# Du traitement automatique des erreurs de ponctuation en français écrit : le cas des correcteurs grammaticaux

Marthe SIMARD

**Abstract.** We present the results of an analysis of the way in which popular grammar checkers handle punctuation errors. By observing the behaviour of these tools when confronted with certain punctuation problems, we can put forward an explanation of their failure in handling that category of spelling mistakes. This also allows us to better understand their diagnosis mistakes in the other writing domains they have to handle. On the other hand, as none of these checkers proved able to detect punctuation errors in a consequent or reliable fashion, we have no intention of asserting their respective qualities. We focus solely on their working mode and try to deduce from their failure some information that could be used in the development of a comma checker.

**Résumé.** Nous présentons les résultats d'une étude du traitement des erreurs de ponctuation par des correcteurs grammaticaux populaires. En examinant comment ces outils se comportent en face de certains problèmes de ponctuation, nous apportons une explication de leur échec dans le traitement de cette catégorie d'erreurs d'orthographe. Nous pouvons également comprendre davantage leurs erreurs de diagnostic dans les autres contextes d'écriture qu'ils sont appelés à analyser. Par ailleurs, comme chacun de ces correcteurs s'est révélé incapable de détecter et de corriger les problèmes de ponctuation de façon constante ou fiable, nous ne posons pas de jugement sur leur valeur respective. Nous nous intéressons exclusivement à leur fonctionnement et cherchons à tirer de leur échec des renseignements exploitables dans le développement possible d'un correcteur de virgules.

**Keywords:** Written French, punctuation, comma, automatic copy-editing, grammar checker, word processing, spelling checker, *Hugo*©, *Antidote*©.

**Mots-clés :** Français écrit, ponctuation, virgule, correction automatique, texteur, traitement de texte, vérificateur d'orthographe, *Hugo*©, *Antidote*©.

✉ Université Laval; Département des Lettres et de Linguistique; Cité universitaire; Québec (Québec) Canada, G1K 7P4.
E-mail : marthe@intouch.bc.ca

## 1. Introduction

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.

### 1.1. Quelques mots sur nos correcteurs grammaticaux

Nous avons retenu deux correcteurs grammaticaux : *Antidote*®[1] développé et mis en marché par la compagnie Druide informatique[2], et

[1] *Antidote*® est une marque déposée de Druide informatique.

[2] Pour toute information technique, consulter le site http://www.druide.com.

la grammaire française de *Word7*®[3], *Hugo*©, développée par Logidisque inc. et mise en marché par Softissimo.

### 1.1.1. *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

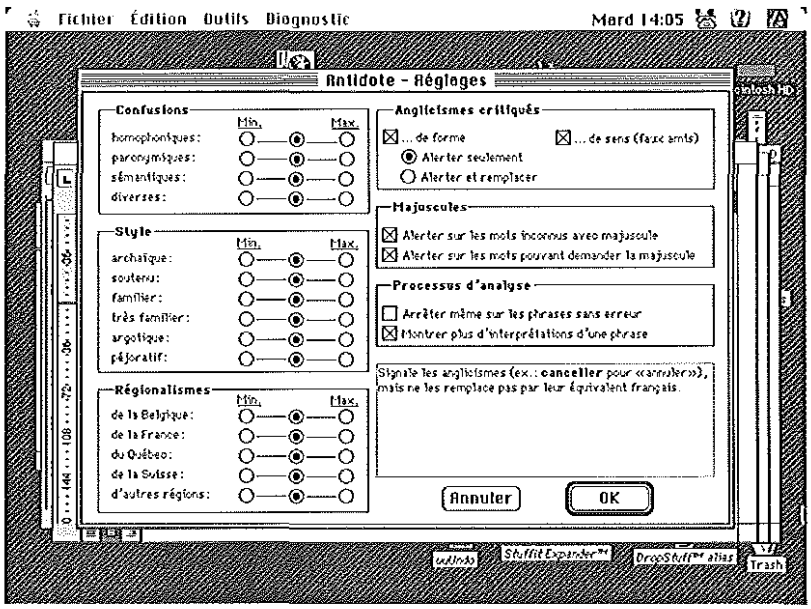


Fig. 1.

[3] *Word* est une marque déposée de Microsoft Corporation.

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*®[4], 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.

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. 1).

*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 ».

### 1.1.2. *Hugo*© dans *Word7*®

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. 2).

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.

---

[4] *Le Correcteur 101* est une marque déposée de Machina Sapiens.

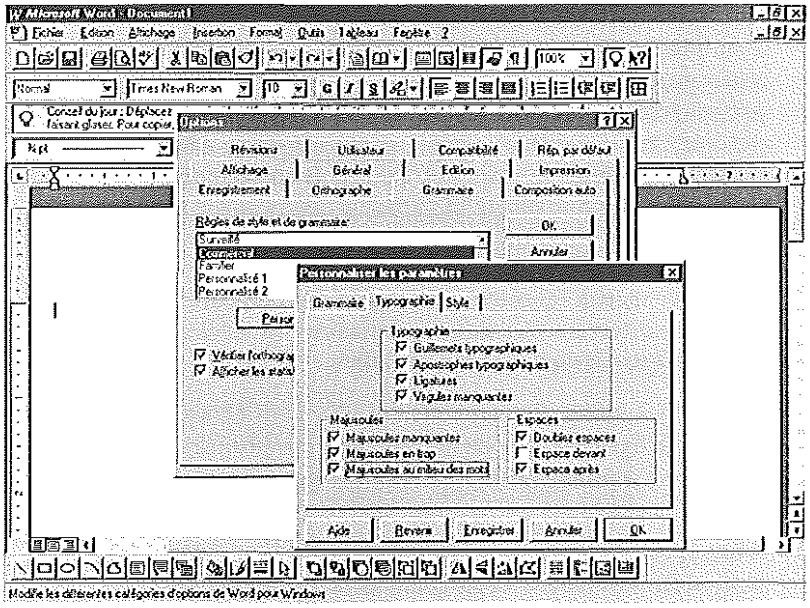

Fig. 2.

## 1.2. Quelques mots sur notre exercice

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 :

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

## 2. Les correcteurs segmentent les textes à partir de leur ponctuation sans la remettre en question

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.

### 2.1. 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. 3a et 3b).

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

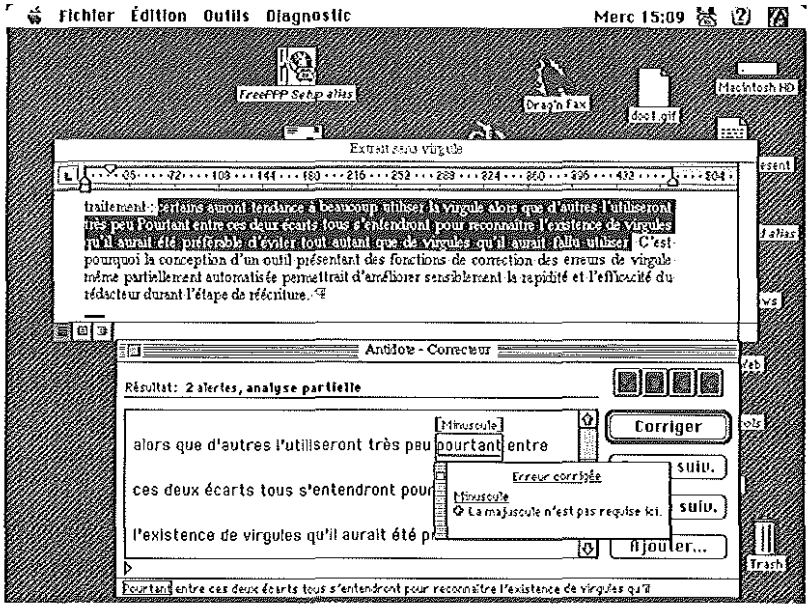


Fig. 3a.– *Antidote*©

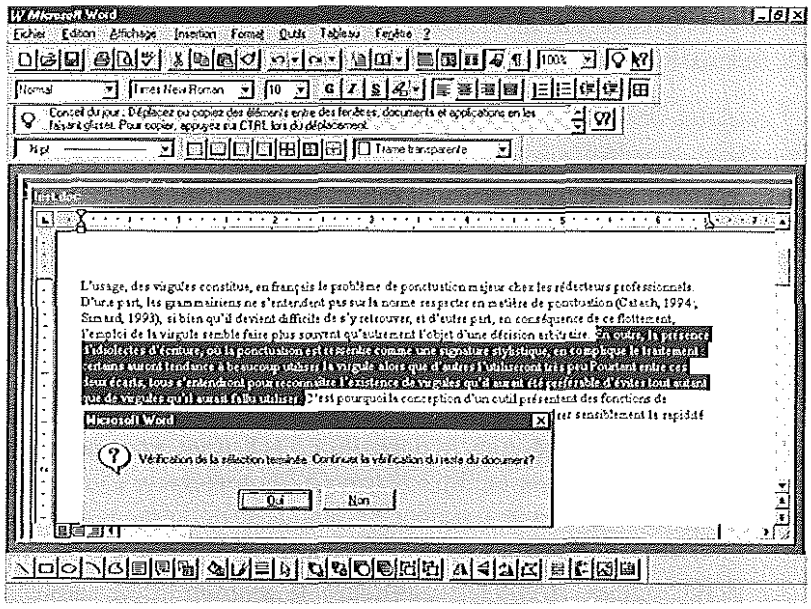


Fig. 3b.– La grammaire de *Word7*©

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.

## 2.2. 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. 4a et 4b).

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.

## 2.3. Introduction d'une virgule placée à tous les deux mots

Chacun de nos correcteurs accepte (2*) [fig. 5a et 5b], 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 :

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

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. 6a et 6b).

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

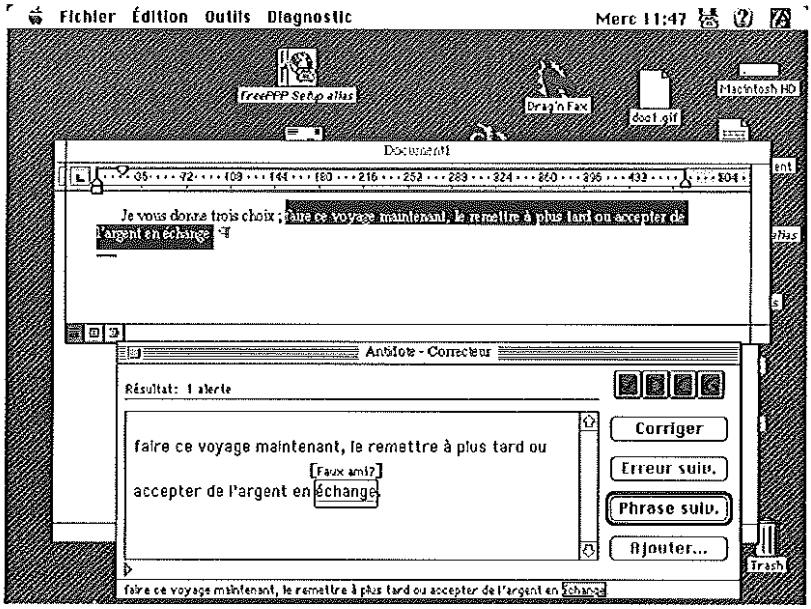


Fig. 4a.– *Antidote*®

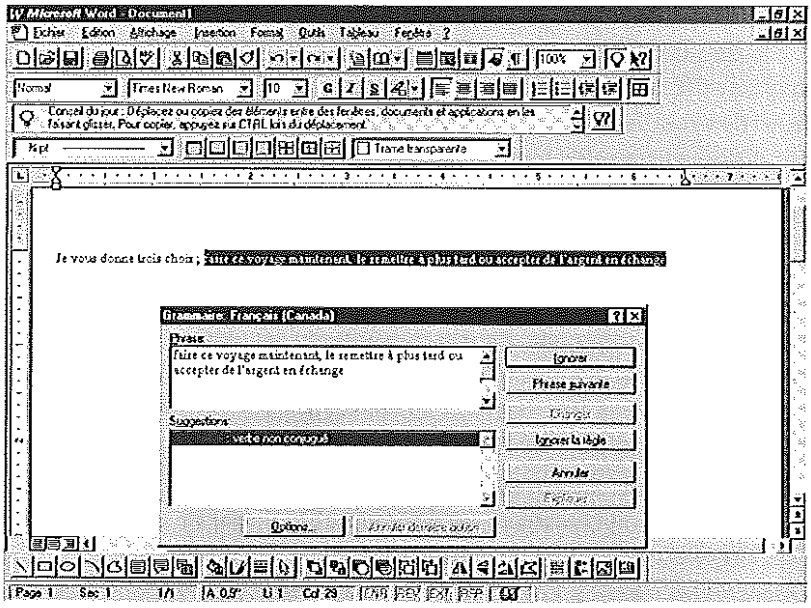


Fig. 4b.– La grammaire de *Word7*®

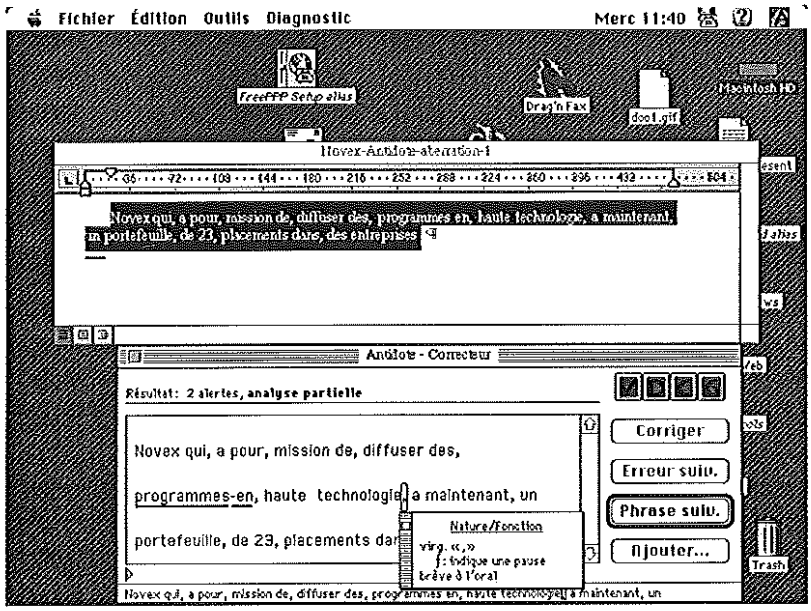


Fig. 5a.– *Antidote*®

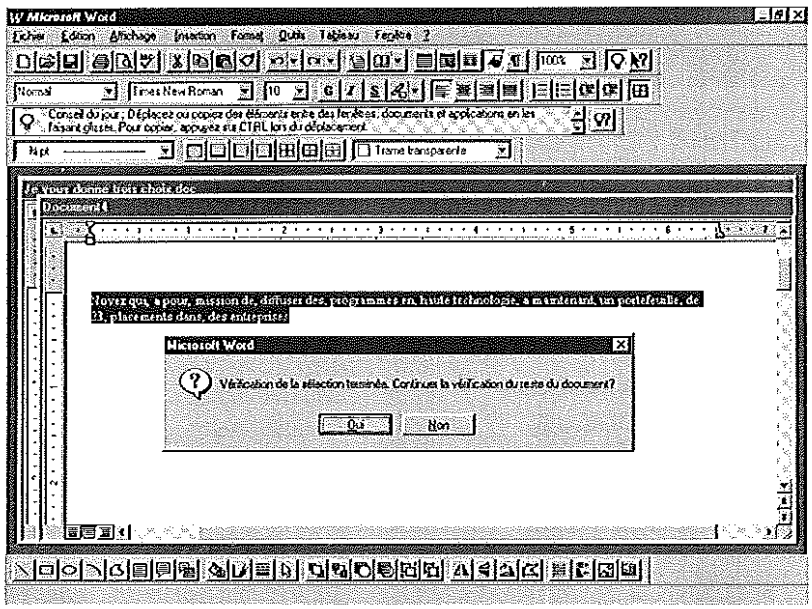


Fig. 5b.– La grammaire de *Word7*®

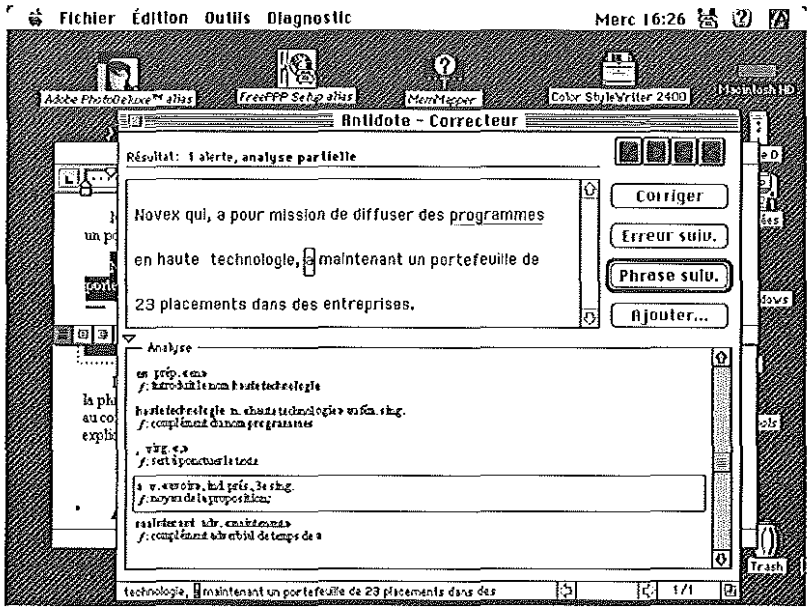


Fig. 6a.– *Antidote*®, analyse détaillée de (2)

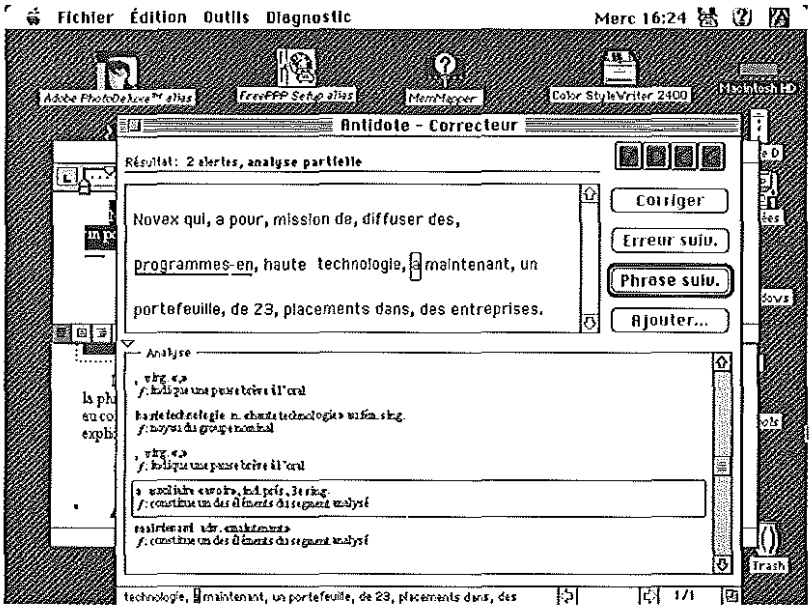


Fig. 6b.– *Antidote*®, analyse détaillée de (2*)

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[5] —, 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.

[5] 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.

## 3. Les correcteurs attribuent un pouvoir de segmentation à certains signes seulement

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.

### 3.1. 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.

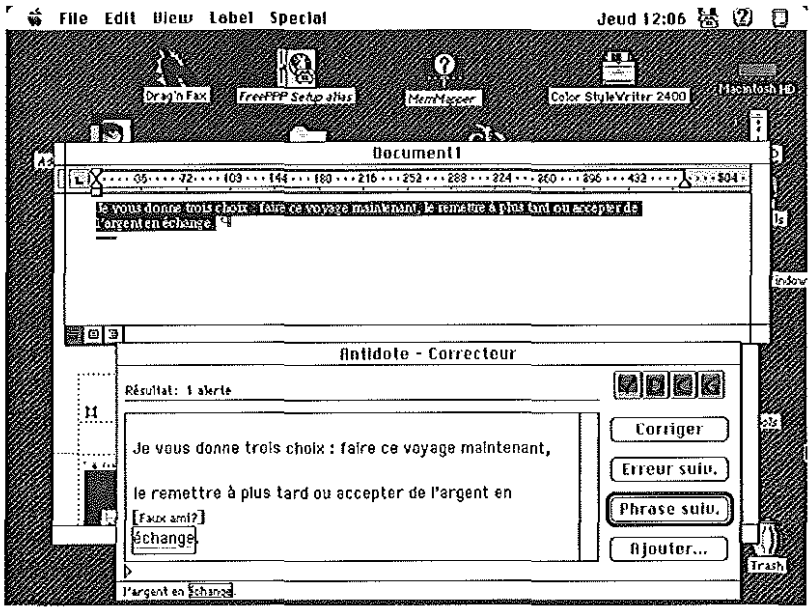


Fig. 7a.– *Antidote*®, suite (3) avec un deux-points

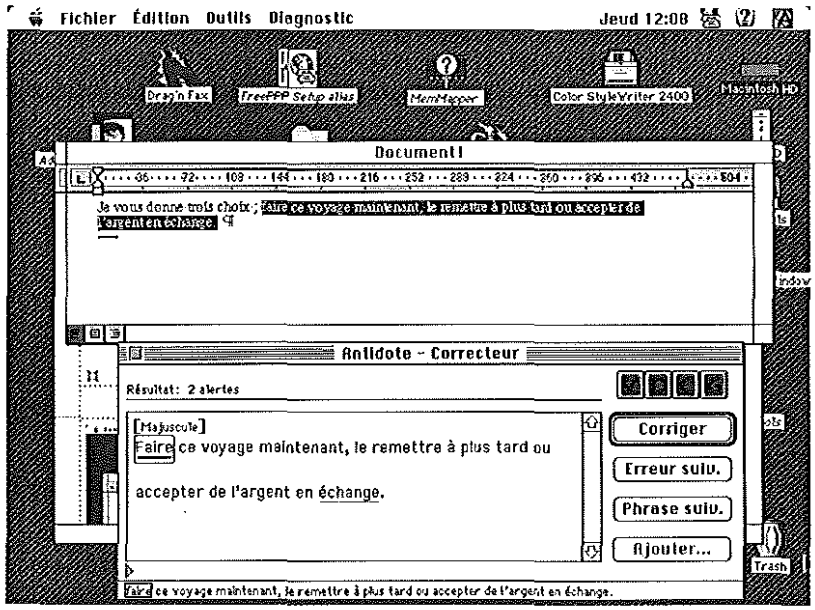


Fig. 7b.– *Antidote*®, suite (3) avec un point-virgule

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. 7a et 7b).

La grammaire de *Word7*® procède de même.

## 3.2. 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. 8a et 8b).

*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. 9a et 9b).

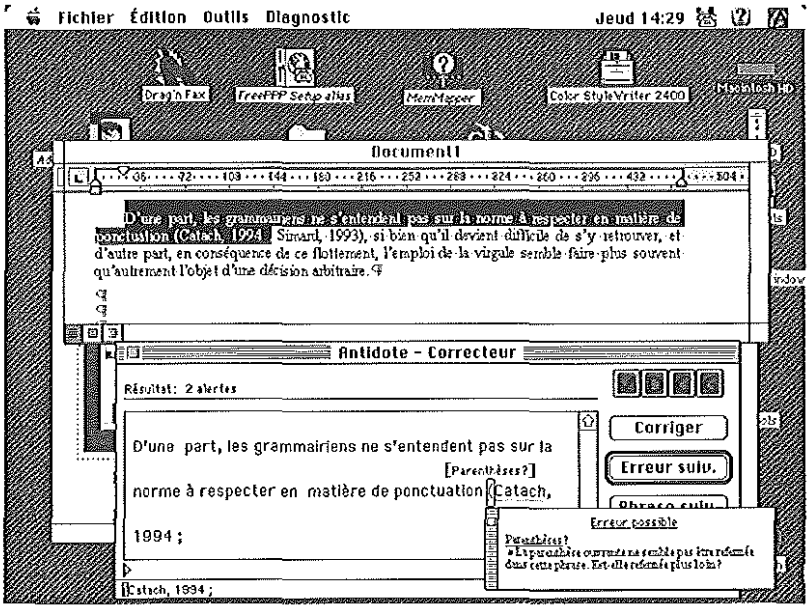


Fig. 8a.– *Antidote*©

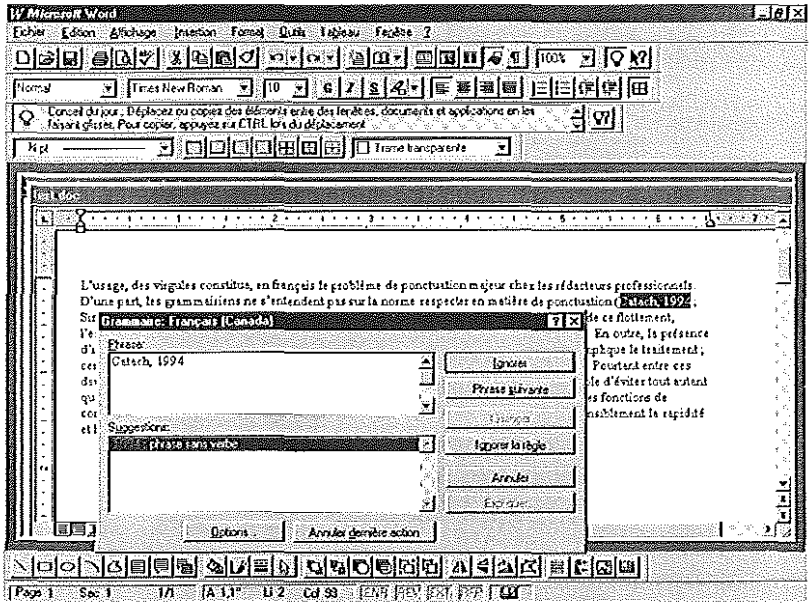


Fig. 8b.– La grammaire de *Word7*©

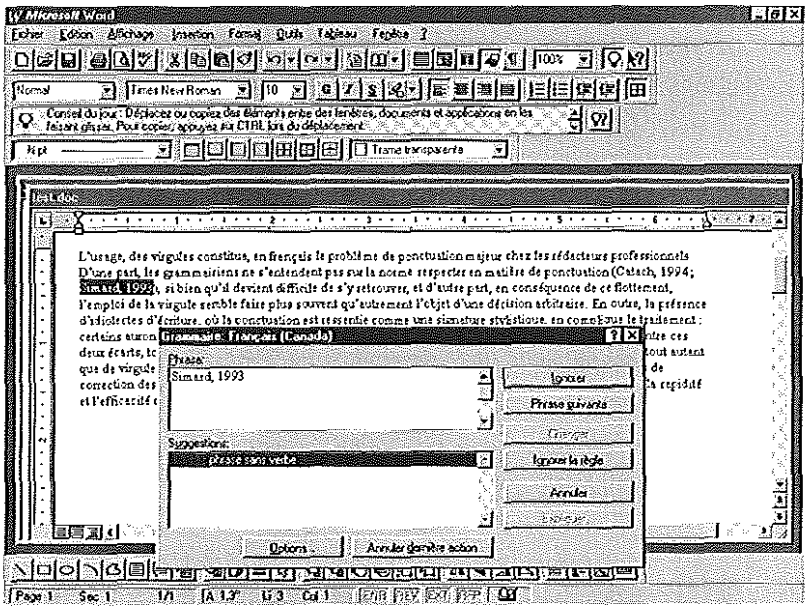


Fig. 9a.– La grammaire de *Word7*®, segmentation du point-virgule à la parenthèse fermante

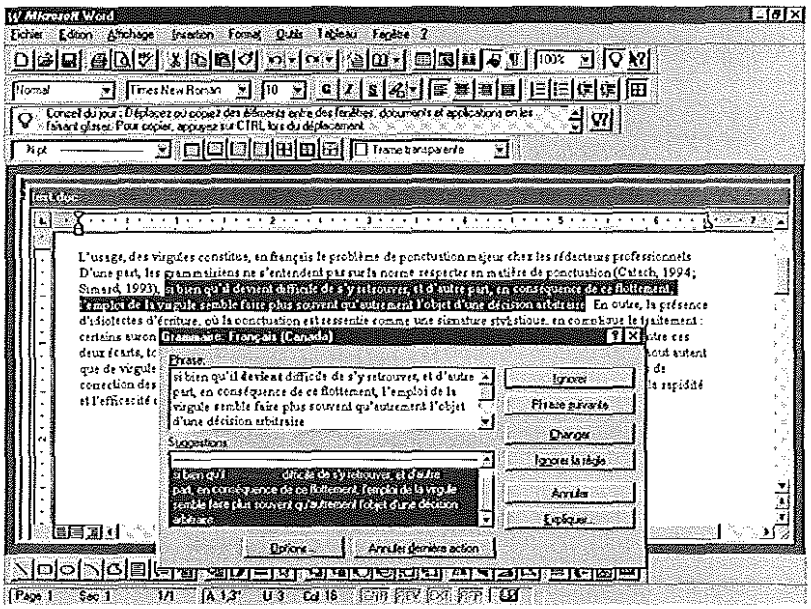


Fig. 9b.– La grammaire de *Word7*®, segmentation de la parenthèse fermante au point assertif

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.

### 3.3. Segmentation par signe de ponctuation : une vue d'ensemble

Le tableau 1 présente une synthèse de la segmentation généralement observée dans nos exemples. Le crochet signale une segmentation autour du signe de ponctuation ; le tiret, une non-segmentation.

Tableau 1

Synthèse de la segmentation observée dans les exemples proposés aux grammaires

| | *Antidote*® | Grammaire de *Word7*® |
|---|---|---|
| Point assertif | ✓ | ✓ |
| Autres points | ✓ | ✓ |
| Point-virgule | ✓ | ✓ |
| Parenthèses et autres signes doubles | – | ✓ |
| Deux-points | – | – |
| Virgule correcte | – | – |
| Virgule 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.

## 4. Les correcteurs ne peuvent généralement pas détecter avec fiabilité des 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*®.

### 4.3.1. Présence d'une virgule non requise

Une virgule non requise est introduite entre le sujet et son verbe (fig. 10a et 10b).

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 ».

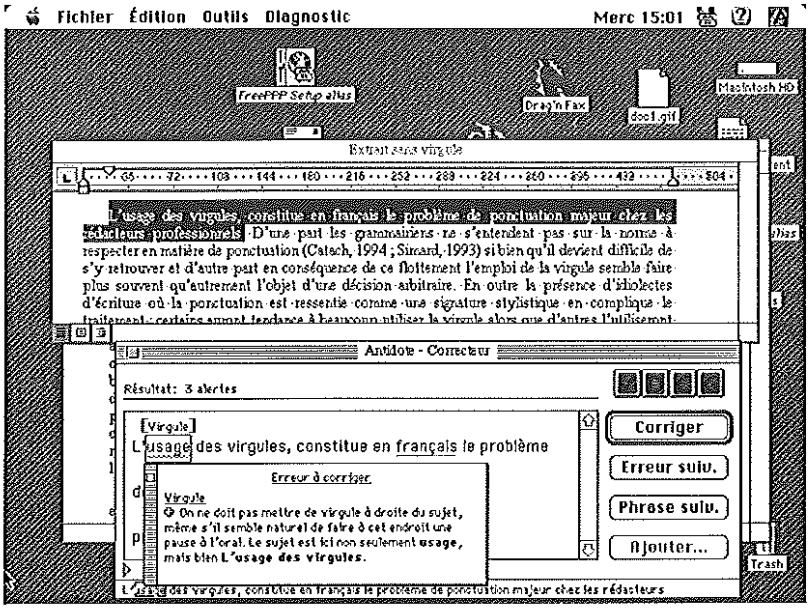


Fig. 10a.– *Antidote*®, virgule non requise introduite entre sujet et verbe, détection réussie

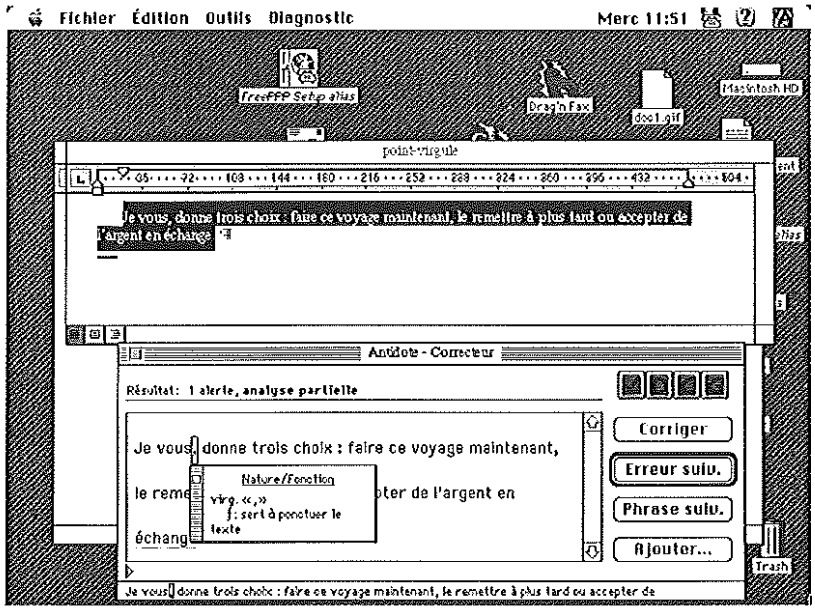


Fig. 10b.– *Antidote*®, virgule non requise introduite entre sujet et verbe, détection échouée

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

Cette fois-ci, *Antidote*® n'est plus en état de reconnaître l'erreur (fig. 11). 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.

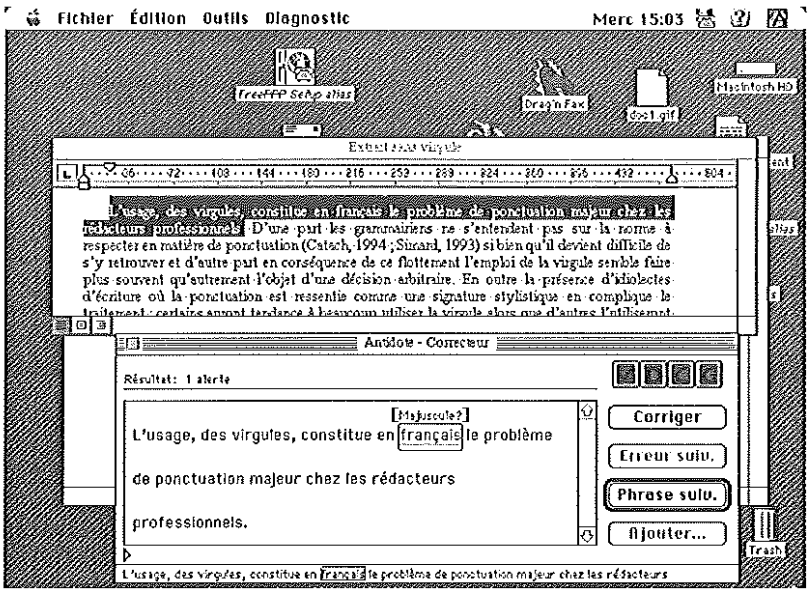


Fig. 11.– *Antidote*®, virgule non requise introduite entre sujet et verbe, détection échouée

## 4.2. 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 et l'un des membres d'une paire de virgules.

Le paragraphe (1) fournit plusieurs occurrences de phrases commençant par un connecteur (fig. 12a et 12b).

*Antidote*® repère la virgule manquante et explique correctement le contexte. Cependant, si la virgule requise après « d'autre part » est effacée, *Antidote*® ne s'en rend pas compte (fig. 13).

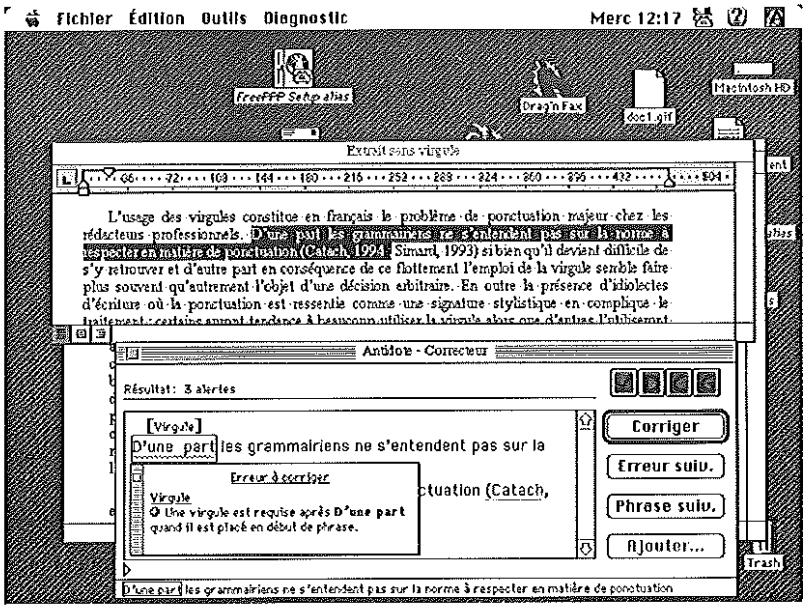


Fig. 12a.– *Antidote*®, virgule effacée après un connecteur placé en initiale de phrase, « D'une part » : détection réussie

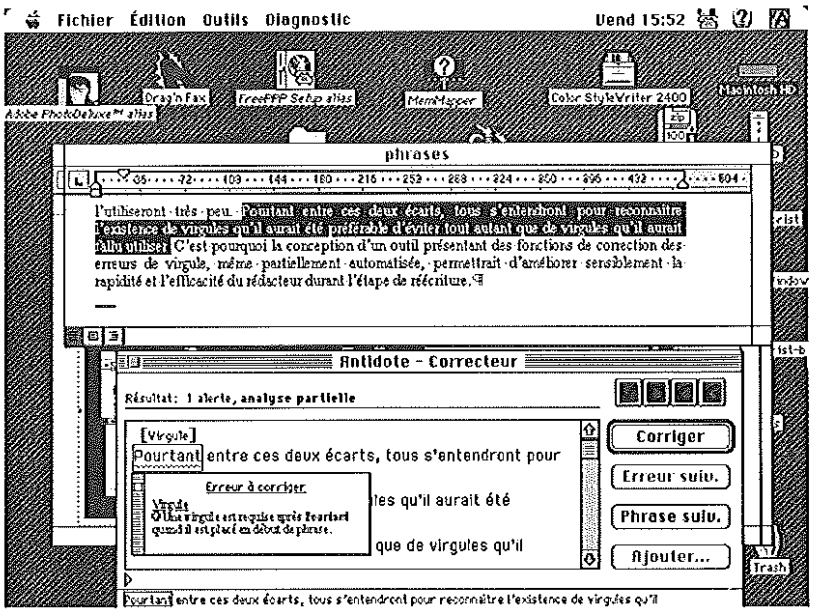


Fig. 12b.– *Antidote*®, virgule effacée après un connecteur placé en initiale de phrase, « Pourtant » : détection réussie

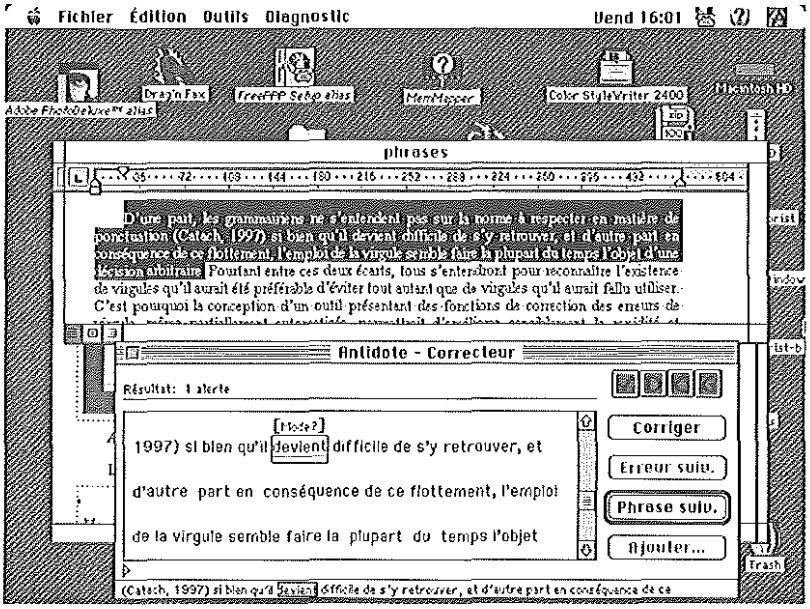


Fig. 13.– *Antidote*®, virgule effacée après « d'autre part » placé en médiane de phrase : détection échouée

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 virgule a été effacé dans la première phrase de (1) [fig. 14a et 14b].

*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.

## 4.3. 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. 15).

*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.

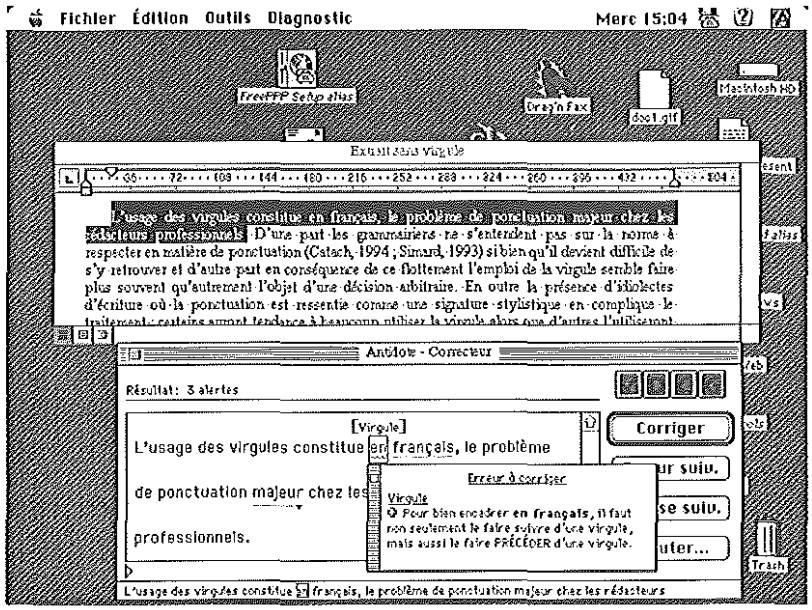


Fig. 14a.– *Antidote*®, effacement de l'un des membres d'une paire de virgules, virgule ouvrante effacée : détection réussie

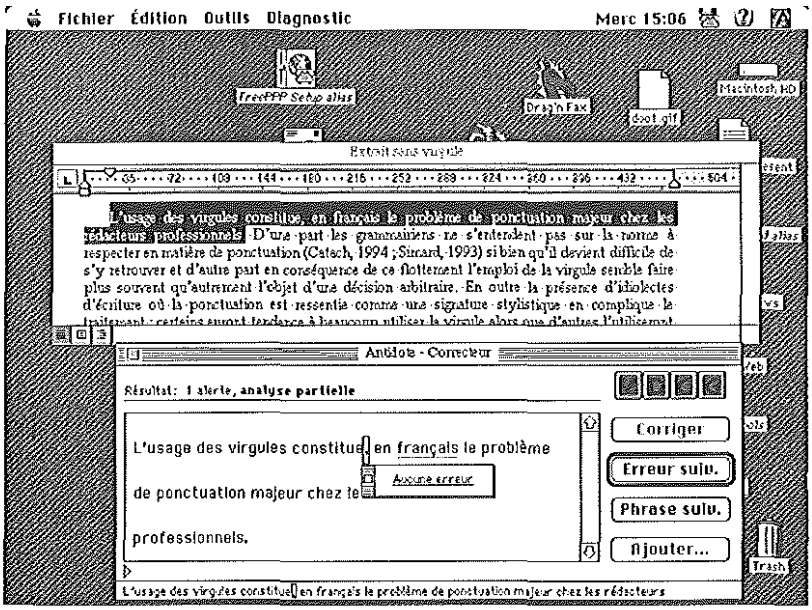


Fig. 14b.– *Antidote*®, effacement de l'un des membres d'une paire de virgules, virgule fermante effacée : détection échouée

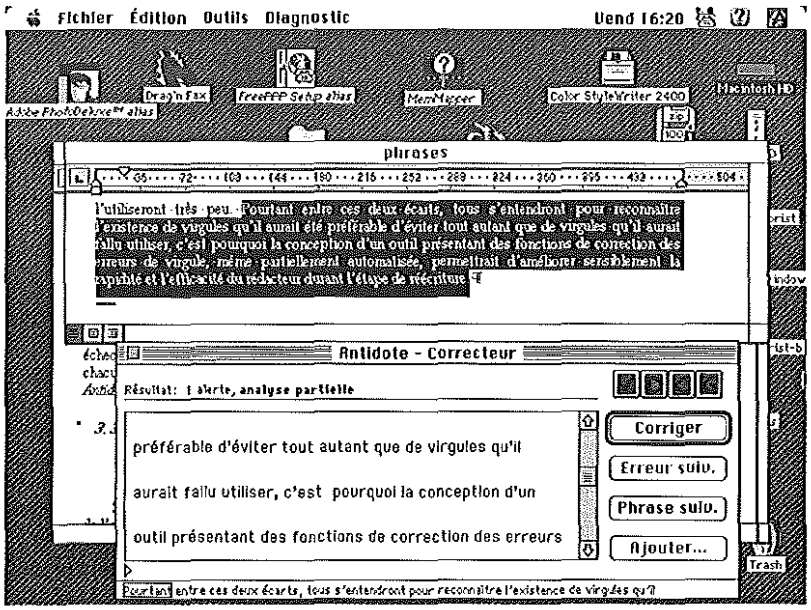


Fig. 15.– *Antidote*©, remplacement du point par une virgule

## 5. Conclusion

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*®.

Voyons l'exemple suivant (fig. 16a et 16b).

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*).

Par ailleurs, selon l'étude statistique du gouvernement du Québec, nous pouvons nous attendre à ce qu'une très forte majorité des textes que les grammaires informatisées sont appelées à réviser comportent un grand nombre d'erreurs de virgules. Nous savons aussi que des virgules erronément placées génèrent une segmentation et que la plupart des erreurs de ponctuation introduites dans nos exemples n'ont pas été repérées par les deux correcteurs examinés. Par conséquent, nous pouvons nous attendre à ce que le pouvoir de diagnostic des grammaires informatisées soit sérieusement diminué par les erreurs de ponctuation.

Or, 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 très 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. Mais nous savons à présent qu'il

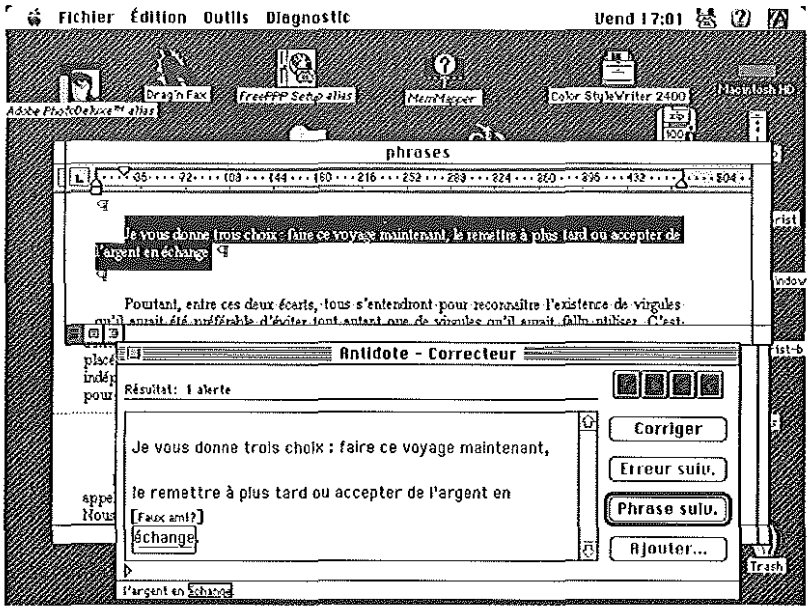


Fig. 16a.– *Antidote*®, diagnostic sur suite ponctuée correctement

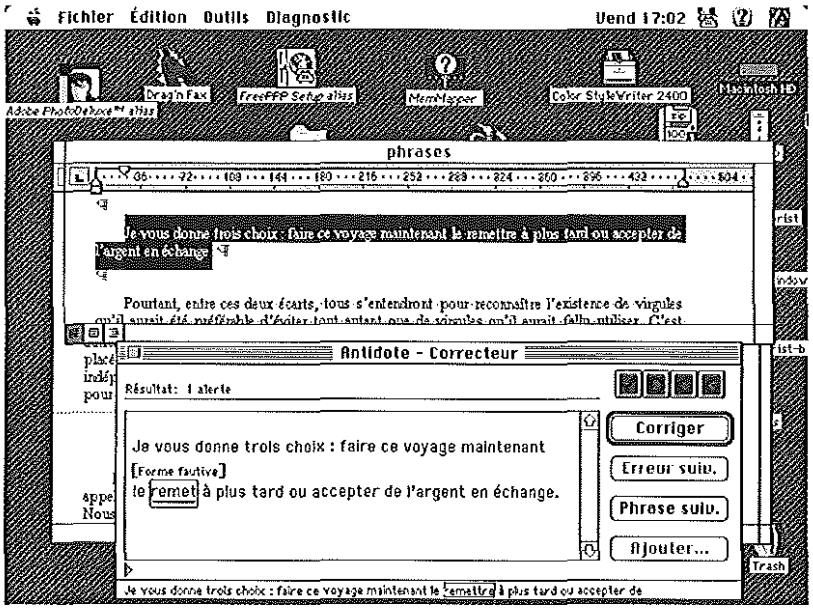


Fig. 16b.– *Antidote*®, diagnostic sur suite incorrectement ponctuée

s'agit là d'une solution risquée pour les correcteurs puisque la ponctuation des textes produits quotidiennement comporte fort probablement trop d'erreurs pour que les signes soient exploitables à cette fin.

Par ailleurs, l'attribution d'une seule catégorie grammaticale aux mots d'un texte (*tagging*) 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) estime que le choix d'étiquettes d'analyses possibles pour un seul mot du vocabulaire français dépasse 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 « artificiel ». En effet, discutant de l'approche répandue chez les linguistes informaticiens de développer des parseurs reconnaissant seulement le « noyau » 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). Or il se trouve que de telles structures se marquent graphiquement par des signes de ponctuation, le plus souvent par des virgules.

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, 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. 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.

**Remerciements**

La recherche dans le cadre de laquelle cet exercice a été effectué est menée sous la direction du professeur Jacques Ladouceur, linguiste-informaticien, du département des Langues et de Linguistique de l'Université Laval (Québec, Canada), et financée par le Fonds pour la Formation de Chercheurs et de l'Aide à la Recherche du Québec et le Conseil de recherches en sciences humaines du Canada.

## Bibliographie

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, « Cambridge Textbooks in Linguistics », 185 p.)

BATES (Madeleine) et WEISCHEDEL (Ralph M.) ed. : 1993, *Challenges in Natural Language Processing* (Cambridge, Massachussets : Cambridge University Press, « Studies in Natural Language Processing », 296 p.)

BÉDARD (Édith) et MAURAIS (Jacques) : 1983, *La Norme linguistique* ([s.l.] : Gouvernement du Québec, Conseil de la langue française) (coll. « L'ordre des mots », Le Robert, Paris).

BLANQUET (Marie-France) : 1994, *Intelligence artificielle et système d'information : le langage naturel* (Paris : ESF Éditeur) (coll. « Systèmes d'information et nouvelles technologies », 269 p.).

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.).

BRISCOE (Ted) : 1994, « Parsing (with) Punctuation etc. » Grenoble : Rank Xerox Research Laboratory, Multilingual Theory and Technology. MLTT-TR-002.

BUTLER (Christopher) : 1985, *Computers in Linguistics* (Oxford, United Kingdom : Basil Blackwell, 266 p.).

CARRÉ (René) *et al.* : 1991, *Langage humain et machine* (Paris : Presses du C.N.R.S., 298 p.).

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. : 1980, *Langue française, La Ponctuation*, n° 45, Février, (148 p.).

CATACH (Nina) : 1990, *Pour une Théorie de la Langue écrite* (Paris : Éditions du Centre national de la recherche scientifique, 259 p.).

CATACH (Nina) : 1994, *La Ponctuation* (Paris : Presses Universitaires de France, « Que sais-je ? », n° 2818, 128 p.).

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) et TAPANAINEN (Pasi) : 1996a, « Rules and Constraints in a Finite-State Grammar » (Grenoble : Rank Xerox Research Laboratory, MultiLingual Theory and Technology. MLTT-024).

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).

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).

CHANOD (Jean-Pierre) : 1993, « Problèmes de robustesse en analyse syntaxique », in *Actes de ILN 93.*

COHEN (Paul R) : 1995, *Empirical Methods for Artificial Intelligence* (Cambridge, Massachusets et London, England : The MIT Press, 405 p.).

DALE (Robert) : 1990, « A Rule-Based Approach to Computer-Assisted Copy-Editing », in *CALL*, vol. 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) (octobre), p. 110–120.

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.).

DRUIDE INFORMATIQUE : 1997, *Antidote®*. Version 1.1.3. Macintosh. CD-ROM. (Montréal, Québec. Canada).

GRISHMAN (Ralph) : 1994, *Computational Linguistics. An Introduction* (Cambridge, Great Britain : Cambridge University Press, « Studies in Natural Language Processing », 193 p.).

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.).

JONES (Bernard) : 1994a, « Can Punctuation Help Parsing? », *Esprit Acquilex-II Working Paper*, n° 29, July (Cambridge University Computer Laboratory. U.K.) (cide@cup.cam.ac.uk.).

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), p. 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.).

MICROSOFT CORPORATION : 1990–1995, *Microsoft Word pour Windows 95*. Version 7.0. IBM-PC ou compatibles. CD-ROM.

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.

NUNBERG (Geoffrey) : 1996, *Lexical Grammar and Text Grammar* (Santa Cruz, California : Sigparse 96).

NUNBERG (Geoffrey) : 1990, *The Linguistics of Punctuation* (Stanford, California : Center for the Study of Language and Information (CSLI), « Lectures Notes », 141 p.).

PERCHERON (Daniel) : 1988, « Un point c'est tout », dans *Traverses 43, Revue du Centre de Création industrielle*, février (Centre Georges Pompidou, édit.), p. 151–157.

REICHGELT (Hans) : 1991, *Knowledge Representation, an AI Perspective* (Norwood, New Jersey : Alex Publishing Corporation, 251 p.).

SABBAH (Gérard) : 1990, *L'intelligence artificielle et le langage. Volume 1. Représentation des connaissances* (2e édition. Paris : Hermes, 357 p.).

SALTON (Gerard) et MCGILL (Michael J.) : 1983, *Introduction to Modern Information Retrieval* (New York : McGraw-Hill), 448 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.

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.

VÉDÉNINA (L. G.) : 1989, *Pertinence de la présentation typographique*, Avant-propos de Nina Catach (Paris : Peeters / Selaf, 153 p.).