Considerations on Parsing a Poorly Punctuated Text in French

Marthe Simard
Université Laval, Département des Lettres et de Linguistique,
Cité universitaire, Québec (Québec), Canada, G1K 7P4
Email : msimard@zone.ca

Abstract

Although a vast normative literature exists to describe the often complex and sometimes contradictory punctuation rules for French, no norm is still widely accepted. The result of this paradox translates in generally poorly punctuated texts, where comma problems represent over 90% of the mistakes (Ministère de l'Enseignement supérieur et de la Science, 1993). Detecting comma delimiter problems poses a fascinating challenge : if punctuation is required to parse a text (Jones, 1995) and if punctuation is generally misused, then how is it possible to parse successfully a poorly punctuated text ? This paper raises a few questions concerning this issue.

1 Introduction

Up to what point would it be possible to automatically detect punctuation problems in non literary French texts produced by unskilled writers ? This is the subject of my Ph.D. I will concentrate on comma delimiter problems (Nunberg, 1991), since, according to a statistical study by the Quebec Ministry of Education (Ministère de l'Enseignement supérieur et de la Science, 1993) of over 20,000 essays written by college level candidates, 93% of these texts presented comma insertion problems.

During 15 years as a lecturer at the Université du Québec à Chicoutimi, I wrote textbooks, exercise manuals and software packages in the field of writing skills, researched and developed original methods to teach French at college level and contributed to program development projects to help unskilled Francophone writers to write better French. Anyone who has had a brush with French grammar knows that writing French can be a tricky business. The leading reference, *Nouvelle Grammaire française*, (Grevisse et Goosse, 1991) lists over 2,500 grammar rules that Francophone writers are supposed to master just to produce a grammatically error free text. French grammar and spelling checkers, style guides and writing books are numerous and widely popular with word processor users who have to know what rules apply and how to apply them in their own daily compositions. Quebec, in fact, produces most of the spelling and grammar checkers currently marketed for French. Two of the best known are *Hugo* (Logidisque), bundled with the French version of Microsoft Word 7 for Windows95, and *Le Correcteur 101* (Machina Sapiens) which is marketed as an expert system capable of detecting and correcting grammar mistakes during text production.

As observed by Robert Dale (1990) for English, today's French spelling and grammar checkers are of no use to detect and correct punctuation problems for French. Although these systems offer the possibility to activate punctuation checking, they fail to detect and correct these problems when confronted with them. However their failure underscores a very interesting issue concerning punctuation detection and correction problems ; a punctuation checking module and parser would have to depend, at least partially, upon punctuation already inserted into the text. That fact poses an apparently insolvable problem : if most of the punctuation marks inserted in a text are misused and if correct punctuation is needed to parse a text, how is it possible to successfully parse a mispunctuated text ?

This paper will address a few of the questions related to this issue.

2 Punctuation Norm and Punctuation Limbo

The first question relates to the definition of a punctuation error. What is a comma insertion mistake in French ? When does it occur ? How can it be recognised and corrected ? These relatively simple questions don't find simple answers in French, mostly because there is no agreement on the norm for punctuation among

French grammarians (Catach, 1994 ; Simard, 1993). Although an abundant literature has been published on this question during (and even before) this century, one cannot affirm that such and such occurrence is a punctuation mistake without having to vigorously defend his or her point of view. Indeed, André Martinet (citated by Bureau, 1978) thought that the debate surrounding punctuation was such a matter of subjectivity that there was no scientific evidence hidden behind the subject. Punctuation was simply the product of accidental and intuitive idiolects that simply did not deserve a "serious" linguistic investigation. This point of view is still more or less widely held among French grammarians. Catach (1994) underlines the fact that modern grammarians have all but abandoned the discussion leaving the question to typographers and editors.

Paradoxially, even without agreement on a norm, the existence of this norm is strongly enforced. From Jacques Damourette (1939) to the latest attempt by Jacques Drillon (1991), many specialists try to explain or impose their view as the norm. Endless lists of rules and principles are described in detail with numerous examples from classical literature aimed to help the ordinary French writer to insert commas, colons and periods without falling into punctuation limbo. In fact, Catach (1994) goes so far as to say that there is such a chasm between punctuation rules and practices that it becomes a challenge not only to understand, let alone to successfully apply any of these sometimes contradictory rules.

Recognising this reality helps understanding the extent of the punctuation problems found in everyday French texts. Even if a user is sensitive to the question, how can one hope to resolve these problems with all the proposed and sometimes contrived rules found in grammars and punctuation guides ? The end result of this lack of normative guiding is easy to grasp : poorly punctuated texts.

On the other hand, punctuation is also a matter of style. Many writers punctuate their texts abundantly, others minimally. What is acceptable ? What is not ? A punctuation checking module should take into account style while checking for syntax problems that a poorly punctuated text can generate. Not an easy task !

3 Defining Comma Insertion Problems

The Ministère de l'Enseignement supérieur et de la Science (1993) with the Office de la langue française du Québec, the official government watchdog group responsible for improving the quality of French among Quebecers, identified three types of comma insertion problems in the surveyed essays :

- misplaced commas as in [1], where a comma is used in place of another punctuation mark, often the period.

  [1] **Le sida est une grave maladie qu'il faut prévenir(,) avec les nouvelles recherches on met au point des médicaments de plus en plus efficaces*. (AIDS is a serious illness that we need to prevent(,) with new technologies we discover new medicines that are more and more efficient).

- absent commas when required as in [2] and [3], [4] and [5], where delimiters are needed to signal an explicative proposition, an apposition or a syntax permutation :

  [2] **Moi, je dis que non car _ lorsque j'étais jeune _ les programmes que je regardais étaient en grand nombre et aussi violents* (I answer no since when I was young the tv shows that I watched were numerous and as violent).

  [3] **Se considérant personne responsable _ on se doit de dénoncer le contenu de la programmation actuelle de nos télédiffuseurs.* (Considering oneself as a responsible human being one has to denounce current tv programming).

  [4] **De plus _ l'adulte n'est pas protégé face au problème*. (Moreover the adult is not protected against the problem).

  [5] **Cet acteur, très populaire auprès des jeunes _ peut avoir une influence qui ne sera pas toujours bénéfique pour certains*. (This actor, very popular amongst youngsters _ may well have an influence that will not always be beneficial for some of them).

- presence of a non required comma as in [6] and [7] where subject and verb or verb and direct complement should not be separated by any punctuation mark :

[6] **Ces salaires de plusieurs millions de dollars(,) ne sont aucunement justifiables à mon avis.* (These multiple million dollar salaries are not at all justifiable in my opinion).

[7] **Il faudrait(,) mettre sur pied des actions ou des programmes afin de venir en aide aux jeunes.* (Actions or programs should be implemented to help youngsters).

Although the norm may vary, these three types of problems seem to be the most widespread among unskilled French writers. Louise Guénette, François Lépine and Renée Lise Roy (1995), from the École des langues vivantes, Laval University, document more or less the same type of problems among their students : commas not required but present, commas required but absent and commas confused with another punctuation mark, mostly the period.

My solution to the problem of norm and punctuation mistakes is to temporarily adopt a norm for the purpose of my research without endorsing it officially. One may argue that such or such occurrence might not be described as a comma insertion mistake but for the programming experiment I want to conduct, any norm will do. Comprised within the rules of this norm, I will try to find a way to detect the problems related to comma delimiters and correct them as much as possible.

4 Ambiguity, Commas and Computer Limitations

The role of commas as an interpretation tool for eliminating ambiguities is well documented in punctuation domain literature (Jones, 1995 ; Catach, 1994 ; Nunberg, 1991 ; Védénina, 1989 ; Fonagy, 1980) as well as in normative reference books (amongst others : Colignon, 1993 ; Drillon, 1991 ; Brun et Doppagne, 1958 ; Damourette, 1939). Misplaced or misused commas can consequently play havoc in a text. That is probably one of the reasons why spelling or grammar checkers that try to detect punctuation problems may meet with difficulties when attempting to parse poorly punctuated sentences. For example, sentence [8]

[8] **Novex, qui a, pour mission, de diffuser, des programmes, en haute technologie, a maintenant, un portefeuille, de 23 placements, dans des entreprises.* (*Novex, whose mission(,) is to market(,) high technology(,) programs, currently has(,) a portfolio(,) of 23 investments(,) in business).

was analysed by *Hugo* (*Word 7*), *Grammatik* (*Word Perfect*, 6.0) and *Le Correcteur 101*. Sentence [8] constitutes in fact an asyntatic phrase where a comma is inserted every two words without respect for French syntax or meaning. While neither *Hugo* nor *Grammatik* reacted to the sentence, *Le Correcteur 101* responded : “Je ne peux pas trouver de complément direct dans cette phrase : diffuser quoi ? diffuser qui ?” (I cannot find a direct complement in this sentence : market what ? market who ?). Assuming that the comma between “programmes” and “en haute technologie” was coordonnating “qui a (pour mission)” and “a maintenant”, the system concluded that the sentence was poorly structured and required manual correction. The most fascinating aspect of this test is the fact that the *Correcteur 101* seemed to have taken into account the inserted commas to interpret the sentence. Although it could not detect the real nature of the problem, it certainly could “see” that the sentence was unreadable. Which brings us to the problem of the detection of comma insertion mistakes where the comma is only a graphic form that needs to be recognised and evaluated under complex syntax and semantic rules.

We know for a fact that computers do not “understand” the simplest sentence. All they can do is add extremely fast and compare two forms to find a match. With such limited capacities, how do we detect comma insertion delimiter problems ? I believe that a good way could be to find linguistic (ex. the number or the variety of words) or graphic (ex. other kinds of delimiters) indicators that would help guide the analysis toward certain problem areas of the text. Assuming that there is a relationship between the mastery level of the text and the text and the mastery level of its punctuation, these indicators could then be used to build a grammar susceptible to successfully parse even a poorly punctuated text.

5 Detecting Comma Insertion Problems

A system capable of detecting comma delimiter problems should be able to perform one task :

- detecting sets of problems as described by the user's norm,

while avoiding two traps :

- detecting false sets of problems ;
- failing to recognize real sets of problems.

Control of the noise (false sets of problems) and the silences (ignored real problems) in grammar and spelling checkers constitutes one of the main challenges (Chandioux, 1996). For instance, sentence [8] was ignored by *Hugo* and *Grammatik* although it contained obvious problems while *Le Correcteur 101* made a false diagnosis to explain its syntax irregularities. Approximation and inefficiency are of no help to the user.

In order to detect comma delimiter problems however, we would need to solve a number of related difficulties : the creation of a knowledge base system modifiable by the user and independant of the set of rules that could apply in any context (Dale, 1990) ; a grammar capable of parsing a text according to a set of punctuation rules and conventions (Jones, 1995), a way of detecting ambiguous contexts where analysis would have to be relayed partially to the user for resolving meaning incertainties, and potentially a probability scale of success in an effort to control noise and silences and to provide validity and constance in the results (Dale, 1990).

Resolving ambiguities constitutes "arguably the single most important problem in natural language processing" (Gazdar and Mellish, 1990). James Allen (in Bates and Weischendel, 1993) shows many typical cases. Carré *et al.* (1991) propose three types of ambiguities. But whatever the case, we can safely assume that our parser on its own could not always perform an efficient analysis without human intervention to describe, underline and interprete difficult contexts and meanings. Consequently, entirely automated detection of comma insertion problems would have to potentially interact with the user to assertain and clarify ambiguities.

6 Particulars Related to Parsing Poorly Punctuated Texts

What kind of linguistic or graphic problems should we expect while parsing a poorly punctuated text ? The firs would be the determination of the limits of the sentence. For the sake of discussion, let us suppose that an Upper case signals the beginning of a sentence and the period signals its end. In a correctly punctuated sentence, these two signals could serve as the basis for sentence punctuation analysis. But we know that one of the common mistakes found in everyday French texts is the confusion of the comma with the period. How would it be possible for our comma delimiter problems checker to recognize first, the limits of sentence [1] and second, the misplaced comma ?

[1] **Le sida est une grave maladie qu'il faut prévenir*(,) *avec les nouvelles recherches on met au point des médicaments de plus en plus efficaces*. (AIDS is a serious illness that we need to prevent(,) with new technologies we discover new medicines that are more and more efficient).

However not all the contexts should introduce these difficulties. Let us look at the example of the inverted subordinate as in [9] :

[9] Bien que les soldats se soient montrés réticents, ils se sont battus avec courage. (Although the soldiers were reticent, they fought with courage).

The occurrence of the conjunctive locution "bien que" at the beginning of the sentence immediately suggests the use of a delimiter to mark the end of the inverted subordonate and signal the beginning of the principale. Yes, but where is the comma placed ? How does one distinguish the end of the subordonate from the beginning of the principale ? Our comma delimiter problem checker should theoritically be able to parse the sentence so that the relevant context could be marked with a comma. But we already know that, particularly for complex sentences, this unique decision requires context and meaning interpretation. We also know that our checker could not operate efficiently under these conditions.

The use of delimiters poses yet another challenge. Sentence [10] presents an explicative relative delimited by a pair of commas while sentence [11] introduces one comma separating elements in an enumeration :

[10] Il a critiqué ses amis, qui l'avaient trahi, avant de quitter la soirée. (He criticised his friends, who had betrayed him, before he left the party).

[11] Il a rencontré sa femme, sa sœur préférée et Simon avant de partir. (He met his wife, his favorite sister and Simon before leaving).

The use of the delimiters in [10] is totally different than the use of the separating comma in [11]. This distinction is particularly meaningful in [12] :

[12] Il a rencontré sa femme, sa sœur préférée(,) et Simon avant de partir (He met his wife, his favorite sister, and Simon before leaving).

where an unrequired comma is placed before the conjunction "et". But let us look at [10] and [12] from the computer's point of view :

[10] Word word word word, word word word, word word word word word.

[12] Word word word word, word word word, word word word word word.

Sentences [10] and [12] are now totally identical. However commas 1 and 2 in [10] are delimiters used correctly in the context of [10] while comma 2 in [12] is unrelated to comma 1 and introduces a punctuation mistake that our comma delimiter problem checker would need to address.

Finally, let us add another twist to this plot. One common mistake surveyed in the Quebec Ministry of Education's corpus is the absence of one of the delimiters required by the context. Let us suppose, for instance, that the user has written [10a] :

[10a] Il a critiqué ses amis, qui l'avaient trahi avant de quitter la soirée. (He criticised his friends, who had betrayed him before he left the party).

This time, sentence [10a] is syntatically correct but translates quite another meaning than the original sentence. The absence of the required closing delimiter introduces a change in the user's message. Would our comma delimiter problem checker be able to detect such a problem. If ever, how ?

Hopefully, my research will provide some insights if not the answers to these questions.

Acknowledgments

This research is funded by a grant from the Quebec government through the *Fonds pour la formation de Chercheurs et l'Aide à la Recherche (F.C.A.R.)* and conducted under the direction of Professor Jacques Ladouceur, computational linguist, Laval University, Quebec.

References

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 University Press, Studies in Natural Language Processing. p 147-148. Cambridge, Royaume-Uni.

Brun, J. and Doppagne, A. [1958]. *La Pontuation et l'art d'écrire*, Beaude, 240 p., Bruxelles, Belgique.

Bureau, Conrad. 1978. *Syntaxe fonctionnelle du français*. Les Presses de l'Université Laval, p. 176, Québec, Québec.

Carré, René *et al.* 1991. *Langage humain et machine.* Presses du C.N.R.S. p. 50-56. Paris, France.

Catach, Nina. 1994. *La ponctuation.* Presses universitaires de France. Que sais-je ? No 2818. p. 43-47. Paris, France.

Chandioux, John. 1996. *À propos de Météo et de détection automatique d'erreurs.* Entrevue. Groupe Chandioux. Bande audio. 90 min. Montréal, Québec.

Colignon, Jean-Pierre. 1993. *Un point, c'est tout ! La ponctuation efficace.* Éditions du Boréal et Presse et Formation, Éditions du Centre de formation et de perfectionnement. p. 13-34. [Montréal, Québec] et [Paris, France].

Dale, Robert. 1990. *A Rule-Based Approach to Computer-Assisted Copy-Editing*. In CALL. Volume 2. p. 59-67.

Damourette, Jacques. 1939. *Traité de la ponctuation française*. Galllimard/inédit. Tel No 177. 144 p. Paris, France.

Fonagy, Ivan. 1980. " Structures sémantiques des signes de ponctuation ". In BLSP No 75. p. 95-129.

Gazdar, Gerald and Mellish, Chris. 1990. *Natural Language Processing In Prolog. An Introduction to Computational Linguistics*. Addison-Wesley Publishing Company. 504 p. Wokingham, England.

Grevisse, Maurice and Goosse, André. 1991. *Nouvelle Grammaire française.* 2ième édition revue. Duculot et Éditions du renouveau pédagogique. 333 p. Paris, France et Saint-Laurent, Québec.

Guénette, Louise and Lépine, François and Roy, Renée Lise. 1995. *Le Français tout compris. Guide d'autocorrection du français écrit.* p. 75-83. Saint-Laurent, Québec.

Jones, Bernard. 1995. *Exploring the Role of Punctuation in Parsing Natural Text.* In Proceedings of the 15th International Conference on Computational Linguistics (COLING-94), p. 421-422, Kyoto, Japan.

Ministère de l'Enseignement supérieur et de la Science. 1993. *Test de français. Épreuves de mars 1993 (Collèges). Analyse détaillée des erreurs.* 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. Québec, Québec.

Nunberg, Geoffrey. 1991. *The Linguistics of Punctuation.* Center for the Study of Language and Information (CSLI) Lecture Notes No 18. p. 9-17 and 52-53. Stanford, California.

Simard, Marthe. 1993. *Étude de la distribution de la virgule dans les phrases de textes argumentatifs d'expression française. Mémoire de maîtrise*. Université Laval. p. 69-85. Québec, Québec.

Védénina, Ludmilla. 1989. *Pertinence de la présentation typographique.* Avant-propos par Nina Catach. Peeters/Selaf. p. 34-39. Paris, France.