Le Jeune dans les Relations des Jésuites, la duchesse d'Aiguillon, nièce du cardinal de Richelieu, et madame de la Peltrie, riche veuve d'Alençon, en Normandie, prêtent leurs fortunes à la fondation des deux premiers monastères féminins en Amérique du Nord: l'un abritant un hôpital administré par des religieuses de la congrégation des Hospitalières de Dieppe, et l'autre, une école destinée à «l'éducation des petites filles, tant des Français que des Sauvages du pays» et ayant à sa tête Marie de l'Incarnation, Ursuline de Tours.

Lorsque, le 1er août 1639, débarquent à Québec ces «amazones du grand Dieu» comme les appelle le père Le Jeune en renvoyant au mythe de ces femmes guerrières héritées de l'Antiquité», elles ont dans leurs bagages de toutes nouvelles constitutions récemment rédigées à Paris en 1631 pour les hospitalières de Dieppe et en 1635 pour les ursulines de Tours. Leur lecture est essentielle à la compréhension de l'histoire de la conservation des archives au monastère des augustines de l'Hôtel-Dieu de Québec, comme à celui des ursulines de la rue du Parloir. Les constitutions régissent tous les aspects de la vie communautaire de ces religieuses appartenant à des ordres cloîtrés et prévoient les lieux physiques de conservation de leurs archives dans des coffres et des voûtes difficiles d'accès, implantant de manière symbolique et quasi sacrée les monastères dans une colonie dont la légitimité repose sur des bases encore fragiles.

Le coffre à trois clefs

Dans un monastère comme dans l'autre, la première référence à une politique de conservation des archives apparaît dans le chapitre des constitutions consacré aux charges de la religieuse dépositaire: «on luy donnera un coffre, dans lequel seront gardés les titres et les papiers de la maison & ce coffre aura trois clefs différentes, l'une pour la mere, l'autre pour elle, & la troisiéme pour l'Assistante, afin que l'une ne puisse l'ouvrir, qu'en la presence de l'autre». Les mêmes précautions de conservation se trouvent exprimées dans

Figure 46

Extrait d'un texte du corpus "Experts"

Le corpus « Experts » compte près de 110 000 mots. Le tableau 15 présente la profession des auteurs des textes du corpus « Experts » de même que le lieu de publication de ces textes.

Tableau 15

Description des rédacteurs du corpus de recherche « Experts »

Catégorie de rédacteurs	Nombre dans le corpus	Textes tirés d'un :			Lieu de publication		
		journal	magazine	recueil	Québec	France	France Québec
Archéologue	1		•		•		
Archiviste	2		•		•		
Chroniqueur	5	•			•		
Critique d'art	1	•			•		
Éditorialiste	2	•			•		
Étudiant (ét. avancées)	1		•		•		
Historien	5	•	•		•		
Humaniste	1	•			•		
Journaliste	7	•			•		
Politicien	6	•			•		
Politicologue	1	•			•		
Professeur	4	•	•		•		
Psychologue	1		•		•		
Recherchiste	1		•		•		
Sociologue	1	•				•	
Correspondant	36	•		•	•	•	•

Les deux tiers des rédacteurs du corpus « Experts » sont des professionnels de l'écriture : correspondants, chroniqueurs, journalistes ou éditorialistes. Les autres rédacteurs du corpus de recherche A utilisent l'écriture dans l'exercice de leur profession: archivistes, historiens, professeurs, etc.

Par ailleurs, la très grande majorité de nos textes ont été tirés de publications québécoises, bien que nous ayons quand même retenu quelques textes publiés en France. L'Annexe 3 liste les textes du corpus « Experts » en les classant selon leur numéro d'identification.

Textes de rédacteurs professionnels: corpus Bissonnette

Le corpus Bissonnette réunit 75 textes d'opinion parus dans *Le Devoir* entre le 26 octobre et le 9 avril 1997. Il s'agit d'éditoriaux portant sur des questions politiques, sociales, économiques ou artistiques, signés par Lise Bissonnette, alors directrice du quotidien. Ils étaient disponibles en format imprimé seulement.

Nous avons utilisé le corpus Bissonnette comme étalon. D'une part, il nous aidait à situer les textes du corpus « Experts » aussi bien que ceux du corpus Moffet par rapport à un niveau élevé de maîtrise linguistique. D'autre part, il permettait de contrebalancer le fait que les textes « Experts » ne représentaient qu'un seul exemple de la production écrite de leurs auteurs. Toutefois, pour monter notre corpus étalon, nous avons préféré aux chapitres d'un seul volume une série de textes relativement courts, écrits au fil des semaines, de façon à nous assurer de la plus grande variété possible dans tous les aspects de l'écriture d'opinion de Mme Bissonnette : changement de thèmes, changement d'événements sociaux ou politiques, changement d'humeur. La variété de textes nous assurait ainsi de monter un corpus étalon structurellement bâti comme les deux autres corpus: textes suivis, relativement courts, défendant une opinion.

Nous avons porté notre choix sur Lise Bissonnette pour au moins trois raisons. Premièrement, Mme Bissonnette était une éditorialiste réputée avec de nombreuses années d'expérience en écriture d'opinion. Deuxièmement, les textes d'une pleine année de sa production de 1997 au *Devoir* nous étaient rendus disponibles grâce à l'un de nos contacts. Troisièmement, la qualité linguistique supérieure de ses textes est bien connue. La page suivante présente un exemple tiré du corpus Bissonnette (Fig. 47).

LE DEVOIR, LE MARDI 28 OCTOBRE 1997

Éducation Québec
Constitution

De piège en piège

Lise Bissonnette

Les opposants à l'amendement constitutionnel qui devrait ouvrir la voie à la laïcisation des structures scolaires au Québec ont pu profiter la semaine dernière de l'ignorance honteuse des membres du Comité mixte des Communes et du Sénat qui gobent n'importe laquelle sornette. Le ministre qui pilote le projet, Stéphane Dion, tancera sans doute ses collègues mais il n'en a pas fini avec les embûches.

Le premier tir de barrage aura été le moins difficile. Le lobby Alliance Québec aura encore une fois trouvé quelques épaules amies ou pleurer son anguisse inextinguible malgré que le gouvernement du Québec s'apprête à donner aux anglophones un système scolaire comme nulle autre minorité ne saurait en rêver, avec un degré de contrôle inégale en toute autre province. On aura beau faire, on n'empêchera jamais la députée de Mont-Royal, Sheila Finestone, de penser que son chez-soi est sordidement menacé de devenir un Maillardville et que nul ne peut être séparatiste sans être aussi ennemi des libertés. Mais M. Dion, qui aura au moins vu jusqu'où peut aller la mauvaise foi en ces milieux faussement opprimés, aura empêché que le débat soit détourné de sa fin.

Le deuxième tir, venu de groupes religieux, est plus complexe.

Fraîchement débarqués dans le dossier, les députés semblent avoir été impressionnés par la Coalition pour la confessionnalité scolaire, un assemblage hétéroclite de groupes canadiens et québécois dominés par des factions *«plus catholiques que les évêques»* pour reprendre le mot de M. Dion, qui est poli. La liste, qui prétend représenter rien moins que 645 000 personnes (pieuse exagération), racle les derniers recoins de la nostalgie, du Mouvement scolaire confessionnel jusqu'aux *Chevaliers du Saint-Sépulcre* en passant par l'Association des veuves de Montréal. Quiconque aurait suivi le débat scolaire au Québec depuis 30 ans ne pouvait les prendre au sérieux, mais ce fut le cas à Ottawa.

L'ignorance du comité mixte accentue les embûches.

La disparition de l'article 93, selon cette Coalition, ferait sauter *«un élément essentiel du compromis historique ayant donné lieu à la création du pays»*, et cet élément essentiel serait *«la protection des droits religieux en éducation»*. Voilà, à sa face même, une interprétation grossièrement abusive de l'article 93. Cet article était plutôt un compromis touchant surtout le Canada central, conclu en réciprocité pour protéger les droits scolaires des minorités religieuses, protestantes au Québec et catholiques en Ontario. Ce n'est que par un accident historique propre à la situation législative du Québec à l'époque de la Confédération si les *«majorités»* religieuses de Montréal et de Québec y ont aussi obtenu protection. L'esprit de la loi constitutionnelle était donc de préserver les droits à la dissidence de quelques-uns et non de consacrer les pratiques religieuses des majorités qui, par définition, détiennent le pouvoir et peuvent se protéger elles-mêmes. On est surpris de voir qu'un constitutionnaliste aussi chevronné que le sénateur Gérald Beaudoin, au lieu de chapitrer les auteurs de thèses aussi farfelues, soit tombé en révérence devant ces extrémistes au point de réclamer rien moins que la tenue d'un *«référendum»* au Québec. Que fait-il de l'unanimité des partis à l'Assemblée nationale qui débat de ces sujets depuis trois décennies, et qui représente l'ensemble des Québécois?

La cause des franco-protestants, qui font partie de cette bizarre coalition pour des raisons conjoncturelles, est plus complexe. Ce sont justement ces groupes qui sont visés par le droit à la dissidence que voulait protéger l'article 93. Dans l'étude de 1996 qui a été à l'origine de la démarche actuelle, les professeurs Jean-Pierre Proulx et José Woehrling, de l'Université de Montréal, proposaient de contourner la difficulté en amendant l'article 93 plutôt qu'en le supprimant entièrement. L'amendement aurait enlevé les protections confessionnelles consenties aux majorités à Montréal et à Québec, et mis fin à *«l'accident historique»*, tout en préservant partout sur le territoire le droit à la dissidence, en pratique réservé ainsi à quelques petits groupes franco-protestants. Tant à Québec qu'à Ottawa, les élus ont préféré la suppression de l'article 93, qui sert mieux la modernisation complète du système scolaire. Mais il n'est pas dit que les franco-protestants ne contesteront pas, devant les tribunaux, le droit des deux parlements à disposer ainsi de leur sort, d'autant que des arguments techniques pourraient leur ouvrir la voie.

Mais le piège le plus dangereux pourrait être tendu par le premier ministre canadien lui-même, M. Chrétien, qui a permis un vote libre aux Communes quand la province de Terre-Neuve a réclamé un amendement semblable, et qui pourrait se sentir obligé de le faire à nouveau. Le Parti libéral du Canada est en effet aux prises, ici, avec des choix idéologiques déchirants, du genre dont le premier ministre préfère habituellement se défiler.

Malgré qu'il porte la Charte canadienne des droits en cocarde, qu'il en fasse l'alpha et l'oméga de la vie publique au Canada, qu'il la sacralise en matière de droits linguistiques qui ne sont même pas des droits fondamentaux, le PLC n'a jamais perdu de sommeil parce que les droits vraiment fondamentaux que sont la liberté de conscience et de religion sont en pratique niés dans l'organisation scolaire québécoise. Pour protéger le statut confessionnel de nos écoles publiques, nos lois scolaires sont en effet truffées de clauses *«nonobstant»* qui ont annihilé la portée des chartes depuis leur adoption, ce dont à peu près personne ne s'émeut. Le débat actuel sur la laïcisation des structures scolaires met en lumière l'extraordinaire hypocrisie d'une société qui se prétend *«multiculturelle»* et qui devrait donc rêver d'institutions publiques communes, mais qui n'hésite pas, au contraire, à imposer les valeurs religieuses de groupes dominants aux nouveaux venus qui ne sont pas de mêmes croyances. Les caucus des différents partis, dont plusieurs comprennent des catholiques aussi impériaux que ceux de la Coalition pour la confessionnalité, risquent d'être fort divisés et même de faire sombrer le projet au moment du vote aux Communes.

S'ils étaient prononcés par des souverainistes québécois, certains des discours qu'on a entendus la semaine dernière à Ottawa sur la nécessité de préserver le caractère *«judéo-chrétien»* de nos sociétés, sur le droit des majorités d'obliger les minorités à s'y ranger, ou sur une immigration qui devrait se contenter de partager *«les privilèges»* obtenus *«par notre histoire et nos combats»*, ou sur les visées particulières de Dieu sur le pays, provoqueraient un scandale généralisé, seraient considérées comme les plus sombres discours du nationalisme *«ethnique»*. Mais on les laisse passer par peur d'affronter des débats moins *«historiques»* et très actuels sur les conditions et les exigences du pluralisme dont nos sociétés aiment se targuer. Le gouvernement du Québec a remis ce débat à une étape ultérieure à la réforme des structures, et confié à un comité le soin d'étudier la place et les modes de l'enseignement religieux à l'école. Le gouvernement du Canada pourrait tenter de l'esquiver par un vote libre qui lui évitera de proposer une direction. L'obstination désormais presque légendaire de M. Dion devrait, espérons-le, empêcher cette catastrophe.

Figure 47

Exemple d'un texte du corpus Bissonnette

Notre sélection représente en fait 75 des 76 textes les plus récents produits par Mme Bissonnette dans la série dont nous disposions[31]. Le 18ième texte de cette série se trouvait l'éditorial portant sur la mort de Lady Diana, déjà utilisé lors de notre étude exploratoire, que nous avons exclu pour des raisons évidentes.

Le corpus Bissonnette compte 75 textes pour un total d'environ 88 0 00 mots. Bien que non sélectionné par échantillonnage systématique, il représente néanmoins un bon indicateur du style d'écriture de l'éditorialiste. L'Annexe 4 décrit le corpus Bissonnette et classe les textes selon le numéro d'identification que nous leur avons attribué.

3.1.3 Élaboration de la grille expérimentale

Dans notre projet de recherche (Simard, 1996c : 12), nous avions planifié identifier des variables linguistiques que nous pourrions associer statistiquement à certains signes de ponctuation. Nous ne connaissions ni le nombre ni la nature de ces variables et nous ne pouvions qu'espérer qu'il existait bel et bien un lien quelconque entre au moins l'une de ces variables et la ponctuation d'un texte. Il nous fallait par conséquent élaborer une grille permettant à la fois de décrire objectivement une production écrite et d'en faire ressortir les caractéristiques.

Après avoir prétraité le texte pour en faciliter l'analyse (numérotation des lignes et évacuation des titres), nous avons entamé la lecture détaillée du corpus exploratoire en entrant nos données à mesure dans le logiciel de base de données *Panorama* (Provue, 1996), qui en facilitait la saisie et la comparaison.

---

31 Notre série commençait avec un éditorial, publié le 26 octobre 1997 et s'achevait avec un autre publié le 6 juin 1996. Nous avions reçu les textes classés selon leur date de publication, du plus récent au plus ancien. Nous n'avons pas modifié cet ordre au moment de notre sélection.

Montage des grilles préliminaires

Nos grilles préliminaires servaient à décrire le texte qualitativement.

Grille exploratoire

Notre première grille était très simple. Elle se limitait à classer les mots du texte selon les parties du discours mais en les regroupant en deux catégories : les mots pleins et les mots-outils. Le tableau 16 fait état des rubriques de cette grille.

Tableau 16

Rubriques grammaticales de la grille exploratoire

Mots pleins	*Mots-outils*
Substantifs;	pronoms relatifs;
Verbes;	pronoms indéfinis;
Adverbes en –ment;	pronoms impersonnels;
Adjectifs épithètes ou participes.	pronoms démonstratifs;
	pronoms interrogatifs;
	conjonctions ou locutions;
	adverbes autre que –ment;
	prépositions ou locutions,;
	adjectifs indéfinis.

Les pronoms personnels et les déterminants autres que les adjectifs indéfinis étaient ignorés selon le postulat que leur emploi – incontournable par rapport à la syntaxe française - ne pouvait varier de façon significative avec le niveau de maîtrise linguistique des auteurs. Chaque mot ou terme complexe du texte était entré manuellement dans la grille. Des sommes et des proportions étaient ensuite calculées pour identifier des valeurs exploitables.

Toutefois, cette grille ne convenait pas pour la description du texte B, qui contenait aussi des erreurs. Nous avons donc ajouté une grille permettant d'identifier ces erreurs selon la catégorie linguistique pertinente. Cette nouvelle grille comportait 10 rubriques :

* Occurrence;
* Orthographe;
* Grammaire;
* Syntaxe;
* Lexique;
* Ponctuation;
* Cohésion textuelle;
* Commentaire
* Ligne

La rubrique « Occurrence » contenait toute partie de texte présentant une erreur. Cette erreur était diagnostiquée dans la rubrique linguistique appropriée, commentée au besoin dans la rubrique « Commentaire » et référencée dans la rubrique « Ligne ». Les doutes en regard du diagnostic de l'erreur étaient résolus en consultant Guénette, Lépine et Roy (1995) et des grammaires normatives comme Grevisse et Goosse (1991). Encore une fois, des calculs de fréquences et de proportions tentaient d'identifier des pistes exploitables. La figure 48 montre un exemple de ces calculs.

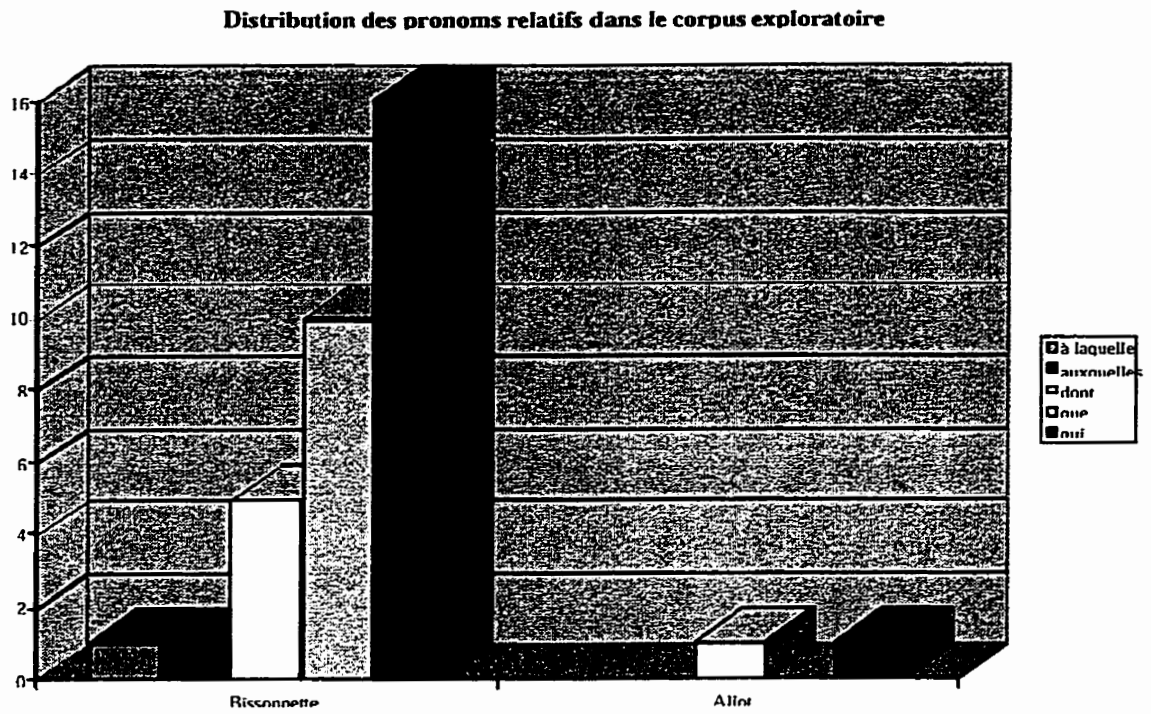


Figure 48

Exemple de traitement du corpus exploratoire

Dans le traitement de nos données préliminaires, nous avons exploré plusieurs voies. L'une de celles-ci consistait à comparer nos listes à celles du *Français fondamental* (1970). En effet, en postulant que la richesse lexicale d'un rédacteur peut se démontrer dans le nombre et la fréquence de mots tombant en-dehors du vocabulaire de base du français, nous pensions que nous pourrions démontrer la pauvreté lexicale d'un texte en identifiant tous les mots ne faisant pas partie du français de base. Dans cet exercice, les mots caractéristiques d'une langue technique, d'un jargon de métier ou empruntés d'une langue étrangère (sans être passés dans le vocabulaire courant selon le dictionnaire) étaient classés sous une rubrique « langage spécialisé ». Cependant, quand il s'est révélé qu'il nous était impossible de trouver une mise à jour récente du document de référence, nous avons abandonné cette voie en raison des problèmes potentiels de validité posés par une comparaison avec des listes vieillies et montées au moyen techniques de recherche désuètes.

Notre étude exploratoire a cependant permis de mettre en relief plusieurs observations prometteuses. Tout d'abord, la détection de l'erreur linguistique

semblait inévitable puisque l'erreur paraissait l'élément déterminant pour distinguer le niveau de contrôle linguistique des deux textes. Ensuite, il semblait exister une différence significative entre la fréquence et la variété des mots-outils selon les textes du corpus. Également, la différence lexicale principale entre les deux textes paraissait se trouver non pas dans l'emploi de mots rares, mais dans le nombre de répétitions abusives de mots ou de structures, même rares. En outre, contrairement au texte A, il n'y avait pas de subjonctifs rares dans le texte B; on y trouvait par contre des passés simples, mais inappropriés. Finalement, le texte B ne présentait pas d'occurrence de tirets intercalaires alors que le texte A en comportait.

La validité de ces observations se limitait aux seuls textes observés. Qu'en serait-il avec des textes plus longs? C'est ce que notre corpus d'entraînement servirait à vérifier.

Grille d'entraînement

Nous avons conservé la même grille (pour faciliter les comparaisons avec les conclusions déjà tirées) mais modifié son véhicule. Pour faciliter le transfert des valeurs de notre base de données à notre tableur, nous nous sommes constitué une nouvelle base au moyen du logiciel *File Maker Pro* (*FMP*; File Maker Pro, 1996; 1999), lequel contient une fonction permettant l'échange de données entre les deux logiciels.

Les données recueillies à partir du corpus d'entraînement ont confirmé plusieurs de nos observations préliminaires sauf une : les textes d'entraînement A présentaient aussi des erreurs, aussi bien linguistiques que de ponctuation. Bien que beaucoup moins nombreuses que dans le texte d'entraînement B (94 contre 231), ces erreurs suggéraient un niveau d'expertise linguistique moins élevé que le texte exploratoire A, malgré le fait qu'il avait été produit sous la supervision d'une équipe de rédaction. Voilà qui jetait un éclairage nouveau sur le problème et donnait à penser que la détection des erreurs devrait faire partie de toute stratégie d'évaluation de niveau de contrôle linguistique.

Grille « Erreurs de ponctuation »

Notre grille « Erreurs de ponctuation » adaptait celle de Guénette, Lépine et Roy (1995) en utilisant toutefois la terminologie de Nunberg (1990), Briscoe (1996a) et Jones (1996c).

Cette grille prévoit 3 catégories principales d'erreurs de ponctuation : les erreurs d'omission de signes, les erreurs de confusion de signes et les erreurs relatives à une ponctuation abusive (signes indus). Nous avons décrit notre corpus de recherche avec cette grille, comme le montre l'extrait de la page suivante (Fig. 49).

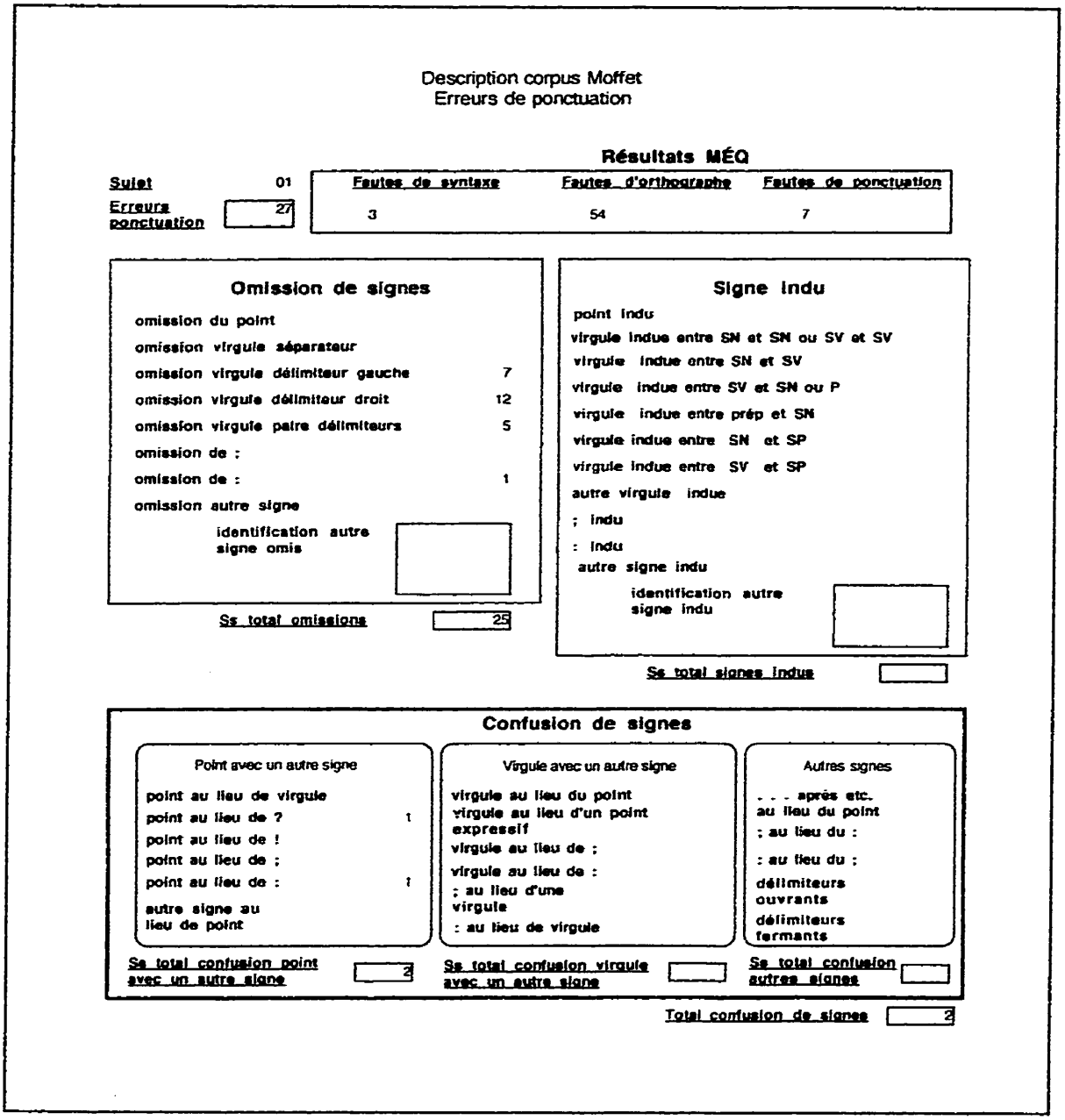


Figure 49

Base de données "Ponctuation" : Sujet Moffet 01

Nous avons bâti notre base de données de façon à générer automatiquement le calcul des résultats pour chaque texte de notre corpus.

Notre grille « Erreurs de ponctuation » a été essentielle à l'étape exploratoire de notre étude. Bien qu'elle ait été appliquée à notre corpus de recherche pour assurer la plus grande certitude possible aux observations qui y étaient notées, elle a

permis d'abord d'évacuer les impasses probables comme celle de la détection directe des erreurs de ponctuation en temps réel. Elle nous a également amenés à établir des objectifs et des limites de recherche tenant compte des habitudes réelles de ponctuation de nos sujets-cibles. Elle a finalement aidé à cerner notre problématique à partir de bases aussi solides que possible, élément non négligeable en l'absence d'une documentation objective appropriée.

Nous n'avons pas tenu compte de la correction du MÉQ en regard des erreurs de ponctuation pour maintenir une base de lecture commune pour l'ensemble des textes. En appliquant en effet la grille de Guénette, Lépine et Roy (1995) pour la ponctuation, nous nous assurions de maintenir une cohérence par rapport au dépouillement du corpus de recherche, puisque notre grille expérimentale constituait une adaptation de celle de Guénette, Lépine et Roy (*op. cit.*) à partir des observations effectuées avec le corpus préliminaire.

Nous avons effectué de nombreuses lectures de la ponctuation du corpus de recherche pour diminuer les risques d'erreurs. Comme les deux tiers de ce corpus étaient des photocopies et qu'en outre, l'échantillon Moffet était écrit à la main, nous pouvions à tout moment commettre des erreurs ou oublier des occurrences. La seule façon de nous prémunir contre ce problème consistait à effectuer des lectures répétitives à tête reposée et entrer systématiquement les données à mesure qu'elles étaient observées. Les chances sont bonnes pour que certaines occurrences aient été manquées, spécialement dans les textes Moffet. Cependant, il est douteux qu'elles soient en nombre suffisant pour invalider les conclusions sur lesquelles nous avons bâti notre problématique.

Précisons enfin que plusieurs catégories de données de notre base « Erreurs de ponctuation » ont été ignorées parce qu'elles tombaient en-dehors du champ de couverture de cette recherche. Elles avaient cependant été notées par souci d'exhaustivité et parce que nous voulions être sûrs de disposer de toutes les données exploitables en cas de rapports possibles entre les erreurs de ponctuation elles-mêmes.

Détermination des rubriques de la grille expérimentale

Nous avons tenté d'identifier sur quoi reposaient les différences entre les textes exploratoire et d'entraînement A. Outre la présence d'erreurs, pouvions-nous repérer des éléments qui nous permettraient de reconnaître le contrôle linguistique en particulier? Par exemple, on ne trouvait pas de tirets intercalaires dans le texte d'entraînement A, non plus que des emplois rares du subjonctif. Mais c'était bien peu. Cependant, notre intuition nous poussait à investiguer ces éléments, de même que tous les autres susceptibles de se trouver dans les textes professionnels aussi bien que non professionnels.

Il nous fallait une grille plus articulée. Montée avec le logiciel de base de données *File Maker Pro* (1996, 1999), notre grille expérimentale comporterait des rubriques permettant de décrire objectivement un texte sous deux angles : les erreurs linguistique et les signaux possibles de maîtrise.

Nous avons alors pris deux décisions importantes. La première: nous n'aurions pas une rubrique pour chaque erreur documentée par Guénette, Lépine et Roy (1995). La deuxième: le choix des rubriques ne tiendrait pas compte de la difficulté de traitement automatique des éléments décrits.

En effet, comme nous cherchions à déterminer s'il existait un rapport entre les signes de contrôle d'un texte et la ponctuation, il ne nous apparaissait pas nécessaire d'en vérifier exhaustivement tous les aspects linguistiques à partir du moment où nous pourrions établir ce rapport. Ce sont d'autres chercheurs qui pourraient peut-être se donner comme objectif de « cartographier » exhaustivement tous les signes possibles, si un tel rapport existait.

Par ailleurs, nous ne pouvions exclure un indice prometteur possible en raison des limites technologiques actuelles. D'une part, ces limites sont constamment repoussées. D'autre part, notre objectif consistait à déterminer les conditions de correction automatique de la ponctuation. Par conséquent, si c'était des limites technologiques qui constituaient le seul empêchement majeur au traitement correctif automatique de la ponctuation, nous établirions du moins les conditions pour qu'une telle correction soit possible dans un environnement technologique plus évolué.

Grille « Erreurs »

Nous nous sommes tournés vers la grille de Guénette, Lépine et Roy (1995). Cette grille organise la langue — et les erreurs associées — en sept catégories (Guénette, Lépine et Roy, 1995 : IX-XXVIII):

* orthographe lexicale;
* grammaire;
* phrase
* ponctuation
* vocabulaire
* style
* texte

Chacune de ces catégories réunit un ensemble d'erreurs possibles. Par exemple, la catégorie « Vocabulaire » comprend neuf types d'erreurs (Guénette, Lépine et Roy, 1995 : XXIII) :

* Erreur sur le sens du mot ou de l'expression;
* Mot ou expression manquant de précision;
* Incompatibilité sémantique entre deux mots ou expressions;
* Combinaison boiteuse;
* Termes inutiles ou redondances
* Barbarisme;
* Altération d'une expression figée;
* Anglicisme;
* Archaïsme.

Nous avons fait une sélection parmi les erreurs listées dans ces catégories selon les observations de notre corpus d'entraînement et bâti une première base de données, la base « Erreurs ». Cependant, dans certains cas, par souci d'économie de moyens (nous devions, faut-il le rappeler, travailler à la main) et en vertu de notre décision de ne pas viser l'exhaustivité, nous avons effectué des regroupements : par exemple, « Accords de SN » pour tous les problèmes d'accords de l'épithète avec le

nom et du nom avec le déterminant; « Homophones » pour tous les problèmes homophoniques autres que *à / a* et *on / ont*.

Grille « Maîtrise »

Les listes de Guénette, Lépine et Roy (1995) révèlent toutefois plus que des erreurs : elles suggèrent également des indices de maîtrise susceptibles d'être présents dans un texte contrôlé. Le tableau 17 illustre ce fait en donnant l'exemple des erreurs du vocabulaire et des indices de maîtrise linguistiques suggérés.

Tableau 17

Erreurs de vocabulaire et éléments de maîtrise suggérés

Erreurs (Guénette, Lépine et Roy, 1995)	*Éléments de maîtrise linguistique associés*
Erreur sur le sens du mot ou de l'expression	Mot juste
Mot ou expression manquant de précision	Mot précis
Incompatibilité sémantique entre deux mots ou expressions	Enchaînement sémantique correct entre les mots ou les expressions
Combinaison boiteuse	Combinaison lexicale juste
Termes inutiles ou redondances	Économie de mots
Barbarisme	Néologisme autorisé par les règles de composition et de dérivation françaises
Altération d'une expression figée	Respect de la formulation des expressions figées
Anglicisme	Mots du français seulement ou emprunts signalés par des guillemets
Archaïsme	Mots du français contemporain seulement ou termes vieillis signalés par des guillemets

Nous avons ainsi établi une liste de signes de « maîtrise » qui pouvaient correspondre aux erreurs de notre grille « Erreurs » tout autant qu'aux indices que nous avions observés dans le corpus préliminaire A.

3.2 Phase expérimentale de la recherche

La phase expérimentale de notre recherche a mené à l'identification d'éléments de méthode pour corriger automatiquement la ponctuation basée sur une approche prédictive par indices de calibrage.

La figure 50 fait la synthèse de notre cheminement méthodologique pendant la phase expérimentale de la recherche et y situe les étapes discutées dans cette section.

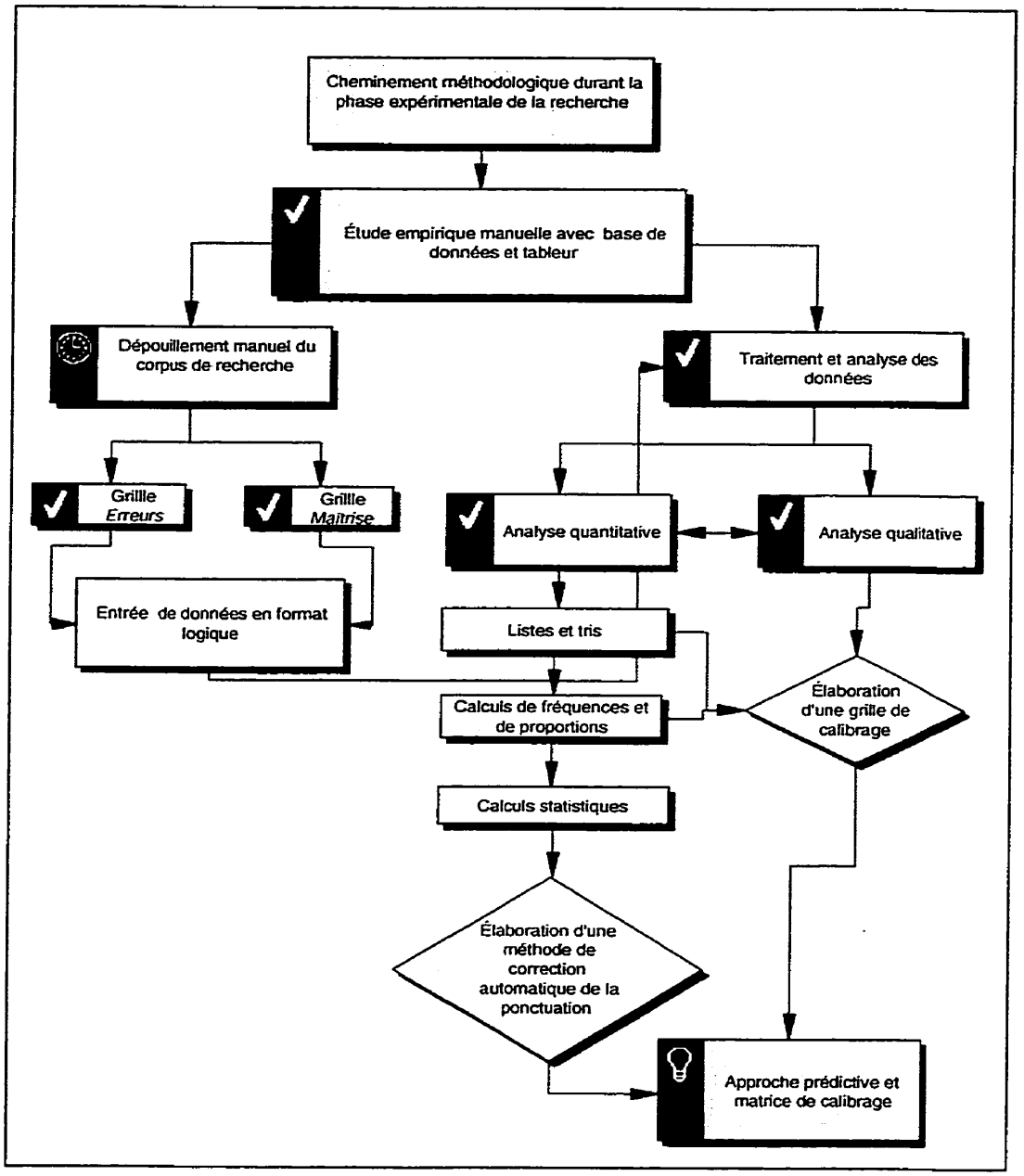


Figure 50

Cheminement méthodologique de la phase expérimentale de la recherche

3.2.1 Dépouillement manuel du corpus de recherche

Les textes du corpus de recherche ont d'abord subi un pré-traitement. Ce pré-traitement consistait en trois opérations : attribuer un numéro d'identification à l'auteur du texte, sauvegarder la copie originale et en compter les mots quand ce décompte n'était pas déjà disponible.

Le numéro d'identification des textes a été attribué arbitrairement. Les textes du corpus Moffet sont numérotés de 1 à 76[32], ceux du corpus Bissonnette, de 1 à 76[33] et ceux du corpus « Experts », de 101 à 175.

Sauvegarder l'état original des textes du corpus de recherche était important pour maintenir leur intégrité. Nous avons photocopié les textes manuscrits et effectué une copie de sauvegarde sur disquette pour les textes électroniques. Le corpus original a été ensuite placé en lieu sûr.

Par ailleurs, nous avons utilisé deux méthodes de comptage de mots. Nous avons compté manuellement les mots des textes non disponibles en format électronique selon la méthode recommandée par le MÉQ pour ses correcteurs[34]. Les mots des textes disponibles en format électronique ont été comptés à l'aide de la fonction « Statistique » du texteur *Word* en excluant les titre et sous-titre, aucune copie du corpus Moffet ne comprenant de titre ou de sous-titre.

Le corpus Moffet a toutefois subi un pré-traitement additionnel. Constitué en fait de 75 examens écrits comptant chacun un certain nombre de feuilles, nous avons dû paginer manuellement chaque feuille du corpus Moffet afin de faciliter la référence des occurrences.

---

32 Une erreur dans l'attribution des numéros d'identification des sujets du corpus Moffet a ajouté artificiellement un numéro 76.

33 C'est-à-dire 1 texte étudié pendant la recherche exploratoire et 75 textes étudiés pendant la recherche expérimentale.

34 La méthode appliquée par le MÉQ est la suivante : compter le nombre de lignes du texte et le nombre de mots dans dix lignes choisies au hasard : la multiplication des deux résultats donne une approximation du nombre de mots du texte

La cueillette de données du corpus de recherche s'est opérée en deux temps. Nous avons annoté manuellement les textes en appliquant systématiquement des codes combinant des abréviations et des couleurs qui traduisaient les rubriques de notre grille expérimentale. Nous avons ensuite transféré le résultat de ces annotations dans la base de données. Pour limiter les risques d'erreurs, tous les textes ont été lus à plusieurs reprises, chaque lecture servant à valider les résultats de la précédente.

Les textes du corpus de recherche ont fait l'objet de deux descriptions : une description externe et une description interne.

Description externe du corpus de recherche

Le tableau 18 met en parallèle les rubriques « Description externe » du corpus de recherche.

Tableau 18

Rubriques de description externe du corpus de recherche

Corpus Moffet	Corpus Bissonnette	Corpus « Experts »
Sujet	Numéro texte	Numéro expert
Pointage « Syntaxe et ponctuation » MÉQ	Titre du texte	Expert
Nombre fautes syntaxe	Référence Devoir	Titre
Nombre fautes ponctuation	Nombre de mots	Titre du texte
Fautes orthographe d'usage et grammaticales		Référence
Nombre de mots		Nombre de mots

La description externe du corpus Moffet enregistre les données relatives à la correction du MÉQ. Le nombre de fautes de syntaxe et de ponctuation détaille le pointage du MÉQ parce qu'il en diffère parfois en raison d'une directive particulière du ministère. Cette directive demande en effet aux correcteurs de distinguer, pour la ponctuation, entre des fautes « mineures » — par exemple, l'absence ou la présence indue d'un point d'interrogation — et « majeures » — par exemple, l'absence ou la

présence indue du point ou du point-virgule (MÉQ, 1998 : 91). Deux erreurs mineures équivalent à une erreur majeure.

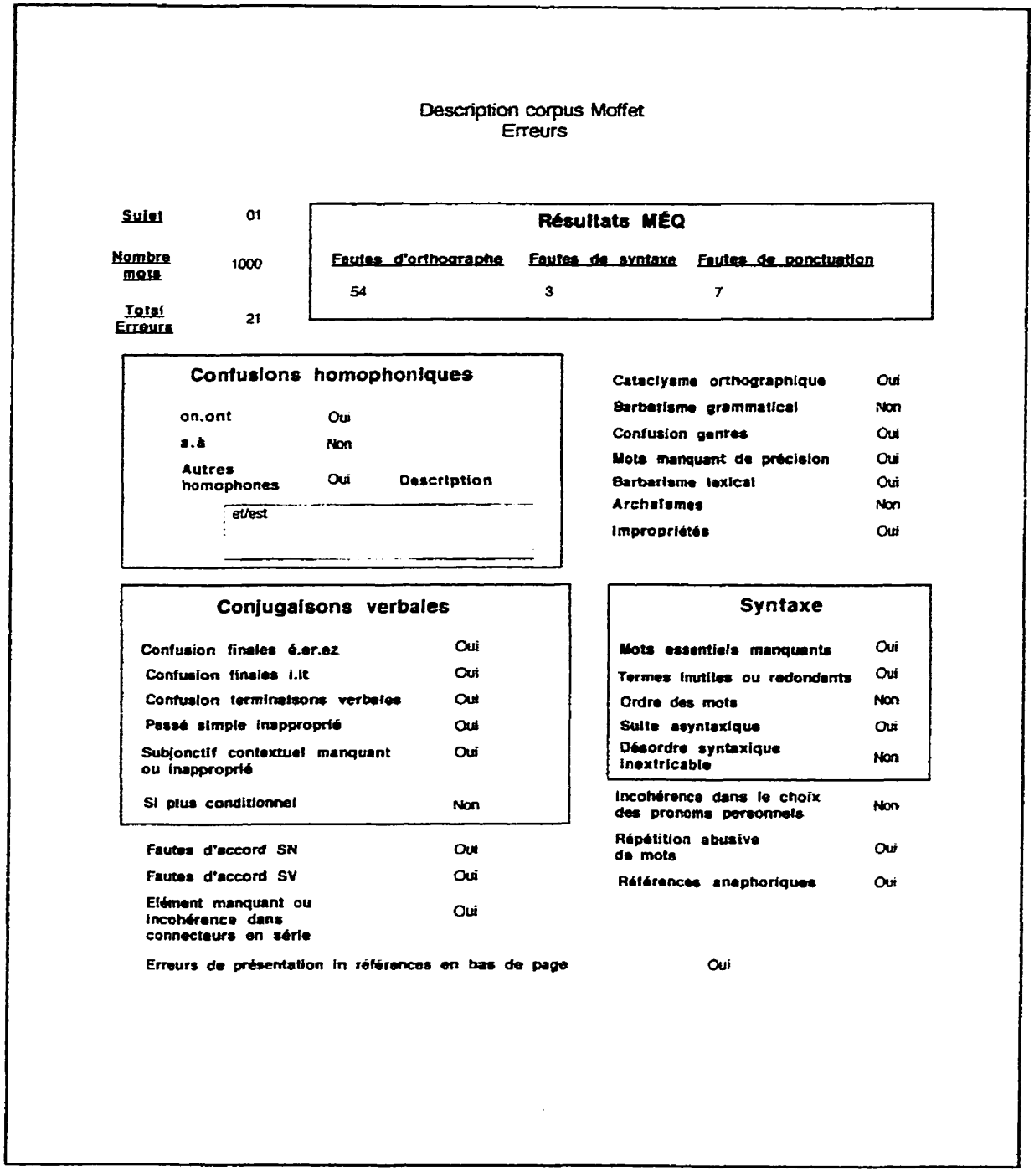
Les rubriques de description externe des corpus Bissonnette et « Experts » se ressemblent. Les auteurs des textes du corpus « Experts » sont identifiés sous l'étiquette « Experts » et leur occupation, sous l'étiquette « Titre ». Autrement, les rubriques de description externe de chacun de ces deux corpus informent sur le titre des textes, leur taille (déterminée par le nombre de mots) et leur source.

Description interne du corpus de recherche

Notre grille expérimentale a servi à décrire la qualité du français des sujets du corpus. La grille de calibrage que nous présentons au chapitre suivant est la version validée de cette grille.

Le format manuscrit du corpus Moffet nous a cependant imposé certaines contraintes sur le plan de la cueillette de données. Il nous a été par exemple impossible de constituer des listes lexicales fiables en raison des risques élevés d'erreurs non repérables. Ce facteur a dicté une décision majeure : utiliser la numérotation logique pour marquer la présence (*oui* = 1) ou l'absence (*non* = 0) des erreurs et des indices de maîtrise linguistique identifiables dans le corpus.

Les pages suivantes présentent les rubriques de notre grille expérimentale « Maîtrise » et « Erreurs » de même qu'un exemple de fiches (Fig. 51 et 52) telles que compilées dans notre base de données expérimentale.



Description corpus Moffet
Erreurs

Sujet	01
Nombre mots	1000
Total Erreurs	21

Résultats MÉQ

Fautes d'orthographe	Fautes de syntaxe	Fautes de ponctuation
54	3	7

Confusions homophoniques

on.ont	Oui	
a.à	Non	
Autres homophones	Oui	Description
		et/est

Cataclysme orthographique	Oui
Barbarisme grammatical	Non
Confusion genres	Oui
Mots manquant de précision	Oui
Barbarisme lexical	Oui
Archaïsmes	Non
Impropriétés	Oui

Conjugaisons verbales

Confusion finales é.er.ez	Oui
Confusion finales i.it	Oui
Confusion terminaisons verbales	Oui
Passé simple inapproprié	Oui
Subjonctif contextuel manquant ou inapproprié	Oui
SI plus conditionnel	Non

Syntaxe

Mots essentiels manquants	Oui
Termes inutiles ou redondants	Oui
Ordre des mots	Non
Suite asyntaxique	Oui
Désordre syntaxique inextricable	Non

Fautes d'accord SN	Oui
Fautes d'accord SV	Oui
Elément manquant ou incohérence dans connecteurs en série	Oui
Incohérence dans le choix des pronoms personnels	Non
Répétition abusive de mots	Oui
Références anaphoriques	Oui

Erreurs de présentation in références en bas de page Oui

Figure 51

Base de données "Erreurs » : Sujet Moffet 01

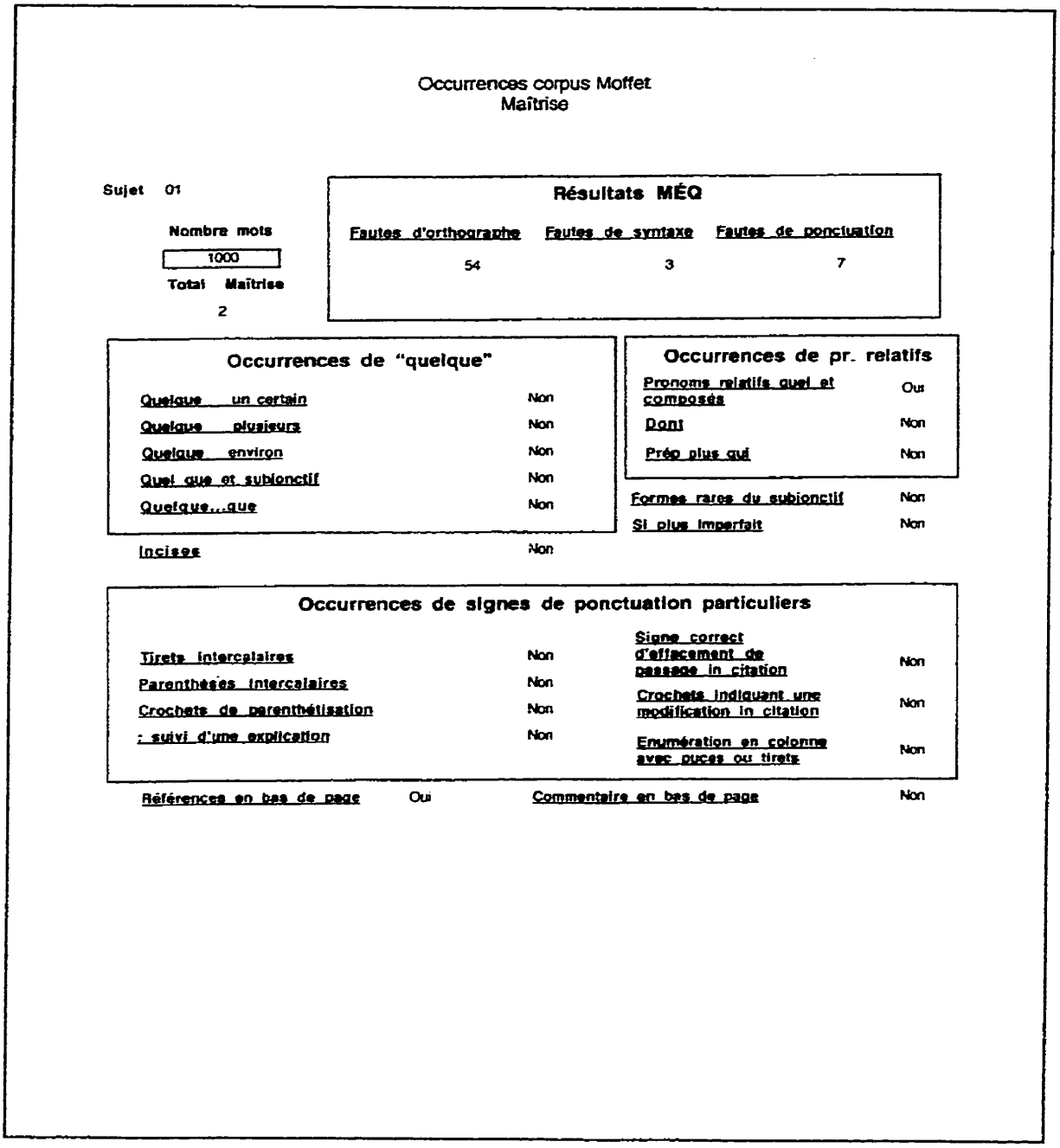


Figure 52

Base de données "Maîtrise : Sujet Moffet 01

3.2.2 Traitement des données

Nos données ont été examinées sous leurs aspects quantitatifs et qualitatifs.

Aspects qualitatifs

L'étude qualitative de notre corpus de validation s'est effectuée sous deux aspects :

* l'étude combinatoire des grilles « Erreurs » et « Maîtrise » comme outil d'évaluation du niveau de contrôle linguistique d'un rédacteur;
* le rapport entre la grille expérimentale et la nature des erreurs de ponctuation possibles d'un rédacteur.

Nous avons élaboré notre grille de calibrage en éliminant les rubriques inefficaces. Par exemple, les rubriques décrivant le traitement des bas de pages ne permettaient pas une comparaison pertinente en raison de la différence entre une dissertation critique, où les citations du texte critiqué génère des références bibliographiques, et un texte d'opinion paru dans un journal ou un magazine ne nécessitant généralement pas ce type de renvois.

Nous avons essentiellement procédé en comparant les listes de variables générées par notre base de données pour les ensembles de textes de notre corpus de recherche, incluant les erreurs de ponctuation. Le tableau 19 répertorie l'ensemble des listes consultées.

Tableau 19

Répertoire des listes provenant des grilles consultées

Grille « Erreurs »	Grille « Maîtrise »	Grille « Erreurs de ponctuation »
Erreurs relatives aux conjugaisons verbales	Indices associés à la typographie	Confusion des signes de ponctuation
Erreurs relatives aux confusions homophoniques	Indices associés à l'emploi de « quelque »	Confusion du point avec un autre signe
Erreurs relatives au style et au vocabulaire	Autres indices de maîtrise	Confusion de la virgule avec un autre signe
Erreurs relatives à la syntaxe française		Signes indus
Autres catégories d'erreurs		Virgules indues
		Omission de signes
		Erreurs de ponctuation avec incidence sur la définition de la phrase en traitement automatique de la langue

La page suivante présente un extrait de l'une de ces listes (Fig. 53).

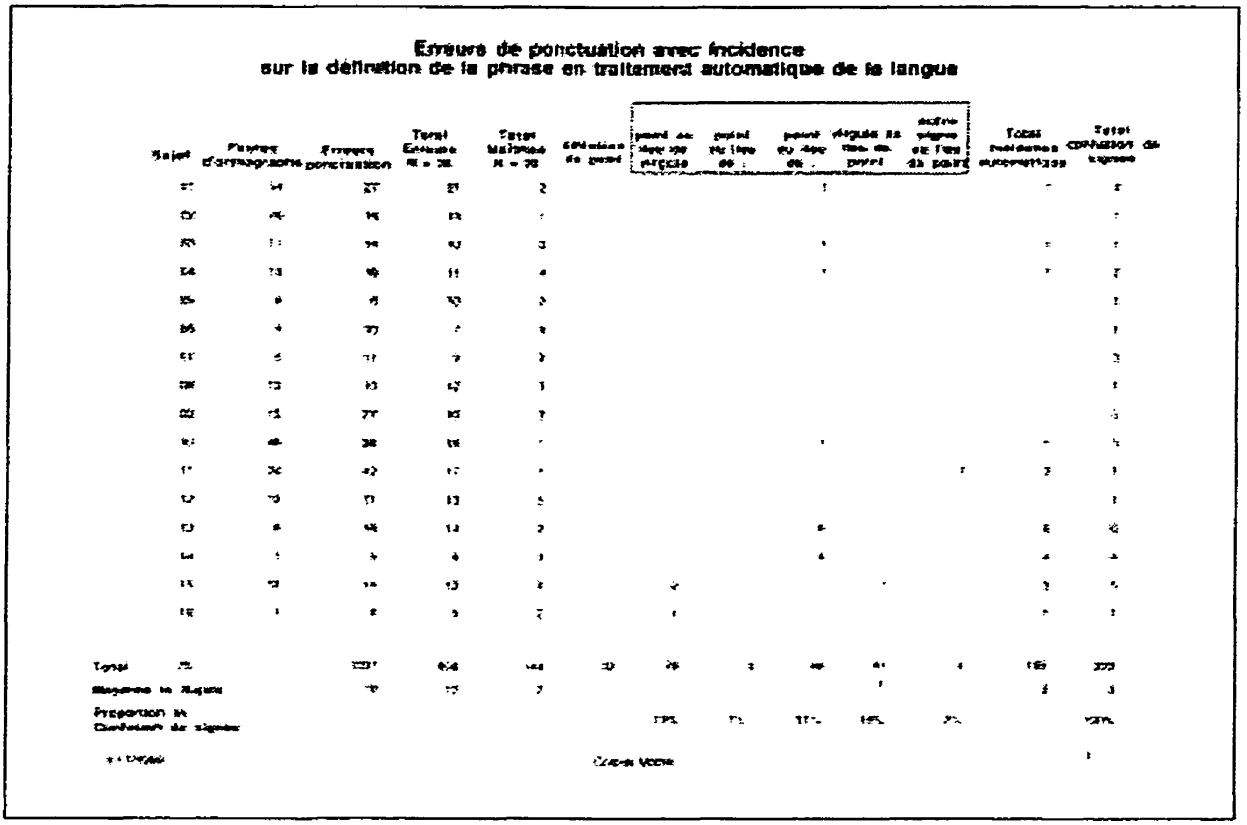


Figure 53

Exemple d'une comparaison de rubriques de la base Moffet

En mettant ainsi en parallèle les résultats des sujets pour chacun des groupes d'éléments de notre grille de recherche, nous avons pu faire ressortir des écarts significatifs, dont nous avons ensuite pu mesurer le pouvoir de prédiction au moyen de la statistique.

En outre, nous avons examiné la nature des combinaisons d'erreurs et d'indices de maîtrise caractéristiques de chacun de nos trois groupes de sujets. En appliquant ces combinaisons à un texte non littéraire, il deviendrait en principe possible de calibrer un correcteur orthographique selon le niveau de contrôle linguistique du rédacteur. La mise au point d'un calibreur orthographique permettrait de rendre les correcteurs « i ntelligents », c'est-à-dire de les rendre capables de « p rédire » le type d'erreurs les plus susceptibles de se produire dans les productions écrites d'un rédacteur de même que « d 'apprendre » des erreurs de ce rédacteur pour améliorer

la vitesse et la validité de la correction. C'est à la validation de ces combinatoires comme outils de calibrage que nous avons consacré la plus grande part de nos efforts de recherche[35].

Aspects quantitatifs

Nous avons soumis nos données à deux types de calculs: des calculs de sommes, moyennes et proportions et des calculs de prévision statistique.

Calculs de sommes, moyennes et proportions

Notre logiciel de base de données *File Maker Pro* (*FMP*) offre des fonctions de calcul de sommes, de moyennes et de proportions. Malgré l'utilisation de la numérotation logique, le calcul de fréquences a été possible pour tout le corpus en faisant porter les sommes logiques sur l'ensemble des textes. L'extrait suivant tiré de la base « Moffet » (Fig. 54) illustre ce calcul. Bien que les résultats de seulement 18 sujets y figurent, ce sont les sommes et les proportions touchant tout le corpus Moffet qui y sont tout de même affichées.

---

[35] Voir chapitre 7 : *Vers une correction automatique calibrée.*

Résultats du corpus Moffet : série "Stylistique et lexique"

Sujet	Fautes d'orthographe	Barbarismes sémantiques	Barbarismes lexical	Confusion sexes	Archaïsmes	Incorrections	Collocations inhabituelles	Mots manquant de précision
01	54	Non	Oui	Oui	Non	Oui	Oui	Oui
02	28	Non	Non	Oui	Non	Oui	Non	Oui
03	11	Non	Non	Non	Non	Oui	Non	Oui
04	14	Non	Non	Non	Non	Oui	Non	Oui
05	8	Non	Oui	Non	Oui	Oui	Non	Oui
06	4	Non	Non	Non	Non	Oui	Non	Oui
07	5	Non	Non	Oui	Non	Oui	Non	Oui
08	13	Non	Non	Non	Oui	Oui	Non	Oui
09	15	Non	Non	Oui	Oui	Oui	Non	Oui
10	48	Non	Non	Non	Non	Oui	Oui	Oui
11	24	Non	Oui	Oui	Non	Oui	Non	Non
12	13	Oui	Oui	Non	Non	Oui	Non	Oui
13	6	Non	Non	Non	Non	Oui	Non	Oui
14	1	Non	Non	Non	Oui	Oui	Non	Oui
15	13	Oui	Oui	Non	Non	Non	Non	Oui
16	7	Oui	Non	Non	Non	Oui	Non	Oui
17	15	Non	Non	Non	Non	Non	Non	Oui
18	16	Oui	Non	Non	Oui	Oui	Non	Oui
Total	75		15	16	12	10	68	4	65
Proportions in Sujet	100%		20%				91%		87%

Figure 54

Fréquences et proportions pour l'aspect "Stylistique et lexique : base « Erreurs » Moffet

Une fois toutes nos données entrées et traitées, nous les avons exportées dans le chiffrier électronique *Excel* (Microsoft, 1998) pour permettre la mise en relief des faits saillants. Notre étude quantitative se portait alors sur la distribution des erreurs linguistiques (ou indices de faiblesse linguistique) et des indices de maîtrise. *Excel* nous a également servi à visualiser certains de nos résultats. La figure 55 montre par exemple un histogramme illustrant le nombre de sujets Moffet (« Fréquence ») ayant présenté un nombre d'erreurs donné (« Résultats »).

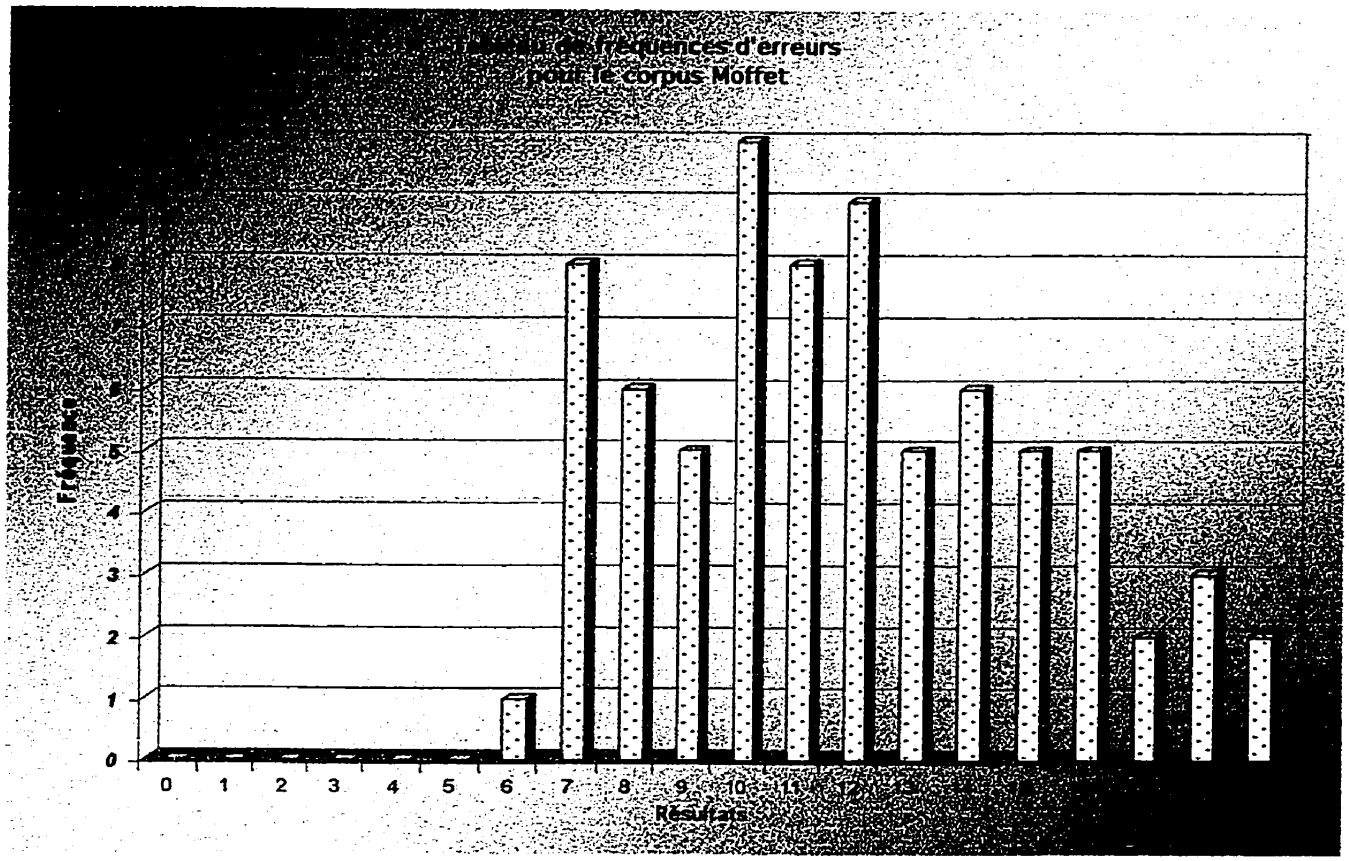


Figure 55

Exemple de traitement par tableur : fréquence d'erreurs dans le corpus Moffet

Excel nous a aussi permis de monter un sommaire de nos résultats afin de comparer nos trois catégories de rédacteurs. La figure 56 montre un exemple d'une feuille de calcul *Excel* mettant en parallèle le nombre d'erreurs des groupes de textes du corpus de recherche (le corpus *Bissonnette* ne comptait aucune erreur).

ERREURS

Experts

	Homophones	Verbe	Vocabulaire	Syntaxe	Style	Accords	Total Experts
TOTAL	2	2	7	12	15	4	42

	Homophones	Verbe	Vocabulaire	Syntaxe	Style	Accords	Total
Moyenne	0	0	0	0	0	0	1
Écart-type	0	0	0	0	0	0	1
Variance	0	0	0	0	0	0	1

Erreurs Moffet

	Homophones	Verbe	Vocabulaire	Syntaxe	Style	Accords	Total Moffet
TOTAL	59	100	192	218	191	123	883
Moyenne	1	1	3	3	3	2	12
Écart-type	1	1	1	1	1	1	3
Variance	1	1	1	1	1	0	12

Figure 56

Exemple d'une feuille de calcul Excel : sommaire des erreurs du corpus de recherche

Calculs de prévision statistique

Outre le calcul de mesures statistiques courantes comme l'écart-type et la variance[36], *Excel* a aussi permis d'effectuer des calculs de prévision statistique au moyen de droites de régression linéaire[37]. Une régression linéaire sert en effet à prédire une valeur inconnue (Y) à partir d'une valeur connue (X), du moment qu'il existe une corrélation entre elles :

> *Nous savons en effet que, s'il existe une relation linéaire positive entre deux variables, une valeur élevée d'une variable aura tendance à être associée à une valeur élevée de l'autre variable. Évidemment, cette tendance sera d'autant plus marquée que la corrélation est forte. De la même façon, s'il existe une relation linéaire négative, une valeur élevée d'une variable aura tendance à être associée à une valeur faible de l'autre variable.* (Allaire, 1998 : 20-1)[38]

---

36 Plusieurs de nos calculs statistiques ne figurent pas dans cette discussion en raison de leur non-pertinence ou de leur redondance, entre autres, le Khi-carré d'indépendance qui synthétise les écarts entre des fréquences observées et des fréquences théoriques (Allaire, 1995 : 17-4), le score Z qui situe une valeur dans une population (*op. cit.* : 12-2) et l'asymétrie qui calcule la symétrie d'une distribution (*op. cit.* : 13-5).

37 Il existe en fait de multiples modèles de régressions, comme les 15 chapitres de Ryan (1997) le démontrent assez. La régression linéaire à un seul régresseur, que nous utilisons ici, constitue en fait le modèle le plus simple.

38 L'absence de relation linéaire entre X et Y indique une corrélation nulle. Dans un tel cas, n'importe quelle valeur de X pourra donc être associée à n'importe quelle valeur de Y et il ne sera pas possible de prédire des résultats à partir d'une droite, bien qu'il reste par ailleurs possible d'effectuer une prédiction (qu'Allaire appelle « la moins pire ») au moyen d'autres formules mathématiques adaptées à ce problème (Allaire, *ibid.*).

Plus le coefficient de corrélation entre les variables X[39] et Y se rapproche de 1 (100%), plus grandes sont les chances de prévoir correctement la valeur inconnue Y, si la valeur X est connue (*loc.cit.*). Allaire (1995) explique ce phénomène par le fait que toutes les coordonnées (X_i Y_i) occupent une place particulière sur une droite qui les traversent.

Une droite de régression se calcule au moyen de la formule

$$Y = bX + a,$$

où Y est une variable dépendante représentant la valeur à prédire, « b » est une constante multiplicative correspondant à la pente de la droite, X est une variable indépendante représentant la valeur connue et « a », une constante additive désignant l'ordonnée à l'origine de la droite (*loc. cit.*; Ryan, 1997 : 4)[40].

Une fois calculée, la droite de régression est considérée comme la « meilleure droite » (en fait, il serait peut-être plus juste de dire "la moins mauvaise") traversant les points d'un diagramme de dispersion (familièrement appelé « nuage de points »).

La figure 57 montre un exemple des régressions bâties avec *Excel.*

---

[39] En réalité, une variable X est elle-même le résultat de l'équation **X = T + E**, où X représente un score observé (le résultat d'une mesure, par exemple, dans notre cas particulier, l'indice de faiblesse linguistique de chaque rédacteur de l'échantillon Moffet); T, le score vrai ou résultat sans aucune erreur de mesure et E, l'erreur de mesure se trouvant dans tout score observé (Allaire, 1995 : 27-1). Autrement dit, dans une analyse de résultats, on doit garder en mémoire que ces résultats contiennent un élément vrai et une marge d'erreur tributaire de l'instrument de mesure utilisé pour les recueillir, dans notre cas, la grille d'indice de faiblesse linguistique que nous avons présentée dans le chapitre 5. Mais, ainsi que le rappelle (Allaire, 1995 : 27-5.), *le score vrai et le score d'erreur de chaque individu* [sont] *des quantités qui ne peuvent être observées* [directement].

[40] Heureusement pour les non-initiés comme nous, *Excel* de Microsoft, entre autres tableurs, permet aujourd'hui d'exploiter automatiquement de telles formules.

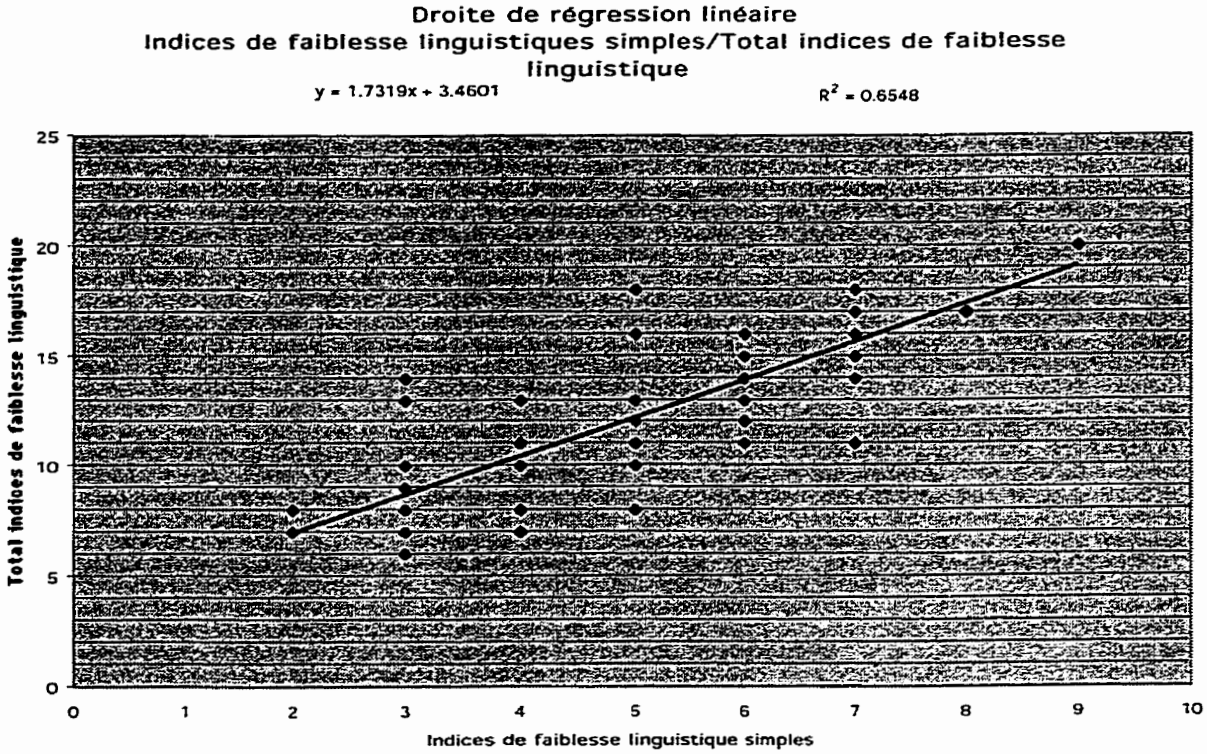
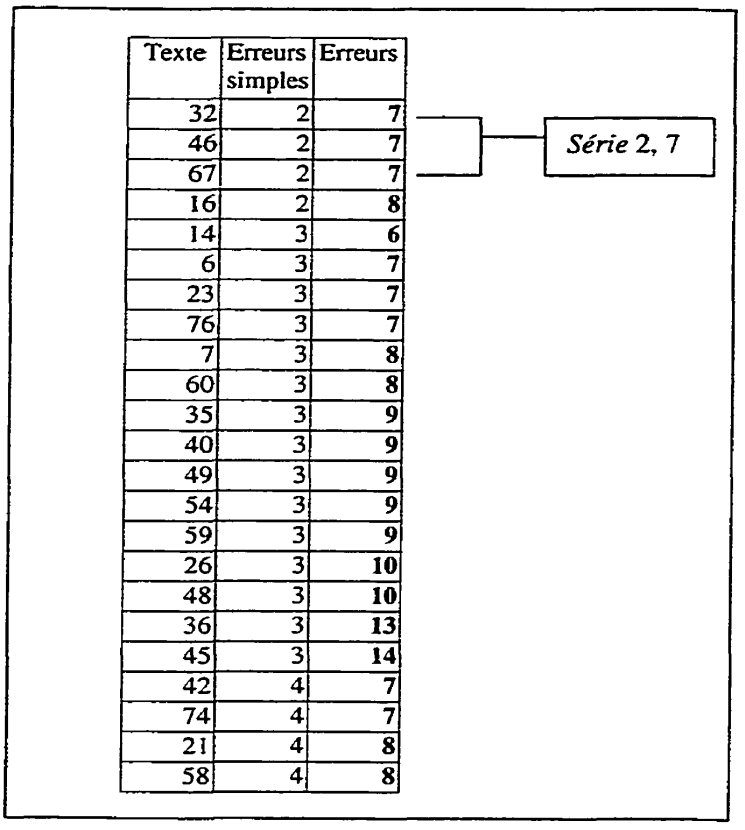


Figure 57

Exemple de régressions linéaires bâties avec Excel

Arrêtons-nous un instant sur cette figure pour mieux comprendre la nature d'un diagramme de dispersion et d'une droite de régression linéaire.

Pour bâtir un nuage de points à partir d'une plage de données XY, *Excel* superpose en un seul point toutes les paires présentant les mêmes valeurs. Chaque point d'un diagramme de dispersion représente donc une série pouvant provenir d'une seule paire de valeurs, de deux paires ou de plusieurs paire de valeurs. La figure 58 présente un extrait de la plage de données ayant généré le diagramme de dispersion de la figure 57.



Texte	Erreurs simples	Erreurs
32	2	7
46	2	7
67	2	7
16	2	8
14	3	6
6	3	7
23	3	7
76	3	7
7	3	8
60	3	8
35	3	9
40	3	9
49	3	9
54	3	9
59	3	9
26	3	10
48	3	10
36	3	13
45	3	14
42	4	7
74	4	7
21	4	8
58	4	8

Figure 58

Extrait de la plage de données associées à la figure 56

Un clic de la souris sur n'importe quel des points d'un diagramme de dispersion (en anglais, *scatter points*) identifiera la série de valeurs représentée par ce point (chaque série est désignée par les valeurs XY qu'elle représente). Il existe donc, sur un diagramme de dispersion, autant de points que de séries de valeurs identiques dans la plage de données XY saisie. En d'autres termes, le nombre de points figurant dans un diagramme de dispersion dépend du nombre de séries différentes se retrouvant dans la plage de données associée. C'est pourquoi la figure 57 montre 35 points alors que la plage de données associée compte réellement 75 paires de valeurs, soit une paire « Erreurs simples » (X) / « Total Erreurs » (Y) par texte.

Une fois un diagramme de dispersion monté, une fonction d'*Excel* permet de calculer automatiquement une droite de régression linéaire et de la représenter sur le diagramme. L'équation de prédiction résultant de ce calcul peut être affiché en même

temps que la droite. L'équation de la droite représentée dans la figure 57 par exemple est

$$Y = 1,7319x + 3,4601.$$

Le membre droit de cette équation constitue en fait la traduction mathématique de la représentation graphique de la droite de la figure 56 : 1,7319x, c'est-à-dire l'ordonnée à l'origine, et 3,4601, c'est-à-dire la pente de la droite. On se rappelle qu'Y, le membre gauche de l'équation, est la valeur à prédire, ici le nombre total d'indices de faiblesse linguistique (ou erreurs linguistiques) si la valeur connue est un nombre X d'indices de faiblesse simples.

En regardant notre diagramme de dispersion (Fig. 57) cependant, nous devons constater deux éléments importants: 1) notre droite pourrait traverser le nuage de points de bien d'autres façons et 2) plusieurs points figurent à l'extérieur de l'agglomération de la plupart des points de notre diagramme, et donc du champ d'action de notre droite. Le pouvoir de prédiction d'une droite dépend en effet de sa capacité à traverser une agglomération de points de la façon la plus parfaite possible.

La méthode de calcul appliquée par *Excel* pour identifier le passage de notre droite est aussi la plus utilisée. Il s'agit de la méthode appelée *critère des moindres carrés* (Allaire, 1998: 20-3), en anglais, *ordinary least squares* ou OLS (Ryan, 1997 : 348), qui prend pour acquis que toutes les valeurs d'une plage de données XY ont un poids égal dans le calcul de la régression. C'est ce calcul qui a produit la droite illustrée dans la figure 57.

À partir de 1972 cependant, les statisticiens se sont mis à critiquer OLS en arguant que toutes les valeurs d'une plage XY n'ont pas un poids égal par rapport à l'agglomération des points qui constituent l'ensemble du modèle (*ibid.*). En fait, il arrive fréquemment que plusieurs valeurs s'écartent passablement d'une agglomération de points, au point de miner de façon significative l'intégrité du calcul de la régression, et partant, le pouvoir de prédiction de la droite. En commentant un exemple qu'il apporte pour illustrer ce problème, Ryan (*op. cit.*: 349) explique:

> *Since this point* [un point sur l'ordonnée situé en-dehors de l'agglomération des points de son diagramme de dispersion] *does not fit the general pattern of the data, nor does it provide any evidence of the need to modify the model, we need* a robust regression approach [nous soulignons] *that will essentially assign a weight of zero to that point.*

Notons que Ryan (*ibid.*) apporte ici deux critères pour nous aider à reconnaître les valeurs polluantes possiblement cachées dans une plage de données: 1) la série qu'elles représentent se situe, sur le diagramme de dispersion, en dehors de l'agglomération de la plupart des points du diagramme et 2) la série suspecte ne suggère pas le besoin de modifier le modèle créé par la dispersion des valeurs de la plage de données.

Y a-t-il quelque chose à faire avec des valeurs possiblement polluantes? Eh bien, nous apprend Ryan (*loc. cit.*), il est possible de soumettre nos données à un calcul qui reviendra à assigner un poids zéro aux paires de valeurs polluantes, et par là, à annuler leur effet négatif sur le pouvoir de prédiction de la droite. C'est pourquoi la régression bâtie au moyen de ce calcul est appelée « robuste ».

Le calcul de régression robuste constitue un domaine de pointe abondamment discuté dans la littérature statistique (Ryan, 1997 : 348). Une régression robuste comporte plusieurs avantages, dont celui de permettre l'évacuation de valeurs polluantes dans la représentation d'une distribution de valeurs (attention ici, non pas dans la banque de données elle-même), ce qui génère un passage plus efficace entre les points d'un diagramme de dispersion.

La figure 59 présente une droite de régression robuste bâtie avec les valeurs du diagramme de dispersion de la figure 57.

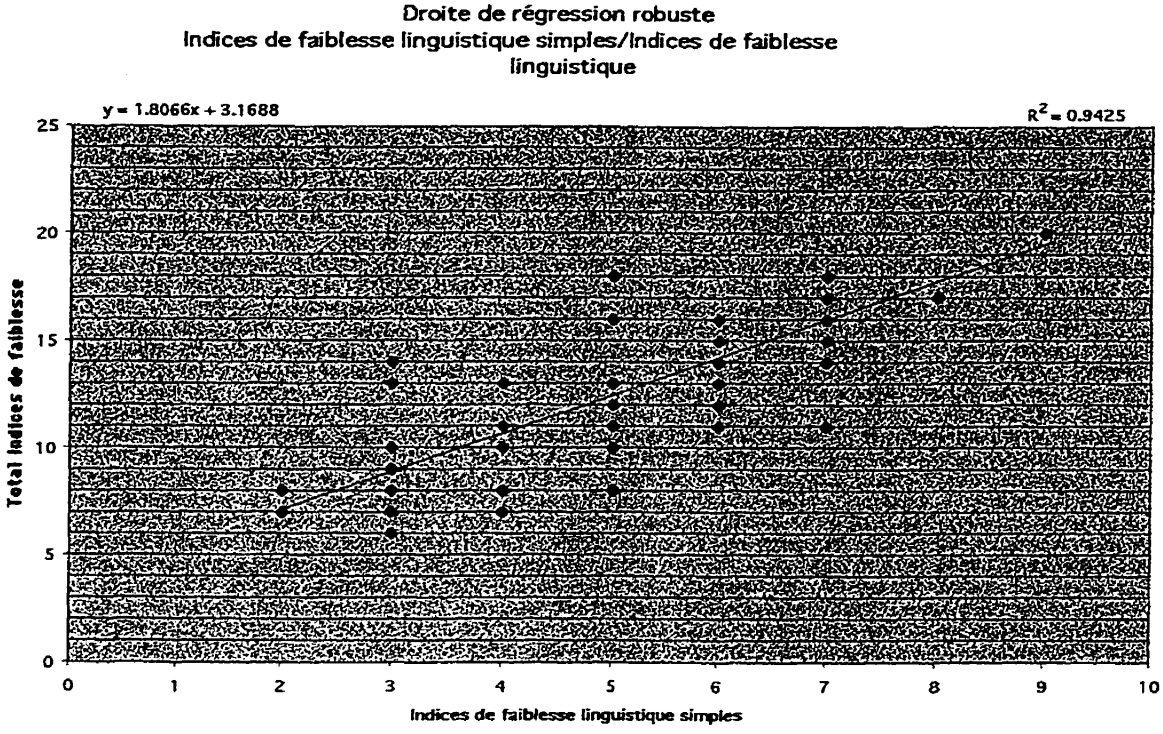


Figure 59

Diagramme de dispersion avec régression robuste

Il n'existe pas une grande différence visuelle entre la droite ordinaire montrée dans la figure 57 et celle-ci. En revanche, la différence est reconnaissable dans l'équation qui la traduit mathématiquement:

$$Y= 1{,}8066x + 3{,}1688$$

comparée à celle de la droite ordinaire,

$$Y = 1{,}7319x + 3{,}4601.$$

Le problème toutefois, avec un calcul de régression robuste, c'est que les statisticiens sont encore à développer des modèles pour identifier mathématiquement les régresseurs robustes dans une plage de données présentant une corrélation. *A fortiori* n'ont-ils évidemment pas encore réussi à monter des moyens informatiques sûrs pour détecter automatiquement les « bruits » dans une séquence de valeurs à cause de la complexité et de la multiplicité

des calculs nécessaires (*ibid* : 365, 387). Malgré ces empêchements, Ryan (*ibid* : 388) recommande de préférer la régression robuste à la régression linéaire usuelle:

> *It is important to realize that the routine use of least squares in regression analysis will often produce poor results. Even though the field of robust regression is far from being well developed, it has been demonstrated that certain robust estimators can perform well when used alone or in tandem. Practitioners should not wait until research has demonstrated which robust estimation procedure is best before considering the use of robust regression instead of least squares.*

Â cette étape-ci de notre discussion, il importe de rappeler que des régressions linéaires, ordinaires ou robustes, s'appuient sur des données qui montrent un coefficient de corrélation. Ce coefficient peut être positif, où à une valeur élevée X correspondra une valeur élevée Y, ou négatif, où à une valeur élevée X correspondra une valeur peu élevée Y. Un coefficient de corrélation peut se situer dans une échelle allant de -1 à +1 (ou -100% à + 100%). Par conséquent, un coefficient de corrélation exprimé en décimale devra se rapprocher le plus possible de 1 ou -1 (selon la nature de la corrélation) pour assurer à notre droite le meilleur (ou le moins mauvais) pouvoir prédictif possible.

Bref, nous disposons de deux formes de calcul pour bâtir des régressions linéaires simples : OLS, qu'*Excel* exécutera automatiquement, et la régression robuste, à laquelle nous décidons de faire appel quand un coefficient de corrélation sera inférieur à 90%. Ce niveau, que nous avons fixé arbitrairement pour l'utilisation des régressions robustes, nous semble cependant justifiable par le fait qu'au-delà de 90%, une régression ordinaire s'appuierait sur une corrélation presque parfaite qu'un algorithme de régression robuste n'améliorerait pas de façon significative.

Une droite de régression peut cependant générer des erreurs (dans le jargon de la statistique, « résidus »)[41]. Les statisticiens possèdent donc tout un arsenal de calculs additionnels pour estimer la validité d'une régression linéaire, en d'autres mots pour estimer la fiabilité des prédictions effectuées avec une telle droite.

---

[41] En fait, Ryan (1996 : 3) introduit sa discussion du modèle de régression linéaire en présentant l'erreur comme élément constitutif dans l'équation $Y = \beta_0 + \beta_1 X + \varepsilon$, où Y est la valeur à prédire; β_0 et β_1, les paramètres à estimer (dans la formule d'une droite de régression, il s'agit de la pente de la droite et de l'ordonnée à l'origine); X, la valeur connue et ε, l'erreur attachée à la prédiction ou valeur résiduelle.

L'une de ces mesures, le coefficient de détermination (R carré ou R^2), calcule la proportion de la variance des valeurs prédites Y expliquée par les valeurs connues X (Allaire, 1995 : 26-5). Bien que comportant plusieurs limites, R^2 reste largement utilisé en statistique (Ryan, 1997 : 12). Pour interpréter le résultat renvoyé par R^2, il importe de se rappeler qu'une prédiction linéaire parfaite donnerait une valeur prédite Y' identique à une valeur observée Y, telle que

$$Y' = Y$$

pour chaque valeur prédite d'une plage de données. Par conséquent, nous cherchons à obtenir une mesure R^2 aussi proche de 1 que possible (*loc. cit.*).

Mais R^2 n'est pas notre meilleur critère de validation. En effet, nous ajoutons à cette formule statistique un critère empirique qui nous permettra d'évaluer nous-mêmes concrètement si nos prévisions se confirment. Nous formulons donc le critère de validation suivant :

> *Si une régression calculée à partir du seul échantillon des 75 rédacteurs occasionnels Moffet est capable de prédire l'expertise linguistique observée dans les 150 textes professionnels de notre corpus, alors nous considérerons notre droite comme valide.*

Nos droites de régression robuste seront bâties manuellement de la façon suivante : nous montons d'abord un diagramme de dispersion à partir de deux plages de données pertinentes et calculons le coefficient de corrélation, la régression linéaire normale et R^2; nous trions ensuite les plages de valeurs associées à cette régression en ordre croissant des valeurs de X et Y; pour attribuer un poids zéro aux valeurs suspectes, nous retirons chaque paire une par une en observant l'effet de ce retrait sur le coefficient de corrélation. Conformément cependant à la méthode du critère des moindres carrés épurés (*least trimmed squares* ou *LTS* [Ryan, *op. cit.* : 362)] qui autorise le nettoyage des valeurs jusqu'à un maximum de 49% des valeurs, nous déterminons que nous ne pourrons attribuer un poids zéro à plus de 39 paires de nos 75 paires de valeurs.

Si l'effet du retrait améliore le coefficient de corrélation (et renforce donc le pouvoir de prédiction de la droite), la valeur est évacuée définitivement de la

représentation graphique (mais attention, pas de la base de données!). Sinon, elle est remise en place et une autre valeur est testée. On se rappelle que chaque point d'un diagramme de dispersion représente une série, laquelle peut être constituée de plusieurs paire de valeurs identiques. Au moment de la vérification des valeurs suspectes cependant, le poids de chaque paire est testé séparément.

La méthode empirique que nous avons appliquée ici est-elle contestable? Sans doute. Nous ne pouvons prétendre à l'expertise nécessaire pour maitriser tous les aspects d'un domaine complexe de la statistique, d'ailleurs en pleine effervescence. Nous ne pouvons non plus affirmer que notre mode de repérage des valeurs suspectes correspond à l'application manuelle des principes mathématiques documentés par Ryan (1997). Il n'est pas sûr non plus que le coefficient de corrélation et de détermination, en raison de leurs limites mêmes, constituent des repères mathématiques valables dans le contexte où nous les appliquons.

Il reste cependant que les régressions robustes calculées au moyen de cette méthode quelque peu monastique nous permettront, comme nous le verrons au chapitre 7, non seulement de traduire les tendances de notre échantillon de rédacteurs occasionnels, mais aussi de prévoir assez correctement des valeurs caractéristiques d'une toute autre population, celle des rédacteurs professionnels, y compris des rédacteurs maître comme Lise Bissonnette, par exemple. Ajoutons finalement que, contrairement aux statisticiens, notre objectif n'est pas de déterminer une formule mathématique universelle pour le calcul de régressions robustes, mais plutôt de trouver un moyen d'identifier, dans notre corpus, le moins mauvais passage possible de notre droite dans l'agglomération des points de nos diagrammes de dispersion.

Notons finalement, et ce point est très important, que tous nos calculs statistiques de prédiction ont été exclusivement effectués à partir des valeurs de notre corpus de textes écrits par des rédacteurs occasionnels. Nous nous en sommes tenus à ce corpus seulement pour une raison de validité évidente : lui seul constituait un échantillon statistiquement représentatif d'une vaste population de textes.

Conclusion

Notre méthodologie est-elle vulnérable à la critique? Peut-être. Nous discutons d'ailleurs certaines de ces limites au chapitre 7, en raison de leur incidence sur la validité de notre matrice de calibrage.

Une méthodologie rigoureuse constitue la base de tout exercice de recherche crédible. C'est pourquoi nous avons fait des efforts, tout au long de notre recherche, pour demeurer sensibles à cet aspect essentiel de notre travail. Par exemple, nous avons préféré à une grammaire listant des problèmes théoriques une grille de lecture développée par des chercheurs indépendants à partir de textes d'étudiants en récupération linguistique (Guénette, Lépine et Roy, 1995 : VI). Cette grille liste les problèmes concrets rencontrés par ces étudiants dans leurs propres textes et nous donnent une base pratique intéressante. Nous avons également limité nos calculs statistiques au seul corpus statistiquement représentatif dont nous disposions. Enfin, nous avons maintenu un doute scientifique tout au long de notre étude afin de demeurer vigilants face à tout problème méthodologique potentiel et d'identifier des stratégies préventives ou, le cas échéant, des solutions opérationnelles applicables dans notre contexte de recherche.

Quatre
Grille de calibrage

Ce que nous désignons comme grille de calibrage est une grille de lecture composée en fait de deux grilles complémentaires : une grille « Erreurs » et une grille « Maîtrise ». La grille « Erreurs » réunit des variables décrivant différents aspects d'un déficit linguistique en français écrit; la grille « Maîtrise » rassemble des variables que nous proposons comme caractéristiques du contrôle linguistique en français écrit.

En conjuguant les résultats de l'application de la grille « Erreurs » avec ceux de l'application de la grille « Maîtrise » pour un texte donné, nous pouvons déterminer le niveau de contrôle linguistique démontré par le rédacteur de ce texte au moment de la rédaction.

Notre grille attribue un nombre logique à chacune de ses variables lors du décodage d'un texte. Ces nombres peuvent être additionnés, d'abord par catégorie d'occurrences, ensuite pour l'ensemble de la grille. En calculant la performance de l'utilisateur pour un ensemble de textes selon les deux aspects « erreur » et « maîtrise », un « portrait » linguistique peut être élaboré (sur une base probabiliste) permettant théoriquement de prévoir, entre autres, si des erreurs sont probables dans les textes futurs et dans quels contextes linguistiques ces erreurs risquent le plus de se produire.

Les pages suivantes décrivent les rubriques de notre grille de calibrage.

4.1 Grille « Erreurs »

La grille « Erreurs » s'inspire de la grille de Guénette, Lépine et Roy (1995). Elle distribue les indices selon 6 catégories et 27 indices ou variables (Tableau 2 0).

Tableau 20

Grille « Erreurs » : catégories et nombre d'indices

Catégories	Nombre d'indices ou variables
Homophones	3
Verbe	6
Vocabulaire	7
Syntaxe	5
Style	4
Accords	2
6 catégories	27 variables

4.1.1 Catégorie « Homophones »

La catégorie « Homophones » s'intéresse à la traduction graphique de mots différents se prononçant de la même façon. Guénette, Lépine et Roy (1995) listent 25 homophones grammaticaux (Guénette, Lépine et Roy, 1995 : 59) et prévoient, avec la rubrique « Autres cas », une case pour toutes les occurrences exclues de cette liste.

Notre grille de calibrage inclut seulement trois indices dans cette catégorie :

* la confusion entre on et ont;
* la confusion entre a et à;
* les autres confusions homophoniques.

Les deux premières variables ont été retenues particulièrement en raison de la confusion entre l'auxiliaire avoir et deux mots-outils fréquents dans la langue courante. La troisième rubrique « Autres confusions homophoniques » est une rubrique de type générique incluant une rubrique « Description » pour archiver l'erreur notée dans le texte. Le tableau 21 introduit des exemples pour chaque variable non générique de la catégorie « Homophones ».

Tableau 21

Indices ou variables de la catégorie "Homophones"

Indice ou variable	Exemple du corpus Moffet occurrence pertinente (sujet : page[42])
Confusion entre *on* et *ont*	*Cependant les extraits on une forte tendance a incités les agriculteurs à laisser tomber leurs vieilles traditions pour aller explorer un monde nouveau, vers l'[suite indécodable]. (02 : 4)*
Confusion entre *a* et *à*	*Ce petit passage dit qu'il regrette ses gestes, que depuis ce temps il pense a elle. (10 : 3)*

La variable générique « Autres confusions homophoniques » a permis de noter, dans le corpus Moffet, les 26 confusions suivantes :

* ce / se;
* ces / s'est;
* ces / ses;
* cet / s'est;
* croit/ croît;
* davantage / d'avantages;
* du / dû
* et / est;
* la / l'a;
* n'ont / non;
* n'y / ni;
* or / hors;
* où / ou;
* par / part;
* parce que / pas ce que;
* peu / peut;
* peut être / peut-être;
* quelques fois / quelquefois;

---

[42] La page réfère à la page du texte même du rédacteur et non à la page du corpus.

* qui / qu'il;
* qui l'a / qu'il a;
* s'en / sans;
* sans / sent;
* si / ci (dans la suite ci-haut);
* si / s'y;
* si tôt / sitôt;
* son / sont

4.1.2 Catégorie « Verbe»

Le bloc « Verbe » décrit des problèmes relatifs à l'emploi du verbe. La grille Guénette, Lépine et Roy comprenait 8 catégories dans ce bloc seulement, chacune présentant de multiples variables pour un total de 48 possibilités. Nous avons réduit cette liste complexe à six éléments seulement, notre but n'étant pas de décrire les erreurs d'un texte de façon exhaustive, mais plutôt d'identifier des erreurs permettant de discriminer des niveaux de rédaction expert et non expert. Dans cet esprit, nous nous sommes plutôt arrêtés à des erreurs fondamentales comme le contrôle des conjugaisons et les problèmes de finales homophoniques. Notre bloc « Verbe » présente les variables suivantes :

* Confusion des finales -é / -er / -ez;
* Confusion des finales –i / -it;
* Confusion de terminaisons verbales;
* Passé simple inapproprié;
* Subjonctif contextuel manquant ou inapproprié;
* *Si* hypothétique suivi d'un conditionnel.

Guénette, Lépine et Roy (1995) introduisent, dans leur bloc « Verbe », une catégorie « Erreurs touchant à la terminaison ». Cette catégorie inclut la confusion des finales –é / -er / -ez, des finales –i / -it — qu'ils regroupent sous la rubrique « Confusion dans les formes verbales » — aussi bien que les erreurs de conjugaison comme telles (Guénette, Lépine et Roy, 1995 : 46). Nous distinguons ces occurrences encore une fois en raison de leur pouvoir possible de discrimination.

Par ailleurs, Guénette, Lépine et Roy (1995) placent les problèmes de subjonctif sous deux rubriques différentes : « Confusion dans l'emploi de l'indicatif et du subjonctif dans les subordonnées » et « Confusion dans l'emploi du subjonctif (ou de l'indicatif) et de l'infinitif dans les subordonnées ». Nous regroupons sous une seule bannière tout problème de subjonctif. Par « subjonctif contextuel », nous entendons un subjonctif requis par le contexte.

L'indice relatif au passé simple mérite un commentaire particulier. Le passé simple ou défini constitue une difficulté particulière pour les rédacteurs non experts. En effet, ce temps de verbe pratiquement disparu de la langue courante (Grevisse, 1980 : 838; Wagner et Pinchon, 1991 : 372) demeure cependant présent dans la langue écrite mais sous certaines conditions seulement:

> *À la 3e personne, le passé défini convient donc aux récits historiques, aux narrations romancées qui ont le ton d'une histoire, au rappel objectif de faits révolus. Il donne à ces relations la couleur dont se teintent les choses lointaines sur lesquelles on n'a plus de prise sinon par la mémoire.*
>
> *Dans un récit à la 1re personne, le passé défini, par effet de style, suggère que le locuteur évoque, sous le* je, *un personnage auquel il n'identifie plus tout à fait sa personne actuelle, présente.* (Wagner et Pinchon, 1991 : 371)

Ces règles rendent l'emploi du passé simple quelque peu hasardeux dans un texte non littéraire. Elles en excluent certainement l'emploi dans une dissertation critique telle que celle que devaient rédiger les participants à l'épreuve de français du ministère de l'Éducation du Québec. Cependant, dans un contexte où la qualité du texte peut avoir une incidence personnelle sérieuse, beaucoup de rédacteurs, par hypercorrection, utilisent le passé défini qu'ils associent à une qualité supérieure de français écrit, en en ignorant toutefois les règles d'emploi. C'est pourquoi nous sommes d'avis que l'occurrence du passé simple — quand cet emploi ne provient pas d'une citation —constitue un autre bon indice de niveau de contrôle.

Le tableau 22 apporte des exemples pour les variables du bloc « Verbe ».

Tableau 22

Indices ou variables de la catégorie "Verbe"

Indice ou variable	Exemple du corpus Moffet, occurrence pertinente (sujet : page)
Confusion des finales *–é / -er / -ez*	*Cependant les extraits on une forte tendance a incités les agriculteurs à laisser tomber leurs vieilles traditions [...] (02 : 4)*
	Tout son texte est bien rimer [...] (10 : 3)
	Un souvenir c'est une image, idée, représentation que la mémoire conserve pour nous rappelez ce qui nous est arrivé de bien ou de mauvais. (62 : 2)
	Allez-y, voyager. (15 : 8)
Confusion des finales *–i / -it*	*Cependant, pour François Paradis, c'est une vie complètement différente qu'il a choisit. (38 : 3)*
	Louis Hémon est un émigrant français qui s'établi au Canada en 1911 [...] (71 : 4)
Confusion entre terminaisons verbales	*Ce qu'ils n'ont pas encore vue ne les intéresses pas.(17 : 5)*
	« Vous ne me trouverai pas au fond des fossets, dans la vase » (20 : 7, citant Germaine Guèvremont)
Passé simple inapproprié	*Cependant, plus les années passèrent et plus les Québécois avaient comme but de changer les choses [...] (31 : 2)*
Subjonctif contextuel manquant ou inapproprié	*J'exposerai d'abord le point de vue de Louis Hémon et de Germaine Guèvremont pour qu'ensuite j'énoncerai mon opinion sur la question. (28 : 2)*
	On peut remarquer que pour les deux auteurs c'était la jeune fille la plus belle qu'il avait vue de leur vie [...] (63 :3)
Si hypothétique plus conditionnel	*Si tout le monde aurait suivi la mode, on aurait peut-être manqué de métier. (68 : 5)*

4.1.3 Catégorie « Vocabulaire »

Le bloc « Vocabulaire » décrit les problèmes lexicaux au sens large. Notre liste comprend les 7 indices suivants :

* Cataclysme orthographique[43];
* Barbarisme grammatical
* Confusion de genres;
* Mots manquant de précision;
* Barbarisme lexical;
* Archaïsme;
* Impropriété.

La catégorie « Vocabulaire » de Guénette, Lépine et Roy (1995) comporte 9 éléments: erreur sur le sens d'un mot ou d'une expression (*ibid.* : 89), mot ou expression manquant de précision (*ibid.* : 90); incompatibilité sémantique entre deux mots ou expressions (*ibid.* : 91); combinaison boiteuse (*loc. cit.*), termes inutiles ou redondances (*ibid.* : 92), barbarisme (*ibid.* : 93), altération d'une expression figée (*ibid.* : 94), anglicisme (*ibid.* : 95) et archaïsme (*loc. cit.*).

Notre bloc « Vocabulaire » se distingue de celui de Guénette, Lépine et Roy sous trois aspects principaux :

* il introduit un problème classé par Guénette, Lépine et Roy comme orthographique (« Cataclysme orthographique »);
* il introduit un problème classé par Guénette, Lépine et Roy comme grammatical (« Confusion de genres »);
* il regroupe 5 types des erreurs classées comme lexicales par Guénette, Lépine et Roy sous la rubrique « impropriétés » (erreur sur le sens d'un mot ou d'une expression, incompatibilité sémantique entre deux mots ou expressions, combinaison boiteuse, altération d'une expression figée, anglicisme).

Notre décision d'inclure deux problèmes orthographique et grammatical dans le bloc « Vocabulaire » peut être discutée, Guénette, Lépine et Roy (1995) séparant nettement ces deux catégories de celle des problèmes lexicaux. Dans un effort de simplification, nous avons pourtant pris la décision de réunir sous la bannière « Vocabulaire » tout problème de graphie, de détermination de genres ou de choix

---

[43] Un cataclysme orthographique survient quand un même mot comprend plusieurs erreurs d'orthographe (Guénette, Lépine et Roy, 1995 : 30)

de mot. Nous pourrions débattre que ces problèmes concernent en fait des aspects différents d'une même difficulté en raison, en autres, de l'effet parfois dévastateur de ce type d'erreur sur l'interprétation et le sens du texte.

En fait, nous avons presque ignoré le problème d'orthographe lexical à cause de notre préoccupation de repérer des indices avec le meilleur pouvoir discriminant possible. Nous n'avons ainsi retenu, parmi les 44 indices lexicaux orthographiques de la grille de Guénette, Lépine et Roy (1995), que l'indice « Cataclysme orthographique », que nous considérions le plus susceptible de porter un pouvoir discriminant sur le plan orthographique.

Nous sommes en effet d'avis que les erreurs d'orthographe lexicale comporte un pouvoir discriminant limité. Les correcteurs détectent les erreurs d'orthographe lexicale en comparant une suite donnée avec les entrées de leur dictionnaire. Or les correcteurs actuels offrent généralement une fonction « dictionnaire personnel », qui permet au rédacteur d'enrichir le dictionnaire original en y ajoutant un nouveau mot de son choix, y compris un nom propre ou un terme spécialisé. Cela veut donc dire que n'importe quel mot — y compris un mot mal orthographié — peut être ajouté au dictionnaire d'un correcteur, si bien qu'en théorie, un rédacteur pourrait se constituer un dictionnaire personnel truffé d'erreurs d'orthographe[44].

L'indice « Mot manquant de précision » a imposé plusieurs décisions importantes. Nous avons d'abord préféré la constitution d'une liste de mots vagues plutôt qu'un jugement appliqué au cas par cas : une liste donne de meilleures chances de rigueur dans l'application de la grille. La détermination des mots de cette liste a imposé à son tour d'autres choix. Nous avons ainsi exclusivement retenu des substantifs, mettant de côté *faire, devoir, pouvoir, avoir* et *être* en raison de leur rôle grammatical essentiel et de leur fréquence dans la composition d'expressions idiomatiques. Avoir à distinguer leurs emplois variés introduisait des risques trop

importants d'erreurs lors de notre correction. C'est pourquoi nous nous en sommes tenus à une liste relativement simple et aisément mémorisable :

* *personne* (au sens d'individu);
* chose;
* gens;
* individu;
* endroit;
* objet;
* situation;
* êtres humains;
* lieu;
* espace;
* être (substantif)

Encore une fois, nous n'avons pas recherché l'exhaustivité, bien que nous nous soyons donné quelques règles pour encadrer notre jugement pour cet indice particulier :

* une seule occurrence d'un mot de notre liste de mots vagues dans tout le texte n'est pas considéré comme une erreur;
* une seule occurrence de plusieurs mots de notre liste de mots vagues nous fait noter l'erreur dès la première occurrence.

Quant à la variable « impropriété », nous l'avons définie comme toute erreur de manipulation d'un mot, y compris les erreurs relatives au ton ou au niveau de langue que Guénette, Lépine et Roy classent dans la catégorie « Style » (*ibid.* : 97). Le tableau 23 présente des exemples pour chaque indice du bloc « Vocabulaire ».

---

[44] En fait, cela doit se produire fréquemment. Le sujet Moffet 20, que nous citons dans le tableau 22, a écrit, en citant Germaine Guèvremont : « *Vous ne me trouverai pas au fond des fossets, dans la vase* » *(20 : 7)* alors qu'il reproduisait une phrase qu'il avait sous les yeux. Voilà un bel exemple de l'effet pervers de la faiblesse orthographique dans la retranscription de suites correctement écrites. Un rédacteur peut ainsi retranscrire un mot lu dans un dictionnaire ou un volume de référence en introduisant une ou des erreurs d'orthographe, ajouter de bonne foi ce nouveau mot avec une graphie erronée dans son dictionnaire personnel, que le correcteur orthographique considère ensuite comme « correcte ».

Tableau 23

Liste des indices ou variables de la catégorie "Vocabulaire"

Indice ou variable	Exemple du corpus Moffet occurrence pertinente (sujet : page)
Cataclysme orthographique	*Pourtant, à l'intérieur des extraits ont voit l'énumération des biens fait et avantage de la terre. (01 : 4)*
Barbarisme grammatical	*Cet espoir semble cependant trop facile lorsquel parle de l'amour et du bonheur. (12 : 4)*
	Il ressent du désespoir de voir tous ces gens vivrent de cette façon (16 : 3)
Confusion de genres	*On compare en quelque sorte une pionnière, Maria Chapdelaine, et un espèce de vagabond, François Paradis. (53 : 2)*
	Elle était une fille qui passait, avec laquelle, le héro avait envie de partager une moment de plaisir [...]. (09 : 5)
Barbarisme lexical	*L'attachement à sa terre ou à son pays font faire de la rétïcence entre chaqu'un. (57 : 5)*
	La terre est, en fait, considérée comme brimeuse de liberté pour les personnages. (73 : 4)
Mots manquant de précision	*Les personnes qui bénéficient d'une chance favorable envers la vie, sont des êtres sachant exprimer une beauté remarquable en essayant seulement de vivre. (13 : 2)*
	Comme l'église avait encore beaucoup de pouvoir, les gens qui avaient le goût de l'aventure étaient bien souvent mal vu. (15 : 3)
Archaïsme	*La mère Chapdelaine n'est point de cet avis [...]. (05 : 4)*
Impropriété	*Elle se manifeste premièrement par le bris de l'attachement traditionnel. (19 : 2) [Erreur sur le sens d'un mot ou d'une expression]*
	Voyons donc la différence des deux œuvres au niveau des personnages témoins de ces « coureurs des bois ». (03 : 2) [Erreur d'emploi d'un mot-outil]
	Une lecture attentive de ces deux extraites dirige mon opinion vers une affirmation négative [...] (08 : 3) [Combinaison boiteuse]

4.1.4 Catégorie « Syntaxe »

Le bloc « Syntaxe » réunit les erreurs relatives à l'organisation de la phrase. Guénette, Lépine et Roy (1995) rassemblent ce type d'erreurs dans un ensemble « Phrase », qui propose 23 indices différents.

Notre catégorie « Syntaxe » regroupe les 5 variables suivantes :

* Mots essentiels manquants;
* Termes inutiles ou redondants;
* Ordre des mots;
* Suite asyntaxique;
* Désordre syntaxique inextricable.

Les deux premiers indices de notre liste constituent les côtés « pile-face » d'un même problème : une erreur de jugement sur l'opportunité d'un mot dans un contexte donné. Un mot ou un groupe de mots peut manquer dans le contexte, par exemple le complément d'objet d'un verbe transitif; un mot ou un groupe de mots peut n'apporter aucune information additionnelle ou répéter une information déjà donnée. Pour considérer un mot manquant comme essentiel, nous nous en sommes tenus aux règles syntaxiques du français comme le déterminant requis dans un groupe nom, un groupe sujet absent, un groupe complément imposé par la nature transitive du verbe, etc .

Nous réunissons sous l'étiquette « Suite asyntaxique » un ensemble d'erreurs identifiées par Guénette, Lépine et Roy (1995 : 62-67) dans leur bloc « Phrase » : les erreurs de construction de phrases interrogatives, les erreurs de mises en relief, les erreurs dans l'emploi du sujet ou des compléments et l'absence de la proposition principale.

Un désordre syntaxique inextricable (Guénette, Lépine et Roy , 1995 : 68) se reconnaît dans une phrase ou un membre de phrase impossible à comprendre. Guénette, Lépine et Roy cite l'exemple suivant (*loc. cit.*) : *Comment ceux-ci en sont-ils venus à une telle réticence envers autrui qu'elle se transforme en racisme entre gens de même race ?*

Le tableau 24 présente les variables de la catégorie « Syntaxe » avec des exemples du corpus Moffet.

Tableau 24

Indices ou variables de la catégorie « Syntaxe »

Indice ou variable	Exemple du corpus Moffet occurrence pertinente (sujet : page)
Mots essentiels manquants	*Pour ce qui est d'une autre différence, est bien sûr, la façon dont la femme est présentée. (09 :6) [« c' » manquant]*
	Cette dissertation critique traitera du fond, de la forme et aussi de la structure de chaque œuvre et terminerons avec une comparaison des deux textes [...] (10 :2) [« nous » manquant]
	D'après moi, il est question des deux côté . (47 : 2) [syntagme prépositionnel manquant]
Termes inutiles ou redondants	*Même si dans ces deux extraits on y retrouve de la valorisation de la terre, cela ne veut pas nécessairement dire que c'était l'unique but des auteurs (05 :4)*
	Pour ce qui en est des héros, les deux sauront livrer leurs messages malgré que celui de Maria Chapdelaine se sentait honteux. (08 :7)
Ordre des mots	*Maintenant, il serait intéressant de remarquer que Louis Hémon, n'ayant vécu qu'au Québec pendant deux ans aime toujours la vie terrienne [...] (40 : 6)*
	[...] nous ne savons pourquoi il oublit de voir le petit pied de la jeune fille nu. (62 : 3)
Suite axyntaxique	*[...] car ces deux livres ont été écrits autour de les mêmes années. (34 : 2)*
	Ceux qui restaient sur une terre, qui étaient-ce? (47 : 5)
	C'est un cousin et une cousine qui se revoit car le père de lui leurs a demander de se marier ensemble [...] (11 : 4)
Désordre syntaxique inextricable	*François dans Maria Chapdelaine et Survenant dans Le Survenant représentent les jeunes qui aiment voir du pays et ensuite suivra l'attachement à la terre par l'autre génération et suivit de la synthèse qui valorise l'attachement à la terre. (45 : 2) [?]*
	Nous pouvons remarquer que le poème de Victor Hugo ressemble beaucoup a ceux de Beaudelaire pourtant il ne sont pas rapprocher point de vue temps mais tout deux écrivent de façon qu'ils expriment l'obscurité indispensable ce qui est obscur et confusément révélé. (62 : 6) [?]

4.1.5 Catégorie « Style »

Guénette, Lépine et Roy (1995) réunissent deux ensembles d'erreurs dans leur bloc « Style » : les maladresses stylistiques et les erreurs de ton ou de niveau de langue (*ibid.* : 95-98). Ils associent 8 erreurs spécifiques à cette catégorie dont la répétition abusive d'un mot ou d'une expression (*ibid.* : 96).

Par ailleurs, Guénette, Lépine et Roy (1995) listent de multiples erreurs dans une catégorie qu'ils appellent « Texte ». Les erreurs relatives au texte touchent tout l'aspect organisation logique des idées et la structure de texte comme les problèmes d'expression des liens logiques (9 problèmes identifiés) et de cohésion textuelle (5 problèmes identifiés).

Nous avons regroupé sous la rubrique « Style » 3 erreurs particulières listées par Guénette, Lépine et Roy (1995) dans leurs blocs « Style » et « Texte »:

* Incohérence dans le choix des pronoms personnels;
* Répétition abusive de mots;
* Références anaphoriques.

Nous avons cependant ajouté un quatrième indice — original — à cette liste :

* Élément manquant ou incohérence dans l'utilisation de connecteurs en série.

La manipulation congruente des pronoms personnels pose problème pour certains rédacteurs. Guénette, Lépine et Roy (1995 : 108) citent deux exemples : le premier illustre la difficulté de maintenir le « nous » narratif tout au long d'un texte; le deuxième, la confusion d'emploi entre les pronoms complément « se » et « nous ». Nous avons, pour notre part, inclus dans cette description la différence de personne dans des emplois reliés d'un pronom personnel et d'un adjectif possessif : par exemple, un « nous » narratif associé à l'adjectif possessif « mon », « ma » ou « mes ».

La répétition abusive de mots ne semble pas à première vue poser de problème d'application. Pourtant, la détermination de « l'abus » demande à être préalablement précisée. Nous basant sur les exemples donnés par Guénette, Lépine et Roy (1995 : 96) tout autant que notre expérience de 15 années de correction de

textes d'étudiants, nous avons considéré comme « répétition abusive » toute reprise identique d'un mot ou d'un groupe de mots dans l'un des trois contextes suivants :

* la même phrase;
* deux phrases conjointes (faisant partie d'un même paragraphe ou de deux paragraphes consécutifs);
* le même paragraphe.

Une référence anaphorique constitue notre troisième indice stylistique. Guénette, Lépine et Roy (1995 : 106-107) appliquent le terme référence anaphorique à trois occurrences particulières :

* un nom ou un syntagme nominal de rappel inadéquat,
* un pronom de rappel inadéquat
* l'absence d'antécédent.

Nous avons respecté cette définition dans l'application de notre propre grille.

Quant à l'indice portant sur l'exploitation maladroite de connecteurs en série, il fait référence à l'emploi d'une catégorie particulière de mots de transition, ce que nous avons appelé — à défaut d'un meilleur terme — les « connecteurs en série ». Des séquences comme *premièrement ...*, *deuxièmement ...*, *troisièmement ...*, *d'abord ...*, *ensuite ...*, *enfin*, *en premier lieu ...*, *en deuxième lieu ...*, *en troisième lieu* ... ou encore *d'une part ...*, *d'autre part ...* sont autant d'exemples de ce que nous désignons comme « connecteurs en série ». Nous avons donc jugé comme une erreur le défaut de maintenir une telle série dans le développement d'une idée, si le rédacteur avait entamé un raisonnement avec le premier terme d'une telle série.

Le tableau 25 introduit les variables de notre catégorie « Style » en citant des exemples tirés du corpus Moffet.

Tableau 25

Indices ou variables de la catégorie "Style"

Indice ou variable	Exemple du corpus Moffet : occurrence pertinente (sujet : page)
Incohérence dans le choix des pronoms personnels	*Nous allons nous éclairer sur cette question en examinant les idées qui sont véhiculées [...]. Une lecture attentive de ces deux extraits dirige mon opinion vers une affirmation négative [...] (08 : 3)*
	En fait, je peux affirmer que dans les extraits de Maria Chapdelaine [...]. En guise de conclusion, nous pouvons affirmer que les auteurs du vingtième siècle [...] (23 : 5)
	Nous n'avons qu'à considérer les lignes 7 à 19 de l'extrait pour s'en apercevoir. (26 : 3)
Répétition abusive de mots	*Par contre, plusieurs mentionnent les désavantages à préférer le bois à la terre. Comme Maria Chapdelaine mentionne [...] (04 : 3)*
	Maria Chapdelaine de Louis Hémon et Survenant de Germaine Guèvremont sont deux textes qui rappelent ces anciennes valeurs. À l'aide de courts extraits de ces deux textes, je démontrerai que [...] [15 : 2)
Références anaphoriques	*Ceci sera tout d'abord prouvé par le thème de la sédentarité qui apporte paix immobile et sécurité, ainsi que par le fait que c'est une continuité familiale qui remonte à bien loin, pour cultiver les terres. (44 :2) [absence d'antécédent]*
	Dans l'extrait, on remarque que François Paradis vend la terre de son père. Ensuite, il raconte son emploi du temps [...] (38 :3) [pronom de rappel inadéquat]
Élément manquant ou incohérence dans l'emploi de connecteurs en série	*Premièrement, dans le texte de Victor Hugo [...]. Deuxièmement, dans le texte de Michel Garneau [...]. Dernièrement, dans le texte de Victor Hugo [...]. (10 :3-5)*
	Tout d'abord, je vais exprimer mes arguments [...]. En second lieu, peut-on affirmer que Guèvremont [...]. Troisièmement, je crois qu'étant donnée la date de parution des romans [...]. (31 :4-6)

4.1.6 Catégorie « Accords »

Nous avons regroupé dans cette série tout problème relatif à l'emploi grammatical des mots variables. Notre intérêt se portant surtout sur le pouvoir discriminant d'un indice, il nous apparaissait peu utile de détailler les multiples variations associées aux problèmes d'accord en français écrit. C'est pourquoi, contrairement à Guénette, Lépine et Roy (1995) qui énoncent près d'une centaine de types d'erreurs de grammaire (*ibid.* : XI-XVII), nous nous limitons à 2 variables seulement dans cette catégorie :

* les fautes d'accord du syntagme nominal;
* les fautes d'accord du syntagme verbal.

Les erreurs du syntagme nominal réunissent toutes les fautes d'accord touchant n'importe quel élément du SN : déterminant, adjectif épithète, nom. Les erreurs du syntagme verbal comprennent toutes les fautes d'accord relatives à l'élément verbal du SV : verbe, auxiliaire, copule, semi-auxiliaire, participe, attribut. Le tableau 26 présente des exemples des deux variables de notre catégorie « Accords ».

Tableau 26

Indices ou variables de la catégorie "Accords"

Indice ou variable	Exemple du corpus Moffet, occurrence pertinente (sujet, page)
Erreurs d'accord du SN	*[..] le désir de laisser la terre pour d'autre horizon [...]. (01 :3)*
	[...] sur ces terre « de petite grandeur » (02 : 3)
	[...] la vie des nomades ou des coureurs des bois comportent plus de désavantage [...]. (04 :5)
	Tout deux traitent d'une ballade dans les bois [...] (11 :3)
	[...] l'impuissance que nous, humain, avons envers cette quête du bonheur. [12 :5)
	[...] il est difficile de trouver beau la vie (13 : 4).
Erreurs d'accord du SV	*Au début du siècle, l'agriculture et l'élevage était deux activités fortement valorisées par la population québécoise. (04 :2)*
	[...] la vie des nomades ou des coureurs des bois comportent plus de désavantage [...]. (04 :5)
	Ces derniers ne sont pas fait pour la terre [...]. (07 : 3)
	Troisièmement, les auteurs de ces œuvres ont vécu dans une période qu'on a appelé le Terroir [...] (25 : 4)

4.2 Grille « Maîtrise »

La grille « Maîtrise » distribue les indices de contrôle linguistique en six catégories (Tableau 27).

Tableau 27

Grille « Maîtrise » : catégories et nombres d'indices

Catégories	Nombre d'indices ou variables
quelque	5
Pronoms relatifs	3
Typographie	7
Verbe	2
Incises	1
5 catégories	18 variables

Notre grille « Maîtrise » avait pour but de noter les occurrences des indices présumés de contrôle linguistique dans notre corpus. Par conséquent, une même occurrence pouvait recevoir deux marques, une pour son rôle indicateur de maîtrise et une autre si elle était mal orthographiée.

Le double système « Erreurs » et « Maîtrise » nous a permis de décrire objectivement l'ensemble des textes du corpus en autorisant la notation de nos indices de contrôle indépendamment de leur qualité de réalisation. Cette distinction nous apparaissait importante. En effet, un sujet pouvait par exemple tenter d'exploiter une forme rare du subjonctif mais se tromper dans la conjugaison. Son effort lui mériterait alors une occurrence dans sa grille « Maîtrise » mais une erreur pour l'indice « Terminaisons verbales » de sa grille « Erreurs ». La connaissance de l'existence de conjugaisons rares reçoit alors une marque comme indice de maîtrise tout autant que la faute grammaticale comme indice d'absence de contrôle.

Par ailleurs, nous ne notions pas une occurrence de contrôle linguistique quand elle se présentait dans une suite citée par le rédacteur.

4.2.1 Catégorie « Quelque »

Les multiples fonctions de *quelque* en français posent toutes sortes de problèmes orthographiques. *Quelque* employé comme déterminant commande un pluriel au sens de « plusieurs » et un singulier au sens de « un certain ». Devant un nombre, *quelque* a une fonction d'adverbe et devient par conséquent invariable. Dans la suite *quelque... que*, *quelque* peut avoir fonction d'adverbe au sens de « si » ou de déterminant au sens de « n'importe quel » ou « peu importe ». Dans le premier cas, il devient invariable; dans le second, il s'accorde avec le nom auquel il se rapporte. Par ailleurs, *quelque* se prononce comme l'adjectif relatif *quel que*. Ce déterminant précède immédiatement *être* au subjonctif ou son pronom personnel sujet et *quel* s'accorde avec le sujet de être.

En raison du haut niveau de contrôle requis par son emploi, nous avons retenu *quelque* comme premier indice possible de contrôle linguistique. Cependant, nous avons exclu les emplois de *quelque* dans les expressions figées suivantes :

* quelque chose;
* quelque temps;
* quelques-uns, quelqu'un;
* en quelque sorte;
* quelque peu;
* quelque part.

Le tableau 28 présente les différents emplois de quelque dans les corpus « Experts » ou Bissonnette.

Tableau 28

Indices ou variables de la catégorie "Quelque"

Indice ou variable	Exemple du corpus « Experts » ou du corpus Bissonnette (expert : page) ou (B : texte)
Quelque déterminant au sens de « un certain »	*Les résultats du 2 juin laisse présager quelque espoir. (106 :1)*
Quelque déterminant au sens de plusieurs	*On peut résumer en quelques mots leur dénominateur commun [...] (104 :1)*
Quelque adverbe au sens de « environ »	*Il s'agit d'un gigantesque projet d'un coût de 3 milliards de dollars canadiens et qui abritera, sur l'emplacement du site de l'Exposition universelle, quelque 15 000 personnes. (103 : 2)*
Quelque ... que (suivie d'un subjonctif)	*Aucun des articles du projet de loi n'autorise les écoles à sélectionner les élèves à partir de quelque critère que ce soit [...] (B : 10)*
Quel que (suivi de *être* au subjonctif)	*En effet, le peuple québécois (quel que soit le sens qu'on donne à cette expression) ne se trouve pas placé sous la domination coloniale. (111 : 1)*
	Mais quelle que soit la teneur de son jugement, ce qui ne fait pas de doute selon plusieurs juristes, il y a fort à parier que le problème existentiel canadien demeurera entier [...] (112 :2)

4.2.2 Catégorie « Pronoms relatifs »

Nous avons retenu les pronoms relatifs en raison des nombreuses erreurs possibles dans l'emploi de certains d'entre eux , particulièrement *dont* et les composés de *quel* comme *auxquels, desquels*, etc. En outre, les composés de *qui*

comme *à qui, pour qui, de qui,* etc. nous ont paru intéressants à cause des risques d'impropriétés liés au choix de la préposition (Guénette, Lépine et Roy, 1995 : 57). Par conséquent, le contrôle et le choix des pronoms relatifs autres que *qui* et *que* pouvaient constituer un signal pertinent de maîtrise.

Les pronoms relatifs dont nous avons noté l'emploi sont *quel* et ses composés, *dont* et *qui* ou *quoi* composé avec une préposition ou une locution prépositive.

Le tableau 29 présente des occurrences notées dans les corpus « Experts » ou Bissonnette.

Tableau 29

Indices ou variables de la catégorie "Pronoms relatifs"

Indice ou variable	Exemple du corpus « Experts » ou du corpus Bissonnette (expert : page) ou (B : texte)
Quel et ses composés	*La loi stipule que le titulaire de ces deux cartes n'est tenu de les présenter que pour les fins pour lesquelles elles sont émises. (102 : 3)*
	Flambeurs et journaliers AVEC[45] l'arrivée d'environ 40 000 nouveaux immigrants par an, auxquels s'ajoutent quelque 38 000 migrants canadiens en provenance de l'est des Rocheuses. [...] (103 : 1)
	[...] la campagne a essentiellement mis en lumière (et les résultats l'ont confirmé) l'état précaire dans lequel se déroulent l'affrontement politique et la vie collective dans ce pays. (106 : 1)
Dont	*D'autre part, le Registre international des transferts d'armes qui pourrait être quelque peu amélioré, puisqu'il n'inclut pas certaines armes (dont les armements légers), ne répertorie pas les stocks existants [...] (160 :1)*
Qui ou *quoi* composé avec une préposition ou une locution prépositive	*Et si c'est une chose d'authentifier numériquement un tel document aux yeux de celui à qui il est destiné, comment ce dernier peut-il prouver à une tierce personne que [...] (115 : 2)*
	Principaux partenaires commerciaux des Algériens avec qui ils entretiennent des rapports « passionnels », les dirigeants français ont été accusés par les islamistes de soutenir la « junte » au pouvoir [...] (138 : 3)

[45] En majuscules dans le texte.

4.2.3 Catégorie « Typographie »

Chanod (1993) a fait ressortir l'importance des structures périphériques dans l'enrichissement syntaxique de la phrase française écrite. Les structures de parenthétisation et les signes de ponctuation qui les signalent — parenthèses, tirets, crochets intercalaires — enrichissent le noyau de la phrase d'information additionnelle. Dans sa théorie de la ponctuation, Jones (1996c : 29), s'appuyant sur Nunberg (1990), décrit la fonction de telles structures périphériques (*clause adjuncts*) :

> *Text clauses are further augmented by the introduction of the category of clausal adjuncts. This category includes colon expansions (3.20), dash interpolations (3.21) and literal parentheticals (3.22). The colon expansion is restricted to being the right-most element of the text clause that contains it, while the other two types of clausal adjuncts can occur anywhere within the text clause except at the beginning, with certain other provisos, for example that they cannot occur adjacent to one another at the same level (3.23).*
>
> *(3.20) I will be frank : there is no way you're going to get the job.*
>
> *(3.21) And — what's more surprising — she left.*
>
> *(3.22) And (not surprisingly), she left.*
>
> *(3.23) *She walked out — who could blame her — (it was during the chainsaw scene, as I recall) and went directly home.*

Postulant que les structures périphériques devraient se trouver plus nombreuses chez des rédacteurs maîtrisant leur langue écrite en raison de leur capacité de construire avec succès des phrases à structure complexe, nous avons également pensé que les marqueurs graphiques de ces structures constitueraient par conséquent un bon indice de contrôle linguistique.

Nous avons ainsi constitué la liste de marqueurs typographiques suivante :

* tirets intercalaires;
* parenthèses intercalaires;
* crochets de parenthétisation;
* deux-points suivi d'une explication

Nous avons ajouté à cette liste des marques de citation plus rares indiquant une connaissance des règles de rédaction de textes documentés :

* signe correct d'effacement de passage dans une citation;

* crochets indiquant une modification du texte original dans une citation;

Nous avons finalement complété l'ensemble des éléments de la catégorie « Typographie » par un dernier indice, postulant la capacité d'un rédacteur expérimenté de mettre en relief, au moyen d'une énumération en colonne, des renseignements se présentant sous forme de liste :

* énumération en colonne avec puces ou tirets;

Au cours de la lecture de notre corpus, nous n'avons posé aucun jugement sur la pertinence ou l'à-propos des structures périphériques signalées — ou non — par le signe graphique approprié.

Le tableau 30 introduit des exemples tirés du corpus « Experts » ou Bissonnette pour chacun des éléments de notre catégorie « Typographie ».

Tableau 30

Indices ou variables de la catégorie "Typographie"

Indice ou variable	Exemple du corpus « Experts » ou du corpus Bissonnette : occurrence pertinente (expert : page) ou (B : texte)
Tirets intercalaires	*Cependant, l'inflation — qui atteignit parfois des taux à deux chiffres — constitua l'un des éléments principaux du système, en allégeant au maximum la charge réelle de la dette. (101 :1)*
Parenthèses intercalaires	*Le plan canadien pour les enfants, Grandir ensemble (3)*[46]*, illustre bien cette tendance (que l'on observe, même si c'est de manière inégale, dans l'ensemble des provinces). (117 :1)*
	L'interprétation contraire (des élections utiles) a été moins entendue. (106 :1)
Crochets de parenthétisation	*Le niveau des taux d'intérêt réels (taux nominaux ajustés de l'inflation) ne pouvait donc que demeurer fort élevé au début de 1992 [cf. tableau*[47]*]. (101 :1)*
	«[...] La France est plus fraternelle dans sa solidarité et son accompagnement que ce qui ressort des propos de monsieur Dion [le ministre fédéral] depuis six mois » (120 : 1)
Deux-points suivi d'une explication	*Les autorités sont donc prévenues : une autre crise de la conscription se prépare au Québec. (144 :1)*
	Pour sa part, L'UEO est sortie de sa torpeur : elle a admis (à part entière ou en tant qu'observateurs ou membres associés) les autres pays de l'Union européenne [...] (136 :2)
Signe correct d'effacement de passage dans une citation	*Dans cette lettre, il reproche à l'élite de craindre que « l'instruction des couches inférieures, et particulièrement [...] l'instruction des Canadiens » ne mène à une révolution. (148 : 1)*
Crochets indiquant une modification dans une citation	*En voulant dénoncer ces propos, le ministre Dion a toutefois quelque peu déraper [sic], affirmant qu'il « est légitime d'essayer d'insister sur les préoccupations de chaque province de ce pays, mais [qu']il n'est pas légitime d'utiliser la sécession comme moyen pour y parvenir ». (108 :1)*
	« Je les jetté par la fenestre de notre chambre [avec tout] ce que se trouva sous ma main » (147 :2) [L'auteur cite Marie de l'Incarnation].
Énumération en colonne avec puces ou tirets	*Mardi, le ministre a refusé net cette recommandation, sous deux prétextes :*
	il faudrait lire « environ deux millions de pages de documents », ce qui semble être au-dessus des capacités intellectuelles de son ministère;
	« la Commission était la mieux placée pour évaluer la crédibilité » des témoins. [...] (B : 05)

[46] Renvoi à une référence bibliographique dans le texte original.

[47] Mis en italique dans le texte original.

4.2.4 Catégorie « Verbe »

Nous nous sommes intéressés à deux éléments seulement dans cette catégorie : les formes rares du subjonctif et *si* hypothétique suivi de l'imparfait.

Wagner et Pinchon (1991) rapportent la vitalité du subjonctif en français moderne. Ils précisent cependant que l'imparfait et le plus-que-parfait du subjonctif, très présents en français classique, ne sont plus en usage dans la langue parlée (*ibid.* : 345) et se retrouvent dans des contextes limités en langue écrite[48], notamment après un verbe principal au passé (*loc. cit.*). Nous avons postulé que cette concordance ne se retrouverait que chez les rédacteurs maîtrisant les règles de conjugaison verbale et de concordance des temps.

La question du *si* hypothétique suscite, chez les rédacteurs moins sensibles à la qualité de leur langue écrite, une erreur fréquente : l'emploi du conditionnel. Nous avons par conséquent postulé que l'emploi de l'imparfait dans une proposition hypothétique introduite par un *si* pourrait signaler une qualité supérieure de contrôle linguistique.

Le tableau 31 présente des exemples de la catégorie « Verbe » tirés des corpus « Experts » et Bissonnette.

---

[48] Pour une lecture fascinante sur les formes rares du subjonctif en français moderne, voir (Grevisse, 1980 : 1407), qui cite les propos — parfois colorés — d'écrivains célèbres, dont Jules Renard, André Gide, Georges Duhamel et Albert Camus, commentant la question. Grevisse termine un article 2750 inusité en notant qu'*une des raisons pour lesquelles l'imparfait du subjonctif décline, c'est l'embarras que bien des gens — et même des écrivains chevronnés — éprouvent à en trouver les formes justes.* (*ibid :* 1408). La citation des fautes de conjugaison de ces temps difficiles dans les écrits de Goncourt et Bernanos conclut sa démonstration.

Tableau 31

Indices ou variables de la catégorie "Verbe"

Indice ou variable	Exemple du corpus « Experts » ou du corpus Bissonnette (expert : page) ou (B : texte)
Formes rares du subjonctif	*Ceci étant dit, je concède que la TGB eût été mieux située place de la Concorde, à ce détail près qu'il eût fallu démolir quelques pavillons monumentaux du XVIII^e^ siècle ou l'obélisque de Louxor [...] (B : 01)*
	Il eût été bien surprenant qu'avec une croissance mondiale de 1% environ en 1991 contre 2,2% en 1990, [...], la sphère financière ait pu être très florissante. (101 :01)
Si hypothétique suivi de l'imparfait	*Mais c'est vrai que si le Québec obtenait la souveraineté, cela pourrait créer ici un effet d'imitation, remarque encore M. Philippe Resnick. (103 : 4)*

4.2.5 Catégorie « Incises »

La proposition incise est décrite par (Grevisse, 1980 : 165) comme une proposition courte, intercalée dans une phrase pour accompagner une citation ou exprimer une sorte de parenthèse. À cause principalement des nombreuses manipulations possibles de l'ordre des mots dans ce type de proposition en dépit de sa brièveté (Wagner et Pinchon, 1991 : 560) et des contextes restreints dans lesquels elle risque d'apparaître (*loc. cit.*), nous avons postulé que l'emploi d'une telle proposition se retrouverait plus fréquemment chez les rédacteurs experts. Par ailleurs, avec le corpus Moffet regroupant des dissertations critiques qui fournissaient justement des contextes favorables à l'emploi des incises, cet indice devenait encore plus intéressant.

Le tableau 32 montre des exemples d'incises tirés des corpus « Experts » et Bissonnette.

Tableau 32

Indices ou variables de la catégorie "Incises"

Indice ou variable	Exemple du corpus « Experts » ou du corpus Bissonnette (expert : page) ou (B : texte)
Incises	*Cette perspective, il est vrai, est contestée par les organisations non gouvernementales [...] (III : 1)* *Elle aimait les pauvres, disait-on, mais elle ne détestait pas la pauvreté. (B : 16)* *« Et si les négociations n'aboutissent pas? », demandera le gouvernement canadien, « qu'est-ce qui se passera? ».(119 : 1)* *Mais ce projet, on le sait, appartient plutôt au premier ministre. (B : 12)*

Conclusion

Notre grille de calibrage constitue une première tentative pour constituer un instrument de mesure objectif pour décrire le profil linguistique d'un rédacteur.

Au cours de la constitution de cette grille, nous avons pris plusieurs décisions importantes, notamment pendant l'étape exploratoire de notre recherche[49]. Nous avons d'abord renoncé à estimer la richesse lexicale par décompte de mots nouveaux et de redondances, postulant que le repérage et le décompte d'une courte liste de mots vagues permettraient de discriminer tout autant sinon mieux. Nous avons ensuite mis de côté toute estimation de la qualité de structure de texte, suggérée dans la catégorie « Texte » de la grille Guénette, Lépine et Roy (1995) en raison principalement du niveau de subjectivité pouvant être impliqué dans ce genre d'analyse. Également, devant deux indices apparemment équivalents pour leur qualité discriminante (par exemple, *tout* et *quelque*), nous avons choisi celui dont les contextes pouvaient se repérer automatiquement le plus facilement. Enfin, nous avons réduit autant que possible l'estimation des niveaux de contrôle à l'exploitation de listes se prêtant bien à une exploitation informatique.

Nos efforts n'ont pas consisté à développer la grille de calibrage idéale, mais à développer la meilleure grille de calibrage possible dans l'état actuel de nos

[49] Voir Chapitre 3. *Méthodologie.*

connaissances. Comme personne avant nous n'a encore pensé d'établir le profil linguistique du rédacteur pour faciliter la correction automatique de ses erreurs, particulièrement de ponctuation, il demeure clair que la présente grille ne constitue qu'une première étape dans ce nouveau domaine de recherche. D'autres chercheurs auront à répéter notre expérience pour valider, et au besoin, améliorer la grille de calibrage proposée ici.

Cinq
Indices de faiblesse linguistique

Les fautes sont traditionnellement associées au manque de contrôle linguistique. Les leçons de grammaire et d'orthographe — et plus tard, les classes de récupération — veulent en effet développer ou accroître la capacité du rédacteur à reconnaître ses erreurs et à les corriger. Par conséquent, il est communément admis que l'incapacité à détecter ou à corriger les fautes constitue une faiblesse majeure chez un rédacteur (de là le besoin — et la popularité — des correcteurs orthographiques).

Dans l'établissement d'un profil linguistique, l'évaluation de la qualité du texte passe donc obligatoirement par le décompte des erreurs détectées[50], l'expérience nous ayant appris que plus les rédacteurs sont faibles, plus leurs textes contiennent d'erreurs. Cependant, le problème de la reconnaissance automatique d'erreurs met à l'avant-scène le problème de la robustesse de l'analyse.

La question de la robustesse des parseurs est abondamment documentée dans la littérature du domaine. Ted Briscoe (1996b) établit par exemple que le traitement des unités tombant en dehors du champ d'analyse (*undergeneration*) constitue l'un des trois aspects fondamentaux du problème de robustesse en analyse syntaxique automatique. Or il se trouve qu'en correction automatique, un analyseur syntaxique doit justement pouvoir reconnaître les suites erronées d'un texte tout en étant capable de remplacer les séquences posant problème par des séquences linguistiquement acceptables (ou à tout le moins les suggérer). Par conséquent, le parseur d'un module de correction automatique devrait être programmé pour

---

[50] En elle-même la détection valide des erreurs d'un texte constitue un aspect seulement partiellement résolu du traitement correctif automatique de la langue (Chandioux, 1996). Nous sommes conscients de cette problématique fort complexe mais choisissons d'assumer, pour les fins de notre recherche, que les techniques de détection d'erreurs linguistiques sont opérationnelles.

reconnaître non seulement les phrases du français mais également toutes les variations erronées de ces phrases. Il n'est donc pas étonnant que la validité du diagnostic posé par les correcteurs sur un texte donné constitue le problème fondamental identifié par des développeurs comme Chandioux (1996) tout autant que par les utilisateurs (Lesage, 1993).

C'est pourquoi l'examen des erreurs de notre corpus a pour but principal l'identification d'erreurs discriminantes, certes, mais surtout d'erreurs discriminantes repérables au moyen d'un analyseur syntaxique aussi simple que possible.

Par ailleurs, certaines erreurs demeurent virtuelles pour la plupart des rédacteurs. En effet, dans une étude d'erreurs faite à partir d'un corpus de textes originaux, l'absence de certaines erreurs ne signifie pas nécessairement qu'une faute ne serait pas commise si le rédacteur avait utilisé le contexte approprié. Dans notre recherche de variables discriminantes, nous devons donc mettre en tête de liste les occurrences ayant apparu en proportion significative dans notre corpus pour augmenter les chances du module de calibrage de trouver les contextes discriminants dans un texte donné.

Quatre questions se posent donc. Est-il vrai que les « experts » ne commettent pas d'erreurs? Certaines erreurs sont-elles plus caractéristiques du niveau de contrôle linguistique que d'autres? Les erreurs les plus discriminantes requièrent-elles un *parseur*? Les erreurs discriminantes se trouvent-elles dans une portion significative de notre corpus?

L'étude de notre corpus révèle la présence occasionnelle d'erreurs dans les textes experts tout en faisant effectivement ressortir plusieurs erreurs discriminantes requérant peu d'analyses syntaxiques complexes dans des contextes largement utilisés.

5.1 Performance « Erreurs » des sujets tous corpus confondus

Les sujets Moffet ont commis 95% des erreurs linguistiques[51] repérées dans l'ensemble de notre corpus (Tableau 33):

Tableau 33

Sommaire du corpus sur le plan de la performance linguistique

	Total Erreurs
Moffet	883
Experts	42
Bissonnette	0
Total Erreurs	925

Les textes du corpus Bissonnette ne contiennent aucune erreur: Comme nous avions pu l'observer dans notre corpus d'entraînement A, les textes experts contiennent aussi des erreurs, bien qu'en nombre significativement moins élevé.

5.1.1 Distribution des erreurs par catégorie

Notre grille « Erreurs » organise les erreurs en 6 catégories[53]. La figure 60 distribue les erreurs du corpus selon ces catégories.

---

51 Par opposition aux erreurs relatives à l'emploi des signes de ponctuation, qui sont plutôt de nature graphique (Nunberg, 1990).

52 Rappelons qu'il s'agit ici de compilation de données exprimées en format logique.

53 Voir Chapitre 4. *Grille de calibrage.*

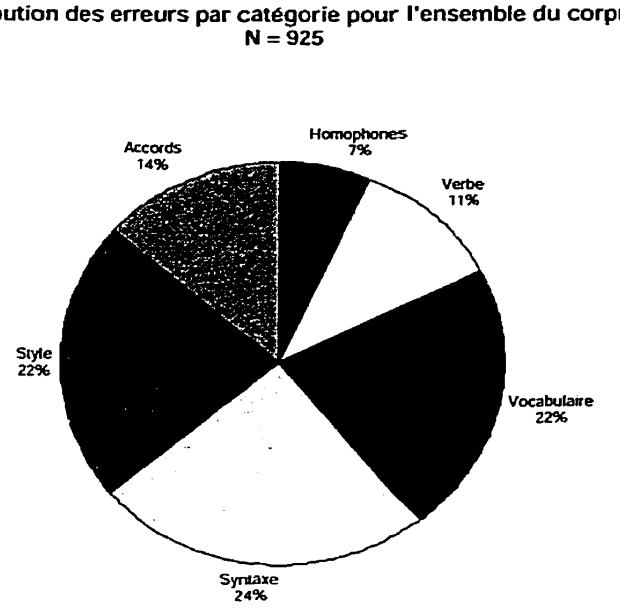


Figure 60

Distribution des erreurs par catégorie pour l'ensemble du corpus

Les problèmes de style, de vocabulaire et de syntaxe constituent près de 75% de toutes les erreurs du corpus. Ces catégories se répartissent en parts à peu près égales, ce qui donne à penser qu'aucun de ces trois groupes de problèmes ne se distingue nécessairement des autres sur le plan de la difficulté si nous considérons ensemble les 225 textes du corpus.

5.1.2 Distribution des erreurs par fréquence

Quelles sont les erreurs les plus souvent notées dans notre corpus de recherche? Le tableau 34 liste ces variables en ordre décroissant, tous groupes de sujets confondus. Précisons que, pour chaque élément de cette liste, le nombre maximal d'occurrences possibles est de 150, soit deux fois 75 textes[54]. Le nombre maximal total possible d'occurrences pour l'ensemble du corpus est par conséquent de 4 050 (27 occurrences x 150 textes).

[54] Le corpus Bissonnette ne compte évidemment pour rien dans ce calcul puisqu'il ne contient pas d'erreurs.

Tableau 34

Liste des erreurs pour l'ensemble du corpus

Variable	Total erreurs	Pourcentage (N = 150)
Mots essentiels manquants	77	51%
Répétition abusive de mots	76	51%
Impropriétés	71	47%
Mots manquant de précision	69	46%
Termes inutiles ou redondants	68	45%
Accords du syntagme nominal (SN)	64	43%
Accords du syntagme verbal (SV)	63	42%
Références anaphoriques	59	39%
Suite asyntaxique	42	28%
Autres homophones que *on / ont* et *à / a*	41	27%
Confusion terminaisons verbales	37	25%
Élément manquant ou incohérence dans l'emploi de connecteurs en série	36	24%
Incohérence dans le choix des pronoms personnels	35	23%
Ordre des mots	30	20%
Confusion –é/-er/-ez	26	17%
Passé simple inapproprié	17	11%
Barbarisme lexical	16	11%
Barbarisme grammatical	16	11%
Confusion on / ont	15	10%
Désordre syntaxique inextricable	13	9%
Confusion de genres	12	8%
Subjonctif manquant ou inapproprié	11	7%
Archaïsmes	10	7%
Confusion –i/-it	7	5%
Confusion à / a	5	3%
Cataclysme orthographique	5	3%
Si hypothétique suivi du conditionnel	4	3%
Total	**925**	**23%**[55]

[55] N= 4 050.

Des « mots essentiels manquants » (catégorie « Syntaxe ») et la « répétition abusive de mots » (catégorie « Style »), constituent les deux erreurs les plus souvent identifiées dans notre corpus, suivies de près par deux erreurs de type lexical, « Impropriétés » et « Mots manquant de précision ». Le conditionnel avec *si* hypothétique, les cataclysmes orthographiques et certains confusions homophoniques sont les occurrences les moins fréquentes pour l'ensemble du corpus.

5.2 Performance « Erreurs » des sujets par corpus

Le graphique 61 distribue les erreurs par groupes de sujets et par catégories.

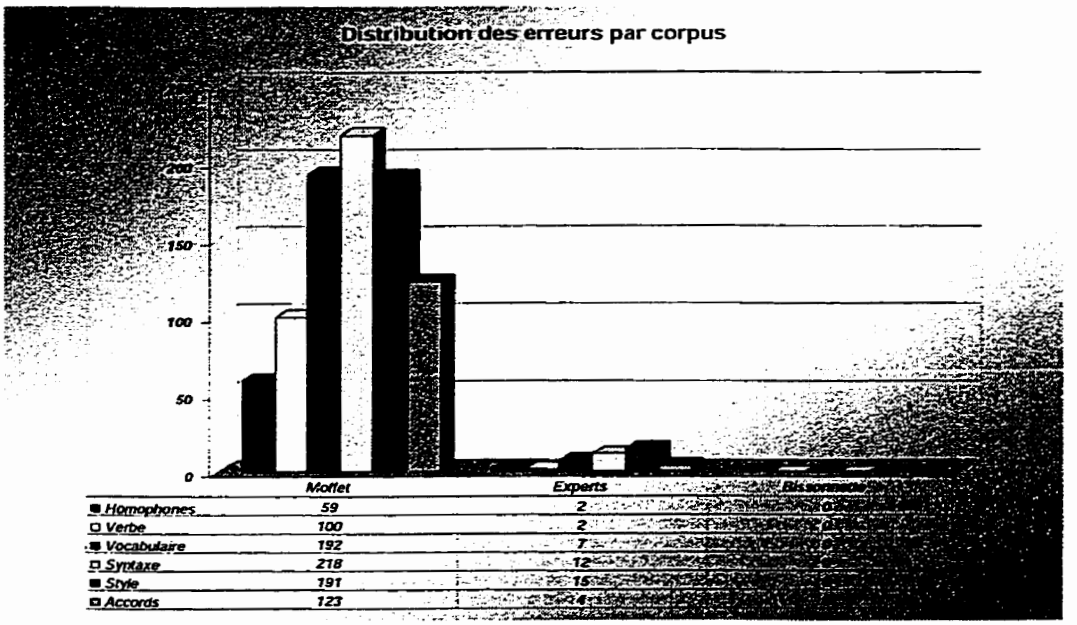


Figure 61

Distribution des erreurs par corpus

Malgré la différence dans le nombre d'occurrences, le style et la syntaxe ont posé les plus grandes difficultés pour les deux groupes de sujets.

5.2.1 Fréquences relatives des catégories d'erreurs dans le corpus Moffet

Le graphique 62 fait ressortir la distribution des erreurs dans le corpus Moffet.

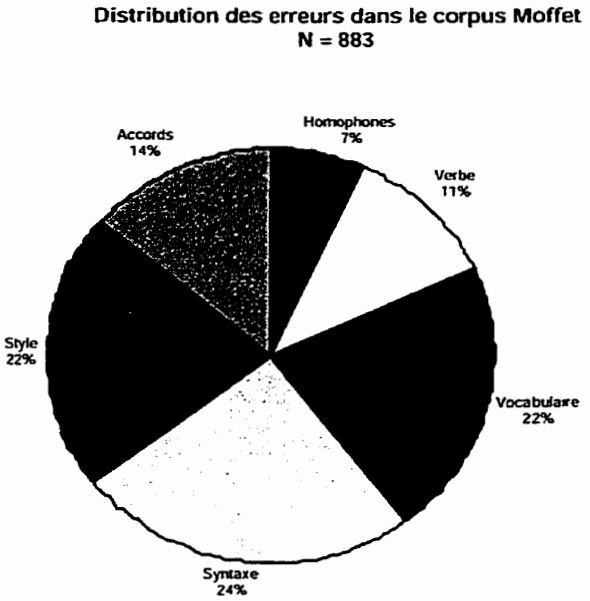


Figure 62

Distribution des erreurs dans le corpus Moffet

Ce sont les problèmes de syntaxe qui paraissent occuper la portion la plus importante du bassin d'erreurs dans le corpus Moffet, bien que les fautes de style et de vocabulaire suivent de près. Les problèmes d'accord, d'emploi du verbe et de confusion homophonique constituent ensemble un peu plus du quart des occurrences.

5.2.2 Fréquences relatives des catégories d'erreurs dans le corpus « Experts »

Le graphique 63 illustre la performance des sujets experts.

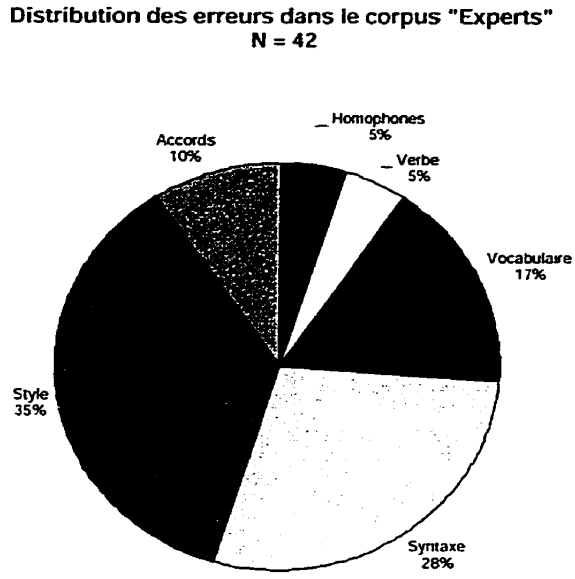


Figure 63

Distribution des erreurs dans le corpus "Experts

Les sujets experts ont fait des fautes dans toutes les catégories linguistiques décrites dans notre grille. Contrastant avec les sujets du corpus Moffet cependant, les rédacteurs experts surprennent par la proportion importante de textes (plus du tiers) montrant des erreurs de style. En fait, les problèmes stylistiques et syntaxiques occupent près des deux tiers (63%) de toutes les erreurs chez les experts en comparaison avec moins de la moitié (46%) chez nos sujets non experts. En revanche, les erreurs de vocabulaire et d'emploi du verbe occupent ensemble moins du quart (22%) des erreurs relevées chez les experts (contre le tiers [33%] chez les rédacteurs occasionnels).

5.3 Performance « Erreurs » des sujets par catégories

Il importe de rappeler qu'en raison du format logique que nous avons adopté pour nos données, comptabiliser une seule occurrence par texte — quand elle se présente — revient en fait à déterminer le nombre de textes ayant présenté cette occurrence dans le corpus.

L'étude des occurrences permet ainsi de faire ressortir les difficultés particulières posées par chacune des catégories. Dans les tableaux d'occurrences, les données sont classées en ordre décroissant d'après les résultats des sujets Moffet.

5.3.1 Catégorie « Syntaxe »

Les erreurs syntaxiques constituent 24% des erreurs du corpus Moffet et 28% de celles du corpus « Experts » (Tableau 3 5).

Tableau 35

Nombre d'erreurs dans la catégorie "Syntaxe"

Erreurs	Moffet	Experts
Mots essentiels manquants	71	6
Termes inutiles ou redondants	65	3
Ordre des mots	29	1
Suite asyntaxique	40	2
Désordre syntaxique inextricable	13	0
SYNTAXE	**218**	**12**

Le jugement sur la nécessité syntaxique d'un mot constitue la difficulté majeure des sujets de notre corpus dans la catégorie « Syntaxe ». En effet, 136 erreurs du corpus Moffet sur 218 (63%), repérées dans cette catégorie, touchent la décision de placer ou non un mot dans une phrase en accord avec les règles de la syntaxe française. Quant aux 12 fautes syntaxiques notées dans le corpus « Experts », 9 d'entre elles concernent le même type de problème.

Cependant, les autres types d'erreurs de la catégorie « Syntaxe » sont pratiquement inexistantes chez les experts mais relativement nombreuses chez les sujets Moffet. Les experts de notre corpus, par exemple, n'ont produit aucune suite incompréhensible alors que 13 textes du groupe Moffet (17%) contiennent au moins une phrase impossible à comprendre.

Le tableau 36 introduit la proportion de textes du corpus présentant les erreurs identifiées dans la catégorie « Syntaxe ».

Tableau 36

Fréquences relatives des problèmes syntaxiques par corpus

Erreurs	Moffet % N = 75	Experts % N = 75
Mots essentiels manquants	95	8
Termes inutiles ou redondants	87	4
Ordre des mots	39	1
Suite asyntaxique	53	3
Désordre syntaxique inextricable	17	0

Notre étude montre une proportion très élevée de textes avec des problèmes syntaxiques sérieux. En effet, il manque des mots essentiels à la phrase française dans 95% des textes Moffet (corpus considéré, rappelons-le, comme statistiquement représentatif de la population du point de vue du critère de la langue[56]). En outre, des mots inutiles sont présents dans près de 9 textes sur 10. Également, plus d'un texte non expert sur deux introduit au moins une suite non conforme à la syntaxe française. Finalement, près de 40% du corpus Moffet confondent l'ordre des mots, erreur qui complique souvent le processus de compréhension de la phrase.

Le pourcentage élevé de textes de notre échantillon non expert avec erreurs syntaxiques importantes est troublant du point de vue de la correction automatique de la langue, et particulièrement de la ponctuation : si la détermination manuelle de la ponctuation passe par une analyse syntaxique adéquate, particulièrement dans l'exploitation des virgules (Jones, 1996c; Simard, 1993; Nunberg, 1990), comment alors transposer cette analyse en correction automatique sans disposer d'un parseur capable de rétablir au besoin l'intégrité syntaxique des phrases à ponctuer?

---

[56] Voir Chapitre 3. *Méthodologie.*

5.3.2 Catégorie « Vocabulaire »

Il importe de rappeler que le corpus Moffet n'est pas représentatif de la population du point de vue du vocabulaire : les sujets de l'échantillon ont eu des résultats meilleurs que la population pour cet aspect de la langue.

Le tableau 37 présente les données de notre corpus pour la catégorie « Vocabulaire ».

Tableau 37

Nombre d'erreurs dans la catégorie "Vocabulaire"

Erreurs	Moffet	Experts
Impropriétés	69	2
Mots manquant de précision	65	4
Barbarisme lexical	16	0
Barbarisme grammatical	15	1
Confusion de genres	12	0
Archaïsmes	10	0
Cataclysme orthographique	5	0
VOCABULAIRE	**192**	**7**

Cette partie de notre grille « Erreurs » permet de poser un jugement sur la qualité lexicale des textes de notre corpus. Les mots mal employés et les mots vagues constituent la majorité des erreurs de vocabulaire des sujets du corpus. Par ailleurs, les erreurs fondamentales comme les barbarismes, les confusions de genres, l'emploi de mots désuets et les cataclysmes orthographiques[57] sont à toutes fins pratiques absents chez les experts, bien que ces erreurs composent ensemble près du tiers (30%) des erreurs de vocabulaire chez les sujets non experts.

Considérant que les résultats de l'échantillon Moffet sont meilleurs que ceux de la population, nous pouvons nous attendre à ce que celle-ci présente une distribution

[57] Mot contenant plus d'une erreur d'orthographe d'usage (Guénette, Lépine et Roy, 1995).

d'erreurs lexicales en nombre encore plus important, avec une majorité d'impropriétés et de mots vagues.

Le tableau 38 affiche le pourcentage de textes par corpus présentant chaque type d'erreurs lexicales.

Tableau 38

Fréquences relatives des erreurs de vocabulaire par corpus

Erreurs	Moffet % N = 75	Experts % N = 75
Impropriétés	92	3
Mots manquant de précision	87	4
Barbarisme lexical	22	0
Barbarisme grammatical	20	1
Confusion de genres	16	0
Archaïsmes	13	0
Cataclysme orthographique	7	0

Les problèmes lexicaux sont pratiquement inexistants dans les textes experts en comparaison avec les textes de l'échantillon Moffet. À cause de la proportion très élevée, dans le corpus Moffet, d'impropriétés et de mots vagues de même que de la présence non significative de ce type d'erreurs dans le corpus « Experts », nous pouvons conclure que les impropriétés et les mots vagues sont des erreurs qui contribuent à caractériser un profil linguistique de rédacteur occasionnel. Malheureusement, les impropriétés demeurent pour l'instant difficilement détectables de façon automatique.

5.3.3 Catégorie « Style »

Les erreurs stylistiques occupent 35% du corpus « Experts » et 22% du corpus Moffet (Tableau 39).

Tableau 39

Nombre d'erreurs dans la catégorie « Style »

Erreurs	Moffet	Experts
Répétition abusive de mots	67	9
Références anaphoriques	55	4
Élément manquant ou incohérence dans l'emploi de connecteurs en série	35	1
Incohérence dans le choix des pronoms personnels	34	1
STYLE	**191**	**15**

Les répétitions abusives de mots et les références anaphoriques constituent 68% des erreurs stylistiques du corpus Moffet. Ces erreurs sont aussi les plus fréquentes dans le petit corpus d'erreurs stylistiques des textes experts.

Le tableau 40 présente le pourcentage de textes par corpus avec des erreurs stylistiques.

Tableau 40

Fréquences relatives des erreurs stylistiques par corpus

Erreurs	Moffet % N=75	Experts % N=75
Répétition abusive de mots	89	12
Références anaphoriques	73	5
Élément manquant ou incohérence dans l'emploi de connecteurs en série	48	1
Incohérence dans le choix des pronoms personnels	45	1

Le tableau 40 fait état de plusieurs éléments utiles pour l'établissement possible d'un profil linguistique :

* la répétition abusive de mots est une erreur stylistique courante chez les rédacteurs occasionnels (9 textes sur 10) mais exceptionnelle chez les rédacteurs experts (1 texte sur 10);

* la référence anaphorique est une erreur très fréquente chez les rédacteurs occasionnels (près des trois-quarts de l'échantillon) mais rare chez les experts (5 textes sur 100);

* le mauvais choix de connecteurs dans une série et l'absence de congruence dans l'emploi des pronoms personnels sont des erreurs se trouvant dans pratiquement un texte sur deux de notre échantillon mais ne retrouvant pas dans notre corpus « Experts » (1 texte sur 100).

Autrement dit, soit par leur fréquence, soit par leur pouvoir discriminatoire, toutes les erreurs stylistiques de notre grille pourraient servir à élaborer un profil linguistique. Cependant, une analyse syntaxique complexe serait requise pour détecter les références anaphoriques. Par conséquent, les fautes stylistiques demandant une analyse syntaxique simple demeurent la répétition abusive de mots et le mauvais choix des connecteurs en série.

5.3.4 Catégorie « Verbe »

Les fautes verbales composent 11% du corpus Moffet mais seulement 5% du corpus « Experts » (Tableau 41).

Tableau 41

Nombre d'erreurs dans la catégorie "Verbe"

Erreurs	Moffet	Experts
Confusion terminaisons verbales	37	0
Confusion –é/-er/-ez	25	1
Passé simple inapproprié	17	0
Subjonctif manquant ou inapproprié	10	1
Confusion –i/-it	7	0
Si hypothétique suivi du conditionnel	4	0
VERBE	**100**	**2**

Comparativement aux autres catégories d'erreurs, les problèmes verbaux apparaissent en moins grand nombre dans l'ensemble des erreurs du corpus. Elles ont également toutes été commises — sauf à deux occasions — par les sujets Moffet. Les fautes de conjugaison constituent l'erreur première de l'échantillon de rédacteurs occasionnels dans la catégorie « Verbe ». Les experts de notre corpus, en revanche, n'ont pas du tout commis cette erreur. Par conséquent, les erreurs de conjugaison peuvent être considérées comme discriminantes.

Cependant, quelle est la proportion de textes Moffet présentant des problèmes de terminaisons verbales?

Tableau 42

Fréquences relatives des erreurs verbales par corpus

Erreurs	Moffet % N = 75	Experts % N = 75
Confusion terminaisons verbales	49	0
Confusion –é/-er/-ez	33	1
Passé simple inapproprié	23	0
Subjonctif manquant ou inapproprié	13	1
Confusion –i/-it	9	0
Si hypothétique suivi du conditionnel	5	0

Les problèmes de conjugaison se retrouvent à toutes fins pratiques dans la moitié de l'échantillon Moffet (Tableau 42).

La présence d'un passé simple inapproprié constitue une erreur intéressante. En effet, les rédacteurs experts ont maintes fois utilisé le passé simple mais en en respectant toujours les règles et limites d'emploi. En revanche, près du quart des sujets de l'échantillon Moffet, associant sans doute le passé simple à l'écriture recherchée, ont vainement tenté d'y avoir recours. Le pouvoir discriminant de cette erreur pourrait apparaître prometteur. Cependant, il n'en est rien du point de vue

d'un module de calibrage puisque celui-ci ne peut pas fonder son diagnostic sur l'interprétation sémantique des contextes. Autrement dit, c'est seulement une fois que le profil linguistique aurait été établi qu'une routine informatique serait en mesure d'associer le passé simple à une erreur d'emploi. Cependant, si le pouvoir discriminant de l'occurrence d'un passé simple ne peut pas être utile à un module de calibrage, il peut l'être pour un module de correction automatique : s'il s'agit d'un rédacteur expert, le passé simple est correctement employé; s'il s'agit d'un rédacteur non expert, les chances sont bonnes pour qu'il s'agisse d'un passé simple inapproprié.

5.3.5 Catégorie « Accords »

Des difficultés d'accords se retrouvent dans une proportion voisine chez les rédacteurs experts (10%) et non experts (14%) (Tableau 43).

Tableau 43

Nombre d'erreurs dans la catégorie "Accords"

Erreurs	Moffet	Experts
Accords du syntagme nominal (SN)	61	3
Accords du syntagme verbal (SV)	62	1
ACCORDS	**123**	**4**

Il est souvent fait grand cas des difficultés des rédacteurs occasionnels à effectuer les accords appropriés. Le nombre peu élevé d'erreurs d'accord dans le corpus indique toutefois que ces préoccupations sont exagérées puisque les sujets Moffet se sont relativement bien débrouillés. Cependant, la différence entre les rédacteurs experts et non experts sur le plan des accords demeure significative.

Le tableau 44 fait état des erreurs proportionnellement au corpus.

Tableau 44

Fréquences relatives des erreurs d'accord par corpus

Erreurs	Moffet % N=75	Experts % N=75
Accords du syntagme nominal (SN)	81	4
Accords du syntagme verbal (SV)	83	1

Les erreurs d'accord se retrouvent dans plus de 80% des textes du corpus Moffet. En revanche, ces erreurs occupent une proportion presque insignifiante du corpus « Experts ». Le pouvoir discriminant des erreurs d'accord apparaît donc élevé. Les fautes d'accord sont intéressantes également en raison de leur haut potentiel d'occurrence directement tributaire des règles morpho-syntaxiques du français écrit.

Néanmoins, les fautes d'accord, malgré leur fort pouvoir de discrimination, posent des difficultés importantes sur le plan du traitement automatique de la langue. En effet, les bruits (fausses détections) et les silences (détections manquantes) abondent dans les diagnostics des fautes d'accord (Chandioux, 1996) des correcteurs grammaticaux si bien que, souvent, de tels diagnostics restent peu fiables. Par conséquent, les erreurs d'accord constituent un outil de calibrage peu intéressant (du moins tant que les correcteurs n'auront pas amélioré leur performance dans ce domaine).

5.3.6 Catégorie « Homophones »

Le tableau 45 montre la distribution des confusions homophoniques dans notre corpus de recherche.

Tableau 45

Nombre d'erreurs dans la catégorie "Homophones"

Erreurs	Moffet	Experts
Confusion *à / a*	5	0
Confusion *on / ont*	15	0
Autres homophones	39	2
HOMOPHONES	**59**	**2**

Ce sont les problèmes posés par d'autres homophones que ceux que nous avions préalablement identifiés[58] qui réunissent le plus grand nombre d'erreurs. Cependant, comme l'ensemble « autres homophones » regroupe 26 occurrences différentes, ce résultat élevé n'est pas surprenant et souligne davantage un autre aspect de la faiblesse lexicale des rédacteurs occasionnels.

Le tableau 46 montre le pourcentage de textes par corpus présentant des confusions homophoniques.

Tableau 46

Fréquences relatives des confusions homophoniques par corpus

Erreurs	Moffet % N = 75	Experts % N = 75
Autres homophones	52	3
Confusion *à / a*	20	0
Confusion *on / ont*	7	0

Selon nos résultats, un texte non expert sur deux présente au moins l'une des confusions homophoniques suivantes : ce / se; ces / ses; cet / s'est; croit/ croît; davantage / d'avantages; du / dû; et / est; la / l'a; n'ont / non; n'y / ni; or / hors; où / ou; par / part; parce que / pas ce que; peu / peut; peut être / peut-être;

---

58 Voir Chapitre 4. *Grille de calibrage*.

quelques fois / quelquefois; qui / qu'il; qui l'a / qu'il a; s'en / sans; sans / sent; si / ci (dans la suite ci-haut); si / s'y; si tôt / sitôt; son / sont.

En revanche, une proportion très réduite de textes experts contient des erreurs homophoniques. Le pouvoir discriminant des confusions homophoniques, bien qu'intéressant à priori, est tempéré néanmoins par la difficulté de les diagnostiquer automatiquement. En effet, ou bien le recours au sens est nécessaire comme dans *quelques fois / quelquefois, ces / ses* ou *croit/ croît,* ou bien c'est le recours à une analyse syntaxique complexe qui l'est comme dans *ce / se, cet / s'est* ou *son / sont.*

Par ailleurs, un texte Moffet sur cinq introduit une confusion entre la préposition *à* et l'auxiliaire avoir, générant, par la même occasion, des ambiguïtés certaines dans l'interprétation de suites syntaxiques du type PRÉP. + Verbe. Les textes de notre corpus « Experts » ne présentent pas cette erreur. La confusion entre *à* et *a* est une autre faute avec un haut potentiel d'occurrence en raison de la fréquence d'emploi de ces deux mots en français. Cependant le recours à une analyse syntaxique complexe serait nécessaire pour détecter correctement la faute, ce qui la rend moins intéressante du point de vue du calibrage.

5.4 Étude statistique

Une étude statistique fait ressortir la marge d'erreurs attendues dans des populations montrant des profils linguistiques différents.

5.4.1 Étude de moyennes

Une étude des moyennes nous permet de dresser le portrait possible d'un texte type sur le plan des erreurs.

La différence entre les sujets de l'échantillon Moffet et ceux du corpus « Experts » est significative : les rédacteurs occasionnels ont réalisé en moyenne 12 erreurs par texte contre 1 seule pour les rédacteurs experts. Le tableau 47 présente la distribution de ces moyennes pour l'ensemble du corpus :

Tableau 47

Moyennes des erreurs par corpus

Corpus	Moyenne	Écart-type
Moffet	11,77	3, 41
Experts	0,56	1, 14
Bissonnette	0	0

En outre, l'écart-type fait ressortir l'homogénéité de nos deux groupes. L'écart-type mesure en effet la dispersion des données autour de la moyenne. Plus ce nombre est élevé, moins significative est la moyenne. Un écart-type de 3, 41 chez les sujets Moffet indique qu'en moyenne, pas plus de 3 erreurs séparent la majorité de l'échantillon de la moyenne générale de 12 fautes. De la même façon, pas plus d'une faute ne sépare la majorité des experts de la moyenne générale d'une erreur.

Cependant, Cohen (1995 : 27) met en garde contre les mesures de moyennes (et de variances) en raison de l'influence indue de données exceptionnelles (*outliers*[59]). Pour déterminer la fiabilité de nos moyennes, il nous faut donc détecter la présence possible d'observations extrêmes. Pour illustrer cette importance, Cohen (*ibid* : 51) a recours à des diagrammes de dispersion (*scatterpoints*) qui permettent de visualiser la distance des données les unes par rapport aux autres[60].

---

59 *An outlier* [...], *is a point that is further from most of points than they are from each other.* (Cohen, 1995 : 51)

60 Cohen (*loc. cit.*) précise aussi qu'il existe des procédures mathématiques pour détecter des données extrêmes dans un ensemble important de données et renvoie à des ouvrages sur ce sujet. Cependant, nous atteignons le même but avec nos graphiques à cause du nombre limité des textes de notre corpus.

La figure 64 situe le nombre d'erreurs par texte dans le corpus Moffet.

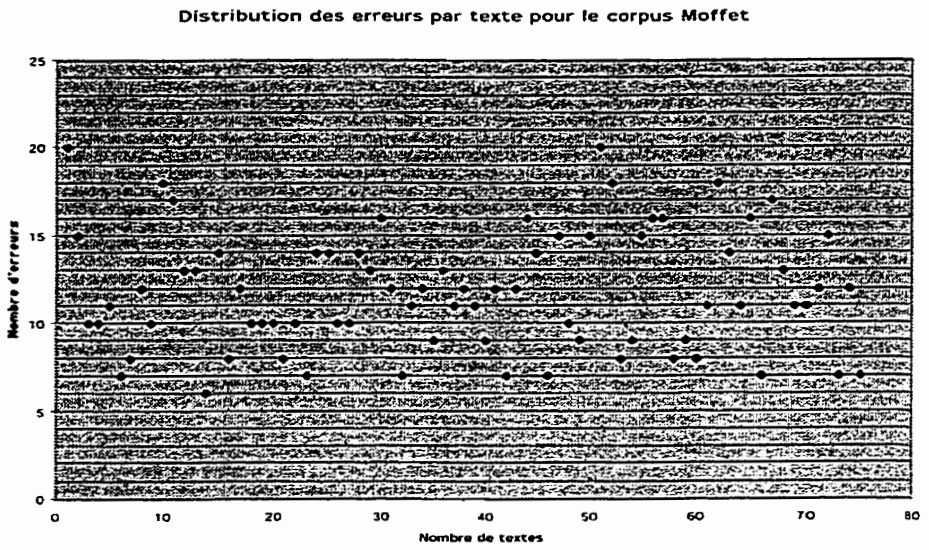


Figure 64

Distribution des erreurs par texte pour le corpus Moffet

Les données du corpus Moffet ne montrent pas de distances extraordinaires entre les points. La moyenne d'erreurs de l'échantillon est donc recevable.

La figure 65 situe les erreurs par texte pour le corpus « Experts ».

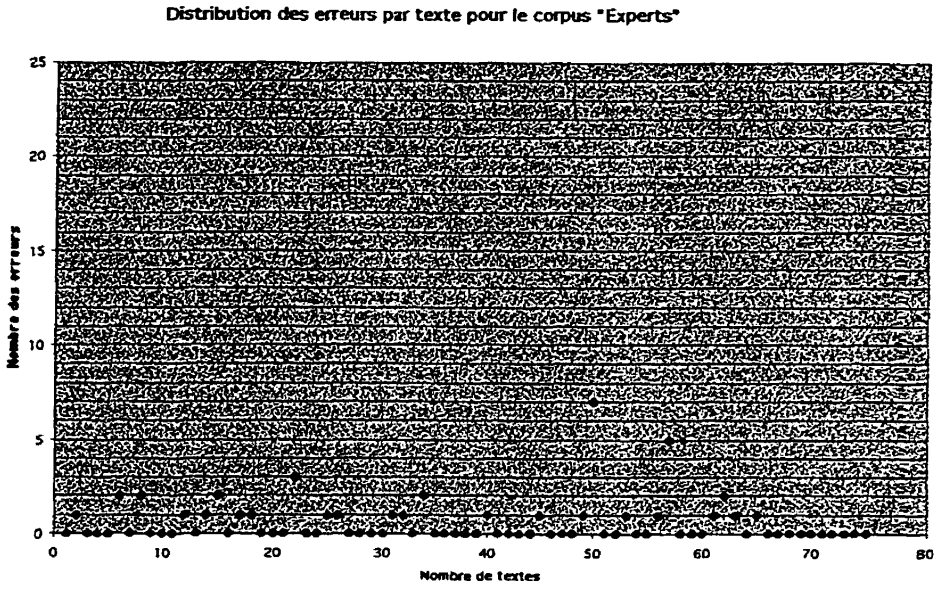


Figure 65

Distribution des erreurs par texte pour le corpus « Experts »

Le corpus « Experts » présente deux données s'écartant davantage de l'ensemble. Cohen (*loc. cit.*) suggère de considérer l'évacuation de telles données si leur présence influence indûment les résultats. Cependant, ajoute-t-il, si l'objectif statistique poursuivi peut être atteint en dépit de la présence de données extraordinaires, leur effacement n'est pas nécessaire.

Le tableau 48 compare le calcul des moyennes des erreurs du corpus « Experts » après évacuation des deux observations possiblement[61] extrêmes.

Tableau 48

Données extraordinaires et analyse des résultats du corpus « Experts »

Corpus « Experts »	Erreurs	Moyenne par texte	Écart-type
Avec observations extrêmes	42	0,56	1,14
Sans observations extrêmes	35	0,93	0,86

61 Comme nous n'avons pas effectué de calculs pour établir mathématiquement que ces données s'écartaient de façon significative des autres, nous parlons d'écarts extraordinaires *possibles*.

L'évacuation des données possiblement extraordinaires n'influence pas les résultats de façon significative. La moyenne d'erreurs par texte est toujours de 1 et l'écart-type modifié ne fera pas de différence au moment de l'établissement de la distribution attendue des erreurs, comme nous le verrons plus loin. Par conséquent, la moyenne d'une seule erreur par texte expert demeure fiable, ce qui revient à dire que, bien que l'erreur soit possible dans un texte rédigé par un sujet expert, nos données indiquent qu'elle est en fait exceptionnelle.

Le premier critère distinctif entre les textes non experts et experts est bel et bien la présence d'erreurs.

Cependant il existe une zone floue entre les deux groupes. En effet, comme le nombre d'erreurs varie de 6 à 20 dans le corpus Moffet et de 0 à 7 dans le corpus « Experts », comment est-il possible de distinguer entre les groupes quand un texte présente 6 ou 7 erreurs ? Nous pouvons obtenir une réponse en utilisant l'écart-type pour situer les résultats sur la courbe normale.

5.4.2 Étude des écarts-types

L'écart-type mesure la dispersion d'une population autour de la moyenne. En utilisant l'écart-type (symbole σ) comme base de calcul, il est possible d'estimer les résultats d'une population, assumant qu'ils sont distribués selon la courbe normale (Cohen, 1995 : 122; Allaire,1998 : 12-1): la moyenne (symbole μ[62]) constituant le centre de la courbe normale, des valeurs se situant entre la moyenne et + 1σ et - 1σ décriront les résultats de 68, 27% de la population; + 2σ et - 2σ, de 95, 45% de la population et + 3σ et - 3σ., de 99, 73% de la population. Allaire (*loc.cit.)* ajoute que *des valeurs situées à trois écarts-types (au-dessous ou au-dessus) de la moyenne sont relativement rares dans une distribution normale.*

---

62 Les statistiques emploient le symbole μ pour désigner la moyenne d'une population et le symbole X surmonté d'une barre horizontale pour désigner celle d'un échantillon. Notre corpus Moffet est un échantillon mais notre corpus « Experts » représente une population au sens statistique. Pour simplifier la présentation de la synthèse de nos données, nous désignerons les moyennes par le symbole μ. Les calculs des moyennes sont les mêmes peu importe qu'il s'agisse d'un échantillon ou d'une population.

Le tableau 49 présente les résultats de ces calculs à partir de nos données. Les chiffres sont arrondis pour maintenant rendre compte de la nature discrète des variables[63]. Les valeurs négatives sont remplacées par la valeur 0, puisqu'il n'est pas possible de faire -1, -2 ou -3 fautes.

Tableau 49

Dispersion de la population d'erreurs autour de la moyenne

Corpus	*-3σ*	*-2σ*	*-1σ*	*μ*	*+1σ*	*+2σ*	*+3σ*
Moffet (σ = 3, 41)	2	5	8	12	15	19	22
Experts (σ = 1, 14) (avec données extraordinaires)	0	0	0	1	2	3	4
Experts (σ = 0, 86) (sans données extraordinaires)	0	0	0	1	2	3	4

Le tableau 49 confirme l'absence d'influence de la donnée exceptionnelle sur les résultats. Ces calculs démontrent en effet qu'en vertu de la courbe normale (Allaire, 1998 : 12-1), 99,73% des textes experts vont présenter de 0 à 3 fautes, avec une prépondérance de textes sans erreurs et 4 erreurs, une occurrence exceptionnelle.

En revanche, 95% des textes de sujets non experts pourront contenir de 5 à 19 fautes. Un texte non expert de 4 ou 3 fautes (entre -2 σ et -3 σ) représentera une occurrence rare — 2,141% de la population (Allaire, *loc. cit.*). Un texte non professionnel avec seulement 2 fautes sera improbable (valeur à -3 σ). Ces calculs sont vérifiables dans notre corpus comme l'illustrent les figures 66 et 67.

---

[63] Allaire (1998 : 3-2) distingue deux types de variables : les variables continues et les variables discrètes. Les variables dites continues sont celles qui peuvent prendre un nombre théoriquement infini de valeurs entre deux points. Allaire donne l'exemple des mesures de température. Les variables discrètes sont celles qui *ne peuvent prendre qu'un nombre limité de valeurs différentes*. L'auteur donne l'exemple du nombre d'enfants dans une famille.

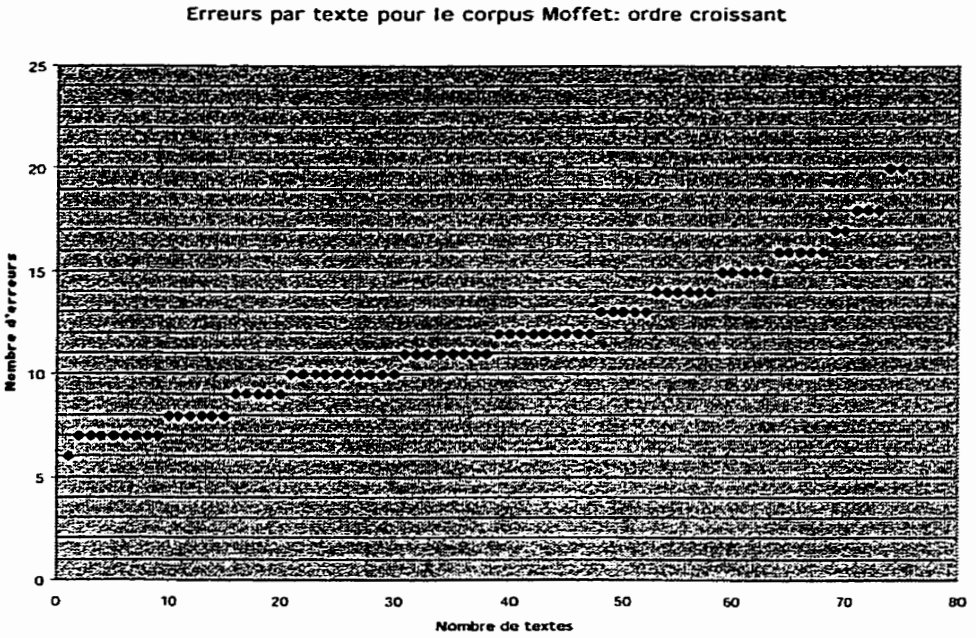


Figure 66

Erreurs par texte dans le corpus Moffet: ordre croissant

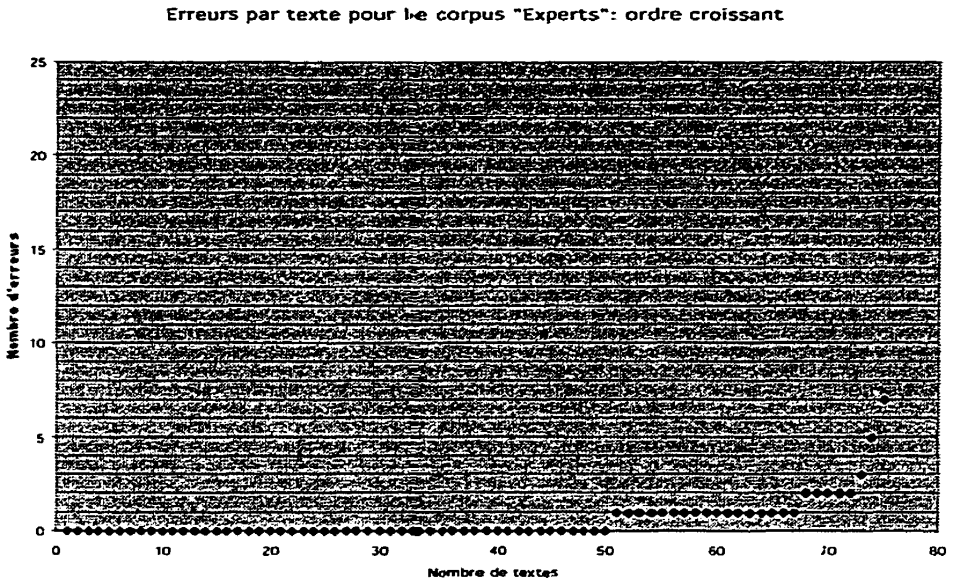


Figure 67

Erreurs par texte dans le corpus « Experts » : ordre croissant

Les figures 66 et 67 montrent également ...

* que les rédacteurs de nos deux corpus constituent des groupes relativement homogènes;
* que deux des rédacteurs du corpus « Experts » — critère qualitatif de sélection : textes publiés sous la supervision d'une équipe de production — présentent 5 et 7 erreurs, soit le profil d'erreurs de rédacteur occasionnel;
* qu'il existe une différence significative sur le plan des erreurs entre les deux groupes.

5.5 Indice de faiblesse linguistique

Un indice de faiblesse linguistique peut être calculé à partir de recherches ciblées d'occurrences prioritaires.

5.5.1 Erreurs « simples »

Le tableau 50 présente la synthèse de notre étude d'erreurs. Il montre la proportion de textes du corpus affichant chacune des variables de notre grille « Erreurs » en les divisant par groupes de sujets. Il ajoute cependant un critère additionnel : le recours à une analyse syntaxique automatique complexe.

Nous créons ainsi deux blocs d'erreurs : les erreurs identifiables au moyen d'une analyse automatique complexe et les autres. Pour notre module de calibrage, nous préférerons les erreurs dont la détection ne requiert pas d'analyse syntaxique automatique complexe.

Les données sont classées au moyen de deux clés de tri : la colonne « Analyse complexe requise » ; la colonne « Moffet ».

Tableau 50

Synthèse des résultats de la grille "Erreurs" selon les groupes de sujets

Erreurs	Moffet %	Experts %	Analyse complexe requise
Répétition abusive de mots	89	12	
Termes inutiles ou redondants	87	4	
Mots manquant de précision	87	4	
Confusion terminaisons verbales	49	0	
Élément manquant ou incohérence dans l'emploi de connecteurs en série	48	1	
Incohérence dans le choix des pr. pers.	45	1	
Passé simple inapproprié	23	0	
Barbarisme lexical	22	0	
Confusion de genres	16	0	
Subjonctif manquant ou inapproprié	13	1	
Archaïsmes	13	0	
Cataclysme orthographique	7	0	
Si hypothétique suivi du conditionnel	5	0	
Mots essentiels manquants	95	8	•
Impropriétés	92	3	•
Accords du syntagme verbal (SV)	83	1	•
Accords du syntagme nominal (SN)	81	4	•
Références anaphoriques	73	5	•
Suite asyntaxique	53	3	•
Homophones (autres que *à / a*; *ont /on*)	52	3	•
Ordre des mots	39	1	•
Confusion –é/-er/-ez	33	1	•
Confusion *à / a*	20	0	•
Barbarisme grammatical	20	1	•
Désordre syntaxique inextricable	17	0	•
Confusion –i/-it	9	0	•
Confusion *on / ont*	7	0	•

Ce classement permet de faire ressortir deux classes d'erreurs : les erreurs que nous appellerons « simples », parce que leur détection requiert une analyse syntaxique automatique simple[64] et les erreurs que nous appellerons « complexes », parce que leur détection requiert une analyse syntaxique complexe[65].

5.5.2 Erreurs discriminantes

Nous reclassons à présent la liste des erreurs simples selon leur pouvoir discriminant mesuré encore une fois au moyen de deux clés de tri : en ordre croissant des résultats des rédacteurs experts et en ordre décroissant selon les résultats des sujets Moffet (Tableau 51).

Tableau 51

Liste des erreurs de type "simple"

Erreurs	Moffet %	Experts %
Confusion terminaisons verbales	49	0
Barbarisme lexical	22	0
Confusion de genres	16	0
Archaïsmes	13	0
Cataclysme orthographique	7	0
Si hypothétique suivi du conditionnel	5	0
Élément manquant ou incohérence dans l'emploi de connecteurs en série	48	1
Incohérence dans le choix des pr. pers.	45	1
Subjonctif manquant ou inapproprié	13	1
Termes inutiles ou redondants	87	4
Mots manquant de précision	87	4
Répétition abusive de mots	89	12

[64] Par exemple, la détection et le comptage de certains mots ou la comparaison d'occurrences particulières avec les unités d'une liste donnée.

[65] Par exemple, la mise en relation des mots dans une phrase et l'attribution de leur fonction syntaxique, la détection et la résolution de problèmes d'interprétation de contextes ambigus.

Le tableau 51 fait état de 12 erreurs identifiables[66] par notre module de calibrage au moyen d'une analyse syntaxique simple. C'est suffisant, compte tenu des caractéristiques établies à partir de l'écart-type chez nos deux groupes de rédacteurs.

Pouvons-nous cependant trouver une utilité à notre second bloc d'erreurs « complexes » ? Oui, puisque ces erreurs proposent en fait une séquence d'analyses à privilégier pour un correcteur orthographique, une fois établi le profil linguistique du rédacteur. En effet, comme notre échantillon Moffet est représentatif de la population des quelque 16 000 textes dont il est tiré, nous pouvons conclure que la distribution des erreurs observées s'appliquent à l'ensemble de cette population et que, par conséquent, la proportion des textes présentant une erreur donnée voisinera celle que nous avons observée chez nos sujets.

5.5.3 Erreurs prioritaires

La liste du tableau 51 doit toutefois être réexaminée à la lumière de notre objectif principal : distinguer des rédacteurs experts des rédacteurs non experts. Il importe en effet de cibler la recherche d'erreurs selon un ordre prioritaire. Cet ordre, nous allons l'établir à partir de trois critères associés directement à l'erreur : son absence ou sa fréquence très basse chez les experts, son haut potentiel d'occurrence selon les caractéristiques de la langue (par exemple, les verbes conjugués, qui doivent apparaître dans tous les textes) et sa fréquence élevée chez les rédacteurs occasionnels. Le tableau 52 fait état de ce reclassement et présente les erreurs dans l'ordre prioritaire que nous suggérons.

---

[66] Nous devons prévoir une provision spéciale pour le passé simple inapproprié, puisque la détermination de cette erreur dépend en fait de la détermination d'appartenance à un groupe particulier de rédacteurs.

Tableau 52

Liste des erreurs prioritaires

Erreurs	Priorité	Moffet %	Experts %
Confusion terminaisons verbales	1	49	0
Élément manquant ou incohérence dans l'emploi de connecteurs en série	2	48	1
Incohérence dans le choix des pr. pers.	3	45	1
Répétition abusive de mots	4	89	12
Mots manquant de précision	5	87	4
Termes inutiles ou redondants	6	87	4
Confusion de genres	8	16	0
Subjonctif manquant ou inapproprié	9	13	1
Archaïsmes	10	13	0
Cataclysme orthographique	11	7	0
Barbarisme lexical	12	22	0
Si hypothétique suivi du conditionnel	13	5	0

Les occurrences repérées dans un texte et conformes à la liste du tableau 52 pourront être comptées et le résultat, sauvegardé pour les fins de l'analyse réalisée dans la deuxième étape de l'exercice de calibrage.

Conclusion

L'exercice que nous venons d'achever a atteint deux objectifs : confirmer la différence significative entre la performance des rédacteurs occasionnels et experts sur le plan des erreurs linguistiques et identifier des erreurs susceptibles de permettre automatiquement la distinction entre les deux groupes à partir d'une classe d'erreurs que nous avons appelée « erreurs simples ».

Le bloc d'erreurs simples ne constitue cependant qu'une étape dans le processus d'élaboration du profil linguistique que nous proposons. La seconde va

chercher un aspect jusqu'à présent encore inexploré dans l'évaluation de textes: l'indice de maîtrise linguistique.

Six
Indices de maîtrise linguistique

Les textes rédigés par des professionnels de l'écriture se reconnaissent facilement de façon intuitive. Toutefois, la recherche d'éléments objectifs sur lesquels se fonde un tel jugement génère toute une série de questions, différentes de celles découlant de la problématique qui a inspiré la présente étude, mais directement reliées à l'application pratique de notre grille de calibrage. Par exemple, sur quels éléments objectifs ce jugement se fonde-t-il? La richesse lexicale ? Le contrôle de l'aspect normatif de la langue ? L'emploi de tournures recherchées ? L'élégance et la solidité de la structure du texte? La logique de l'organisation des idées ? Sur l'ensemble de ces critères ou sur quelques critères en particulier ?

Nous pourrions penser trouver toutes nos réponses dans les grammaires normatives. Il n'en est rien. La norme mise de l'avant dans les grammaires normatives se compose en fait de plusieurs centaines de règles — et de notes —classées selon des critères variant avec l'école linguistique qui l'inspire : Grevisse (1980), par exemple, formule 2770 règles en français écrit; Wagner et Pinchon (1991), 734. Le français vu par Wagner et Pinchon serait-il plus simple que celui de Grevisse ? Nous pourrions croire que oui. Cependant, la grille de lecture de Wagner et Pinchon s'organise différemment (comme en témoignent les tables des matières des deux grammaires[67]).

Documenter un texte selon les règles normatives du français écrit ne conduit donc pas nécessairement à une banque universelle d'indices de maîtrise. Toutefois, si les grammaires prescriptives ne peuvent pas réellement constituer une base d'observations fonctionnelles en raison de leur diversité d'approche, les signaux

---

[67] *Éléments de la langue, La Proposition, Les Parties du discours, Les Propositions subordonnées* pour Grevisse (1980 : 1513-1518); *Préliminaires, Le Substantif et ses déterminants, L'Adjectif qualificatif, Les Pronoms, Le Verbe, Les Adverbes, Les Conjonctions de coordination, Les Prépositions, La Phrase* chez Wagner et Pinchon (1991 : 684-687).

d'application de ces règles dans des textes experts, eux, le peuvent bien davantage. De tels signaux peuvent être inférés à partir d'une grille de relecture corrective comme celle de Guénette, Lépine et Roy (1995)[68].

Cependant, un texte peut-il être décrit de façon discriminante à l'aide de critères de maîtrise linguistique ? Nous défendons l'idée que oui : le nombre et la nature des indices de maîtrise linguistique peuvent permettre de distinguer aussi efficacement les rédacteurs faibles des autres que la nature et le nombre de fautes de français.

Il demeure néanmoins que le processus traditionnel de révision de textes est largement fondé sur la détection des erreurs. Les enseignants n'ont pas, par exemple, le réflexe de signaler les marques de maîtrise linguistique mais celles de l'absence de maîtrise. Un professeur détectera et commentera les fautes de ses étudiants, mais se contentera d'une remarque encourageante pour un texte bien rédigé. De la même façon, les correcteurs grammaticaux sont programmés pour détecter les erreurs et non les indices de maîtrise. En fait, l'utilité même du concept d'indice de maîtrise n'apparaît nulle part dans les grammaires normatives du français, bien que celui de la norme y soit très présent.

Nous sommes d'avis que les marques d'indice de maîtrise linguistique constituent en fait un aspect capital de la correction automatique de textes. Jusqu'à présent, les correcteurs orthographiques ont exploité un système tout azimut de révision : le texte d'un professionnel de l'écriture peu susceptible de contenir des erreurs, comme nous venons de le démontrer, sera révisé de la même façon que celui d'un rédacteur inefficace. Plus même. Comme les dictionnaires des correcteurs sont composés d'un lexique de base, les mots plus recherchés seront signalés comme possiblement erronés[69] : au rédacteur de les introduire lui-même dans son dictionnaire personnel. En ce sens, les rédacteurs professionnels sont aussi

---

[68] Voir Chapitre 3 *Méthodologie*.

pauvrement desservis par les correcteurs actuellement disponibles que les rédacteurs occasionnels : dans un cas, le texte comprendra de nombreuses fausses détections, autant de « bruits » irritants; dans l'autre cas, des détections manquantes empêcheront le repérage de failles linguistiques parfois importantes.

Notre recherche voulait répondre à la question fondamentale suivante:

* Existe-t-il des indices de maîtrise linguistique automatiquement identifiables dans les textes experts ?

Cette question s'accompagne à son tour de quatre questions corollaires:

* Ces indices de maîtrise se présentent-ils en nombre significatif par rapport aux textes non experts?
* Quels sont les indices de maîtrise les plus caractéristiques du niveau de contrôle linguistique du rédacteur?
* La reconnaissance des indices de maîtrise les plus discriminants requiert-elle une analyse syntaxique complexe?
* Les indices discriminants se retrouvent-ils dans une portion significative de notre corpus expert?

L'étude de notre corpus confirme la présence significative de signes de maîtrise linguistique discriminants repérables de façon automatique dans les textes d'expression française.

6.1 Performance « Maîtrise » des sujets tous corpus confondus

Les sujets experts ont réalisé la très grande majorité (785 : 86%) des indices de maîtrise repérés dans l'ensemble de notre corpus (Tableau 53).

---

[69] En fait, les correcteurs orthographiques repèrent tous les mots ne faisant pas partie de leur dictionnaire. C'est pourquoi ils peuvent signaler les fautes d'orthographe. Cependant, ils signalent du même coup tous les noms propres et les mots rares ou considérés comme tels par leurs développeurs. Par exemple, le correcteur de notre texteur — Word98 —, a indiqué le mot *prescriptive*, que nous avons employé dans le paragraphe précédent, comme possiblement erroné. Il signale également le mot *texteur* paraissant dans la phrase précédente (et dans celle-ci).

Tableau 53

Sommaire du corpus sur le plan de la maîtrise linguistique

Corpus	Indices de maîtrise
Moffet	127
Experts	390
Bissonnette	395
Total Maîtrise	912

Les textes des sujets experts et ceux de Lise Bissonnette comportent un nombre à peu près semblable d'indices de maîtrise. Voilà qui est intéressant, surtout parce que le corpus « Experts » rassemble des textes rédigés par 75 auteurs différents alors que le corpus Bissonnette réunit 75 textes rédigés par une seule et même personne. Le nombre très voisin d'indices de maîtrise repérés dans les deux corpus donne par conséquent à penser que 1) la constance dans le nombre d'indices de maîtrise dans une variété de textes rédigés par un seul auteur peut constituer à elle seule un indice de maîtrise[70] et 2) un seul texte — pourvu qu'il comporte un nombre suffisant de mots (par exemple, 1 0 00 mots) — peut déjà donner un bon indice de la qualité de la production linguistique d'un rédacteur.

6.1.1 Distribution des indices de maîtrise par catégorie

Notre grille « Maîtrise » répartit les indices de maîtrise en 5 catégories[71]. La figure 68 distribue les indices de maîtrise du corpus selon ces catégories.

---

70 Nous n'avons cependant pas pu utiliser la constance comme indice de maîtrise parce que nous ne disposions que d'un seul texte pour chacun des rédacteurs occasionnels du corpus Moffet.

71 Voir Chapitre 4. *Grille de calibrage.*

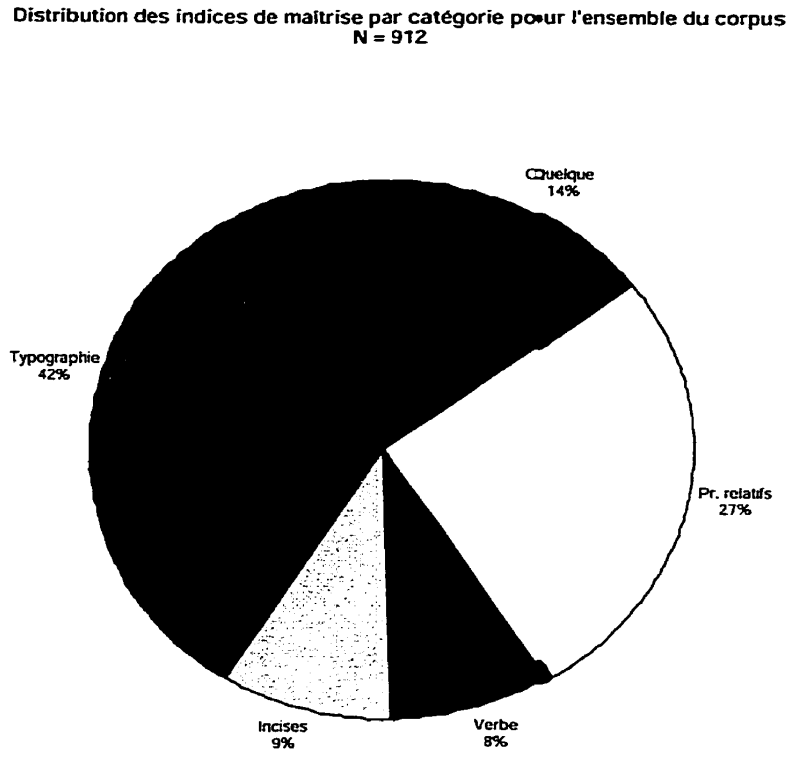


Figure 68

Distribution des indices de maîtrise par catégorie pour l'ensemble du corpus

Les indices typographiques et l'emploi des pronoms relatifs constituent près de 70% des indices de maîtrise identifiés dans l'ensemble du corpus.

6.1.2 Distribution des indices de maîtrise par fréquence

Quels sont les indices de maîtrise les plus souvent repérés dans notre corpus? Le tableau 54 liste ces variables en ordre décroissant, tous groupes de sujets confondus. Précisons que, pour chaque élément de cette liste, le nombre maximal d'occurrences possibles est de 225, c'est-à-dire trois fois 75 textes. Le nombre maximal total possible d'occurrences d'indices de maîtrise pour l'ensemble du corpus est par conséquent de 4 0 50 (18 occurrences x 225 textes).

Tableau 54

Liste des indices de maîtrise pour l'ensemble du corpus

Variable	*Total indices de maîtrise*	*Pourcentage (N = 225)*
Dont	140	62%
Parenthèses intercalaires	137	61%
Tirets intercalaires	90	40%
Quel et ses composés	89	40%
Deux-points suivi d'une explication	86	38%
Incises	84	37%
Quelque au sens de plusieurs	79	35%
Si suivi de l'imparfait	39	17%
Formes rares du subjonctif	38	17%
Signe correct d'effacement de passage dans une citation	36	16%
Quelque au sens de un certain	30	13%
Qui ou *Quoi* précédé d'une préposition ou d'une locution prépositive	15	7%
Quelque au sens de environ	11	5%
Crochets de parenthétisation	10	4%
Crochets indiquant une modification dans une citation	9	4%
Quel ... que suivi d'un verbe d'état au subjonctif	8	4%
Énumération en colonne avec puces ou tirets	8	4%
Quelque ... que suivi du subjonctif	3	1%
Total	**912**	**23%**[72]

Les emplois de *dont* et des parenthèses intercalaires représentent les deux indices les plus fréquemment observés dans l'ensemble du corpus. En revanche les suites *quel...que* et *quelque...que* n'ont été repérées que rarement (moins de 5% de toutes les occurrences possibles pour ces suites). Les énumérations en colonne avec

[72] Total / 4 050.

puces ou tirets constituent également une réalisation peu fréquente dans notre corpus.

6.2 Performance « Maîtrise » des sujets par corpus

Le graphique de la figure 69 distribue les indices de maîtrise par groupes de sujets et par catégories.

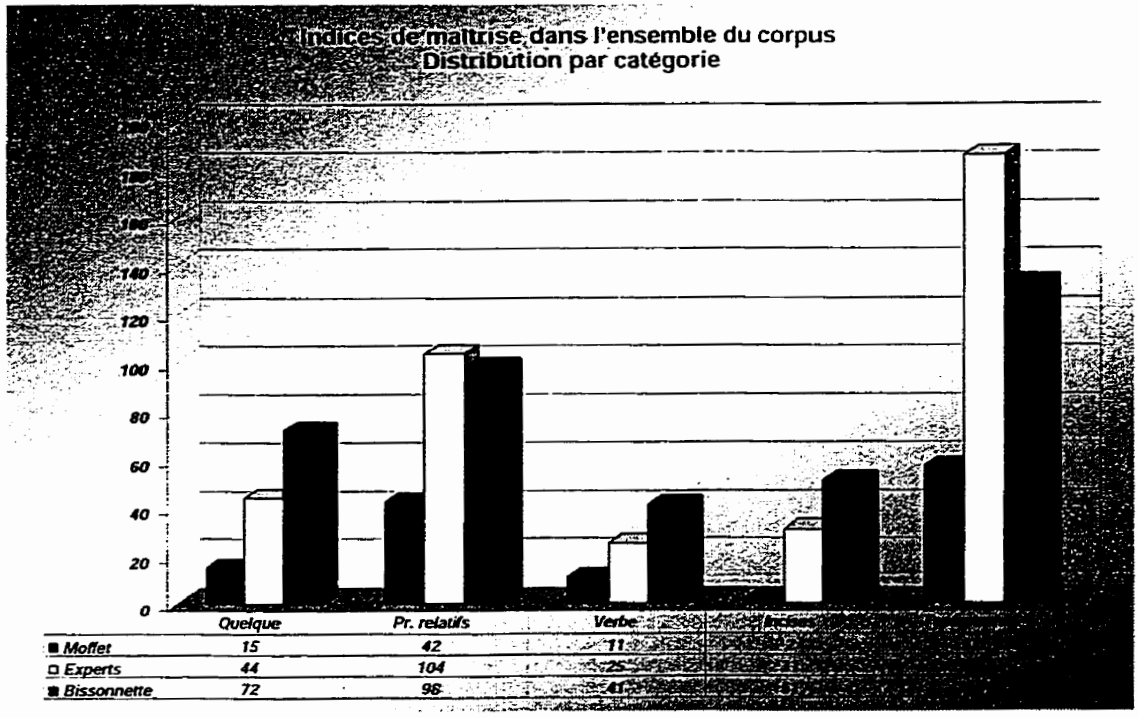


Figure 69

Distribution des indices de maîtrise par corpus

La figure 69 fait ressortir la représentation importante des indices typographiques (376 occurrences) et des pronoms relatifs (244 occurrences) dans l'ensemble du corpus. Ce graphique fait également ressortir que, dans toutes les catégories, les textes du corpus Moffet comportent un nombre beaucoup moins élevé d'indices de maîtrise que ceux des corpus « Experts » et Bissonnette.

6.2.1 Fréquences relatives des catégories d'indices de maîtrise dans le corpus Moffet

Le graphique 70 fait ressortir la distribution des indices de maîtrise dans le corpus Moffet.

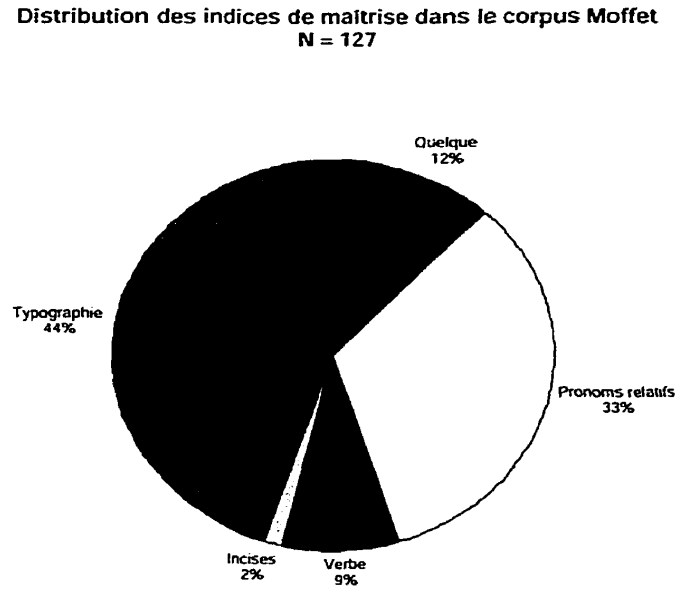


Figure 70

Distribution des indices de maîtrise dans le corpus Moffet

Le corpus Moffet présente une distribution semblable à l'ensemble du corpus pour deux des catégories d'indices. Les indices typographiques et les pronoms relatifs constituent en effet la majorité (77%) des indices de maîtrise observés chez les sujets non experts. Les incises, avec seulement 2%, représentent une tournure exceptionnelle dans le corpus Moffet.

6.2.2 Fréquences relatives des catégories d'indices de maîtrise dans le corpus « Experts »

Le graphique 71 illustre la performance des sujets experts.

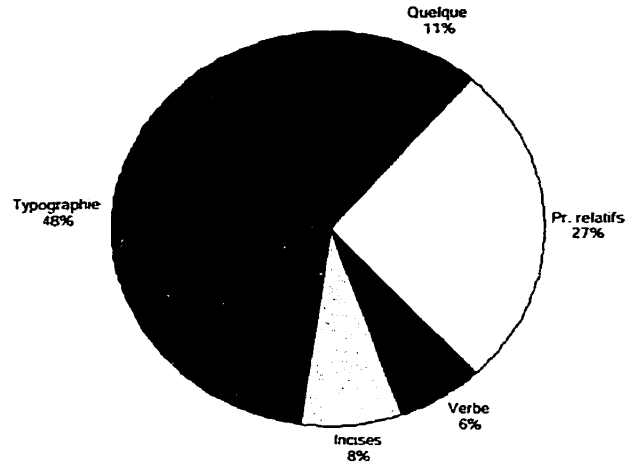


Figure 71

Distribution des indices de maîtrise dans le corpus "Experts

Les indices de maîtrise repérés dans le corpus « Experts » se distribuent de façon similaire à celle du corpus « Moffet » : une majorité d'indices dans les catégories « Typographie » et « Pronoms relatifs » (75%) et une répartition semblable des autres indices à l'exception des incises qui apparaissent en proportion plus importante dans les textes experts.

6.3.3 Fréquences relatives des catégories d'indices de maîtrise dans le corpus Bissonnette

La figure 72 montre la répartition des indices de maîtrise dans les textes de Lise Bissonnette.

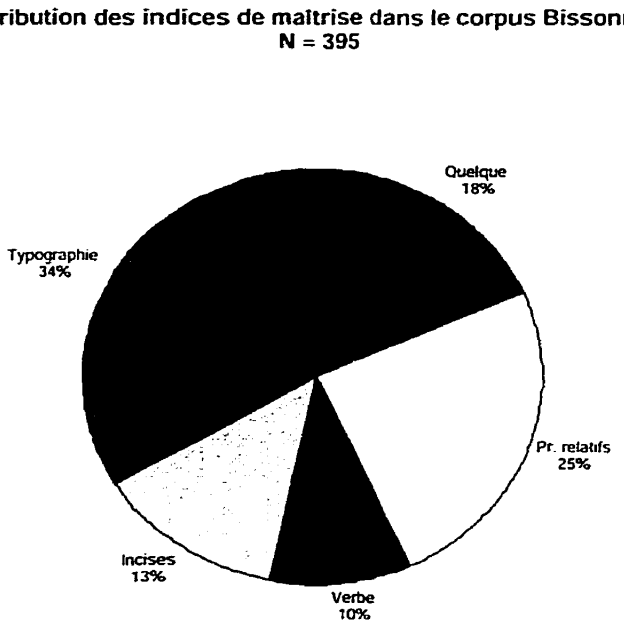


Figure 72

Distribution des indices de maîtrise dans le corpus Bissonnette

Bien que la représentation des indices typographiques et des pronoms relatifs soient toujours majoritaire (59%) dans le bloc d'indices de maîtrise identifiés dans les textes Bissonnette, celle des catégories « Incises » et « Quelque » augmente en importance.

6.3 Performance « Maîtrise » des sujets par catégories

L'étude des occurrences[73] des signaux de maîtrise du corpus permet de présenter leur distribution selon leurs catégories. Cette fois-ci, les textes Bissonnette font partie de la comparaison.

Dans les tableaux d'occurrences, les données sont classées en ordre décroissant d'après les résultats des sujets du corpus « Experts ».

---

73 Il importe de rappeler qu'en raison du format logique que nous avons adopté pour nos données, comptabiliser une seule occurrence par texte — quand elle se présente — revient en fait à déterminer le nombre de textes ayant présenté cette occurrence dans le corpus.

6.3.1 Catégorie « Typographie »

Notre corpus compte 376 indices typographiques. Le tableau 55 les distribue selon les trois groupes de textes.

Tableau 55

Nombre d'indices de maîtrise dans la catégorie "Typographie"

Indices de maîtrise	Moffet	Experts	Bissonnette
Parenthèses intercalaires	23	67	47
Deux-points suivi d'une explication	1	47	38
Tirets intercalaires	2	46	42
Crochets de parenthétisation	0	10	0
Signe correct d'effacement de passage dans une citation	26	6	4
Crochets indiquant une modification dans une citation	4	5	0
Énumération en colonne avec puces ou tirets	1	5	2
TYPOGRAPHIE	**57**	**186**	**133**

Le corpus « Experts » compte trois fois plus de marqueurs typographiques que le corpus Moffet; le corpus Bissonnette, deux fois plus.

Cependant, le tableau 55 fait ressortir trois autres éléments intéressants :

* Les parenthèses intercalaires sont les marqueurs typographiques les plus représentés dans l'ensemble du corpus bien que leur nombre va jusqu'à presque tripler chez les rédacteurs experts.

* Les deux-points introduisant une explication (par opposition aux deux-points introduisant une énumération ou une citation) de même que les tirets intercalaires sont fréquents chez les experts mais exceptionnels, sinon pratiquement absents, chez les non-experts.

* Les données relatives à la représentation du signe d'effacement de passages dans une citation (points de suspension encadrés par des crochets) différencient les types de textes argumentatifs de notre corpus, les sujets Moffet rédigeant une dissertation critique souvent enrichie de citations tirées des textes analysés et les sujets experts rédigeant plutôt des chroniques, des lettres d'opinion ou des éditoriaux où les citations sont rares.

Le tableau 56 introduit la proportion de textes du corpus présentant les indices de maîtrise linguistique de la catégorie « Typographie ».

Tableau 56

Fréquences relatives de l'indice « Typographie » par corpus

Indices de maîtrise	Moffet % N=75	Experts % N=75	Bissonnette % N=75
Parenthèses intercalaires	31	89	63
Deux-points suivi d'une explication	1	63	51
Tirets intercalaires	3	61	56
Crochets de parenthétisation	0	13	0
Signe correct d'effacement de passage dans une citation	35	8	5
Crochets indiquant une modification dans une citation	5	7	0
Énumération en colonne avec puces ou tirets	1	7	3

Dans notre liste des indices de maîtrise typographiques, seulement trois marqueurs paraissent réellement discriminants :

* la fréquence des parenthèses intercalaires;
* la présence des deux-points suivis d'une explication;
* l'occurrence de tirets intercalaires.

Les autres indices, particulièrement le signe d'effacement de passage dans une citation, sont trop dépendants de catégories particulières de textes pour être fiables dans un module de calibrage. Quant aux crochets de parenthétisation, ils sont absents aussi bien dans les textes Moffet que les textes Bissonnette. Leur valeur discriminante est donc nulle.

La reconnaissance des deux-points explicatifs pourraient requérir une analyse syntaxique complexe. En effet, même si les autres emplois des deux-points pourraient se reconnaître sans analyse (une citation s'introduit, soit par une alternance de police

de caractères, soit par des guillemets; une énumération se marque par une série de syntagmes généralement courts séparés par des virgules), il existe quand même des chances d'erreur. Par conséquent, seules les occurrences de parenthèses et de tirets intercalaires seront considérées comme des indices typographiques de maîtrise simples.

6.3.2 Catégorie « Quelque »

Notre corpus de recherche réunit 131 occurrences de *quelque* (Tableau 57).

Tableau 57

Nombre d'indices de maîtrise dans la catégorie "Quelque"

Indices de maîtrise	Moffet	Experts	Bissonnette
Quelque au sens de plusieurs	14	31	34
Quelque au sens de environ	1	5	5
Quel ... que suivi d'un verbe d'état au subjonctif	0	5	3
Quelque au sens de un certain	0	3	27
Quelque ... que suivi du subjonctif	0	0	3
QUELQUE	**15**	**44**	**72**

Le seul emploi de *quelque* repéré dans le corpus Moffet est le déterminant au sens de *plusieurs*. Cet emploi apparaît toutefois dans deux fois plus de textes experts que de textes non experts. En revanche, la tournure *quelque...que* est pratiquement absente de notre corpus.

Quant aux autres emplois, les textes experts les présente en nombre restreint. Une exception cependant : l'emploi du déterminant *quelque* au sens de *un certain*, dont l'occurrence est significativement plus élevée dans le corpus Bissonnette. Nous nous trouvons peut-être en présence de l'un des emplois préférés de l'auteure. Dans notre corpus de recherche, Lise Bissonnette réalise presque une occurrence de

quelque par texte, soit quatre fois plus que les sujets Moffet et une fois et demie de plus que les sujets du corpus « Experts ».

Le tableau 58 affiche le pourcentage de textes par corpus présentant chaque emploi identifié de *quelque*.

Tableau 58

Fréquences relatives de l'indice « Quelque» par corpus

Indices de maîtrise	Moffet % N = 75	Experts % N = 75	Bissonnette % N = 75
Quelque au sens de plusieurs	19	41	45
Quelque au sens de environ	1	7	7
Quel ... que suivi d'un verbe d'état au subjonctif	0	7	4
Quelque au sens de un certain	0	4	36
Quelque ... que suivi du subjonctif	0	0	4

La répartition des occurrences de *quelque* en pourcentages montre deux types d'indices discriminants pour cette catégorie :

* la fréquence d'emploi du déterminant *quelque* au sens de *plusieurs;*
* la présence, dans un texte, de n'importe quel autre emploi de *quelque* à l'exception de *quelque...que.*

Cependant, le déterminant *quelque* au sens de *un certain* ne peut pas constituer un indice discriminant acceptable aux fins de calibrage. En effet, un module de calibrage pourrait toujours repérer une occurrence de *quelque* au singulier dans le texte d'un rédacteur occasionnel si celui-ci, utilisant le déterminant au sens de *plusieurs*, négligeait la marque du pluriel. Par conséquent, l'emploi de

quelque au sens de *un certain* n'est discriminant que dans la mesure où le profil linguistique du rédacteur est déjà établi[74].

Le repérage automatique du mot *quelque* peut s'effectuer simplement sans recours à une analyse syntaxique complexe. Cependant, l'interprétation des sens de *quelque* s'avère plus compliqué. C'est pourquoi nous considérerons *quelque* comme un indice complexe.

6.3.3 Catégorie « Pronoms relatifs »

Notre corpus contient 244 occurrences de pronoms relatifs autres que *qui* ou *que*. Le tableau 59 montre la distribution par corpus de ces occurrences selon les catégories d'emploi.

Tableau 59

Nombre d'indices de maîtrise dans la catégorie « Pronoms relatifs»

Indices de maîtrise	Moffet	Experts	Bissonnette
Dont	22	54	64
Quel et ses composés	17	40	32
Qui ou *Quoi* précédé d'une préposition ou d'une locution prépositive	3	10	2
PRONOMS RELATIFS	**42**	**104**	**98**

Le nombre de pronoms relatifs est élevé dans notre corpus. Les emplois de *dont* et des composés de *quel* sont plus nombreux chez les experts que chez les non-experts. Toutefois les pronoms relatifs *qui* ou *quoi* composés avec une préposition se retrouvent en nombre égal dans les textes Moffet et Bissonnette, ce qui invalide la qualité discriminante de cet emploi.

---

[74] Nous avons vu, dans le chapitre 5. *Indice de faiblesse linguistique*, que le passé simple représentait un autre exemple d'emploi dont l'intégrité n'est confirmée qu'après l'établissement du profil linguistique du rédacteur.

Le tableau 60 présente le pourcentage de textes par corpus avec des pronoms relatifs autres que *qui* ou *quoi.*

Tableau 60

Fréquences relatives de l'indice « Pronoms relatifs » par corpus

Indices de maîtrise	Moffet % N=75	Experts % N=75	Bissonnette % N=75
Dont	29	72	85
Quel et ses composés	23	53	43
Qui ou *Quoi* précédé d'une préposition ou d'une locution prépositive	4	13	3

Les fréquences d'apparition de *dont* et des composés de *quel* sont très élevées dans les textes experts, plus du double que chez les non-experts. Ces deux indices constituent donc des indices discriminants utiles, d'autant plus que leur reconnaissance exclut le recours à une analyse syntaxique complexe.

6.3.4 Catégorie « Verbe »

Nous n'avions retenu que deux indices de maîtrise dans la catégorie « Verbe » : l'emploi de l'imparfait avec le *si* hypothétique et la présence de formes rares du subjonctif.

Notre corpus comprend 77 indices de maîtrise verbale. Le tableau 61 répartit ces indices par groupes de sujets.

Tableau 61

Nombre d'indices de maîtrise dans la catégorie "Verbe"

Indices de maîtrise	Moffet	Experts	Bissonnette
Si suivi de l'imparfait	9	19	11
Formes rares du subjonctif	2	6	30
VERBE	**11**	**25**	**41**

Les textes Bissonnette et Moffet présentent un nombre comparable d'occurrences de l'imparfait avec le *si* hypothétique. En revanche, les formes rares du subjonctif sont exceptionnelles dans le corpus Moffet, rares chez les experts mais relativement fréquentes chez Bissonnette.

Le tableau 62 présente la représentation des indices du verbe par pourcentages.

Tableau 62

Fréquences relatives des indices de maîtrise verbaux par corpus

Indices de maîtrise	Moffet % N = 75	Experts % N = 75	Bissonnette % N = 75
Si suivi de l'imparfait	12	25	15
Formes rares du subjonctif	3	8	40

Les indices de maîtrise de la catégorie « Verbe » ne représentent pas un potentiel discriminatoire réellement intéressant.

En fait, l'imparfait avec le *si* hypothétique ne comporte aucun élément discriminant puisque l'un de nos deux corpus experts ne présente pas d'occurrence en proportion significativement supérieure à celle du corpus Moffet.

En outre, les formes rares du subjonctif, bien qu'en proportion deux fois plus importante chez les sujets experts, représente moins de 10% du corpus « Experts ». Il est vrai que près d'un texte du corpus étalon sur deux comporte au moins une réalisation d'un subjonctif rare, ce qui revient à peut-être signaler la possibilité que la présence de cette variable pourrait être associable à un niveau possiblement plus élevé de maîtrise[75].

---

75 Nous n'avons pas fait de recherche spécifique pour confirmer cette hypothèse. La fréquence élevée de subjonctifs rares dans les textes Bissonnette peut ne représenter qu'un autre trait signature de l'auteure.

6.3.5 Catégorie « Incises »

Notre corpus présente 84 incises (Tableau 63).

Tableau 63

Nombre d'indices de maîtrise dans la catégorie "Incises"

Indices de maîtrise	Moffet	Experts	Bissonnette
Incises	2	31	51
INCISES	**2**	**31**	**51**

La présence d'incises constitue l'un des indices de maîtrise les plus discriminants de notre liste avec une quasi-absence dans les textes Moffet et une forte occurrence dans les textes des corpus « Experts » et Bissonnette.

Le tableau 64 montre la proportion du corpus de recherche avec des incises.

Tableau 64

Fréquences relatives des incises par corpus

Indices de maîtrise	Moffet %	Experts %	Bissonnette %
Incises	3	41	68

La proportion des occurrences d'incise est significativement plus importante chez les rédacteurs professionnels que non professionnels de notre corpus. En effet, près de 7 textes du corpus étalon sur 10 présente au moins une incise. De la même façon, 4 textes experts sur 10 en contiennent une. Par contre, les incises n'apparaissent que de façon très exceptionnelle chez les sujets Moffet.

Toutefois, malgré leur pouvoir discriminant, les incises ne se reconnaissent automatiquement qu'au moyen d'une analyse syntaxique assez puissante pour repérer les tournures syntaxiques souvent complexes qui les caractérisent.

6.4 Étude statistique

6.4.1 Étude de moyennes

Une étude des moyennes nous permet de dresser le portrait possible d'un texte type sur le plan des indices de maîtrise.

La différence entre les sujets de l'échantillon Moffet et ceux du corpus « Experts » est importante : les rédacteurs non experts ont réalisé en moyenne 2 indices de maîtrise par texte contre 5 pour les rédacteurs experts, c'est-à-dire 2 1/2 fois plus. Le tableau 65 présente la distribution de ces moyennes pour l'ensemble du corpus.

Tableau 65

Moyennes des indices de maîtrise par corpus

Corpus	Moyenne	Écart-type
Moffet	1, 69	1, 17
Experts	5, 20	1, 76
Bissonnette	5, 27	1, 64

Les groupes de sujets sont homogènes. La mesure de dispersion autour des moyennes est en effet peu élevée: 1 chez les non-experts et 2 chez les rédacteurs experts. Nous devons cependant valider ces mesures par la recherche de données extraordinaires risquant d'influencer indûment nos résultats (Cohen, 1995 : 27).

La figure 73 fait état de la dispersion des observations dans le corpus Moffet.

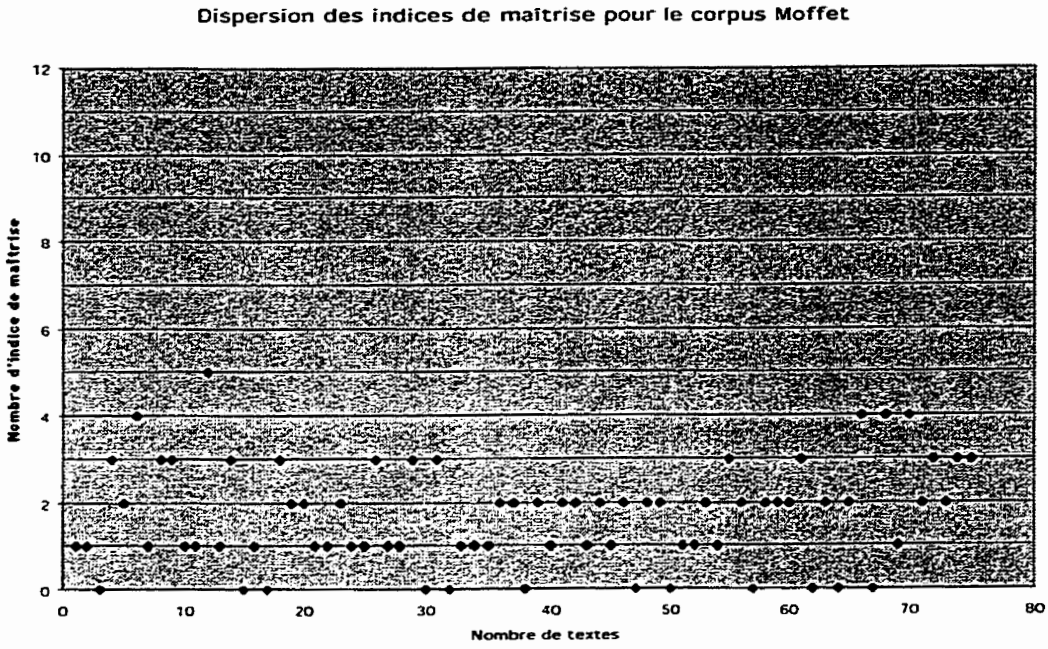


Figure 73

Dispersion des indices de maîtrise pour le corpus Moffet

Les données du corpus Moffet ne montrent pas de distances extraordinaires entre les points, ce qui valide la moyenne des indices de maîtrise par texte.

Les figures 72 et 73 situent les indices de maîtrise par texte pour les corpus « Experts » et Bissonnette.

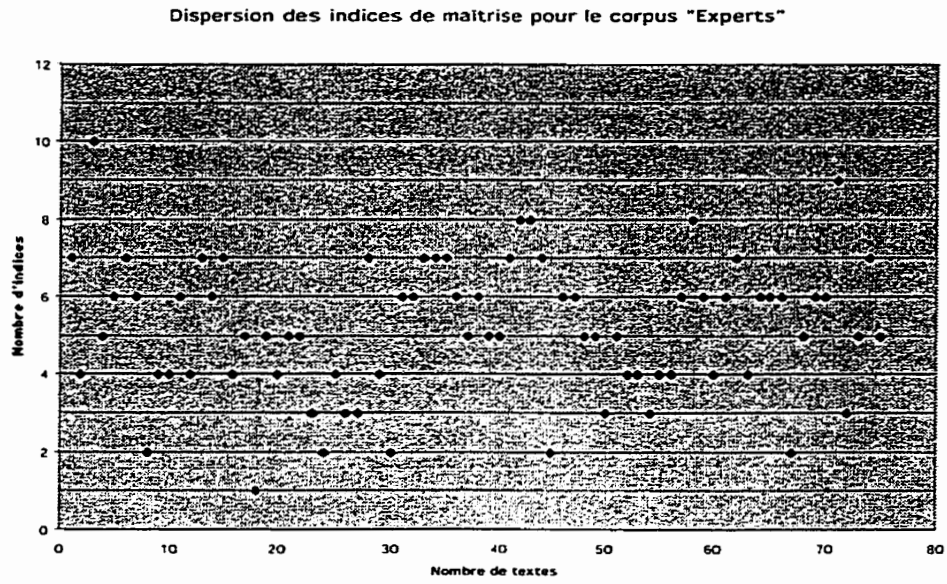


Figure 74

Dispersion des indices de maîtrise pour le corpus « Experts »

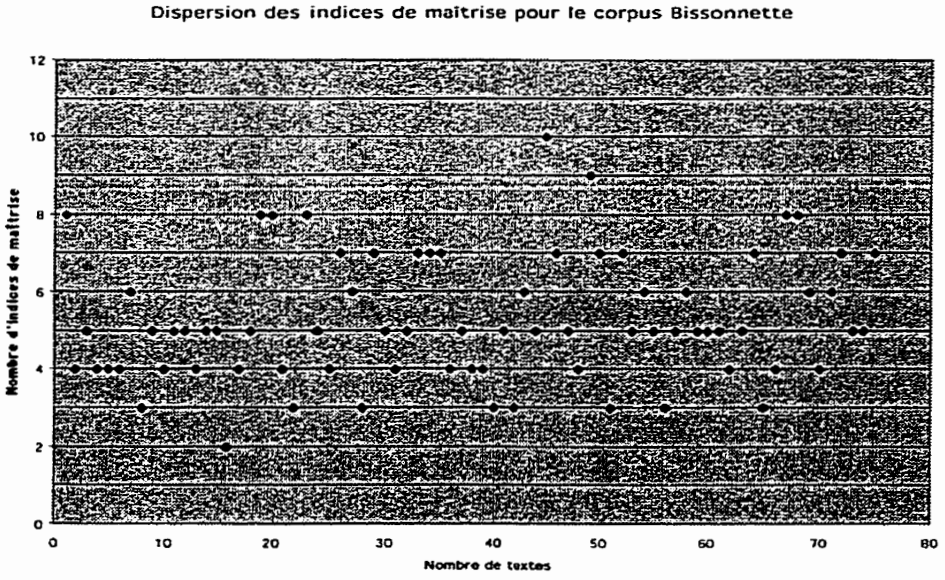


Figure 75

Dispersion des indices de maîtrise pour le corpus Bissonnette

Les corpus « Experts » et Bissonnette ne présentent pas d'observations extraordinaires selon la définition de Cohen (*loc. cit.*).

Nos moyennes validées, nous pouvons maintenant nous arrêter à l'étude des écarts-types.

6.4.2 Étude des écarts-types

Le tableau 66 présente les résultats des calculs des écarts-types[76] à partir des moyennes d'indices de maîtrise pour nos trois groupes de sujets. Les chiffres sont arrondis pour rendre compte de la nature discrète[77] des variables. Encore une fois, les valeurs négatives sont remplacées par 0.

Tableau 66

Dispersion de la population d'indices de maîtrise autour de la moyenne

Corpus	-3σ	-2σ	-1σ	μ	+1σ	+2σ	+3σ
Moffet (σ = 1, 17)	0	0	1	2	3	4	5
Experts (σ = 1, 76)	0	2	3	5	7	9	10
Bissonnette (σ = 1, 64)	0	2	3	5	7	9	10

Le tableau 66 fait ressortir deux éléments importants:

* 99,73% des sujets Moffet pourront réaliser de 0 à 4 indices de maîtrise alors que 99,73% des sujets experts en réaliseront de 2 à 9.

* Les valeurs extrêmes égales à + ou - 3 σ (0 indice de maîtrise pour les textes des corpus « Experts » et Bissonnette; 5 indices de maîtrise pour les textes de l'échantillon Moffet) sont improbables dans chacune des populations (Allaire, *loc. cit.*).

---

76 L'écart-type mesure la dispersion d'une population autour de la moyenne. En utilisant l'écart-type (symbole σ) comme base de calcul, il est possible d'estimer les résultats d'une population, assumant qu'ils sont distribués selon la courbe normale (Cohen, 1995 : 122; Allaire,1998 : 12-1): la moyenne (symbole μ^{76}) constituant le centre de la courbe normale, + 1σ et - 1σ décriront les résultats de 68, 27% de la population; + 2σ et - 2σ, de 95, 45% de la population et + 3σ et - 3σ., de 99, 73% de la population. Allaire (*loc.cit.)* ajoute que *des valeurs situées à trois écarts-types (au-dessous ou au-dessus) de la moyenne sont relativement rares dans une distribution normale.*

77 Voir chapitre 5. *Indice de faiblesse linguistique.*

Nous pouvons donc conclure que ...

* si 5 indices de maîtrise constituent un nombre moyen chez les experts, il représentera une occurrence rarissime chez les sujets non experts;

* l'absence d'indices de maîtrise sera exceptionnelle chez les experts mais probable chez les non-experts;

* au moins deux zones partagées entre experts et non-experts existent dans les deux groupes de textes, mais pour une portion différente de la distribution normale.

Ces conclusions sont vérifiables dans notre corpus comme le montrent les figures 76, 77 et 78.

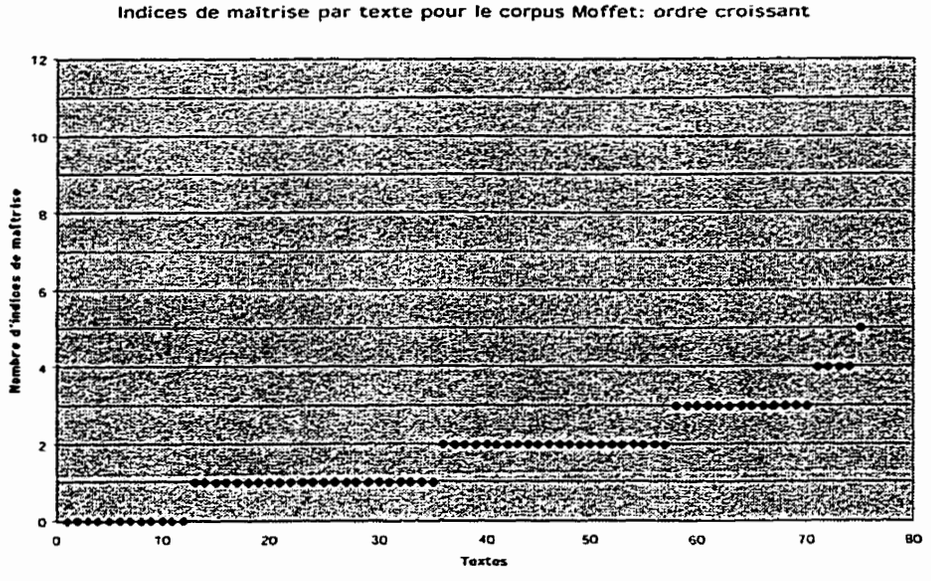


Figure 76

Indices de maîtrise par texte dans le corpus Moffet: ordre croissant

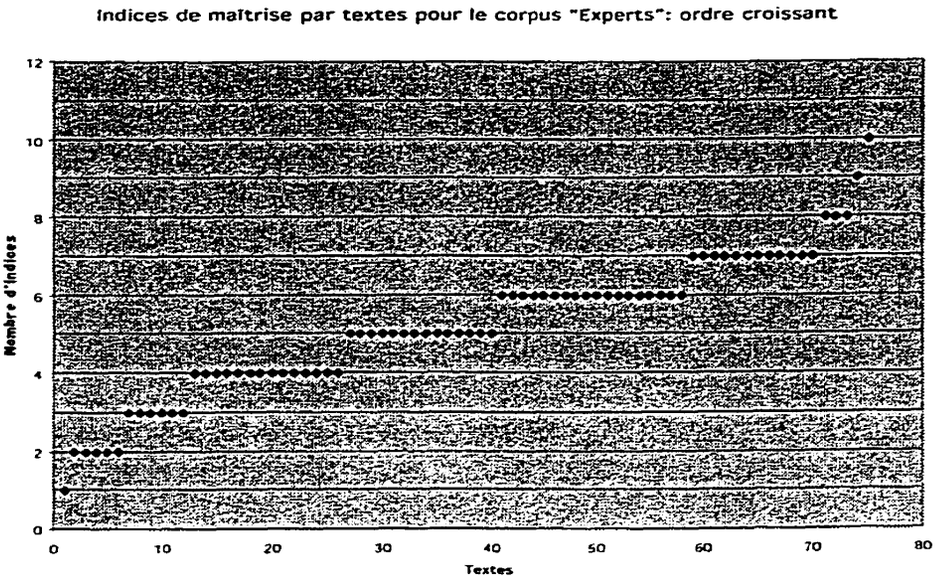


Figure 77

Indices de maîtrise par texte dans le corpus « Experts » : ordre croissant

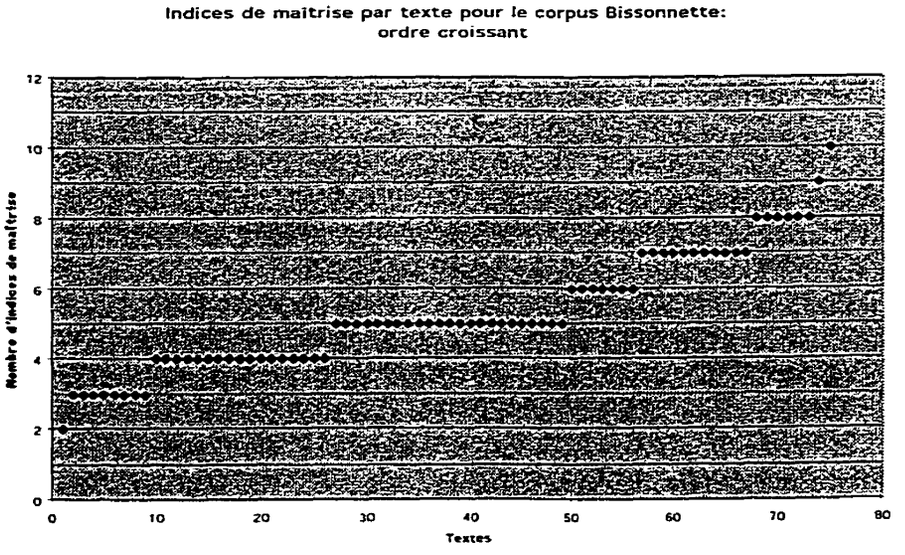


Figure 78

Indices de maîtrise par texte dans le corpus Bissonnette: ordre croissant

Les figures 76, 77 et 78 montrent en outre

* que les groupes sont homogènes;
* qu'il existe une différence significative entre les sujets experts et non experts sur le plan des indices de maîtrise;
* qu'il n'existe pas de différence notable entre les textes experts et Bissonnette.

Les zones ambiguës représentées par les valeurs à chacune des extrémités de la courbe normale chez les rédacteurs occasionnels appelleront une méthode de désambiguïsation pour permettre le calibrage des textes tombant dans cette catégorie.

6.5 Indice de maîtrise linguistique

Un indice de maîtrise linguistique peut être calculé à partir de recherches ciblées d'occurrences prioritaires.

6.5.1 Indices de maîtrise « simples »

Le tableau de la page suivante (Tableau 67) présente la synthèse de l'examen de nos indices de maîtrise. Il montre la proportion de textes du corpus affichant chacune des variables de notre grille « Maitrise » en les divisant par groupes de sujets. Il ajoute cependant le critère du recours à une analyse syntaxique complexe[78].

Nous créons ainsi deux blocs d'indices de maîtrise : les indices identifiables au moyen d'une analyse syntaxique complexe et les autres. Encore une fois, nous placerons en priorité les indices détectables sans analyse syntaxique élaborée.

Les données sont classées au moyen de deux clés de tri : la colonne « Analyse complexe requise»; la colonne « Experts ».

---

78 Voir chapitre 5. *Indices de faiblesse linguistique.*

Tableau 67

Synthèse des fréquences relatives de la grille "Maîtrise" selon les corpus

Indices de maîtrise	Moffet %	Experts %	Bissonnette %	Analyse complexe requise
Parenthèses intercalaires	31	89	63	
Dont	29	72	85	
Tirets intercalaires	3	61	56	
Quel et ses composés	23	53	43	
Si suivi de l'imparfait	12	25	15	
Crochets de parenthétisation	0	13	0	
Qui ou Quoi précédé d'une préposition ou d'une locution prépositive	4	13	3	
Signe correct d'effacement de passage dans une citation	35	8	5	
Formes rares du subjonctif	3	8	40	
Crochets indiquant une modification dans une citation	5	7	0	
Énumération en colonne avec puces ou tirets	1	7	3	
Quel ... que suivi d'un verbe d'état au subjonctif	0	7	4	
Quelque ... que suivi du subjonctif	0	0	4	
Deux-points suivi d'une explication	1	63	51	•
Quelque au sens de *plusieurs*	19	41	45	•
Quelque au sens de *environ*	1	7	7	•
Quelque au sens de *un certain*	0	4	36	•
Incises	3	41	68	•

Le tableau 67 présente tous les indices étudiés au cours de notre recherche. Cependant, nous venons d'identifier des variables montrant un pouvoir discriminant faible ou nul: les crochets de parenthétisation, les crochets indiquant une modification dans une citation, *qui* ou *quoi* précédé d'une préposition ou d'une locution prépositive, le signe d'effacement de passage dans une citation, *si* hypothétique suivi de l'imparfait, *Quelque ... que* suivi du subjonctif et, finalement, l'énumération en colonnes avec puces ou tirets. Ces variables doivent donc être évacuées avant de poursuivre notre analyse. Le tableau 68 montre la liste des indices de maîtrise ayant montré un pouvoir discriminant exploitable.

Tableau 68

Liste des indices de maîtrise avec pouvoir discriminant

Indices de maîtrise	Moffet %	Experts %	Bissonnette %	Analyse complexe requise
Parenthèses intercalaires	31	89	63	
Dont	29	72	85	
Tirets intercalaires	3	61	56	
Quel et ses composés	23	53	43	
Formes rares du subjonctif	3	8	40	
Quel ... que suivi d'un verbe d'état au subjonctif	0	7	4	
Deux-points suivi d'une explication	1	63	51	•
Quelque au sens de *plusieurs*	19	41	45	•
Incises	3	41	68	•
Quelque au sens de *environ*	1	7	7	•
Quelque au sens de *un certain*	0	4	36	•

Ce classement permet de faire ressortir nos deux groupes de signaux de maîtrise : les indices de maîtrise « simples sans recours à une analyse syntaxique

complexe et les indices de maîtrise « complexes requérant une analyse syntaxique complexe.

6.5.2 Indices de maîtrise discriminants

Nous reclassons à présent la liste des indices de maîtrise simples selon leur pouvoir discriminant mesurée au moyen de deux clés de tri : en ordre décroissant selon les résultats des rédacteurs du corpus "Experts" et en ordre croissant des résultats des sujets Moffet (Tableau 69).

Tableau 69

Liste des indices de maîtrise de type "simple"

Indices de maîtrise	Moffet %	Experts %	Bissonnette %
Parenthèses intercalaires	31	89	63
Dont	29	72	85
Tirets intercalaires	3	61	56
Quel et ses composés	23	53	43
Formes rares du subjonctif	3	8	40
Quel ... que suivi d'un verbe d'état au subjonctif	0	7	4

Les parenthèses intercalaires, *dont*, les tirets intercalaires de même que les composés de *quel* constituent les indices les plus discriminants de cette liste.

Toutefois, plusieurs indices intéressants ne figurent plus dans notre liste. Si la reconnaissance automatique des incises devenait possible par exemple, la détermination automatique du niveau de maîtrise d'un texte serait plus définitive. En effet, nous avons trouvé que les occurrences d'incises, de tirets intercalaires et de deux-points suivis d'une explication dans un même texte constituent des signes assez sûrs que ce texte est de niveau professionnel.

6.5.3 Indices de maîtrise prioritaires

Comme nous l'avions fait pour les erreurs prioritaires, nous allons aussi établir une liste d'indices de maîtrise prioritaires. Cette fois, nos trois critères de tri seront

* son absence ou sa fréquence très basse chez les sujets Moffet;
* son haut potentiel d'occurrence selon les caractéristiques de la langue;
* sa fréquence élevée chez les rédacteurs experts.

Le tableau 70 fait état de ce reclassement et présente les indices selon un ordre prioritaire possible.

Tableau 70

Liste des indices de maîtrise prioritaires

Indices de maîtrise	Priorité	Moffet %	Experts %	Bissonnette %
Tirets intercalaires	2	3	61	56
Parenthèses intercalaires	3	31	89	63
Dont	4	29	72	85
Quel et ses composés	5	23	53	43
Quel ... que suivi d'un verbe d'état au subjonctif	6	0	7	4
Formes rares du subjonctif	7	3	8	40

Le tableau 70 fait état de 6 indices de maîtrise reconnaissables par notre module de calibrage sans recours à une analyse syntaxique complexe. Dans cette liste, les plus discriminants sont les tirets intercalaires et les parenthèses intercalaires suivis de *quel* et ses composés et finalement *dont.* Les autres indices ne sont pas apparus dans une proportion de textes experts très élevée, bien qu'ils demeurent discriminants du fait de leur rareté, ou même de leur absence, dans les textes de rédacteurs occasionnels.

Notre liste d'indice de maîtrise prioritaires permet d'attribuer un indice de maîtrise linguistique au texte examiné. Cet indice s'établira sur 6.

Conclusion

Nous venons de confirmer la différence significative entre la performance des rédacteurs experts et non experts sur le plan de la maîtrise linguistique et d'identifier des variables susceptibles de permettre automatiquement la distinction entre les deux groupes sur le plan des indices de maîtrise.

Pour établir automatiquement un profil linguistique, nous tiendrons donc également compte de l'indice de maîtrise.

Sept
Vers une correction automatique calibrée

Corriger automatiquement la ponctuation dans le texte d'un rédacteur occasionnel semble un exercice voué à l'échec. D'une part, les fautes de ponctuation dans les textes non professionnels ne sont associées à aucun système de règles apparemment prévisibles; d'autre part, les analyseurs syntaxiques ont besoin de la ponctuation pour repérer automatiquement les frontières de phrase et segmenter le texte en unités analysables. Voilà un bon exemple de la proverbiale quadrature du cercle.

Comme nous le verrons dans le présent chapitre, le calibrage constitue, d'après nous, la meilleure solution à ce problème. En effet, non seulement cette solution est-elle originale — elle permet d'éviter les embûches de l'analyse automatique de suites complexes, sources de nombreux problèmes de détection — mais elle est également réalisable dans le contexte informatique actuel : elle s'appuie largement sur l'une des forces de l'ordinateur, le traitement de formules mathématiques.

Le calibrage des correcteurs constitue une voie de recherche originale débouchant sur des perspectives d'application prometteuses en correction automatique de la langue écrite, y compris dans celui de la ponctuation.

7.1 Matrice de calibrage

Ce que nous appelons *matrice de calibrage* est en fait une grille permettant à un correcteur de savoir à quel profil linguistique appartient l'utilisateur.

Le calibrage d'un correcteur est une idée entièrement nouvelle. Jusqu'à présent, les détecteurs de fautes envisagent chaque texte de l'utilisateur indépendamment de ses autres productions et corrigent à l'aveuglette, sans le « connaître ». Mais comme les humains, qui peuvent faire un meilleur travail d'encadrement en connaissant les forces et les faiblesses des personnes qu'ils veulent aider, les correcteurs pourraient mieux réussir à traiter les textes qui leur sont soumis s'ils pouvaient prévoir les forces et les faiblesses linguistiques de ceux qui les rédigent. Ils pourraient alors concentrer leurs efforts de correction sur les contextes « faibles » des utilisateurs et ignorer les autres.

À première vue, cette idée semble impossible à réaliser. Comment, en effet, peut-on « présenter » (au sens humain du terme) un rédacteur à son texteur? Les formules de politesse d'usage ne sont d'aucun secours, bien sûr. Ce dont nous avons besoin, c'est une formule mathématique qui constituerait un bon équivalent. Or c'est justement ce que les indices de faiblesse et de maîtrise linguistiques que nous venons de décrire dans les chapitres précédents nous permettent de développer.

7.1.1 Indices de calibrage

Il existe deux grandes populations de rédacteurs[79] : ceux qui contrôlent leur français écrit et les autres. Il est courant de penser que, pour être considéré faire partie de la première population, il suffit à un rédacteur de ne pas faire de fautes. L'indice de maîtrise, que nous venons de décrire, démontre qu'il n'en est rien. La qualité d'un texte se mesure, certes, par l'absence d'erreurs, mais elle se mesure aussi par la présence d'éléments distinctifs de maîtrise. Les indices de maîtrise linguistique et les indices de faiblesse constituent donc ensemble une banque d'indices de calibrage permettant d'établir le profil linguistique d'un rédacteur.

---

[79] Pour des raisons de cohérence et de clarté, nous devons maintenir, dans ce chapitre, le terme "rédacteur" pour désigner les ensembles de textes catégorisés. Nous sommes conscients que cette décision peut générer une ambiguïté quant à nos intentions réelles, qui ne sont pas, faut-il le rappeler, de jeter du discrédit sur l'intelligence ou les habiletés de quiconque non plus que de défendre l'idée que des personnes soient catégorisables de quelconque façon. Il reste que l'alternative, comme de désigner les classes de textes par des lettres ou des chiffres, si elle apparaît plus neutre et scientifiquement plus appropriée, ne fait rien pour maintenir la clarté du propos quand il s'agit d'élaborer et de documenter une nouvelle approche.

Le tableau 71 réunit nos indices de calibrage[80].

Tableau 71

Banque d'indices de calibrage et ordre de détection

Séquence de repérage	Indices de faiblesse linguistique	Indices de maîtrise linguistique
1	Confusion terminaisons verbales	Tirets intercalaires
2	Répétition abusive de mots	Parenthèses intercalaires
3	Mots manquant de précision	Dont
4	Incohérence dans le choix des pronoms personnels	Quel et ses composés
5	Élément manquant ou incohérence dans l'emploi de connecteurs en série	Quel ... que suivi d'un verbe d'état au subjonctif
6	Termes inutiles ou redondants	Formes rares du subjonctif
7	Confusion de genres	
8	Subjonctif manquant ou inapproprié	
9	Archaïsmes	
10	Cataclysme orthographique	
11	Barbarisme lexical	
12	Si hypothétique suivi du conditionnel	

Il importe de rappeler trois aspects importants de cette grille.

Premièrement, les indices de calibrage sont recherchés par le détecteur dans un ordre de saisie spécifique. Cet ordre a été fixé, on s'en souviendra, à partir de deux critères : le pouvoir discriminant de l'indice et la facilité d'exécution informatique de la détection.

Deuxièmement, les indices de calibrage retenus ne constituent pas tous les indices possibles du français. Il peut en exister d'autres. Ceux dont nous avons noté la présence dans notre corpus sont ceux que nous avions identifiés au terme de notre recherche exploratoire.

---

[80] Voir chapitres 5 *Indice de faiblesse linguistique* et 6 *Indice de maîtrise linguistique.*

Troisièmement, ce que nous appelons *indices de calibrage* sont les indices linguistiques simples, c'est-à-dire les indices dont la détection peut être effectuée automatiquement sans recours à des procédés complexes d'analyse automatique.

7.1.2 Populations de rédacteurs

Notre recherche a permis de décrire statistiquement le nombre et la nature des indices de calibrage que nous pouvons nous attendre à trouver dans chacune des deux populations de rédacteurs.

Distinction selon l'indice de faiblesse linguistique

Les populations de rédacteurs se distinguent statistiquement sur le plan du nombre d'erreurs commises. Le tableau 72 présente le nombre d'erreurs possible selon l'étude des écarts-types (symbole σ) des corpus Moffet, « Experts » et Bissonnette.

Tableau 72

Dispersion des erreurs autour de la moyenne selon la courbe normale

Corpus	*-3σ*	*-2σ*	*-1σ*	*μ*	*+1σ*	*+2σ*	*+3σ*
Moffet (σ = 3)	2	5	8	12	15	19	22
Experts (σ = 1)	0	0	0	1	2	3	4
Bissonnette (σ = 0)	0	0	0	0	0	0	0

Ces données s'interprètent selon la distribution d'une courbe normale. Rappelons que, selon la courbe normale, la moyenne (μ) +1σ et -1σ décrit les résultats de 68, 27% de la population; μ+2σ et -2σ, les résultats de 95, 45% de la population et μ+3σ et -3σ, ceux de 99,73% de la population. Les valeurs situées à +3σ et -3σ sont *relativement rares* (Allaire,1998 : 12-1).

L'échelle d'erreurs décrivant la population de rédacteurs occasionnels se situe entre 3 (>2 erreurs) et 21 (<22 erreurs). Selon la courbe normale, la très grande majorité de cette population (95, 45%) réalisera de 5 à 19 erreurs. Bien que

possibles, les valeurs extrêmes des textes de rédacteurs occasionnels (3 et 4 erreurs; 20 et 21 erreurs) se présentent rarement (2,141% pour chacun de ces sous-groupes). Notons que les textes de rédacteurs occasionnels contiennent toujours des erreurs.

Notre décision d'inclure le corpus Bissonnette dans le tableau 72 nous permet toutefois d'établir une distinction entre les textes professionnels présentant un nombre restreint d'erreurs et un texte professionnel n'en présentant pas du tout. Nous appellerons dorénavant « professionnel » un texte comptant de 1 à 3 erreurs et un texte « maître », un texte n'en comptant aucune.

Il nous est à présent possible de décrire des classes de rédacteurs à partir de l'échelle de valeurs attendues pour les populations étudiées (Tableau 73) présentant un indice i.

Tableau 73

Profils des populations de rédacteurs sur le plan des erreurs

Classes de rédacteurs[81]	Échelle de valeurs descriptives (indice de faiblesse linguistique)
Occasionnel	$5 \leq i \leq 19$
Intermédiaire	$i = 4$
Professionnel	$1 \leq i \leq 3$
Maître	$i = 0$

Nous établissons à un minimum de 5 erreurs et à un maximum de 19 les balises de l'indice de faiblesse linguistique globale d'un texte, selon les valeurs du tableau 72. Les niveaux professionnels d'écriture sont séparés pour tenir compte de la distinction entre un texte professionnel sans erreur (du calibre de ceux de Lise Bissonnette et de 50% de ceux des rédacteurs de notre corpus « Experts ») et un texte avec un nombre limité d'erreurs. En outre, selon les valeurs du tableau 72, un texte professionnel pourra contenir exceptionnellement 4 erreurs. Nous établissons donc un niveau intermédiaire pour tenir compte de cette éventualité. Ce qui distinguera cependant

---

[81] Voir note 79.

les rédacteurs intermédiaires des rédacteurs professionnels sera le nombre et la nature des signaux de maîtrise.

Distinction selon l'indice de maîtrise linguistique

Les populations de rédacteurs se distinguent aussi sur le plan du nombre des signaux de maîtrise. Le tableau 74 distribue les indices de maîtrise sur une courbe normale selon nos corpus.

Tableau 74

Distribution des indices de maîtrise autour de la moyenne selon la courbe normale

Corpus	-3σ	-2σ	-1σ	μ	+1σ	+2σ	+3σ
Moffet (σ = 1)	0	0	1	2	3	4	5
Experts (σ = 2)	0	2	3	5	7	9	10
Bissonnette (σ = 2)	0	2	3	5	7	9	10

Les deux groupes de textes professionnels de notre corpus de recherche présentent une échelle identique de valeurs en dépit du fait que le corpus "Experts" réunit les productions de 75 auteurs différents alors que le corpus étalon rassemble 75 textes d'une seule et même rédactrice. Nous pouvons donc conclure que les textes « Experts » et Bissonnette font partie de la même population[82].

Nous venons d'opérer, on s'en souviendra, une distinction entre les rédacteurs « professionnels » et les rédacteurs « maître » sur le plan des indices de faiblesse linguistique. Les valeurs du tableau 74 toutefois nous montrent que cette distinction ne peut s'effectuer au chapitre des indices de maîtrise. Autrement dit, un rédacteur professionnel sera reconnaissable au nombre d'indices de maîtrise et au fait que son texte présentera un nombre limité d'erreurs alors qu'un rédacteur professionnel de

---

82 Voilà qui nous permet de valider un point important : le fait que nous n'ayons disposé que d'un seul texte par auteur pour les groupes Moffet et « Experts » n'a pas eu d'incidence sur nos résultats. Un seul texte d'au moins 800 mots (la directive du MÉQ aux rédacteurs Moffet que nous avons adoptée comme seul critère quantitatif pour notre corpus « Experts ») suffit pour donner, sinon une représentation, du moins une forte indication du profil linguistique de son auteur.

niveau maître produira un texte comptant un nombre comparable d'indices de maîtrise mais aucune erreur.

Nous pouvons maintenant opposer les textes de rédacteurs occasionnels à ceux des rédacteurs professionnels selon une échelle de valeurs de maîtrise distribuées encore une fois sur la courbe normale (Tableau 75).

Tableau 75

Distribution des indices de maîtrise pour les populations

Corpus	-3σ	-2σ	-1σ	-μ	+1σ	+2σ	+3σ
Rédacteurs occasionnels (σ = 1)	0	0	1	2	3	4	5
Rédacteurs professionnels et maître (σ=2)	0	2	3	5	7	9	10

L'échelle de marqueurs de maîtrise décrivant la population de rédacteurs occasionnels se situe entre 0 et 4 signaux de maîtrise. Un indice de maîtrise de 5 est improbable. En revanche, les textes de rédacteurs professionnels présenteront un indice de maîtrise variant de 1 à 9. Un indice de maîtrise de 0, improbable chez un rédacteur professionnel (valeur à -3 σ), se réalisera dans 15, 73% (valeurs entre -1 σ et -3 σ) des textes non professionnels. Les chevauchements nous aideront à graduer les niveaux d'expertise du texte.

Le tableau 76 introduit les paramètres caractérisant les populations de rédacteurs sur le plan de l'indice de maîtrise linguistique.

Tableau 76

Profil des populations de rédacteurs sur le plan des signes de maîtrise

Classes de rédacteurs	Échelle de valeurs descriptives (indice de maîtrise linguistique)
Occasionnel	$0 \leq i \leq 2$
Intermédiaire	$3 \leq i \leq 4$
Professionnel et maître	$i \geq 5$

Ces paramètres sont établis en situant les valeurs du tableau 75 selon les chevauchements observés. Nous plaçons au niveau occasionnel un texte montrant 2 indices de maîtrise (soit la moyenne d'indices de maîtrise observée dans le corpus Moffet) ou moins, parce que nous savons que, selon la courbe normale, seulement une faible tranche (13,590%[83]) de la population de rédacteurs professionnels pourra produire des textes montrant seulement 2 indices de maîtrise. Dans le cas d'une telle éventualité, cette valeur pourra être interprétée en utilisant l'indice de faiblesse linguistique globale du texte : si un texte est de niveau professionnel, son indice de faiblesse linguistique globale sera de 3 ou moins.

Par ailleurs, nous avons également établi un niveau intermédiaire pour le profil des populations sur le plan des signes de maîtrise. Ce niveau réunira les textes avec un indice de maîtrise de 3 ou 4, soit moins que la moyenne de la maîtrise globale d'un texte professionnel (5), mais plus que celle d'un texte de niveau occasionnel (2). Le niveau de maîtrise linguistique intermédiaire pourra ainsi réunir des rédacteurs experts aussi bien que des rédacteurs occasionnels. Ce qui distinguera les deux groupes sera la faiblesse linguistique globale du texte.

7.1.3 Statistiques de prédiction

Des calculs statistiques de prédiction permettent de traduire en formules mathématiques les profils déterminés au moyen de la détection des indices de

[83] Voir chapitre 3, *Méthodologie*, tableau 3.

calibrage. C'est à partir des résultats de ce calcul que nous pouvons à présent prédire le niveau d'expertise linguistique du texte.

Prédiction de l'indice de faiblesse linguistique globale

Rappelons que le détecteur de notre module de calibrage repère les erreurs ne requérant pas d'analyse syntaxique complexe. Mais il reste toujours les autres erreurs de notre grille. Jusqu'à quel point pouvons-nous savoir si le texte comporte par exemple des impropriétés, des suites asyntaxiques, des problèmes d'accords, etc., justement toutes les erreurs difficiles à détecter automatiquement? Nous aimerions pouvoir prédire le nombre total d'erreurs à partir des seules erreurs simples du texte. Nous appliquerons donc à ce texte une formule de régression linéaire simple pour en calculer la faiblesse linguistique globale.

Le coefficient de corrélation entre le nombre d'erreurs simples et le nombre d'erreurs total des textes de notre corpus est de 81% (0,8092). Conformément à notre décision d'appliquer un algorithme de régression robuste à toute plage de valeurs dont le coefficient de corrélation est inférieur à 90%[84], nous cherchons le meilleur passage possible[85] entre les séries de valeurs de nos plage de données « erreurs simples » / « toutes les erreurs » et obtenons l'équation

$$y = 1{,}8066x + 3{,}1688.$$

La figure 79 montre le passage robuste de cette droite dans l'ensemble des séries des valeurs « Erreurs simples » (x) et « Total erreurs » (y).

---

[84] Voir chapitre 3 *Méthodologie.*

[85] Voir chapitre 3 *Méthodologie.*

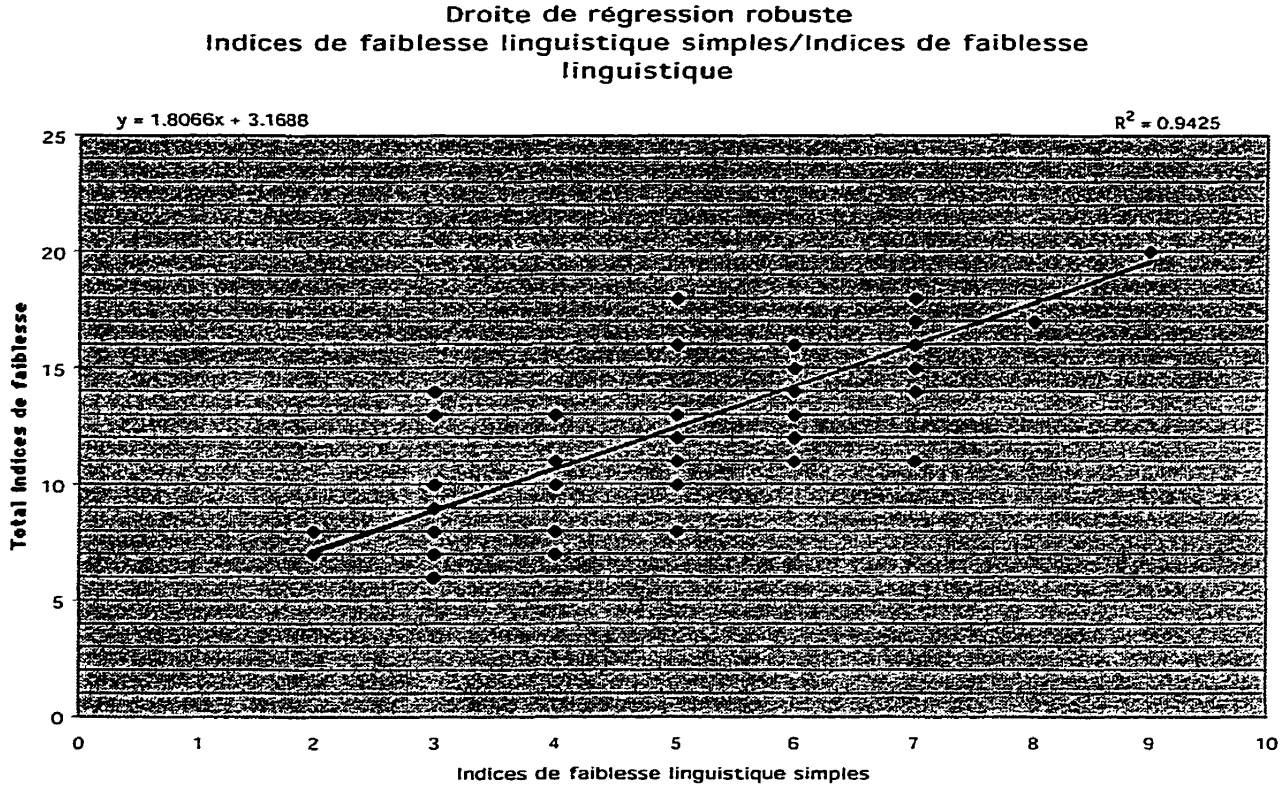


Figure 79

Droite de régression robuste «Indices de faiblesse linguistique/Total indices de faiblesse linguistique»

Avec un coefficient de détermination (R^2)[86] de 94% (0.9425) la formule de prévision statistique de cette droite prédit le total des erreurs linguistiques du texte, y compris les erreurs complexes. Ainsi, à un résultat de 0 erreur, la formule de prédiction prévoit un nombre total probable de 3 fautes, ce qui tombe dans la zone de maîtrise (de 0 à 3) établie à partir de notre étude des indices de calibrage de faiblesse linguistique. Cette valeur représente l'indice de faiblesse linguistique globale du rédacteur, autrement dit le nombre total d'erreurs linguistiques probablement comprises dans le texte.

Le tableau 77 met en parallèle les prédictions de notre droite robuste avec celles d'une droite calculée avec OLS.

Tableau 77

Estimations comparées des valeurs de faiblesse linguistique

Équation de prédiction par OLS $y = 1,7319x + 3,4601$ Corrélation = 0,8092 $R^2 = 0,6548$		Équation de prédiction robuste $y = 1,8066x + 3,1688$ Corrélation = 0,9469 $R^2 = 0,9425$	
Faiblesse observée Indices simples(x)	Faiblesse prédite Total indices (y)	Faiblesse observée Indices simples(x)	Faiblesse prédite Total indices (y)
-2	0	-2	0
-1	2	-1	2
0	3	0	3
1	5	1	5
2	7	2	7
3	9	3	9
4	10	4	10
5	12	5	12
6	14	6	14
7	16	7	16
8	17	8	18
9	19	9	19
10	21	10	21
11	23	11	23
12	24	12	25

Les valeurs prédites au moyen d'OLS sont assez comparables avec celles qui sont prédites avec l'équation robuste, à deux exceptions près (marquées en caractères gras). L'algorithme de régression robuste améliore cependant de façon non négligeable le coefficient de corrélation entre les indices de faiblesse simples (x) et le total des indices de faiblesse (y) de même que le coefficient de détermination (R^2). Un indice de faiblesse linguistique globale (valeur prédite) de 3[87] désignera un rédacteur

---

[86] On se rappelle que R^2 ou R carré évalue le pouvoir de prédiction de la droite. Voir chapitre 3 *Méthodologie*.

[87] Une valeur inférieure à 3 est impossible à obtenir avec l'équation de prédiction calculée à partir des indices de faiblesse simples parce que la détection d'indices de calibrage simples ne prévoit pas un résultat négatif.

professionnel. Avec un indice de faiblesse globale de 5[88], nous considérerons avoir affaire à un rédacteur de niveau intermédiaire. À partir d'un indice de faiblesse globale de 6, nous aurons affaire à un rédacteur occasionnel.

Prédiction de l'indice de maîtrise linguistique globale

Nous avons montré l'utilité des indices de maîtrise pour dresser le profil du niveau d'expertise linguistique d'un texte. Nous allons vouloir à présent prédire la maîtrise linguistique globale d'un texte à partir de ses seuls indices de maîtrise simples.

Le coefficient de corrélation entre le nombre d'indices de maîtrise simples et le nombre total d'indices de maîtrise est de 89% (0.8944). Avec un tel coefficient, nous pourrions nous déclarer satisfaits et conserver la formule de régression linéaire calculée automatiquement par *Excel* au moyen de la formule du critère des moindres carrés (OSL). Cependant, nous avons décidé que nous tenterions d'améliorer le passage de notre droite entre les séries de nos plages de valeurs quand le coefficient de corrélation se montrerait inférieur à 90%. Il devient en effet important de nous assurer du moins mauvais passage possible entre les séries de valeurs d'indices de maîtrise particulièrement quand on se rappelle que celles-ci montrent très peu de variation au sein de notre groupe de rédacteurs occasionnels (Fig. 80[89]) par comparaison à notre groupe de rédacteurs professionnels.

En appliquant un algorithme de régression linéaire robuste à notre plage de valeurs pour les indices de maîtrise du corpus Moffet, nous obtenons le passage entre nos séries de valeurs d'une droite (Fig. 80) décrite par l'équation

$$y = 1{,}1778x - 0{,}0564.$$

---

88 En réalité, nous devrions considérer les valeurs 4 et 5 comme caractéristiques de ce niveau particulier d'expertise linguistique. Mais comme notre équation de prédiction ne peut générer le chiffre 4 à partir du nombre d'indices de faiblesse simples, nous nous en tenons à la valeur 5.

89 La figure 80 compte en effet seulement 8 séries de valeurs de maîtrise linguistique pour les 75 paires du corpus Moffet.

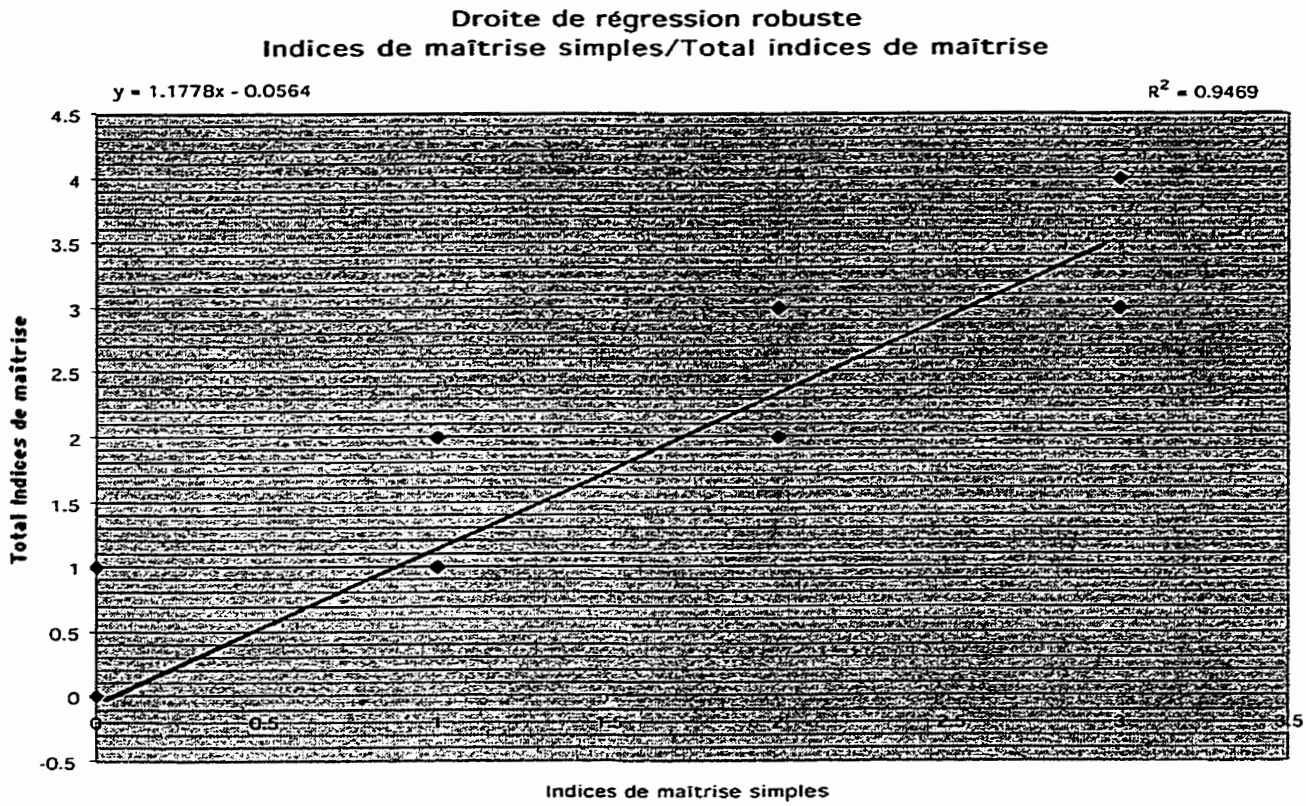


Figure 80

Régression linéaire robuste «Indices de maîtrise simples/ Total indices de maîtrise»

Encore une fois, il importe de rappeler que les points d'un diagramme de dispersion ne représentent pas chacun une seule paire de valeurs mais plusieurs paires de valeurs identiques. Le tableau 78 montre, par exemple, la distribution des paires de valeurs de maîtrise observées dans l'échantillon Moffet. Ce tableau fait ressortir « l'identité » réelle des points du diagramme de dispersion de la figure 80. Le point 0,0 par exemple est mis pour 22 textes, le point 0,1 pour 5, ainsi de suite. Voilà ce qu'il ne faut jamais perdre de vue dans un diagramme de dispersion. Chaque point peut en effet y représenter une seule paire de valeurs (c'est d'ailleurs le cas du point 3,4), deux paires de valeurs (ainsi du point 3,3) ou plusieurs paires de valeurs comme les points 0,0 et 1,1 l'illustrent bien.

Tableau 78

Variation des paires de valeurs de maîtrise observées dans le corpus Moffet

Combinaison de valeurs		Nombre de textes Moffet présentant cette combinaison
Total indices de maîtrise simples (x)	Total indices de maîtrise (y)	
0	0	22
0	1	5
1	1	24
1	2	9
2	2	9
2	3	3
3	3	2
3	4	1
	Total	75

Il s'ensuit que le *poids* attribué à chaque point d'un diagramme de dispersion influence le passage de la droite relativement au nombre d'individus représentés par ce point. Il s'ensuit également que les écarts majeurs par rapport à l'agglomération des points d'un diagramme de dispersion sont souvent le fait de paires de valeurs uniques ou représentant un nombre très limité de paires de valeurs. En attribuant une valeur zéro à ces valeurs extrêmes comme le permet l'algorithme de régression robuste, nous pouvons rendre mieux compte de la distribution réelle des valeurs de l'échantillon et assurer à notre droite un meilleur pouvoir de prédiction.

Le tableau 79 compare les résultats prédits au moyen des deux équations. Les valeurs prédites font ressortir une amélioration, non seulement dans le coefficient de corrélation, mais également dans la qualité de la prédiction de la droite robuste. Contrairement à la droite ordinaire, cette nouvelle droite permet en effet de faire la distinction entre les textes de notre corpus « Experts » et ceux de notre corpus étalon en prédisant un indice de maîtrise globale supérieur à 5 (l'indice moyen du groupe

de rédacteurs professionnels de notre corpus de recherche) quand le nombre d'indices de maîtrise simples dépasse 4.

Tableau 79

Estimations comparées des valeurs de maîtrise linguistique

Équation de prédiction par OLS *y = 1,0433x + 0,2019* *Corrélation = 0,8944* *R^2 = 0,8*		*Équation de prédiction robuste* *y = 1,1778x - 0,0564* *Corrélation = 0,9469* *R^2 = 0,9469*	
Maîtrise observée Indices simples(x)	Maîtrise prédite Total indices (y)	Maîtrise observée Indices simples(x)	Maîtrise prédite Total indices (y)
0	0	0	0
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	6
6	6	6	7

Un indice de maîtrise globale supérieur à 5 accompagnera généralement un indice de faiblesse linguistique globale de 0 et désignera un rédacteur maître comme Lise Bissonnette par exemple. Un indice de maîtrise globale de 4[90], avec un indice de faiblesse linguistique globale de 1 à 3, désignera un rédacteur professionnel. Un indice de maîtrise globale de 3 ou 4 couplé à un indice de faiblesse linguistique globale de 4 ou 5 constitueront des balises marquant des textes de niveau linguistique intermédiaire ou quasi-professionnel. Un indice de maîtrise globale de 2 ou moins couplé à un indice de faiblesse globale supérieur à 5 désignera des rédacteurs occasionnels.

---

[90] Comme il est impossible de générer la valeur prédite 5 à partir de l'équation robuste, nous considérerons un indice de 4 comme la borne séparant le niveau expert de maîtrise du niveau professionnel. Rappelons toutefois que nous proposons une matrice de calibrage qui combine les deux indices de faiblesse et de maîtrise linguistiques pour établir un profil d'expertise probable.

Prédiction du nombre d'erreurs de ponctuation

Notre corpus indique que le nombre d'erreurs de ponctuation est significativement plus élevé chez le rédacteur occasionnel que chez le rédacteur professionnel[91] : 1339 (corpus Moffet) contre 212 (corpus « Experts ») et 51 (corpus Bissonnette).

Par ailleurs, la distribution des erreurs de ponctuation de notre corpus nous montre également que les textes professionnels ont peu de chance de contenir des erreurs de ponctuation de type I (tableau 80), c'est-à-dire avec impact sur la définition automatique de frontière de phrase.

Tableau 80

Distribution des erreurs de ponctuation de type I dans le corpus de recherche

Données	Moffet	Experts	Bissonnette
Erreurs type I (chiffres absolus)	189	3	0
μ[92]	3	0	0
σ[93]	3	0	0

En revanche, un calcul de distribution des erreurs de ponctuation de type I selon la courbe normale montre que leur nombre pourra se situer, pour 99,73% de la population de rédacteurs occasionnels, entre 0 (μ-1σ, μ-2σ ou μ-3σ) et 12 (μ+1σ, μ+2σ ou μ+3σ), la valeur 12 (μ+3σ) constituant une occurrence exceptionnelle.

Les textes professionnels ne contiendront pas non plus d'erreurs de confusion de signes de type II, c'est-à-dire sans incidence sur la définition automatique de la phrase (tableau 81). Un texte non professionnel toutefois pourra contenir des erreurs de confusion de signes variant de 0 (1σ, μ-2σ ou μ-3σ) à 10 (μ+1σ, μ+2σ ou μ+3σ), 10 (μ+3σ) constituant une occurrence exceptionnelle.

---

91 Voir Chapitre 1 *Objectifs et problématique*, tableau 4 *Moyennes et écarts-types pour les résultats en ponctuation des rédacteurs du corpus.*

92 Moyenne.

93 Écart-type.

Tableau 81

Distribution des erreurs de confusion de signes de type II dans le corpus de recherche

Données	Moffet	Experts	Bissonnette
Confusion de signes (chiffres absolus)	72	8	0
μ	1	0	0
σ	3	0	0

Par conséquent, il devrait exister une corrélation positive entre le nombre d'erreurs de ponctuation d'un texte et le nombre d'erreurs linguistiques tout court. Nous voudrons utiliser cette corrélation pour bâtir des régressions linéaires simples en vue de prédire le nombre d'erreurs de ponctuation du texte.

Prédiction des erreurs de ponctuation de type I

La droite d'estimation d'erreurs de ponctuation de type I permet de dépister les textes susceptibles de présenter des erreurs de ponctuation avec incidence sur la définition automatique des frontières de phrase[94]. De telles erreurs ont des effets négatifs importants sur la capacité d'un correcteur grammatical de repérer automatiquement les erreurs linguistiques tout autant que sur celle des humains de comprendre les portions du texte qui les renferment. La formule de régression robuste sur laquelle s'appuie notre droite (Fig. 81[95]) permettra ainsi au calibreur d'estimer s'il se trouve en présence d'un texte difficile à segmenter et à combien d'erreurs de ponctuation de ce type il peut probablement s'attendre.

---

94 Voir Chapitre 1. *Objectifs et problématique,* « Impact des erreurs de ponctuation sur la segmentation automatique du texte ».

95 Rappelons que les points d'un diagramme de dispersion ne représentent pas nécessairement une seule paire de valeurs comme le tableau 78 le démontre pour les points du diagramme de la figure 80.

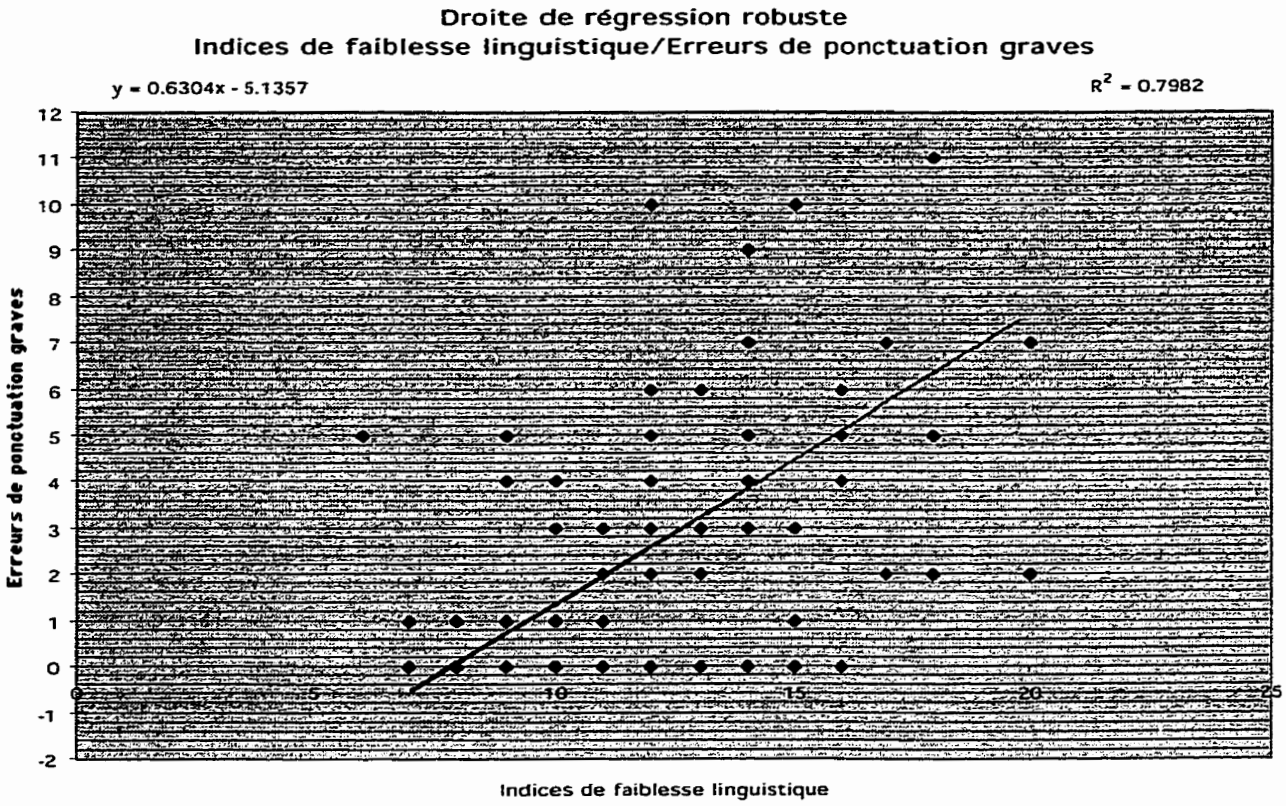


Figure 81

Droite de régression robuste Indices de faiblesse linguistique / Erreurs de ponctuation graves

Cette droite se calcule à partir de deux séries de variables : la série de variables indépendantes *Indices de faiblesse linguistique* et la série de variables dépendantes *Erreurs de ponctuation graves.* Se fondant sur un coefficient de corrélation positif de 0,8934 (91%) et un R^2 de 0,7982 (80%), elle se traduit par l'équation

$$y = 0{,}6304x - 5{,}1357,$$

où la valeur à prédire y est le nombre d'erreurs de ponctuation de type I (avec incidence sur la définition automatique de la frontière de la phrase). Le tableau 82 met en parallèle les valeurs prédites avec des équations de régression linéaire ordinaire et robuste.

Tableau 82

Estimations comparées des valeurs d'erreurs de ponctuation de type I

Équation de prédiction par OLS *$y = 0,4144x - 2,3586$* *Corrélation = 0,4921* *$R^2 = 0,2421$*		*Équation de prédiction robuste* *$y = 0,6304x - 5,1357$* *Corrélation = 0,8934* *$R^2 = 0,7982$*	
Faiblesse globale observée (x)	Erreurs ponctuation type I prédites (y)	Faiblesse globale observée (x)	Erreurs ponctuation type I prédites (y)
0	-2	0	-5
1	-2	1	-5
2	-2	2	-4
3	-1	3	-3
4	-1	4	-3
5	0	5	-2
6	0	6	-1
7	1	7	-1
8	1	8	0
9	1	9	1
10	2	10	1
11	2	11	2
12	3	12	2
13	3	13	3
14	3	14	4
15	4	15	4
16	4	16	5
17	5	17	6
18	5	18	6
19	6	19	7
20	6	20	7
21	6	21	8
22	7	22	9
23	7	23	9
24	8	24	10
25	8	25	11

Le tableau 82 montre une différence majeure entre les valeurs prédites avec l'équation calculée au moyen d'OLS et l'équation robuste. En effet, alors que le maximum d'erreurs de type I prédites par l'équation calculée avec OLS est de 8, le maximum prédit par l'équation robuste est de 11. Or cette valeur rejoint le maximum observable dans la population[96] selon les valeurs du tableau 80. Les valeurs prédites avec l'équation robuste se comparent aussi avec celles de l'échantillon Moffet, qui se situent également entre 0 et 11. Par conséquent, nous pouvons dire que la prédiction robuste rend mieux compte des données de l'échantillon — et de la population qu'il représente — que la prédiction ordinaire. Voilà qui n'est pas étonnant si on considère l'amélioration importante du coefficient de corrélation après l'application de l'algorithme de régression robuste.

Le tableau 82 livre également un autre renseignement précieux : le niveau que nous appellerons « de sécurité » pour une opération de segmentation automatique de texte. On se souviendra que les erreurs de ponctuation graves ou de type I sont celles qui interfèrent avec la définition automatique des frontières de phrase. L'équation robuste fixe ainsi à 8 erreurs le niveau de sécurité d'un texte. En d'autres mots, un texte qui compte 9 erreurs ou plus est susceptible de présenter un certain nombre d'erreurs de ponctuation graves, qui risquent d'interférer avec une opération de segmentation automatique de texte.

Dans notre matrice de calibrage, nous allons utiliser cette borne pour opérer une nouvelle subdivision, mais cette fois, parmi nos rédacteurs occasionnels. Les rédacteurs dont les textes présentent un indice de faiblesse linguistique globale de 8 ou moins[97], avec une valeur prédite d'erreurs de ponctuation graves de 0, seront appelés « occasionnels ». Les autres seront appelés « occasionnels faibles ». Une opération de segmentation automatique pour les textes des rédacteurs occasionnels faibles sera considérée à risque, ce risque augmentant avec le niveau de faiblesse linguistique globale enregistrée.

---

96 Voir tableau 80 et le paragraphe qui le suit.

97 Nous sommes conscients qu'il s'agit là de valeurs prédites et que, dans la réalité, certains textes de moins de 8 erreurs pourraient présenter des erreurs de ponctuation graves alors que des textes de 9 erreurs et plus pourraient ne pas en présenter du tout.

Prédiction du nombre d'erreurs de ponctuation de type II

Les erreurs de ponctuation de type II sont toutes les erreurs de ponctuation n'ayant pas d'incidence sur la définition automatique des frontières de phrase. Tombent dans cette catégorie les omissions de signes autres que le point, les confusions entre des signes autres que le point et l'introduction de signes abusifs autres que le point. Nous cherchons donc à bâtir une régression linéaire simple pour prédire le nombre d'erreurs de ponctuation de type II à partir des erreurs linguistiques du texte.

Notre corpus d'erreurs de ponctuation de type II distinguent les groupes de textes sous au moins deux aspects significatifs. : la nature de ce type d'erreurs et leur dispersion par rapport à leurs moyennes respectives.

Les erreurs de ponctuation de type II des textes Moffet tombent en effet dans les trois catégories prévues par notre grille : l'omission de signes, la confusion de signes et l'introduction abusive de signes, avec une majorité significative (77%) d'erreurs d'omission de signes (Fig. 82).

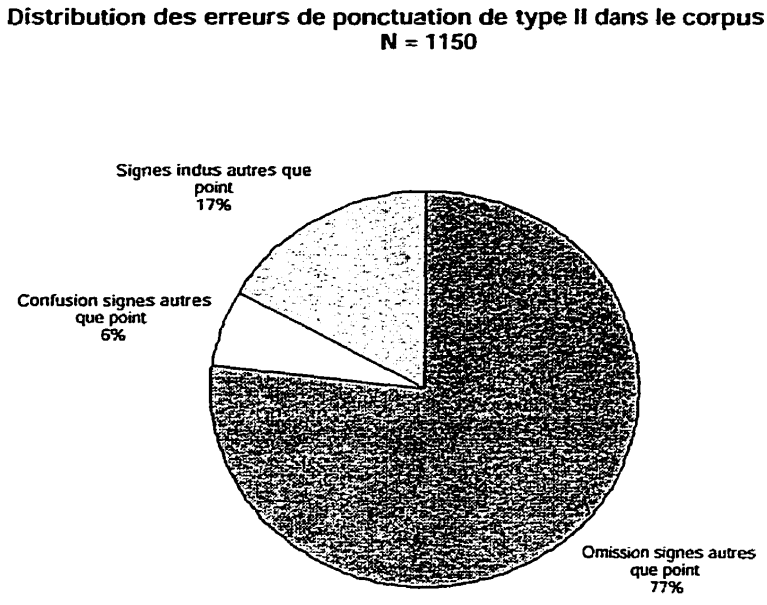


Figure 82

Distribution des erreurs de ponctuation de type II dans le corpus Moffet

En revanche, les textes du corpus « Experts » ne comprennent qu'exceptionnellement d'autres erreurs que des erreurs d'omission de signes[98], dont 98% sont d'ailleurs des virgules manquantes. En fait, les erreurs de ponctuation autres que des signes manquants sont si rares que le nombre de ces erreurs par catégorie (confusion de signes : 9[99]; signes indus : 31) dans le corpus « Experts » donne une moyenne de 0. Quant au corpus Bissonnette, il n'en compte aucune.

Les groupes de textes se distinguent aussi selon la dispersion des erreurs de ponctuation autour de leur moyenne. Avec un écart-type de 10 et une moyenne de 15 (tableau 83), les textes Moffet pourraient théoriquement compter de 0 à 44 erreurs de ponctuation de type II. Par contre, la distribution des erreurs de ponctuation sur la courbe normale pour les textes professionnels montre une variation peu importante : de 1 à 5 erreurs pour les textes « Experts; de 0 à 3 erreurs pour les textes Bissonnette (tableau 83).

---

98 Voir chapitre 1 *Objectifs et problématique*, tableau 1 *Place des virgules manquantes dans les erreurs de ponctuation.*

99 Voir tableau 81.

Tableau 83

Distribution des erreurs de ponctuation de type II par corpus

Corpus	-3σ	-2σ	-1σ	μ	+1σ	+2σ	+3σ
Moffet (σ = 10)	0	0	5	15	25	35	45
« Experts » (σ = 1)	0	1	2	3	4	5	6
Bissonnette ((σ = 1)	0	0	0	1	2	3	4

Ces données nous aident d'ailleurs à ajouter un critère pour raffiner notre nouvelle subdivision entre les rédacteurs « professionnels » et les rédacteurs « maître ». Les textes « professionnels » contiendront au moins une erreur de ponctuation de type II par texte (et pourront en contenir jusqu'à 5); les textes « maître » pourront, une fois sur deux (la moitié de la distribution des textes Bissonnette [tableau 83] renferme une valeur 0), ne contenir aucune erreur de ponctuation.

Cependant, il n'est pas possible de donner une estimation complètement fonctionnelle du nombre d'erreurs de ponctuation de type II au moyen d'une régression linéaire simple[100], même robuste, à partir de notre corpus de textes non professionnels. En effet, malgré une corrélation positive de 84% (0,8403) après l'application de l'algorithme de régression robuste, la droite (Fig. 83[101]), représentée par l'équation

$$Y = 1{,}604x - 3{,}7616,$$

prédit une seule erreur de ponctuation de type II pour un texte professionnel présentant 3 erreurs linguistiques et aucune, pour un texte avec 2 erreurs linguistiques ou moins.

---

100 Comme il est facile de le constater en ouvrant l'essai de Ryan (1997), il existe bien des formes de régressions, peut-être mieux adaptées à ce problème : entre autres, les régressions multiples, multivariées (il ne s'agit pas de la même chose (ibid., 118], polynomiales-trigonométriques, logiques ou les *ridge regression*, qui font appel à une forme de calcul appelée *quadratic fit*. Toutefois, nos connaissances limitées en statistiques et le temps requis pour nous familiariser avec de tels calculs nous empêchent, à cette étape-ci de notre travail, d'explorer cette avenue.

101 Comme pour les autres diagrammes de dispersion, les points de la figure 83 ne représentent pas nécessairement chacun une seule paire de valeurs.

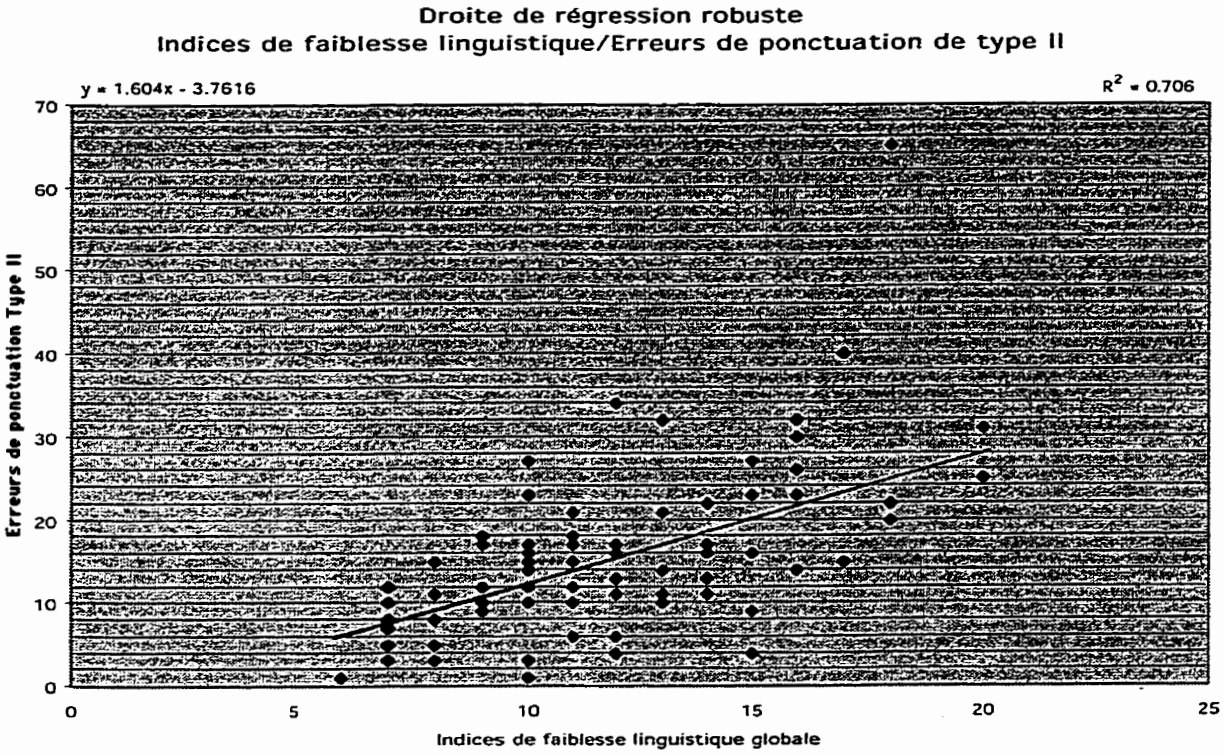


Figure 83

Droite de régression robuste « Indices de faiblesse linguistique globale / Erreurs de ponctuation de type II

Le tableau 84 met en parallèle les valeurs prédites avec les équations ordinaire et robustes.

Tableau 84

Estimations comparées des valeurs d'erreurs de ponctuation de type II

Équation de prédiction par OLS $y = 1,7868x - 5,7028$ Corrélation = 0,5987 $R^2 = 0,3584$		Équation de prédiction robuste $y = 1,604x - 3,7616$ Corrélation = 0,8403 $R^2 = 0,706$	
Faiblesse globale observée (x)	Erreurs ponctuation type II prédites (y)	Faiblesse globale observée (x)	Erreurs ponctuation type II prédites (y)
0	-6	0	-4
1	-4	1	-2
2	-2	2	-1
3	0	3	1
4	1	4	3
5	3	5	4
6	5	6	6
7	7	7	7
8	9	8	9
9	10	9	11
10	12	10	12
11	14	11	14
12	16	12	15
13	18	13	17
14	19	14	19
15	21	15	20
16	23	16	22
17	25	17	24
18	26	18	25
19	28	19	27
20	30	20	28
21	32	21	30
22	34	22	32
23.	35	23	33
24	37	24	35
25	39	25	36

Les valeurs prédites par la droite robuste se rapprochent quand même davantage des valeurs observées dans notre corpus de textes professionnels. En effet, la droite robuste prévoit l'occurrence d'une erreur de ponctuation pour un texte de 3 erreurs, la valeur maximale pour la faiblesse globale du texte caractérisant le niveau professionnel, alors que la droite ordinaire n'en prévoit aucune.

Nous pourrions à la rigueur arguer que, puisque le tableau 83 démontre qu'un texte professionnel comporte au moins une erreur de ponctuation, notre droite robuste arrive à donner une estimation acceptable du nombre d'erreurs de ponctuation de type II attendues dans un tel texte. Cependant, nous savons que les textes professionnels avec 1 ou 2 erreurs linguistiques comportent également au moins une erreur de ponctuation de type II et qu'un texte maître sur deux en comporte aussi une, une fois sur deux.. Par conséquent, nous ne pouvons être que partiellement satisfaits des résultats générés par notre droite.

Par ailleurs, les valeurs prédites par notre régression robuste pour les textes de niveau occasionnel se rapprochent plus des valeurs réellement observées dans l'échantillon Moffet que celles générées par la régression ordinaire. En effet, si nous excluons le résultat du sujet Moffet 62, avec ses 65 erreurs de ponctuation, résultat qui s'éloigne de façon significative de l'échelle de résultats des autres rédacteurs (de 1 à 40), la droite robuste permet de prédire assez convenablement le nombre des erreurs de ponctuation de type II chez les rédacteurs occasionnels.

Bref, notre droite robuste nous donne des valeurs partiellement utilisables. Un module de calibrage devrait donc inclure une routine de correction pour tenir compte des valeurs attendues (selon la distribution du tableau 83) quand le nombre d'erreurs du texte est inférieur à 2.

7.1.4 Grille de calibrage

Le module de calibrage détermine chaque profil linguistique selon une grille montée d'après les observations de notre corpus (Tableau 85). Cette grille classe les rédacteurs selon les 5 niveaux de compétence que nous venons d'établir : occasionnel faible, occasionnel, intermédiaire, professionnel et maître.

Tableau 85

Grille de calibrage

Classe de rédacteurs	*Profil linguistique*	*Erreurs de ponctuation prédites*	*Impact sur la correction automatique du texte*
occasionnel faible	Indice de faiblesse linguistique : >8; Indice de maîtrise linguistique : < 3.	Ponctuation incohérente. de 1 à 11 erreurs de type I; de 11 à 36 erreurs de type II réparties ainsi: 77% d'omissions de signes; 17% de signes indus; 6% de confusions de signes.	Segmentation automatique risquée; Nombreuses erreurs de détection probables; Correction manuelle préférable.
occasionnel	Indice de faiblesse linguistique : : $5<i\leq 8$. Indice de maîtrise linguistique : 3	Ponctuation incohérente. 0 erreur de type I; de 5 à 9 erreurs de ponctuation de type II réparties ainsi: 77% d'omissions de signes; 17% de signes indus; 6% de confusions de signes	Segmentation automatique possible et fiable. Correction automatique possible.
intermédiaire	Indice de faiblesse linguistique : 4 ou 5 Indice de maîtrise linguistique : 3 ou 4	Ponctuation possiblement cohérente; 0 erreur de type I; de 3 à 6 erreurs de ponctuation de type II réparties ainsi: 77% d'omissions de signes; 17% de signes indus; 6% de confusions de signes	Segmentation automatique possible et fiable. Correction automatique possible.
professionnel	Indice de faiblesse linguistique : $1\leq i<4$ Indice de maîtrise linguistique : ≥4	Ponctuation cohérente; 0 erreur de type I; un nombre limité d'erreurs de ponctuation de type II, surtout des virgules manquantes. Ponctuation de type II abusive ou confusions de signes de type II possibles, mais exceptionnels.	Segmentation automatique possible et fiable; Révision de la ponctuation utile. Correction limitée au repérage de coquilles.
maître	Indice de faiblesse linguistique : 0; Indices de maîtrise linguistique ≥4	Ponctuation cohérente. 0 erreur de type I; 1 virgule manquante, généralement à fonction de délimiteur, un texte sur deux. Aucune confusion de signes. Pas de ponctuation abusive.	Segmentation automatique possible et fiable; Révision de la ponctuation inutile; Correction de texte limitée au repérage de coquilles.

Cette grille de calibrage classe les niveaux de compétence des rédacteurs au moyen d'indicateurs quantitatifs, les indices de calibrage de faiblesse et de force linguistique. Elle calcule ces indices en format logique mais devrait également tenir compte du nombre d'occurrences pour ajouter du poids à l'information, selon le postulat que plus un indice se retrouve souvent dans un même texte, plus il est significatif[102]. Ces niveaux de compétence constituent en fait des bornes permettant de classer les rédacteurs sur une ligne allant de l'absence de contrôle à la maîtrise linguistique (Fig. 84).

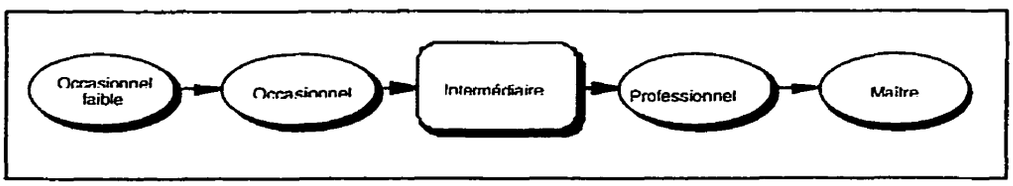


Figure 84

Ligne de compétence linguistique écrite

Le profil linguistique associé à chaque niveau de compétence permet de situer ensuite le rédacteur dans la population qui partage avec lui ce profil et de prédire la nature et la distribution des erreurs de ponctuation qui lui sont caractéristiques.

7.2 Éléments de méthode

Dans la pratique informatique actuelle, les utilisateurs et leurs texteurs n'entretiennent généralement pas de « relations ». Cette équation est toutefois appelée à changer avec l'introduction de ce que nous appelons la *clé de calibrage*, un dispositif permettant d'appliquer une correction automatique adaptée au portrait linguistique du rédacteur. Un même traitement de texte pourra ainsi disposer d'autant de clés de calibrage que d'utilisateurs, si bien qu'il pourra théoriquement offrir une aide adaptée aux forces et aux faiblesses linguistiques de chacun.

À défaut de présenter ici une méthode de correction automatique complète et fonctionnelle de la ponctuation (ce qui constituerait en soi l'objet d'une autre thèse),

---

[102] Notre recherche ne comporte pas d'étude particulière de ce phénomène en raison du format manuscrit de l'échantillon Moffet. Nous sommes convaincus cependant que le nombre d'occurrences d'un même indice devrait être comptabilisé dans une matrice de calibrage.

nous allons plutôt élaborer une simulation pour faire ressortir comment le calibrage pourrait être exploité dans le futur.

La figure 85 illustre la configuration possible d'un correcteur enrichi d'une clé de calibrage.

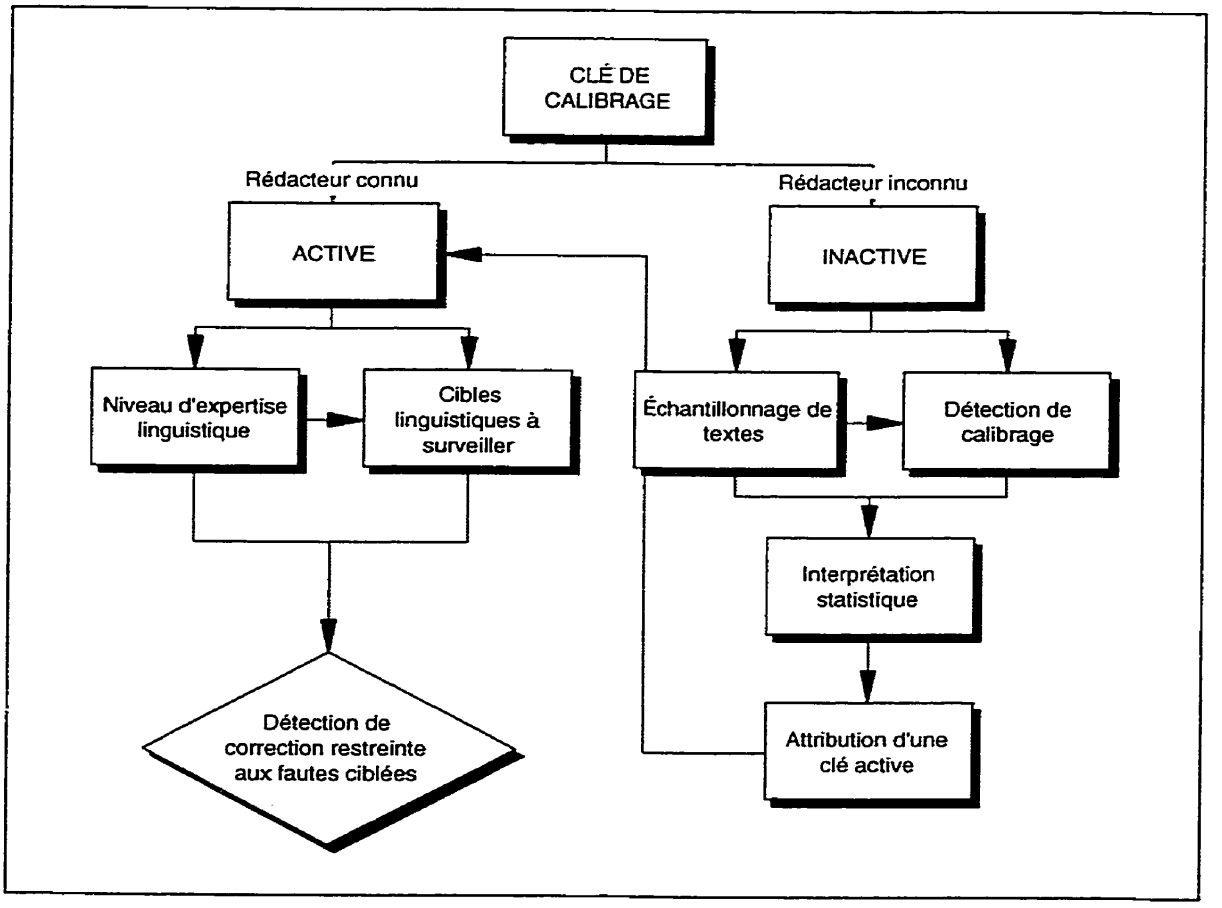


Figure 85

Scénario de calibrage

La clé de calibrage, une suite alphanumérique déterminée au terme de l'opération de calibrage, intervient dès le démarrage du texteur et génère deux réponses possibles: ou bien la clé est active — ce qui signale que le profil linguistique de l'utilisateur a déjà été déterminé —, ou bien la clé est inactive: le profil linguistique de l'utilisateur n'a pas encore été déterminé.

7.2.1 Clé de calibrage active

Une clé de calibrage active associe un utilisateur à des cibles de correction spécifiques. Connaissant en effet, grâce à la matrice de calibrage, le niveau de compétence probable du rédacteur de même que les contextes d'erreurs associés (par exemple, la répétition abusive de mots, les impropriétés, etc.), le correcteur identifie ces contextes comme des cibles à surveiller au moment de la saisie du texte. Quand l'une d'entre elles se présente, le correcteur la signale au rédacteur et propose des solutions au besoin. En revanche, les cibles ne figurant pas dans le profil actif sont ignorées par le correcteur.

Les textes des rédacteurs intermédiaires reçoivent le même traitement que ceux des rédacteurs professionnels.

7.2.2 Clé de calibrage inactive

Une clé de calibrage inactive génère une opération de calibrage. Cette opération s'effectue en trois temps :

* L'échantillonnage de textes;
* La détection d'erreurs pour fins de calibrage;
* L'interprétation statistique des résultats.

Échantillonnage de textes

Pour effectuer son analyse, le détecteur du module de calibrage a besoin d'un texte argumentatif ou informatif suivi de 800 à 1 500 mots, rédigé par l'utilisateur mais non corrigé.

Un texte de 800 à 1 500 mots est-il suffisant pour générer un calibrage valide? Oui, si nous en croyons notre corpus. Les 75 textes du corpus Bissonnette ont, par exemple, bien fait ressortir comment la performance écrite se fonde généralement sur des habitudes linguistiques individuelles se reproduisant d'un texte à l'autre (ainsi, chez Lise Bissonnette, une utilisation fréquente de l'adjectif indéfini *quelque* au sens de *un certain* et un attachement évident pour l'imparfait ou le plus-que-parfait du

subjonctif[103]). En revanche, les messages courts (par exemple, les courriels ou les mémos) — que nous avions d'abord essayé d'utiliser[104] — sont souvent trop courts pour offrir les contextes utiles en nombre suffisant.

Le critère du texte suivi est par ailleurs important à cause de la nature de plusieurs de nos indices de calibrage. Par exemple, la cohérence d'emploi des connecteurs en série, la répétition abusive de mots, les références anaphoriques sont autant d'erreurs significatives qui peuvent plus facilement se repérer dans un texte suivi. Un texte suivi, en outre, favorise l'organisation logique de paragraphes, la mise en relief des idées et une syntaxe davantage axée sur la délivrance d'un message élaboré. Les textes rédigés en style télégraphique sont à éviter. Les messages électroniques, les notes , les rappels ou les mémos ne se prêtent pas à une détection de calibrage efficace.

Détection pour fins de calibrage

La détection pour fins de calibrage consiste à rechercher les indices de calibrage simples.

Dans un exercice de calibrage, les indices de faiblesse et de maîtrise linguistiques constituent deux voies distinctes mais complémentaires pour décrire le rédacteur. En effet, dans l'estimation du portrait linguistique du rédacteur, chaque catégorie devient le pendant de l'autre, chacune autorisant un exercice de validation contribuant à minimiser les erreurs d'interprétation de résultat.

Le repérage des indices simples s'effectue par des lectures successives ciblées au moyen de routines visant à repérer chaque erreur dans l'ordre suggéré. Des dispositifs de comptage de fréquences — un compteur d'indices (comptés en format logique pour permettre la comparaison avec les données de notre étude) et des compteurs d'occurrences[105], un par indice — recueillent et accumulent les données quantitatives pertinentes. Ces données sont ensuite traduites en indice, lequel sert de

---

103 Voir Chapitre 6. *Indices de maîtrise.*

104 Voir Chapitre 3. *Méthodologie.*

105 D'autres recherches sont nécessaires pour déterminer le poids particulier imputable aux fréquences réelles d'occurrences dans l'évaluation du niveau d'expertise écrite d'un rédacteur.

tremplin aux calculs de prévision statistiques de notre matrice de calibrage. Au terme de l'opération, le profil du rédacteur est établi. L'utilisateur valide sa clé de calibrage en lui attribuant un code alphanumérique de son choix et le correcteur calibre sa lecture en fonction de cette clé.

7.3 Limites du processus de calibrage

Nous pouvons actuellement formuler au moins 2 limites au processus de calibrage tel que nous venons de le décrire: la validité de l'extrapolation de l'échantillon à la population et la pertinence du niveau de compétence « intermédiaire».

7.3.1 Validité de l'extrapolation de l'échantillon à la population

Nous avons franchi le pas de l'extrapolation en établissant des populations de rédacteurs selon des niveaux de compétence conclu à partir d'un échantillon. Si nous sommes confiants dans la validité de notre extrapolation pour la population de rédacteurs occasionnels d'où cet échantillon a été tiré, nous le sommes moins pour celle des rédacteurs professionnels. En effet, nous savons que notre corpus de textes professionnels n'a pas été tiré de façon aléatoire en respectant les règles de la statistique. Par conséquent, nous ne pouvons affirmer que ces textes sont représentatifs de la population de rédacteurs professionnels pour laquelle nous proposons quand même des profils particuliers.

Nous avons également extrapolé à la population de rédacteurs occasionnels d'expression française à partir des 16 200 candidats à l'université ayant passé l'épreuve de français du Gouvernement du Québec en mars 1998. Cette dernière extrapolation est-elle valide? Comme rien ne nous prouve que ces rédacteurs sont représentatifs des rédacteurs occasionnels de la communauté francophone internationale, rien ne nous prouve non plus que notre matrice de calibrage serait valide avec des corpus provenant d'autres régions francophones du monde.

Dans tous les cas, nos conclusions débouchent sur d'autres recherches. Il faudra d'abord certainement tenter de reproduire nos conclusions à partir de

nouveaux corpus, d'abord en constituant un corpus de textes professionnels selon les règles de la statistique et ensuite, en vérifiant la validité de l'extrapolation à la population de rédacteurs occasionnels internationale d'expression française.

7.3.2 Pertinence du niveau de compétence intermédiaire

Pour caractériser le passage d'une population de rédacteurs à l'autre, nous avons proposé un niveau de compétence que nous avons appelé « intermédiaire». Mais ce niveau est-il réellement pertinent? Marque-t-il un groupe de textes présentant des caractéristiques spécifiques ou bien regroupe-t-il des textes de rédacteurs se trouvant accidentellement à un niveau autre que le leur?

Le niveau intermédiaire cherche en effet à décrire les textes dont la lecture de performance aboutit à une interprétation ambiguë. Par exemple, un texte professionnel pourra exceptionnellement présenter plus d'erreurs (4 ou 5) que le nombre d'erreurs attendu de la population de textes professionnels selon notre étude (0 à 3), mais démontrer quand même un nombre d'indices de maîtrise supérieur au niveau de rédacteur occasionnel (0 à 2), et inversement. Bien que très rares en effet dans une distribution normale, des mesures se situant à plus ou moins 3 écarts-types de la moyenne peuvent toutefois se produire, puisque cette distribution ne couvre pas 100% de toutes les mesures possibles d'une courbe normale, mais 99,73% (Allaire, 1995 : 12-1).

Rien ne nous dit par conséquent qu'un texte classé par notre matrice de calibrage au niveau de compétence intermédiaire ne constitue pas en fait une production exceptionnelle de la part d'un rédacteur appartenant à l'une ou l'autre de nos deux populations. Dans un tel scénario, le niveau de compétence intermédiaire serait associé à une zone d'incertitude (Fig. 86).

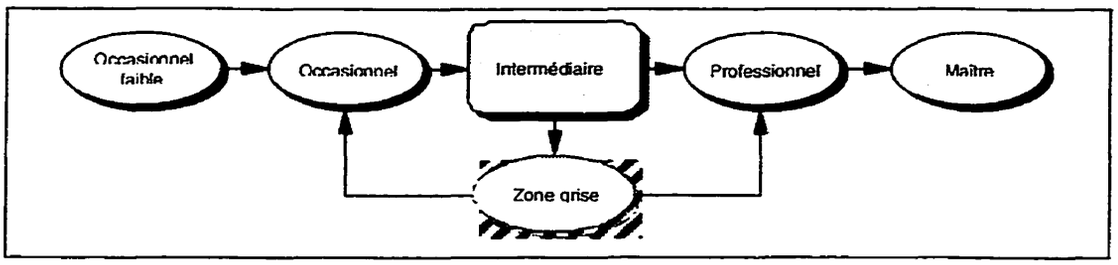


Figure 86

Zone d'incertitude possible du niveau de compétence intermédiaire

Mais peut-être cette zone grise ne reflète en rien la réalité. Peut-être existe-t-il en effet une troisième population de rédacteurs, ni tout à fait professionnels, ni tout à fait occasionnels.

La question se pose probablement parce que nous avons proposé une lecture de calibrage à partir d'un seul texte. Nous pourrions répondre à cette question et éviter l'ambiguïté en appliquant un scénario où nous fonderions notre profil linguistique sur un corpus de textes suivis plutôt que sur un seul, dans le cas où les résultats tombent dans la catégorie intermédiaire.

Nous pourrions par exemple imaginer une période d'entraînement à la matrice de calibrage. Cette période, définie par un nombre X de textes, verrait les données accumulées dans la matrice avec chaque lecture additionnelle de calibrage. C'est seulement au terme de la période d'entraînement que la clé de calibrage serait émise et validée par cet utilisateur.

Conclusion

Notre matrice de calibrage effectue quatre prévisions statistiques au moyen de formules de régression linéaire robuste (tableau 86).

Tableau 86

Formules statistiques de calibrage

Droite	Corrélation	R^2	Formule de prévision
Indices de faiblesse linguistique simples / Faiblesse linguistique globale	0,9708	0,9425	y = 1,8066x + 3,1688
Indices de maîtrise linguistique simples / Maîtrise linguistique globale	0,8944	0,8	y = 1,1778x - 0,0564
Faiblesse linguistique globale / Erreurs de ponctuation de type I	0,9082	0,8248	y = 0,5864x - 4,6302
Faiblesse linguistique globale / Erreurs de ponctuation de type II	0,8403	0,706	y = 1,604x - 3,7616

La mise en relation des trois premiers calculs permet une estimation assez juste de l'expertise linguistique du texte en le situant par rapport à une population de rédacteurs classée selon 5 niveaux de compétence. Le dernier calcul génère des prédictions partiellement utilisables en ce qu'elles permettent d'estimer le nombre d'erreurs de ponctuation de type II pour des textes présentant plus de 3 erreurs, mais pas d'estimer celui des textes qui comptent 2 erreurs et moins.

La première formule attribue un indice de faiblesse linguistique globale au texte en calculant le nombre total d'erreurs à partir du seul nombre d'erreurs simples détectées lors de la routine de diagnostic. Reposant sur une corrélation presque parfaite de 97%, la valeur prédite est solide. Bien que cette nouvelle valeur ne représente en fait qu'une estimation, sa validité nous apparaît quand même supérieure à celle que nous pourrions obtenir en tirant des conclusions à partir d'un analyseur syntaxique que nous lancerions à la chasse aux impropriétés et aux références anaphoriques.

La deuxième formule calcule l'indice de maîtrise linguistique globale à partir du seul nombre d'indices de maîtrise simples détectés automatiquement dans le texte. Avec une corrélation de 89%, elle permet une plus grande certitude dans l'établissement du profil linguistique du texte en permettant un critère de dépistage additionnel pour les textes de niveau professionnel.

Le troisième calcul désigne les textes présentant des faiblesses de ponctuation majeures. S'appuyant sur une corrélation de 91%, elle signale au correcteur grammatical les textes qu'il ne pourra pas segmenter en s'appuyant sur la ponctuation assertive originale.

Le quatrième calcul estime partiellement le nombre d'erreurs de ponctuation de type II, c'est-à-dire sans incidence sur la définition automatique des frontières de phrase, pour les textes de plus de 3 erreurs.

Comme nous travaillons à partir d'un échantillon statistiquement significatif, nous considérons nos formules de calibrage comme valides pour l'ensemble de la population de rédacteurs occasionnels d'où cet échantillon est tiré. Théoriquement, notre matrice de calibrage aurait pu ainsi estimer automatiquement le profil linguistique des 16 2 00 candidats à l'université s'étant soumis à l'épreuve de français du ministère de l'Éducation du Québec en mars 1998. Elle pourrait également estimer le niveau d'expertise d'un texte qu'on projetterait publier, puisqu'elle permet également de reconnaître les textes de qualité professionnelle, y compris ceux de calibre supérieur comme les textes de Lise Bissonnette.

Conclusion

Nous savons déjà comment les outils automatiques d'aide à l'écriture actuellement en circulation génèrent parfois plus d'erreurs qu'ils n'en corrigent. Nous avons également fait ressortir qu'une majorité très significative des erreurs de ponctuation repérées dans notre corpus sont des erreurs d'omission de signes. Ce que nous venons de voir, c'est comment un modèle mathématique de prévision robuste pourrait permettre à un correcteur d'améliorer la production écrite des utilisateurs en tenant compte de leur niveau personnel d'expertise de la langue française.

Le calibrage débouche sur des applications intéressantes. Avant d'envisager de telles perspectives cependant, il importe de jeter un regard critique sur deux aspects limitatifs potentiels importants : l'aspect de la validité méthodologique de notre recherche et celui de son pouvoir d'extrapolation.

L'aspect méthodologique de notre étude impose une question d'objectivation fondamentale : avons-nous réussi à éliminer ce que Cohen (1995 : 79) appelle les *spurious effects* ou sources courantes de conclusions fallacieuses dans le domaine de la recherche en intelligence artificielle? Bien qu'il pose ce problème pour les expériences en intelligence artificielle, nous pensons que cette réflexion est également pertinente dans un exercice comme celui que nous venons de compléter.

Cohen souligne en effet quatre erreurs méthodologiques à éviter en intelligence artificielle. Il explique (*ibid.* : 80) par exemple que des erreurs importantes d'interprétation de résultats peuvent se produire dans le cas où des comparaisons de performance sont effectuées à partir de tests, ou bien trop faciles pour poser un défi significatif dans le domaine de la performance mesurée (*ceiling effect*), ou bien trop difficiles pour être réussis même par un expert du domaine (*floor effect*). On pourrait toujours arguer que l'épreuve imposée, au Québec, aux candidats et candidates à l'université constitue un exercice ne pouvant pas mesurer la performance écrite et

que, par conséquent, toute conclusion tirée à partir de l'examen de ces épreuves n'est pas valide. Au-delà de toutes discussions théoriques ou idéologiques sur la validité des épreuves de français du MÉQ, il faut bien avouer que la seule façon définitive d'apaiser ces doutes, à laquelle nous avons pensé, consiste à répéter notre expérience en appliquant notre instrument de mesure à une nouvelle population et un nouvel échantillon.

Cohen rapporte aussi (*ibid* : 84) deux autres risques méthodologiques insidieux. L'un porte sur une mésestimation du rôle de la chance dans la détermination des résultats; l'autre, sur l'effet de l'ordre des questions dans un test de performance. Nous ne croyons pas notre expérimentation vulnérable à ces deux problèmes. En effet, la chance n'a pas joué un grand rôle dans l'épreuve de français du MÉQ (du moins dans celle que nous avons utilisée) : rédiger un texte d'au moins 800 mots sur un thème prédéterminé à partir d'extraits fournis à l'étudiant au moment de l'épreuve ne semble pas laisser une grande place au hasard. En outre, en choisissant d'appliquer notre propre grille, nous avons mis tous les efforts pour éviter les biais — si tant est que de tels biais auraient existé — dont certains candidats auraient pu être victimes de la part des réviseurs du Ministère. Également, en montant notre grille de lecture à partir d'une grille s'appuyant, non pas sur une école linguistique particulière, mais sur un corpus de textes produits par des rédacteurs occasionnels, la grille de Guénette, Lépine et Roy (1995), nous croyons avoir contourné assez le problème du biais théorique accompagnant souvent de tels outils de mesure. Finalement, en effectuant plusieurs lectures successives de chacun des textes, nous pensons avoir réussi à limiter les effets des oublis ou des erreurs de traitement, qui auraient pu intervenir au hasard et limiter la validité de nos résultats.

Quant à l'effet de l'ordre des questions sur le résultat des épreuves, ou de notre lecture de ces épreuves, le problème ne se pose pas, puisque le test de français du MÉQ n'est pas un questionnaire.

Il reste bien entendu toute la question de notre corpus de textes professionnels. Nous savons déjà que, n'ayant pas été tiré selon les règles de l'échantillonnage statistique, ce corpus ne peut déboucher que sur des conclusions indicatives. C'est

d'ailleurs pourquoi nous avons préféré élaborer notre étude mathématique en nous limitant à notre échantillon statistiquement significatif. Nous savons aussi que notre critère de sélection de textes professionnels — tout texte informatif ou argumentatif publié de longueur comparable à celle des sujets Moffet — peut porter à discussion. Nous sommes aussi tout à fait conscients que le choix d'un texte parmi d'autres, dans un journal ou un magazine, pourrait être discutable. Cependant, comme nous travaillions avec l'hypothèse que l'expertise linguistique se manifeste sous forme d'indices repérables dans tout texte produit par des rédacteurs maîtrisant le français écrit, nous n'étions pas préoccupés nécessairement par un échantillonnage statistique pour ce groupe de sujets. Encore une fois, il importerait de confirmer ou d'infirmer les résultats de notre recherche à partir d'un autre bassin de textes publiés.

Le deuxième aspect à examiner d'un œil critique concerne le pouvoir d'extrapolation de notre étude. En effet, jusqu'à quel point peut-on attribuer à la population de rédacteurs occasionnels en général les conclusions de notre étude des indices de contrôle linguistique? Nous sommes assez sûrs que ces conclusions s'appliquent au moins à la population des textes produits par les quelque 16 200 rédacteurs occasionnels d'où notre échantillon a été tiré (Moffet, 1998). Nous ne pouvons toutefois pas prétendre que l'écriture de ces 16 000 personnes est représentative de celle de la francophonie. Par conséquent, tout un champ d'études consisterait, d'une part, à confirmer ou infirmer notre démarche auprès de nouvelles populations et, d'autre part, à effectuer un portrait linguistique automatiquement détectable de populations de rédacteurs occasionnels provenant d'autres régions de la francophonie.

Nous voulions, par la présente recherche, étudier les conditions pouvant mener à l'automatisation de la correction des erreurs de ponctuation. Nous pouvons aujourd'hui apporter certains éléments de réponse.

L'application automatique de la ponctuation constitue le champ de recherche à développer. En effet, les erreurs de ponctuation les plus fréquentes se trouvant des erreurs d'omission de signes, notamment de virgule, corriger les erreurs de ponctuation revient à introduire des signes absents. En outre, la problématique réelle

de la révision automatique de la ponctuation ne vise pas tant à corriger les erreurs de ponctuation qu'à repérer les textes où les erreurs de ponctuation les plus sérieuses sont susceptibles de se produire. Dans ces textes, corriger la ponctuation demande qu'on efface toute la ponctuation présente et qu'on la remplace par la ponctuation appropriée. Encore une fois, c'est l'introduction automatique des signes de ponctuation, particulièrement du point assertif, qui constitue la difficulté majeure.

Par ailleurs, la définition usuelle de l'erreur de ponctuation — tout écart par rapport à une norme — nous apparaît contre-productive en correction automatique de la ponctuation. En effet, comme nous l'avons fait ressortir dans notre problématique, plus que le nombre d'erreurs de ponctuation, un élément capital séparait nos rédacteurs occasionnels de nos rédacteurs experts sur le plan de la ponctuation : ils appliquaient les signes de façon cohérente, ce qui nous permet d'inférer l'influence d'un système commun de règles, d'une école normative. C'est pourquoi, nous arrivons au terme de cette recherche à une définition de l'erreur de ponctuation s'adaptant davantage à la réalité observée dans notre corpus.

Nous proposons ainsi qu'un texte pourra être considéré avoir échoué sur le plan de la ponctuation quand aucune école normative n'y sera reconnaissable. Dans le cas où une école prescriptive est toutefois reconnaissable, toute incohérence par rapport à cette norme constituera une erreur de ponctuation. Selon nos observations en effet, les rédacteurs occasionnels faibles appliquent les signes de ponctuation de façon incohérente : ils ne suivent aucune règle précise et ne privilégient donc pas d'école prescriptive particulière. Voilà ce que nous considérons comme un échec en ponctuation. Quant aux autres rédacteurs, il nous apparaît peu important qu'ils choisissent la méthode de ponctuer recommandée par Drillon (1991), Colignon (1993) ou Ramat (1989) s'ils appliquent les règles de leur grille préférée de façon cohérente. Dans cette optique, le défi du grammairien défendant une grille de ponctuation devient donc, non pas de défendre un standard possible sur la scène linguistique internationale, mais plutôt de démontrer que cette grille repose sur des

règles d'application suffisamment opérationnelles pour être reprises avec cohérence par des rédacteurs occasionnels[106].

Mais notre recherche soulève plusieurs questions nouvelles.

Peut-être n'est-il pas nécessaire, par exemple (ou même réaliste), de détecter systématiquement toutes les erreurs d'un texte. Nous considérons traditionnellement toutes les erreurs grammaticales présentant un « poids » égal sur la « droite » symbolique de la correction linguistique. Mais peut-être n'est-ce pas le cas. Un rédacteur confondant par exemple *c'est, ces, ses, s'est, sais* ou *sait* générera des problèmes d'interprétation sérieux aussi bien pour un correcteur humain que pour le correcteur d'un traitement de texte. Tel ne sera pas le cas pour un rédacteur ne présentant pas du tout ce problème. Avons-nous donc raison de traiter tous les problèmes morpho-syntaxiques de français écrit sur le même pied? Peut-être vaudrait-il mieux limiter les interventions d'un correcteur informatique aux champs qu'il peut bien contrôler et réduire d'autant les risques d'hypercorrections et de détections manquantes. Peut-être même y a-t-il des textes qu'il serait préférable de ne pas tenter de corriger automatiquement.

Jusqu'à présent, les correcteurs grammaticaux ont proposé une correction universelle à partir d'un repérage d'erreurs tout azimut. Nous croyons, considérant les problèmes posés par l'analyse automatique robuste (Briscoe, 1996b; Dale, 1996), que cette avenue constitue probablement une impasse et que le calibrage présente de meilleures chances de réussite.

Bien entendu, tout cela reste encore à débattre. Il reste que notre approche, ne représentant qu'une ébauche dans un champ de recherche complètement nouveau, ouvre la porte à des applications pratiques aussi complètement nouvelles, dont la

---

106 À cet égard, les grilles de ponctuation se fondant sur la syntaxe nous apparaissent beaucoup plus utiles que celles qui s'appuient sur des critères aussi aléatoires et intangibles que, par exemple, la respiration. Nous sommes conscients d'ouvrir ici la proverbiale boîte de Pandore en faisant exactement ce que nous nous étions juré de ne pas faire : nous mêler de querelles théoriques n'ayant jamais trouvé de résolution satisfaisante jusqu'à aujourd'hui. Cependant, en finissant cette recherche, nous ne pouvons nous empêcher de nous demander s'il ne faudrait pas voir un lien entre la popularité de la thèse prosodique, toujours reproduite partout dans les traités de ponctuation et les grammaires — et donc les manuels scolaires —, et l'incohérence observée dans la ponctuation de notre corpus de rédacteurs occasionnels.

mesure automatique du rendement linguistique de rédacteurs et l'adaptation des correcteurs grammaticaux aux besoins individuels des utilisateurs.

BIBLIOGRAPHIE

AÏT-MOKHTAR, Salah et CHANOD, Jean-Pierre. 1997. "Incremental Finite-State Parsing ". Dans *Proceedings of ANLP'97*. Washington. March 31st to April 3rd. p. 72-79. http://www.xrce.xerox.com/publis/mltt/mlttart.html

ALLAIRE, Denis. 1998. *Comprendre la statistique. Manuel d'autoformation. Volume 1.* Prix Adrien-Pouliot, Association mathématiques du Québec. Québec : Presses de l'Université du Québec.

ALLEN, James F. 1993. "Natural language, knowledge representations, and logical form" in Bates, Madeleine and Weischedel, Ralph M., ed . *Challenges in Natural Language Processing*. Cambridge, Royaume-Uni : Cambridge University Press, Studies in Natural Language Processing. p 147-148.

ALLWOOD, Jens et ANDERSSON, Lars-Gunnar et DAHL, Östen. 1995. *Logic in Linguistics*. (Logik for Lingvister). Cambridge, Great Britain : Cambridge University Press. 185 p. (coll. « Cambridge Textbooks in Linguistics »).

ARMSTRONG, Susan, Ed. 1994. *Using Large Corpora*. Cambridge, Massachusetts et London, England : The MIT Press, 349 p.

AUTHIER, Jacqueline. 1979. "Parler avec des signes de ponctuation ou : de la typographie à l'énonciation", O.R.L.A.V., Mélanges de syntaxe et sémantique, no 21 (novembre), Paris : Université de Paris VIII, Centre de Recherche, p. 76 - 87.

BAKER, Sheridan. 1973. *The Practical Stylist*, Third Edition, New York : Thomas Y . Crowell Company, 182 p.

BARKO, I. 1977. "Contribution à l'étude de la ponctuation française au XVIIe siècle. Problèmes de méthode, La ponctuation de Racine". Dans *La Ponctuation : recherches historiques et actuelles*. CATACH, Nina, édit. Paris et Besançon : CNRS et Groupement de recherches sur les textes modernes, p. 59-126.

BATES, Madeleine et WEISCHEDEL, Ralph M., ed. 1993. *Challenges in Natural Language Processing*. Cambridge, Massachussets : Cambridge University Press, 296 p. (coll. « Studies in Natural Language Processing »).

BEAUD, Michel et LATOUCHE, Daniel. 1988. *L'Art de la thèse. Comment préparer et rédiger une thèse, un mémoire ou tout autre travail universitaire.* [Saint-Laurent], Québec : Boréal, 169 p.

BÉDARD, Édith et MAURAIS, Jacques. 1983. *La Norme linguistique*, [s.l.] : Gouvernement du Québec, Conseil de la langue française, 850 p. (coll. «L'ordre des motes», Le Robert, Paris) .

BESSON, Robert. 1987. *Guide pratique de la communication écrite, avec exercices et corrigés*. Paris : Éd. Casteilla, 192 p.

BLANQUET, Marie-France. 1994. *Intelligence artificielle et système d'information : le langage naturel*. Paris : ESF Éditeur, 269 p. (coll. « Systèmes d'information et nouvelles technologies »).

BOUCHARD, Jacques B. et HUDON, Jean-Denis et LAVOIE, Thomas. 1992. *Guide de présentation d'un travail de recherche*, 5e édition revue et augmentée. Chicoutimi : Module des lettres, Maîtrise en études littéraires, Maîtrise en linguistique, Université du Québec à Chicoutimi, 92 p.

BRISCOE, Ted et CARROLL, John. 1994. "Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars." IN *Using Large Corpora*. ARMSTRONG, Susan, ed. Cambridge, Massachusetts et London, England : The MIT Press, p. 25 - 59.

BRISCOE, Ted. 1994. "Parsing (with) Punctuation etc..." Grenoble : Rank Xerox Research Laboratory, Multilingual Theory and Technology. MLTT-TR-002.

BRISCOE, Ted. 1996a. "The Syntax and Semantics of Punctuation and its Use in Interpretation". In *Punctuation in Computational Linguistics. Proceedings*. Santa Cruz : Sigparse 96, p. 1-8.

BRISCOE, Ted. 1996b. "Robust Parsing". In *Survey of the State of the Art in Human Language Technology*. Ronald A. Cote, Editor in Chief. CSLU Home Page : National Science Foundation European Commission, chapter 3.7.

BRUN, J. et DOPPAGNE, A. [1958]. *La Ponctuation et l'art d'écrire*. Bruxelles, Belgique : Beaude, 240 p. (coll. «Bien écrire et bien parler») .

BUREAU, Conrad. 1978. *Syntaxe fonctionnelle du français*. Québec, Canada : Les Presses de l'Université Laval, 246 p.

BUTLER, Christopher. 1985. *Computers in Linguistics*. Oxford, United Kingdom : Basil Blackwell, 266 p.

CAJOLET-LAGANIÈRE, Hélène *et al.* 1983. *Rédaction technique*. Sherbrooke, Canada : Éditions Laganière, 281 p.

CALLAMAND, Monique. 1987 et 1989. *Grammaire vivante du français. Français, langue seconde*. [s.l.] : Librairie Larousse et Clé international, 252 p.

CARRÉ, René *et al.* 1991. *Langage humain et machine*. Paris : Presses du C.N.R.S., 298 p.

CARROLL, J et BRISCOE, E. J et GROVER, C. 1991. « A Development Environment for Large Natural Language Grammars ». *Technical Report 233*. Cambridge, United Kingdom : Cambridge University Computer Laboratory.

CATACH, Nina *et al.* 1980b. *L'Enseignement de l'orthographe (L''alphabet phonétique international, la typologie des fautes, la typologie des exercices), Formation initiale et continue*. Paris : Éditions Fernand Nathan, 96 p. (coll. «Dossiers Didactiques Nathan») .

CATACH, Nina, édit. 1977. *La Ponctuation : recherches historiques et actuelles.* Paris et Besançon : CNRS et Groupement de recherches sur les textes modernes, 272 p.

CATACH, Nina, édit. 1979. *La Ponctuation : recherches historiques et actuelles, Fascicule deux.* Paris et Besançon : CNRS et Groupement de recherches sur les textes modernes, 292 p.

CATACH, Nina, édit. 1980a. *Langue française, La Ponctuation*, nº 45, 128 p. Février.

CATACH, Nina, édit. 1990. *Pour une Théorie de la Langue écrite.* Paris : Éditions du Centre national de la recherche scientifique, 259 p.

CATACH, Nina. 1980c. "La ponctuation", Dans *Langue française. La Ponctuation.* CATACH, Nina, édit.. nº 45, p. 16 - 27. Février.

CATACH, Nina. 1987. "Rôle historique de la ponctuation : la virgule et les propositions incidentes au XVIIIe siècle". Dans *Langages*, nº 88, p. 31 - 40. Décembre.

CATACH, Nina. 1989. *Les délires de l'orthographe.* Paris : Plon, 349 p.

CATACH, Nina. 1994. *La Ponctuation.* Paris : Presses Universitaires de France, 128 p. (coll. « Que sais-je ? », no 2818).

CENTRE GEORGES POMPIDOU, édit. 1988. "Le Génie de la ponctuation". Dans *Traverses 43, Revue du Centre de Création industrielle,* 170 p. Février.

CHANDIOUX, John. 1996. *À propos de Météo et de détection automatique d'erreurs.* Entrevue. Montréal, Québec : Groupe Chandioux. Bande audio. 90 min..

CHANOD (Jean-Pierre) : 1993, PROBLÈMES DE ROBUSTESSE EN ANALYSE SYNTAXIQUE dans *Actes de ILN 93.*

CHANOD, Jean-Pierre et TAPANAINEN, Pasi. 1994. "Tagging French - Comparing a Statistical and a Constraint-based Method". Grenoble : Rank Xerox Research Laboratory, MultiLingual Theory and Technology. MLTT-016. http://www.xrce.xerox.com/publis/mltt/mltttech.html

CHANOD, Jean-Pierre et TAPANAINEN, Pasi. 1996a. "Rules and Constraints in a Finite-State Grammar". Grenoble : Rank Xerox Research Laboratory, MultiLingual Theory and Technology. MLTT-024. http://www.xrce.xerox.com/publis/mltt/mltttech.html

CHANOD, Jean-Pierre et TAPANAINEN, Pasi. 1996b. "A Lexical Interface for Finite-State Syntax". Grenoble : Rank Xerox Research Laboratory, MultiLingual Theory and Technology. MLTT-025. http://www.xrce.xerox.com/publis/mltt/mltttech.html

CHARNIAK, Eugene. 1993. *Statistical Language Learning.* Cambridge, Massachusetts et London, England : The MIT Press, 170 p.

CHEVALIER, Jean-Claude *et al.* 1964. *Grammaire Larousse du français contemporain.* Paris : Larousse, 494 p. (coll. «Références Larousse Langue française»).

CLAS, André et HORGUELIN, Paul A. 1979. *Le Français, langue des affaires*, 2e édition, Préface de Robert Dubuc. Montréal, Canada : McGraw-Hill Éditeurs, 391p.

COHEN, Paul R. 1995. *Empirical Methods for Artificial Intelligence.* Cambridge, Massachusets et London, England : The MIT Press, 405 p.

COLIGNON, Jean-Pierre. 1988a. *La Ponctuation (art et finesse).* Paris : Éditions Éole, 96 p.

COLIGNON, Jean-Pierre. 1988b. "La ponctuation : un point, ce n'est pas tout !". Dans *Traverses 43, Revue du Centre de Création industrielle.* CENTRE GEORGES POMPIDOU, édit. p 71 - 79 . Février.

COLIGNON, Jean-Pierre. 1993. *Un point, c'est tout ! La ponctuation efficace.* Montréal, Canada : Boréal, 119 p.

DAGNAUD-MACÉ, Pierre et SYLNÈS, Georges. 1978. *Le Français sans faute.* Préface d'André Rougerie. Paris : Hatier, 159 p. (coll. «Profil Formation» nº 311 / 312) .

DALE, Robert. 1990. "A Rule-Based Approach to Computer-Assisted Copy-Editing". In *CALL.* Volume 2. p. 59-67.

DALE, Robert. 1991. "Exploring the Role of Punctuation in the Signalling of Discourse Structure". In *Proceedings of a Workshop on Text Representation and Domain Modelling : Ideas from Linguistics and AI.* Berlin, Allemagne : Technical University Berlin, p. 110-120. Octobre.

DALE, Robert. 1996. "Computer Assistance in Text Creation and Editing". In *Survey of the State of the Art in Human Language Technology.* Ronald A. Cote, Editor in Chief. CSLU Home Page : National Science Foundation European Commission, chapter 7.5.

DAMOURETTE, Jacques. 1939. *Traité moderne de ponctuation.* Paris : Larousse, 144 p.

DAOUST, François, LAROCHE, Léo et OUELLET, Lise *et al.* 1994. *Le projet SATO-CALIBRAGE.* Montréal : Centre de Recherche en Cognition et Information ATO-CI et Université du Québec à Montréal, 191 p. (« Cahier de recherche : 3 »).

DAVID, Michel. 1984. *Dis-moi.* Montréal et Toronto, Canada : Guérin, 315 p. et 39 fiches (coll. «Clé» dirigée par Anne-Marie Connolly) .

DE BRAY, Alain et THERRIEN, Michel. 1980. *Nouveau Code grammatical.* Montréal, Canada : Breault et Bouthilliers, 283 p.

DEFAYS, Jean-Marc et ROSIER, Laurence et TILKIN, Françoise, Éds. 1998. *À qui appartient la ponctuation?* Préface de Marc Wilmet. Paris et Bruxelles: Éditions Duculot, 465 p. (Coll. Champs linguistiques — Recueils).

DEMANUELLI, Claude. 1987. *Points de repère, Approche interlinguistique de la ponctuation français-anglais.* Travaux LVIII. Paris : Centre Interdisiciplinaire d'Études et de Recherches sur l'Expression contemporaine, Université de Saint-Étienne, 279 p.

DISTER, Anne. 1998. « Problématique des fins de phrase en traitement automatique du français ». Dans *À qui appartient la ponctuation?.* DEFAYS, Jean-Marc, ROSIER, Laurence et TILKIN, Françoise, éds. Paris et Bruxelles: Éditions Duculot, p 437-447.

DOPPAGNE, Albert. 1978. *La Bonne Ponctuation : clarté, précision, efficacité de vos phrases,* Paris - Gembloux : Duculot, 112 p.

DOPPAGNE, Albert. 1984. *La Bonne Ponctuation : clarté, précision, efficacité de vos phrases.* Deuxième édition revue. Paris - Gembloux : Duculot, 112 p.

DRILLON, Jacques. 1991. *Traité de la ponctuation française.* Paris : Éditions Gallimard, 472 p. (coll. «Tel») .

DRUIDE INFORMATIQUE. 1997. *Antidote.* Version 1.1.3. Macintosh. CD-rom. Montréal, Québec. Canada.

DUBOIS, Jean *et al.* 1973. *Dictionnaire de linguistique.* Paris : Larousse, 516 p.

DUBOIS, Jean et DUBOIS - CHARLIER, Françoise. 1970. *Éléments de linguistique française : syntaxe.* Paris : Librairie Larousse, 295 p. (coll. «Langue et langage») .

DULIÈRE, André. 1988. *Les Secrets de la langue française.* Lausanne, Suisse et Montréal, Canada : Guérin littérature, 396 p.

FISCHER, Maurice et HACKQUARD, Georges. 1959. *À la Découverte de la grammaire française.* [Paris] : Librairie Hachette, 538 p.

FONAGY, Ivan. 1980. "Structures sémantique des signes de ponctuation". Dans *BSLP*, nº 75, p. 95-129.

GALICHET, Georges. 1967. *Grammaire structurale du français moderne.* Montréal, Canada : Éditions HMH, 248 p.

GAZDAR, Gerald and MELLISH, Chris. 1990. *Natural Language Processing In Prolog. An Introduction to Computational Linguistics.* Wokingham, England : Addison-Wesley Publishing Company, 504 p.

GEORGIN, René. 1952. *Difficultés et finesses de notre langue.* Nouvelle édition revue et augmentée. Paris : Éditions André Bonne, 336 p.

GOBBE, Roger et TORDOIR, Michel. 1986. *Grammaire française.* Adapté pour le Québec par Pierre Filion. Saint-Laurent, Canada : Éditions du Trécarré.

GODAERT, Paul. 1975. *Rédiger dans les Affaires.* Louvain, Belgique : Vander, 343 p.

GREVISSE, Maurice et GOOSSE, André. 1991. *Nouvelle Grammaire française.* 2e édition revue. Paris-Louvain-la-Neuve : Duculot et Éditions du renouveau pédagogique inc., 377 p.

GREVISSE, Maurice. 1980. *Le Bon Usage.* 11e édition. Préface de Paul Robert. Paris-Gembloux : Duculot et Éditions du renouveau pédagogique, 1519 p.

GRISHMAN, Ralph. 1994. *Computational Linguistics. An Introduction.* Cambridge, Great Britain : Cambridge University Press, 193 p. (coll. « Studies in Natural Language Processing »).

GUÉNETTE, Louise et LÉPINE, François et ROY, Renée Lise. 1995. *Le français tout compris. Guide d'autocorrection du français écrit.* Saint-Laurent, Canada : Les Éditions du Renouveau Pédagogique, 114 p.

HARRIS, Zellig S. 1976. *Notes du cours de syntaxe.* Traduit de l'anglais par Maurice Gross. Paris : Éditions du Seuil, 237 p.

HARRISON, P., ABNEY, S., BLACK, E., FLICKENGER, D., GDANIEC, C., GRISHMAN, R., HINDLE, D., INGRIA, B., MARCUS, M., SANTORINI, B. et STRZALKOWSKI, T. 1991 . "Evaluating syntax performance of parser/grammars of English." In *Proceedings of the Workshop on Evaluating Natural Language Processing Systems*. ACL.

HUMPHREYS, R. Lee. 1992-1993. "Book Reviews : Geoffrey Nunberg. The Linguistics of Punctuation". In *Machine Translation*. Volume 7, No 3, p. 199-201.

IDE, Nancy M. 1987. *Pascal for the Humanities*. Philadelphia, Pennsylvania : University of Pennsylvania Press, 375 p.

JONES, Bernard. 1994a. *Can Punctuation Help Parsing*? Esprit Acquilex-II Working Paper No. 29. Cambridge University Computer Laboratory. U.K. cide@cup.cam.ac.uk. July.

JONES, Bernard. 1994b. "Exploring the Role of Punctuation in Parsing Natural Text". In *Proceedings of the 15th International Conference on Computational Linguistics*, Kyoto, Japan : COLING-94, p. 421-425.

JONES, Bernard. 1995. "Exploring the Variety and Use of Punctuation". In *Proceedings of the 17th Annual Cognitive Science Conference*. Pittsburgh, Pensylvania : CogSci 1995, p. 619-624.

JONES, Bernard. 1996a. "Towards Testing the Syntax of Punctuation". In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*. Santa Cruz, California : Association for Computational Linguistics, 363-365.

JONES, Bernard. 1996b. "Towards a Syntactic Account of Punctuation". In *Proceedincs of the 17th International Conference of Computational Linguistics* (COLING-96). Copenhagen, Denmark : Coling-96.

JONES, Bernard, 1996c. *What's The Point? A (Computational) Theory of Punctuation*. Edinburgh, Scotland : University of Edinburgh, 163 p.

JONES, Bernard, 1996d. *Proceedings of the ACL/SIGPARSE International Meeting on Punctuation in Computational Linguistics*. Technical Report HCRC/WP2, Human Communications Research Centre, University of Edinburgh, UK.

KETTUNEN, Kimmo. 1996. "Low-Level Typographical Spellchecking: A Proposal". Dans *Computers and the Humanities*. 30: 77-84. http://www.wkap.nl/oasis.htm/106233

LAROUSSE, édit. 1977. *La Linguistique*. Paris : Larousse, 255 p. (coll. «Encyclopoche Larousse»).

LAURENCE, Jean-Marie. 1976. *Grammaire française*. Montréal, Canada : Guérin, 565 p.

LE GAL, Étienne. 1933. *Apprenons à ponctuer. Pourquoi, Comment il faut ponctuer*. Paris : Librairie Delagrave, 118 p.

LÉONARD, L. 1965. *La Pratique de la rédaction. Classes du premier cycle des lycées et collèges d'enseignement général*. Paris : Bordas, 287 p.

LEPAPE, Pierre. 1988. "Pour une poignée de virgules". Dans *Traverses 43, Revue du Centre de Création industrielle*. CENTRE GEORGES POMPIDOU, édit., p. 5 - 9 . Février.

LESAGE, René *et al.*. 1993. *Enquête sur l'état d'utilisation des outils informatisés d'aide à la rédaction dans les organisations. Document B-22.* Québec, Québec : Centre francophone de recherche en informatisation des organisations (CEFRIO), 117 p.

MICROSOFT CORPORATION. 1990-1995. Microsoft Word pour Windows 95. IBM-PC ou compatibles. CD-rom.

MILLER, J. 1985. *Semantics and Syntax (Parallels & Connections)*. Cambridge, Great Britain : University Press, 262 p.

MINISTÈRE DE L'ÉDUCATION. 1998. *Épreuve de français. Langue et littérature. Guide de correction. Année scolaire 1997-1998. Document de travail.* Gouvernement du Québec. Direction de l'enseignement collégial. Québec : www.meq.gouv.qc.ca/ens-coll/Eprv_uniforme/mfrancais.htm. 118 p.

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA SCIENCE. 1993. *Test de français. Épreuves de mars 1993 (Collèges). Analyse détaillée des erreurs.* Québec, Canada : Gouvernement du Québec. Direction générale de l'enseignement collégial. Direction de la recherche et du développement. Service des études et du développement des collèges, p. 30-33.

MOFFET, Jean-Denis. 1998. *Démonstration de la représentativité statistique du corpus Moffet.* Gouvernement du Québec. Direction générale de l'enseignement collégial. Direction de la recherche et du développement. Service des études et du développement des collèges. 7 juillet. 11 p.

MULLER, Charles. 1992a. *Initiation aux méthodes de la statistique linguistique.* Paris : Honoré Champion Éditeur, 185 p. (« Collection Unichamp : 32»).

MULLER, Charles. 1992b. *Principes et méthodes de statistique lexicale.* Paris : Honoré Champion Éditeur, 205 p. (« Collection Unichamp : 33»).

NUNBERG, Geoffrey. 1990. *The Linguistics of Punctuation.* Stanford, California : Center for the Study of Language and Information (CSLI), 141 p. («Lectures Notes»).

NUNBERG, Geoffrey. 1996. *Lexical Grammar and Text Grammar*. Santa Cruz, California : Sigparse 96.

OFLAZER, Kemal. 1996. "Error-tolerant Finite-state Recognition with Applications to Morphological Analysis and Spelling Correction." In *Computational Linguistics*, Volume 22, Number 1, p. 73-91. March.

PAILLET, Jean-Pierre et DUGAS, André. 1973. *Principes d'analyse syntaxique*. Montréal, Canada : Les Presses de l'Université du Québec, 223 p.

PALMER, David. 1994. SATZ – *An Adaptive Sentence Segmentation System.* Berkeley, CA : University of California, Computer Science Division (EFCS). December. Report No. UCB/CSD-94-846.

PALMER, David D. et HEARST, Marti A. 1997. "Adaptive Multilingual Sentences Boundary Disambiguation". Dans *Computational Linguistics* . Volume 23, Number 2. June: 241-267.

PASQUES, Liselotte. 1977. "Ponctuation à l'écrit, arrangement rythmique à l'oral d'un conte de Marcel Jouhandeau lu par l'Auteur". Dans *La Ponctuation : recherches historiques et actuelles.* CATACH, Nina, édit. Paris et Besançon : CNRS et Groupement de recherches sur les textes modernes, p. 189 - 222.

PERCHERON, Daniel 1988. "Un point c'est tout". Dans *Traverses 43, Revue du Centre de Création industrielle.* CENTRE GEORGES POMPIDOU, édit., p. 151 - 157. Février.

PERROT, Jean. 1980. "Ponctuation et fonctions linguistiques". Dans *Langue française, La Ponctuation.* CATACH, Nina, édit. nº 45, p. 67 - 76. Février.

PURNELLE, Gérald. 1998. « Théorie et typographie : une synthèse des règles typographiques de la ponctuation » Dans *À qui appartient la ponctuation?.* DEFAYS, Jean-Marc, ROSIER, Laurence et TILKIN, Françoise, éds. Paris et Bruxelles: Éditions Duculot, p 211-221.

RAMAT, Aurel. 1989. *Grammaire typographique.* 4e édition mise à jour. Saint-Lambert, Canada : Aurel Ramat éditeur, 93 p.

REICHGELT, Hans. 1991. *Knowledge Representation, an AI Perspective.* Norwood, New Jersey : Alex Publishing Corporation, 251 p.

RICHARDSON, Stephen D. 1994. "Bootstrapping Statistical Processing into a Rule-based Natural Language Parser." In *The Balancing Act : Combining Symbolic and Statistical Approaches to Language. Proceedings of the Workshop.* Las Cruces, New Mexico : New Mexico State University, 96-103. July. MSR-TR-95-48.

RICHAUDEAU, François. 1973. *Le langage efficace. Communiquer, Persuader, Réussir.* Paris : C.E.P.L., 300 p. (coll.«Marabout service» MS360) .

RIGAULT, André. 1971. *La Grammaire du français parlé. Recherches. Applications.* Paris : Librairie Hachette, 175 p.

RYAN, Thomas P. 1997. *Modern Regression Methods.* New York : Weily Inter-Science, 515 p. (Wiley Series in Probability and Statistics. Includes disk).

SABBAH, Gérard. 1990. *L'intelligence artificielle et le langage. Volume 1. Représentation des connaissances.* 2º édition. Paris : Hermes, 357 p.

SALKOFF, Morris. 1973. *Une grammaire en chaîne du français. Analyse distributionnelle.* Paris : Dunod Éditeur, 199 p. (coll. «Monographies de linguistique mathématique») .

SALTON, Gerard et McGILL, Michael J. 1983. *Introduction to Modern Information Retrieval.* New York, New York : McGraw-Hill, 448 p.

SAMPSON, Geoffrey. 1992. "Book Reviews : Geoffrey Nunberg : The Linguistics of Punctuation". In *Linguistics 30.* 467-475.

SAY, B. and AKMAN, V. 1997. "Current Approaches to Punctuation in Computational Linguistics". Dans *Computers and the Humanities* . 30: 457-469. http://www.wkap.nl/oasis.htm/140181.

SAY, B. and AKMAN, V. 1997. "An Information-Based Treatment of Punctuation in Discourse Representation Theory". Dans *Mathematical & Computational Analysis of Natural Language* . Carlos Martin-Vide, ed. Amsterdam et Philadelphia :JohnBenjamin. http://www.cs.bilkent.edu.tr/~akman/papers.html

SENSINE, Henri. 1930. *La Ponctuation en français*. Paris : Payot, 144 p.

SHAN, Harry. 1963. *Punctuate it Right!* New York, New York : Harper & Row, Publishers, 176 p.

SIMARD, Marthe. 1993. *Étude de la distribution de la virgule dans les phrases de textes argumentatifs d'expression française. Mémoire de maîtrise.* Québec, Canada : Université Laval. p. 69-85.

SIMARD, Marthe. 1996a. « Considerations on Parsing a Poorly Punctuated Text in French ». Dans *Sigparse 96. Punctuation in Computational Linguistics. Proceedings.* Bernard Jones, éd., Juin. Santa Cruz : University of California at Santa Cruz, CA., p. 67-72.

SIMARD, Marthe. 1996b. *La correction automatique de la ponctuation à fonction de délimiteur : perspectives et limites.* Projet de thèse de doctorat présenté à M. Jacques Ladouceur dans le cadre du programme de doctorat en linguistique. Québec : Département des langues et de linguistique. Faculté des lettres. Université Laval.

SIMARD, Marthe. 1997. « Du traitement automatique des erreurs de ponctuation en français écrit : le cas des correcteurs grammaticaux ». Dans *Revue Informatique et Statistique dans les Sciences humaines.* Liège, Belgique : Centre informatique de Philosophie et Lettres, Laboratoire d'Analyse statistique des Langues anciennes, Université de Liège. 33° année, n^{os} 1 à 4, p. 319 à 351.

SIMARD, Marthe et LADOUCEUR, Jacques. 1998. «Est-il possible de corriger automatiquement les erreurs de virgule?». Dans *À qui appartient la ponctuation?*. DEFAYS, Jean-Marc, ROSIER, Laurence et TILKIN, Françoise, éds. Paris et Bruxelles: Éditions Duculot, p 449-462.

THERRIEN, Michel. 1987. *Aide-mémoire grammatical.* Boucherville : Vézina Éditeur .

THIMONNIER, René. 1970 et 1974. *Code orthographique et grammatical.* Préface de Georges Matoré. Verviers : Librairie Hatier et Marabout, 442 p .

TOURNIER, Claude. 1977. "Essai de définition de la ponctuation et de classement des signes". Dans *La Ponctuation : recherches historiques et actuelles.* CATACH, Nina, édit. Paris : CNRS et Groupement de recherches sur les textes modernes, p. 223-243.

VAARLOOT, J. 1977. "Faisons le point". Dans *La Ponctuation : recherches historiques et actuelles.* CATACH, Nina, édit. Paris : CNRS et Groupement de recherches sur les textes modernes, p. 11 - 28.

VAILLOT, R. et MAÎTRE, R. 1969. *Grammaire fonctionnelle.* Paris : Librairie Eugène Belin, 295 p.

VÉDÉNINA, L. G. 1980. "La triple fonction de la ponctuation dans la phrase : syntaxique, communicative et sémantique". Dans *Langue française, La Ponctuation*. CATACH, Nina, édit. nº 45, p. 60 - 66. Février.

VÉDÉNINA, L. G. 1989. *Pertinence de la présentation typographique*. Avant-propos de Nina Catach. Paris : Peeters / Selaf, 153 p.

WAGNER, Robert Léon et PINCHON, Jacqueline. 1991. *Grammaire du Français classique et moderne*. [Paris] : Hachette, 688 p. (coll. «HU, Langue française, Hachette supérieur»).

WARD, Nigel. 1994. *A Connectionist Language Generator*. Norwood, New Jersey : Ablex Publishing Corporation, 298 p.

Annexe 1
Grille descriptive de la ponctuation du corpus

Difficulté d'application objective d'une grille normative

La détermination objective de l'erreur de ponctuation passe par l'application d'une grille nécessairement prescriptive. Voilà une contradiction qui soulève en soi deux problèmes fondamentaux : la qualité opérationnelle de la grille et un dépouillement objectif du corpus.

Qualité opérationnelle de la grille

Nous cherchions à décrire la ponctuation de notre corpus de la façon la plus exhaustive possible. C'est pourquoi nous avons inclus, dans notre grille, toutes les erreurs de ponctuation documentées par Guénette, Lépine et Roy (1995), et non pas seulement les erreurs de virgule.

Notre grille est en fait une base de données constituée de rubriques tombant dans deux grandes catégories : les rubriques de description externe et celles de description interne.

Rubriques de description externe

Outre un numéro de code assigné arbitrairement aux sujets, les champs de description externe précisent des données complémentaires utiles : pour les sujets non experts, leurs résultats aux épreuves de français du MÉQ; pour les sujets experts, le titre du texte examiné et le nombre de mots.

Rubriques de description interne

Les champs de description interne sont principalement des champs numériques[107] répartis en trois classes correspondant aux types d'erreurs de ponctuation identifiées par Guénette, Lépine et Roy (1995): l'omission de signes de ponctuation, l'occurrence indue de signes de ponctuation et la confusion de signes de ponctuation.

Omission de signes de ponctuation

L'omission de signes de ponctuation compte 10 rubriques, dont un champ alphanumérique et un champ de calcul, permettant de préciser la nature et la fréquence des omissions observées :

* omission du point;
* omission virgule séparateur[108];
* omission virgule délimiteur gauche;
* omission virgule délimiteur droit;
* omission virgule paire délimiteurs
* omission point-virgule;
* omission deux-points;
* omission autre signe;
* identification autre signe omis [champ alphanumérique];
* sous-total omissions [champ de calcul].

Occurrence indue de signes de ponctuation

La classe « signes indus » compte 13 rubriques incluant un champ alphanumérique et un champ de calcul . Contrairement à la grille de Guénette, Lépine et Roy cependant, certains des champs de description interne adoptent une terminologie générativiste pour décrire les contextes où sont relevées les erreurs :

* point indu;

---

107 Les données qui peuvent y être entrées ne peuvent être que des nombres.

108 (Jones, 1996c; Simard, 1993; Nunberg, 1990). Une virgule à fonction de séparateur permet d'éviter les redondances en se substituant à un mot ou à un syntagme: « J'ai mis au monde deux enfants et ma sœur, quatre » [virgule = *en a mis au monde*]; « La ferme comptait une écurie, une porcherie et une petite poulaillerie » [virgule = *et*].

* virgule indue entre SN et SN ou SV et SV *[entre deux syntagmes de nature semblable]*;
* virgule indue entre SN et SV *[entre le sujet et son verbe]*;
* virgule indue entre SV et SN ou P [entre le verbe et son complément d'objet, que ce complément soit un syntagme ou une proposition];
* virgule indue entre prép. et SN [entre la préposition et le complément qu'elle introduit];
* virgule indue entre SN et SP [entre le groupe nominal et son groupe prépositionnel];
* virgule indue entre SV et SP [entre le groupe verbal et son groupe prépositionnel];
* autre virgule indue;
* point-virgule indu;
* deux-points indu;
* autre signe indu;
* identification autre signe indu *[champ alphanumérique]*
* sous-total signes indus *[champ de calcul]*

Confusion de signes

La catégorie d'erreurs « Confusion de signes » se distribue en trois sous-catégories :

* confusion du point avec un autre signe;
* confusion de la virgule avec un autre signe;
* confusion d'autres signes avec des marques autres que le point ou la virgule.

Le tableau 87 présente la synthèse des erreurs décrites dans chacune de ces catégories; les italiques désignent les rubriques de calcul.

Tableau 87

Liste des erreurs de confusion de signes

Confusion du point avec un autre signe	*Confusion de la virgule avec un autre signe*	*Confusion d'autres signes entre eux*
Point au lieu de virgule	Virgule au lieu du point	Points de suspension après etc. au lieu du point
Point au lieu du point d'interrogation	Virgule au lieu d'un point expressif [? ou ! ou ...]	Point-virgule au lieu du deux-points
Point au lieu du point d'exclamation	Virgule au lieu de point-virgule	Deux-points au lieu du point-virgule
Point au lieu du deux-points	Virgule au lieu du deux-points	Délimiteurs ouvrants [i.e. (ou [ou {]
Autre signe[109] au lieu du point	Point-virgule au lieu d'une virgule	Délimiteurs fermants [i.e.) ou] ou }]
—	Deux-points au lieu d'une virgule	—
Sous-total erreurs confusion du point avec un autre signe	Sous-total erreurs confusion de la virgule avec un autre signe	Sous-total confusion autres signes entre eux
		Total Confusion de signes

Dépouillement objectif du corpus

Pour assurer la plus grande objectivité possible au dépouillement de notre corpus, nous avons tenté d'appliquer une méthodologie la plus rigoureuse possible[110].

Les lectures répétées des mêmes textes ont permis de dépouiller le corpus avec suffisamment de précision pour assurer à notre analyse une validité acceptable. Bien que les oublis soient encore possibles, nous ne croyons pas que leur nombre soit assez important pour modifier nos résultats de façon significative.

Les rubriques de calcul de notre base de données ont contribué par ailleurs à une meilleure objectivité lors du dépouillement. En effet, il nous a été possible de comptabiliser automatiquement les erreurs relevées dans un seul texte aussi bien que dans l'ensemble du corpus. Comme il n'était plus nécessaire pour nous de nous arrêter à ce détail en cours de

[109] Sauf la virgule. Nous avons considéré les erreurs impliquant la virgule comme une catégorie à part.

dépouillement, nous avons ignoré les résultats quantitatifs tout au long de l'exercice. Cette ignorance nous a ainsi évité la tentation de prendre des décisions susceptibles de favoriser une catégorie de textes plutôt qu'une autre.

---

[110] Voir Chapitre 3. Méthodologie.

Annexe 2
Textes du corpus Moffet selon le numéro d'identification des sujets

Sujet	*Pointage Syntaxe et ponctuation MéQ*	*Fautes d'orthograph e d usage et grammatical (MÉQ)*	*Fautes de syntaxe seulement (pointage non traité)*	*Fautes de ponctuation seulement (pointage non traité)*	*Nombre de mots du texte*
01	6	54	3	7	1000
02	4	26	0	8	1000
03	18	11	15	6	904
04	16	14	11	11	934
05	11	8	10	2	933
06	2	4	2	1	1050
07	1	5	0	3	854
08	5	13	2	7	1056
09	15	15	9	6	1390
10	15	48	11	8	844
11	15	24	9	12	810
12	10	13	8	2	1000
13	8	6	6	2	1000
14	4	1	3	3	960
15	2	13	2	0	844
16	21	7	7	5	486
17	8	15	6	5	835

Sujet	*Pointage Syntaxe et ponctuation MéQ*	*Fautes d'orthographe d usage et grammatical*	*Fautes de syntaxe seulement (pointage non traité;)*	*Fautes de ponctuation seulement (pointage non traité)*	*Nombre de mots du texte*
18	9	16	8	2	912
19	4	4	4	0	810
20	9	12	7	4	1119
21	4	2	3	2	960
22	3	8	2	2	965
23	4	4	3	2	1111
24	20	10	17	6	990
25	7	6	7	3	827
26	10	1	6	9	923
27	9	6	5	8	897
28	25	18	19	12	851
29	6	10	3	3	950
30	26	26	23	6	807
31	5	10	3	5	891
32	11	5	9	4	822
33	3	8	2	3	875
34	18	8	16	5	1000
35	3	4	3	0	990
36	9	4	8	2	815
37	12	8	10	5	1150
38	7	8	5	5	852
39	5	3	4	3	1070
40	8	8	7	3	1151
41	9	6	8	3	900
42	6	3	5	3	865
43	7	10	4	7	823
44	6	9	5	3	835
45	12	16	9	6	819
46	6	2	5	2	813
47	20	7	18	4	946
48	10	3	8	4	873
49	5	7	4	3	726
50	11	36	7	8	954
51	34	29	28	12	981
52	15	18	14	8	853
53	2	6	3	6	942
54	9	3	7	5	958
55	2	11	2	1	844
56	11	29	7	3	914
57	24	31	17	14	1008
58	4	0	3	1	918
59	5	12	3	4	1010
60	1	5	1	1	1200
61	6	9	6	0	825
62	29	44	23	6	977
63	10	11	5	5	905
64	5	8	3	2	913
65	4	17	2	5	839

Sujet	Pointage Syntaxe et ponctuation MéQ	Fautes d'orthographe d usage et grammatical	Fautes de syntaxe seulement (pointage non traité)	Fautes de ponctuation seulement (pointage non traité)	Nombre de mots du texte
67	4	5	3	2	833
68	26	11	23	3	839
69	7	14	4	6	915
70	4	2	4	1	1040
71	10	3	8	4	1452
72	5	18	3	5	974
73	8	12	3	10	887
74	3	2	0	3	733
75	8	9	4	8	998
76	2	3	2	0	1000

Annexe 3
Textes du corpus « *Experts* » selon leur numéro d'identification

Code	*Expert*	*Profession*	*Titre du texte*	*Référence*	*# mots*
101	Bourguina, Henri	Correspondant	*Des marchés financiers dans l'expectative*	(1981-1996) "Dossier 1993: La nouvelle donne de l'économie mondiale", Dossiers 1993, Éditions La Découverte-CEDROM-SNi: 93Aj230	1151
102	Venne, Michel	Correspondant et éditorialiste	*Une carte inutile. Le Québec n'a nul besoin d'une nouvelle carte d'identité*	Le Devoir, Éditorial, lundi 24 mars 1997	1016
103	Beauge, Florence	Correspondante spéciale	*Les ambitions pacifiques de Vancouver*	Le Monde diplomatique, Août 1996, pp 12-13	4794
104	Bouchard, Lucien	Politicien, premier ministre du Québec	*Calgary nous rapetisse, nous comprime et nous réduit*	Le Devoir, mercredi 17 septembre 1997	2423
105	Bovet, Philippe	Correspondant spécial	*Le Nunavut, ultime redécoupage du Canada?*	Le Monde diplomatique, Août 1997, p. 8	1835
106	Bozzini, Luciano	Sociologue	*Des élections et une campagne électorale bien précieuses*	Le Devoir, 16 juin 1997	1032
107	Chossudovsky, Michel	Correspondant spécial	*L'éclatement annoncé de la Confédération canadienne*	Le Monde diplomatique, décembre 1995, p. 25	2480

Code	*Expert*	*Profession*	*Titre du texte*	*Référence*	*# mots*
108	CornellierManon	Journaliste	*"Le gouvernement du Canada pourrait dire la même chose", répond Stéphane Dion*	Le Devoir, Le monde, mardi 30 septembre 1997	870
109	Dion, Jean	Correspondant	*Ménage à cinq*	Le Devoir, Politique, lundi 22 septembre 1997	1062
110	Dion, Stéphane	Politicien, mnistre canadien des affaires gouverne-mentales	*Nier et invoquer la pertinence du droit*	Le Devoir, mercredi 14 août 1997	1233
111	Emmanuelli, Claude	Professeur de droit international	*Aux frontières du réel*	Le Devoir, mardi 23 septembre 1997	1323
112	Lachapelle, Guy	Professeur de sciences politiques	*Le verdict sera-t-il impartial?*	Le Devoir, Agora, mardi 23 septembre 1997	1222
113	Landry, Bernard	Politicien, vice-premier ministre du Québec	*Tourner le dos aux principes démocratiques*	Le Devoir, Agora, mercredi 14 août 1997	1442
114	Morin, Claude	Politicien, ancien ministre, Québec	*Lettre ouverte aux premiers ministres anglophones*	Le Devoir, Agora, vendredi 5 septembre 1997	879
115	Munger, Benoît	Journaliste	*Écrire des secrets*	Le Devoir, novembre 1993	1517
116	O'Neill, Pierre	Correspondant spécial, sondeur	*Entre la méfiance et la confiance*	Le Devoir, vendredi 19 septembre 1997	827
117	Parazelli, Michel	Correspondant spécial	*De la pauvreté traitée comme une maladie*	Le Monde diplomatique, décembre 1995, p. 25	1545
118	Ramonet, Ignacio	Correspondant spécial	*Québec et mondialisation*	Le Monde diplomatique, avril 1996, p. 1	857
119	Parizeau, Jacques	Politicien, ancien premier ministre du Québec	*La déclaration unilatérale est indispensable*	Le Devoir, Agora, mardi 16 septembre 1997	1848
120	Rioux, Christian	Journaliste	*Chirac réitère son appui au Québec*	Le Devoir, Le monde, mardi 30 septembre 1997	1089
121	Sansfaçon, Jean-Robert	Éditorialiste	*Nulle lueur en vue*	Le Devoir, Politique, mardi 2 septembre 1997	1316

Code	*Expert*	*Profession*	*Titre du texte*	*Référence*	*# mots*
122	Salwyn, André	Chroniqueur, haute technologie	*La vidéoconférence prête à prendre son essor*	Le Devoir, Internet, lundi 21 juillet 1997	1021
123	Coulon, Jocelyn	Journaliste	*Les certitudes d'un ministre*	Le Devoir, Perspectives, lundi 17 février 1997	831
124	Leduc, Louise	Chorniqueuse, culture	*Déluge de petits ensembles*	Le Devoir, Rentrée culturelle, samedi 23 août 1997	1284
125	Francoeur, Louis-Gilles	Journaliste	*Montréal, Far West du CFC*	Le Devoir, mardi 9 septembre 1997	1434
126	Lemieux, Louis-Guy	Historien	*Un jardin d'Éden*	Le Soleil, dimanche 11 mai 1997	1421
127	Bélair, Michel	Chroniqueur, haute technologie	*La simeuse saga du DVD*	Le Devoir, Bits Bauds et Pixels, mardi 2 août 1997	896
128	Champagne, Maurice	Humaniste	*Ili est minuit moins trois sur la réserve du peuple québécois*	Le Devoir, jeudi 23 octobre 1997	1493
129	Trudel, Rémy	Politicien, ministre des Affaires municipales du Québec	*Un engagement à respecter*	Le Devoir, mardi 30 septembre 1997	1359
130	David, Michel	Journaliste	*Entre l'URSS et Charlie Brown*	Le Soleil, jeudi 28 août 1997	820
131	Rheault, Ghislaine	Journaliste	*Le roi des tarlas*	Le Soleil, jeudi 28 août 1997	851
132	Binette, Pierre	Professeur, histoire et sciences politiques	*La roulette russe du partitionisme*	Le Devoir, samedi 27 septembre 1997	1429
133	Defert, Daniel	Correspondant	*Le sida, quinze ans après son identification*	in "Questions stratégiques 1996', L'État du monde, CEDROM-SNI: 96A8	1887
134	Bertrand, Maurice	Correspondant	*Peut-on transformer le monde par les organisations internationales?*	in "Dossier 1995: L'état des organisations internationales", Dossiers 1995, L'état du monde CEDROM-SNI: 95A247	1829
135	Chemillier-Gendreau, Monique	Correspondante	*Droit international, droit des États, droit des peuples*	in "Dossier 1995: L'état des organisations internationales", Dossiers 1995, L'état du monde, CEDROM-SNI: 95A248	1563

Code	*Expert*	*Profession*	*Titre du texte*	*Référence*	*# mots*
136	Ghebali, Victor-Yves	Correspondant	*Sécurité collective ou défense collective? - Le choix de l'Europe*	in "Dossier 1995: L'état des organisations internationales", Dossiers 1995, L'état du monde, CEDROM-SNI: 95A249	1487
137	Smouts, Marie-Claude	Correspondante	*Bretton Woods et l'ordre économique mondial*	in "Dossier 1995: L'état des organisations internationales", Dossiers 1995, L'état du monde, CEDROM-SNI: 95A250	1606
138	Habib, Ali	Correspondant	*Algérie:. Une "guerre sans chiffre"*	in "États", L'état du monde, CEDROM-SNI:Algérie96	1368
139	Pochoy, Michel	Correspondant	*La tension au Cachemire*	in "Conflits et tensions 1996", L'état du monde, CEDROM-SNI: 96A16	987
140	Urjewicz, Charles	Correspondant	*La nation tchéchène aux prises avec l'Histoire*	in "Conflits et tensions 1996", L'état du monde, CEDROM-SNI: 96A14	1458
141	Camroux, David	Correspondant	*Australie. Consolidation et redressement*	in "États", L'état du monde, CEDROM-SNI: Australie96	1338
142	Foucher, MIchel	Correspondant	*L'Union européenne à l'heure des élargissements*	in "Questions stratégiques 1996", L'état du monde, CEDROM-SNI: 96A7	1653
143	Sévigny, André	Historien à Parcs Canada, Gestion du patrimoine culturel, Québec	*Ces militaires qui ont peuplé la Nouvelle-France (1683-1715)*	"Vie sociale", Cap-Aux-Diamants, numéro 43, automne 1995, p. 10-13	2199
144	Lavoie, Elzéar	Professeur, Faculté des lettres, département d'histoire, Laval	*Le plébiscite de 1942: la déprime des défaites*	"Politique", Cap-Aux-Diamants, numéro 29, printemps 1992, p. 14-17	2651
145	Vachon, Christian	HIstorien de l'art	*Les recueils de Pierre-Georges Roy*	"Patrimoine", Cap-Aux-Diamants, numéro10, été 1987, p. 45-47	1201
146	Laroche Joli, Ginette	Historienne de l'art	*Des vitraux "Made in Quebec"*	"Arts", Cap-Aux-Diamants, numéro 7, automne 1986, p. 7-9	1553

Code	*Expert*	*Profession*	*Titre du texte*	*Référence*	*# mots*
147	Turgeon, Christine C.	Archiviste	*Des coffres et de constitutions: archives des premières communautés religieuses*	"Archives", Cap-Aux-Diamants, numéro 31, automne 1992, p. 22-25	2370
148	Lambert, James	Historien, dictionnaire biographique du Canada	*Daniel Wilkie: un humaniste au service de l'éducation*	"Portrait", Cap-Aux-Diamants, numéro 7, automne 1986, p. 17-20	1950
149	Dion, Jean-Noël	Archiviste, Séminaire de Saint-Hyacinthe	*Un politicien maudit: T.- D. Bouchard*	"Portrait", Cap-Aux-Diamants, numéro 30, été 1992, p. 38-40	1563
150	Gaumond, Michel	Archéologue au ministère de la Culture et des Communicatio ns	*Au coeur du Vieux-Québec: Le Cavalier du Moulin*	"Urbanisme", Cap-Aux-Diamants, numéro 37, printemps 1994, p. 26-27	932
151	Élémond, André	Politicologue	*Les élections à l'heure des médias*	"Journalisme", Cap-Aux-Diamants, numéro 30, été 1993, p. 42-45	2013
152	Lortie, André	Recherchiste, Radio-Québec	*Hôtel Roberval: tout le monde descend!*	"Bâtiment", Cap-Aux-Diamants, numéro 33, printemps 1993, p. 54-57	1605
153	Castonguay, Jacques	Psychologue et historien militaire, ancien recteur de l'académie militaire royal de Saint-Jean	*L'artillerie: une présence indissociable de l'histoire du Québec*	"Urbanisme", Cap-Aux-Diamants, numéro 43, automne 1995, p. 24-26	1465
154	Boutin, Jean-François	Étudiant-chercheur en didactique du français (littérature), Université Laval	*Rencontre avec onze écrivaines et écrivains*	"Éducation et francophonie", Québec français, Volume XXIV, numéros 1 et 2, 1996	3641
155	Baillargeon, Stéphane	Critique d'art	*Firida la noire*	"Rentrée culturelle", Le Devoir, samedi 23 août 1997	1164
156	Cauchon, Paul	Chroniqueur	*Un pavé dans la mare*	"Ordinateur et éducation", Le Devoir, lundi 25 août 1997	872
157	Tremblay, Odile	Chroniqueurse	*La manne hollywoodienne*	Le Devoir, mercredi 17 septembre 1997	822

Code	*Expert*	*Profession*	*Titre du texte*	*Référence*	*# mots*
158	Bourg, Dominique	Correspondant	*Les ruptures dangereuses de l'écologie non humaniste*	(1981-1996) "Questions stratégiques 1997", Éditions La Découverte- CEDROM-SNi: 97Aj11	1172
159	Moreau Defarges, Philippe	Correspondant	*Quelles sont les racines des conflits "post-guerre froide" ?*	(1981-1996) "Conflits et tensions 1997", Éditions La Découverte- CEDROM-SNi: 97Aj17	1153
160	Adam, Bernard	Correspondant	*Contrôler la prolifération et l'accumulation des armements légers*	(1981-1996) "Questions stratégiques 1997", Éditions La Découverte- CEDROM-SNi: 97Aj10	1548
161	Badie, Bertrand	Correspondant	*Mondialisation, les termes du débat*	(1981-1996) "Mondialisation, état des lieux", Éditions La Découverte- CEDROM-SNi: 94A238	1553
162	Barrère, Martine	Correspondante	*Environnement et développement - Le local, le global; le présent, le futur*	(1981-1996) "Mondialisation, état des lieux", Éditions La Découverte- CEDROM-SNi: 94A242	1304
163	Boyer, Robert	Correspondant	*Économie et finances internationales - Le temps des nations n'est pas fini*	(1981-1996) "Mondialisation, état des lieux", Éditions La Découverte- CEDROM-SNi: 94A239	1300
164	Moreau Defarges, Philippe	Correspondant	*Quelles sont les racines des conflits "post-guerre froide"?*	(1981-1996) "Conflits et tensions 1997", Éditions La Découverte- CEDROM-SNi: 97A17	1130
165	Delouvin, Patrick	Correspondant	*Libertés sans frointières? - Le cas dudroit d'asile en Europe*	(1981-1996) "Mondialisation, état des lieux", Éditions La Découverte- CEDROM-SNi: 94A243	1207
166	Donnet, Pierre-Antoine	Correspondant	*Tibet, une civilisation en péril*	(1981-1996) "Conflits et tensions 1997", Éditions La Découverte- CEDROM-SNi: 97A16	1301
167	Dorronsoro, Gilles	Correspondant	*Les mouvements islamistes au Kurdistan*	(1981-1996) "Conflits et tensions 1996", Éditions La Découverte- CEDROM-SNi: 96A18	946
168	Gentelle, Pierre	Correspondant	*Géopolitique interne de la Chine*	(1981-1996) "États 1997", Éditions La Découverte- CEDROM-SNi: 97E30	925

Code	*Expert*	*Profession*	*Titre du texte*	*Référence*	*# mots*
169	Haski, Pierre	Correspondant	*La situation palestinienne, clef de la paix au Proche-Orient*	(1981-1996) "Conflits et tensions 1997", Éditions La Découverte- CEDROM-SNi: 97A13	1514
170	Jean, François	Correspondant	*Tchétchénie, une guerre totale*	(1981-1996) "Conflits et tensions 1997", Éditions La Découverte- CEDROM-SNi: 97A15	1225
171	Mattelart, Armand	Correspondant	*Communication-monde - Culture mondiale ou système baroque?*	(1981-1996) "Mondialisation, état des lieux", Éditions La Découverte- CEDROM-SNi: 94A241	1212
172	Massiah, Gustave	Correspondant	*La planète des villes - Vers une civilisation urbaine*	(1981-1996) "Mondialisation, état des lieux", Éditions La Découverte- CEDROM-SNi: 94A244	1082
173	Mink, Georges	Correspondant	*Fin des certitudes et montée des inquiétudes*	(1981-1996) "États", Éditions La Découverte- CEDROM-SNi: Pologne92	1370
174	Bach, Daniel C.	Correspondant	*Le cap de 1992*	(1981-1996) "États", Éditions La Découverte- CEDROM-SNi: Nigéria92	1611
175	Meyer, Éric	Correspondant	*Une nouvelle donne?*	(1981-1996) "États", Éditions La Découverte- CEDROM-SNi: Sri Lanka96	1298

Rubriques de description interne

Les champs de description interne sont principalement des champs numériques[107] répartis en trois classes correspondant aux types d'erreurs de ponctuation identifiées par Guénette, Lépine et Roy (1995): l'omission de signes de ponctuation, l'occurrence indue de signes de ponctuation et la confusion de signes de ponctuation.

Omission de signes de ponctuation

L'omission de signes de ponctuation compte 10 rubriques, dont un champ alphanumérique et un champ de calcul, permettant de préciser la nature et la fréquence des omissions observées :

* omission du point;
* omission virgule séparateur[108];
* omission virgule délimiteur gauche;
* omission virgule délimiteur droit;
* omission virgule paire délimiteurs
* omission point-virgule;
* omission deux-points;
* omission autre signe;
* identification autre signe omis [champ alphanumérique];
* sous-total omissions [champ de calcul].

Occurrence indue de signes de ponctuation

La classe « signes indus » compte 13 rubriques incluant un champ alphanumérique et un champ de calcul . Contrairement à la grille de Guénette, Lépine et Roy cependant, certains des champs de description interne adoptent une terminologie générativiste pour décrire les contextes où sont relevées les erreurs :

---

107 Les données qui peuvent y être entrées ne peuvent être que des nombres.

108 (Jones, 1996c; Simard, 1993; Nunberg, 1990). Une virgule à fonction de séparateur permet d'éviter les redondances en se substituant à un mot ou à un syntagme: « J'ai mis au monde deux enfants et ma sœur, quatre » [virgule = *en a mis au monde*]; « La ferme comptait une écurie, une porcherie et une petite poulaillerie » [virgule = *et*].

Annexe 4
Textes du corpus *Bissonnette* selon leur numéro d'identification

Numéro du texte	*Titre*	*Référence* Le Devoir	*Nombre de mots*
01	*Pour un Quartier latin*	Les Arts, les samedi 25 et dimanche 26 octobre 1997, p. B 3	1440
02	*De piège en piège*	mardi 28 octobre 1997, p. A 6	1168
03	*Atmosphères*	Les Arts, samedi 18 et dimanche 29 octobre 1997, p. B 3	1360
04	*L'intolérable*	mardi 21 octobre 1997, p. A 8	882
05	*Encore du théâtre*	jeudi 16 octobre 1997, p. A 8	1064
06	*Images d'après-guerre*	Les Arts, samedi 11 et dimanche 12 octobre 1997, p. B 3	1224
07	*L'instinct de la Cour*	vendredi 10 octobre 1997, p. A 12	1064
08	*À l'écoute*	Les Arts, samedi 4 et dimanche 5 octobre 1997, p. B 3	1188
09	*Temps libre*	Les Arts, samedi 27 et dimanche 28 octobre 1997, p. B 3	1360
10	*Un défilé d'intérêts*	jeudi 25 septembre 1997, p. A 6	1120
11	*Une image de liberté*	Les Arts, samedi 20 et dimanche 21 septembre 1997, p. B 3	1224
12	*Revoir l'esquisse*	jeudi 18 septembre 1997, p. A 6	1260
13	*Les travaux forcés*	mardi 16 septembre 1997, p. A 6	1200
14	*La moyenne bibliothèque*	Les Arts, samedi 13 et dimanche 14 septembre 1997, p. B 3	1024
15	*Lire, écrire, et plus*	jeudi 11 septembre 1997, p. A 6	1300
16	*La sainte et le truand*	mardi 9 septembre 1997, p. A 6	1024
17	*Une journée chez V.L.B.*	Les Arts, samedi 6 et dimanche 7 septembre 1997, p. B 3	1320
18	*Après le rêve*	mercredi 3 septembre 1997, p. A 6	1024

Numéro du texte	*Titre*	*Référence* Le Devoir	*Nombre de mots*
19	*Les résistants*	vendredi 29 août 1997, p. A 8	1008
20	*La semence de la colère*	mercredi 27 août 1997, p. A 6	1456
21	*Dans les marges*	lundi 25 août 1997, p. A 6	1050
22	*Un intellectuel dans la cité*	vendredi 22 août 1997, p. A 8	1152
23	*Difficile déminage*	mercredi 20 août 1997, p. A 6	1072
24	*Le déclin d'un mythe*	vendredi 15 août 1997, p. A 8	1176
25	*Les vestiges de Montfort*	jeudi 14 août 1997, p. A 6	1040
26	*Opposition ou reddition?*	mardi 12 août 1997, p. A 6	1312
27	*Un acte manqué*	samedi 9 et dimanche 10 août 1997, p. A 8	1176
28	*Vers un Ordre des enseignants*	lundi 7 juillet 1997, p. A 6	1152
29	*De déshonneur en déshonneur*	vendredi 4 juillet 1997, p. A 8	1944
30	*Choc en vue*	jeudi 3 juillet 1997, p. A 6	1152
31	*Hors saison*	Les Arts, samedi 21 et dimanche 22 juiln 1997, p. B 3	1240
32	*Le passé présent*	Les Arts, samedi 11 et dimanche 12 juin 1997, p. B 3	896
33	*Le théâtre de l'absurde*	lundi 16 juin 1997, p. A 6	1184
34	*Le cabinet de la continuité*	jeudi 12 juin 1997, p. A 6	1206
35	*Les rumeurs canadiennes*	mercredi 25 juin 1997, p. A 6	1176
36	*Les rumeurs québécoises*	jeudi 26 juin 1997, p. A 8	1344
37	*La responsabilité des élus*	vendredi 27 juin 1997, p. A 8	1440
38	*En terrain miné*	Les Arts, samedi 7 et dimanche 8 juin 1997, p. B 3	1280
39	*L'itinéraire réformiste*	mardi 10 juin 1997, p. A 8	928
40	*Une transition*	mercredi 4 juin 1997, p. A 8	1104
41	*Le Canada réel*	mardi 3 juin 1997, p. A 1	704
42	*Objection de conscience*	Les Arts, samedi 31 mai et dimanche 1er juin 1997, p. B 3	1240
43	*Le symbole et la chose*	vendredi 30 mai 1997, p. A 8	1152
44	*Le Bloc québécois, pour le principe*	mercredi 28 mai 1997, p. A 6	2160
45	*Le mauvais oeil*	samedi 24 et dimanche 25 mai 1997, p. B 3	1476

Numéro du texte	*Titre*	*Référence* Le Devoir	*Nombre de mots*
46	*Les choix du 2 juin (1). La question nationale*	mardi 20 mai 1997, p. A 10	2064
47	*Les choix du 2 juin (2). Des mutations de valeurs*	mercredi 21 mai 1997, p. A 6	2056
48	*Nature, culture*	Les Arts, samedi 17 et dimanche 18 mai 1997, p. B 3	1008
49	*Notre berceau*	Les Arts, samedi 10 et dimanche 11 mai 1997, p. B 3	1080
50	*L'effet Parizeau*	jeudi 31 octobre 1996, p. A 1	736
51	*L'anasthésie*	mercredi 7 mai 1997, p. A 8	1040
52	*L'espoir avec réserves*	Les Arts, samedi 30 avril et dimanche 1 mai 1997, p. B 3	1080
53	*Un cauchemar*	Les Arts, samedi 26 et dimanche 27 avril 1997, p. B 3	1120
54	*Au ban de la morale*	lundi 14 avril 1997, p. A 6	1072
55	*La propagande*	vendredi 17 janvier 1997, p. A 8	972
56	*L'utopie Charest*	jeudi 20 mars 1997, p. A 6	944
57	*Tous pour une: la réponse*	Les Arts, samedi 19 et dimanche 20 avril 1997, p. B 3	1008
58	*L'ombre du passé*	mercredi 16 avril 1997	1104
59	*Jusques à quand?*	Les Arts, samedi 12 et dimanche 13 avril 1997, p. B 3	928
60	*Une réédition de 1982*	mercredi 9 avril 1997, p. A 6	984
61	*La TGBQ (bis)*	Les Arts, samedi 5 et dimanche 6 avril 1997, B 3	952
62	*Une aministie*	jeudi 3 avril 1997, A-6	938
63	*Ode au bon sauvage*	Les Arts, samedi 29 et dimanche 30 1997, B 3	1280
64	*Tous pour une*	jeudi 3 avril 1997, A-6	974
65	*Un exercice démagogique*	vendredi 14 mars 1997, A-10	1104
66	*Le handicap du PLQ*	mardi 11 mars 1997, A-8	1206
67	*Cherchez l'erreur*	Les Arts, samedi 8 et dimanche 9 mars 1997	1008
68	*Le PLQ en mal de crédibilité*	vendredi 7 mars 1997, A-10	1104
69	*Au coeur de la capitale*	jeudi 6 mars 1997, A-6	966
70	*FIFA*	Les Arts, samedi 1er et dimanche 2 mars 1997, B 3	1152
71	*La famine oubliée*	Les Arts, samedi 15 et dimanche 16 mars 1997. B 3	1152
72	*Un chef pour le Bloc québécois*	vendredi 28 février 1997, A-10	1350

Numéro du texte	*Titre*	*Référence* Le Devoir	*Nombre de mots*
73	*Une partition planétaire*	Les Arts, samedi 22 et dimanche 23 février 1997, B 3	1080
74	*Un monde coupé du monde*	mercredi 26 février 1997, A-10	1136
75	*En décomposition*	vendredi 21 février 1997, A-10	924
76	*Pour des esprits libres*	Les Arts, samedi 15 et dimanche 16 février 1997, B 3	1008