Ponctuations fortes abusives

Laurence Danlos, Benoît Sagot

TALN 2010, Montréal, 19–23 juillet 2010

Ponctuations fortes abusives

Laurence Danlos & Benoît Sagot
Alpage, INRIA Paris–Rocquencourt & Université Paris 7
Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France
laurence.danlos@linguist.jussieu.fr, benoit.sagot@inria.fr

Résumé. Certaines ponctuations fortes sont « abusivement » utilisées à la place de ponctuations faibles, débouchant sur des phrases graphiques qui ne sont pas des phrases grammaticales. Cet article présente une étude sur corpus de ce phénomène et une ébauche d'outil pour repérer automatiquement les ponctuations fortes abusives.

Abstract. Some strong punctuation signs are "wrongly" used instead of weak punctuation signs, leading to graphic sentences which are not grammatical sentences. This paper presents a corpus study of this phenomenon and a tool in the early stages to automatically detect wrong strong punctuation signs.

Mots-clés : pseudo-phrase, phrase averbale, analyse syntaxique et sémantique.

Keywords: pseudo-sentence, verbless utterance, syntactic and semantic analysis.

1 Introduction

Pratiquement tout système de TAL pour l'écrit commence par une segmentation en phrases du texte à traiter, segmentation qui s'appuie entre autres sur les ponctuations fortes[1]. Les phrases ainsi obtenues, qualifiées de graphiques, ne correspondent pas toujours à des phrases grammaticales[2]. Citons (Grevisse, 2007, page 124) : « Les écrivains contemporains emploient parfois le point (au lieu de la virgule) pour détacher de la phrase un membre auquel ils veulent donner un relief particulier ». Grevisse illustre ce phénomène par l'exemple donné en (1).

(1) On avait donné dans le Nord un grand coup de pied dans la fourmilière, et les fourmis s'en allaient. Laborieusement. Sans panique. Sans espoir. Sans désespoir. Comme par devoir. [Saint Exupéry, Pilote de guerre]

C'est ce phénomène, que nous appelons (abusivement) « ponctuation forte abusive », qui est étudié dans cet article[3]. Nous appelons « pseudo-phrase » une phrase précédée d'une ponctuation forte abusive. Ainsi (1) enchaîne cinq pseudo-phrases (de type adverbial). Notre objectif à terme est de mettre au point un outil

[1] Les ponctuations fortes sont le point, les points d'interrogation et d'exclamation, le point-virgule et les points de suspension et les deux-points.

[2] La notion de phrase grammaticale varie selon les théories. Nous n'entrerons pas dans une discussion sur ce sujet, nous contentant de considérer en première approximation que cette notion correspond à celle « d'unité rectionnelle » introduite dans (Berrendonner, 2002) et (Benzitoun *et al.*, 2010).

[3] Dans les études sur l'oral, où les signes de ponctuation cèdent la place aux marqueurs prosodiques, ce phénomène correspond *grosso modo* à « l'épexégèse », terme introduit dans (Bailly, 1944).

de détection des pseudo-phrases afin de pouvoir les analyser syntaxiquement et surtout sémantiquement. Ainsi en (1), on veut pouvoir dire que l'adverbe *laborieusement* porte sur le départ des fourmis.

Si les pseudo-phrases sont identifiées, la prise en compte des ponctuations fortes abusives en vue de l'analyse syntaxique et sémantique peut se faire de façon simple : il suffit de réétiqueter en ponctuation faible la ponctuation forte abusive[4], puis de « recoller » la pseudo-phrase à la phrase qui la précède pour obtenir une phrase grammaticale, dont l'analyse syntaxique et sémantique peut se faire de façon standard.

Les pseudo-phrases ne comportant souvent pas de verbe fini[5], leur identification pose le problème de les distinguer des « phrases averbales » qui doivent elles recevoir une analyse syntaxique et sémantique autonome (modulo les phénomènes anaphoriques) : ce ne sont pas des membres détachés de la phrase qui précède, elles n'occupent aucune fonction syntaxique dans cette phrase. Pour illustrer la difficulté de distinguer les pseudo-phrases des phrases averbales, considérons le paradigme en (2) où chaque seconde phrase graphique ne comporte qu'un GN. En (2a), ce GN occupe la fonction d'apposition par rapport au dernier GN de la phrase qui précède, il s'agit d'une pseudo-phrase. En (2b), ce GN n'a aucune fonction à l'intérieur de la phrase qui précède, il s'agit d'une phrase averbale. (2c) a deux lectures : soit le voyage de Luc était un échec total — on a alors affaire à une apposition et à un point abusif —, soit le récit de Luc était un échec total — on a alors affaire à une phrase averbale.

(2) a. [Nous avons été obligés de vendre une voiture.] Une jeep. [L'Est Républicain]

b. [Nous avons été obligés de vendre une voiture.] Un échec total.

c. [Luc a raconté son voyage en Islande.] Un échec total.

Les phrases averbales ont fait l'objet de nombreuses études, pour le français citons (Laurens, 2007). En revanche, nous ne connaissons pas de travaux systématiques sur les pseudo-phrases et les ponctuations fortes abusives. Le travail présenté ici est donc une ébauche de l'étude de ce phénomène en vue de son traitement automatique. Nous sommes partis d'un corpus, duquel nous avons extrait des phrases graphiques candidates à être des pseudo-phrases (Section 2). Une étude manuelle des résultats de cette extraction nous a permis de proposer une classification automatique des pseudo-phrases en s'appuyant sur leur forme (Section 3).

2 Étude sur corpus

Nous avons commencé notre travail sur les ponctuations fortes abusives par une étude s'appuyant à la fois sur des connaissances linguistiques et sur un corpus, afin d'étudier la possibilité de les identifier automatiquement, et notamment de les différencier des phrases averbales, d'en ébaucher une classification et de déterminer leur fréquence.

Nous sommes partis d'un corpus journalistique extrait d'un quotidien régional, *L'Est Républicain*. Le corpus a été préalablement segmenté en phrases graphiques et tokenisé par la chaîne de traitements de surface SXPipe (Sagot & Boullier, 2008)[6] puis étiqueté automatiquement à l'aide de MElt (Denis & Sagot,

[4]Il est important de conserver telle quelle la ponctuation utilisée et de ne changer que sa catégorie, afin de ne pas perdre les nuances sémantiques ou pragmatiques apportées par la ponctuation forte abusive.

[5]Néanmoins, certaines pseudo-phrases commencent par une conjonction de subordination ou un pronom relatif et comportent au moins un verbe fini. Ces pseudo-phrases, qui constituent la classe CONJ, seront discutées à la Section 2

[6]SXPipe n'est utilisé ici que comme tokeniseur. Pour reprendre les termes dans lesquels cette chaîne est décrite par (Sagot & Boullier, 2008), seules les deux premières des cinq phases de traitement sont appliquées.

2009). MElt est un étiqueteur morphosyntaxique entraîné sur le Corpus Arboré de Paris 7 (Abeillé *et al.*, 2003), dans sa variante à 29 étiquettes telle qu'utilisée dans (Candito *et al.*, 2009). Le résultat est un ensemble de 20 millions de tokens, composant 1 212 659 phrases graphiques se terminant par au moins une ponctuation forte.

Les phrases qui commencent par une conjonction de subordination[7] ou par un pronom relatif, avec à droite de cet élément au plus un verbe fini, étant nécessairement des pseudo-phrases (3a-b), nous les avons extraites du corpus (par des expressions régulières) et nous les avons classées dans la **Classe CONJ**. Ceci concernait 175 pseudo-phrases. Néanmoins, nous avons affiné nos expressions régulières de façon à intégrer des exemples comme (3c) dans la Classe CONJ tout en éliminant des exemples comme (3d) : ces phrases comportent deux verbes finis dont le second est la tête d'une subordonnées (relative) en (3c) et d'une principale en (3d) — (3d) est une phrase grammaticale construite sur le patron *CONJ* P_2, P_1 avec une subordonnée antéposée. Les subordonnées peuvent être distinguées des principales par le fait qu'elles commencent par un élément de la classe SUB comprenant les pronoms relatifs, les conjonctions de subordination et le complémenteur *que*.

(3) a. [Dimanche, je ferai un dessin et je marquerai dessus : bonne fête maman.] Parce que ma maman c'est la plus jolie…

b. [Qui ont été contrôlés même si non aménagés.] Parce que fréquentés.

c. [Et il compte fort sur le Vosgien pour amener plus de service aux hôteliers indépendants.] Tandis que Michel Philippe prend la tête de l'association régionale de l'Est de la France, qui compte quarante établissements.

d. [Son œuvre est plus proche de la peinture.] Parce que la pagination est importante, le livre est découpé en chapitres (...).

Ayant écarté les phrases de la classe CONJ de notre corpus, nous avons alors extrait toutes les phrases qui ne comportaient pas de verbe fini ou qui comportaient un verbe fini apparaissant à droite d'un élément de la classe SUB (et donc un verbe fini tête d'une subordonnée). Un examen manuel du sous-corpus obtenu nous a permis de délimiter et d'écarter quatre sous-ensembles :

– les phrases qui, à cause d'erreurs de l'étiqueteur MElt, comportaient un verbe fini ; c'est en particulier le cas pour les phrases comportant certains impératifs ;
– les phrases commençant par un nombre ou contenant un deux-points qui correspondent presque toutes à des énumérations (résultats sportifs ou électoraux) ; nous avons considéré de telles phrases comme non pertinentes pour l'étude ;
– les phrases dont la phrase précédente se termine par un point d'interrogation ou un point-virgule, que nous réservons pour une étude ultérieure plus approfondie ;
– les phrases qui sont clairement des phrases averbales, à savoir :
  – celles commençant par une forme fléchie du lexème *quel*, ou par des séquences de mots particulières telles que *gare à*, *au programme*, *en avant*, *bon anniversaire*… ;
  – celles de la forme *N ADJ* ou *ADJ N* ; en effet, nous faisons l'hypothèse que les pseudo-phrases qui ne comportent qu'un GN incluent toutes un déterminant préfixant le GN, comme en (2a), et que donc les phrases de forme *N ADJ* ou *ADJ N* sont des phrases averbales. Cette hypothèse est justifiée par

[7] Nous avons utilisé la liste des conjonctions de subordination dressée dans le lexique LEXCONN (Roze *et al.*, 2010) dont nous avons enlevé certains éléments (*preuve/pourvu que*) qui introduisent des phrases averbales lorsque précédés d'une ponctuation forte.

les faits suivants : le déterminant est obligatoire dans la pseudo-phrase (2a), facultatif dans la phase averbale (2b), et (2c) sans déterminant n'a que la lecture averbale[8].

Le résultat est un corpus de 9 276 phrases graphiques candidates à être des pseudo-phrases (Classe CONJ exclue). Il reste donc à distinguer parmi ces candidates lesquelles correspondent effectivement à des pseudo-phrases, et à classer ces dernières. Pour mener à bien ce travail, nous avons associé à toute phrase candidate la phrase qui précède.

3 Classification des pseudo-phrases

Suite à un premier examen manuel du corpus de phrases candidates, nous les avons classées automatiquement en fonction des tests suivants appliqués séquentiellement :

Classe COORD : le premier mot est l'une des conjonctions de coordination suivantes : *et*, *ou*, *mais*, *car*, *puis* (590 candidats) ;

Classe PRÉP : la phrase candidate commence par un syntagme prépositionnel (1 232 candidats) ;

Classe PAS : la phrase candidate commence par le mot *pas* (202 candidats) ;

Les phrases restantes sont rassemblées dans une classe par défaut (**Classe RESTE**).

Classe COORD Les phrases de cette classe sont quasiment toutes des pseudo-phrases, sauf certaines familles :
– celles comme *Et pour cause*, *Mais au contraire*, *Et pourtant* ;
– des exclamatives en *ConjCoo+quel*, comme *Et quel suspens !* ou *Mais quelle efficacité lorsqu'il s'agit de brouter la pelouse sans effort (…)* ;
– la tournure *et+GN* de type *Et la solidarité, chauffard !* ;
– quelques cas particuliers comme *Et aujourd'hui, place au jeu*.
Les pseudo-phrases trouvées sont coordonnées avec un premier terme dans la phrase précédente qui est généralement situé à la fin de la phrase précédente et qui occupe des fonctions diverses, telles qu'objet direct (4a), objet indirect (4b), complément oblique (4c), modifieur (4d).

(4) a. [Des défis qui engagent son avenir.] Et son devenir.
b. [Les enfants n'ont jamais manqué de rien.] Et surtout pas d'amour.
c. [C'est essentiel pour sa renommée.] Et pour les retombées liées au tourisme.
d. [La communauté de communes (…) [invite] le public à venir nombreux à ce spectacle extraordinaire.] Et entièrement gratuit.

Classe PRÉP Dans leur majorité, les phrases de cette classe sont bien des pseudo-phrases. On trouve cependant un certain nombre de phrases averbales (5). On constate que la plupart d'entre elles sont construites sur le modèle *GP, GN* — qui pourrait être filtré automatiquement.

(5) a. Après les cravates, les shorts.

[8] Citons quelques phrases averbales de forme *N ADJ* trouvées dans le corpus : *entrée libre*, *dépaysement garanti*, *inscription gratuite*, *pari tenu*, *peine perdue*, *mission accomplie*, *frissons garantis*. Et quelques phrases averbales de forme *ADJ N* : *joli coup*, *triste journée*, *vaste programme*, *sévère constat*, *mauvais présage*, *sacrée soirée*, *vaste débat*, *bel objectif*.

b. Chez les vétérans, scénario inverse.

Les pseudo-phrases de la Classe PRÉP sont principalement des compléments obliques, des compléments locatifs, des arguments nominaux ou des adjoints (6). Nous n'avons pas trouvé d'exemple d'objet indirect.

(6) a. [La jeune togolaise a travaillé quatre ans.] Sans salaire.
b. [L'ASGE a désormais une histoire.] Avec ses joies, multiples.
c. [Un patrimoine dont on devrait tous être les héritiers.] Sans exception.
d. [Dès maintenant, la mobilisation est de mise.] Pour l'amour des mots.
e. [Dominique va vivre ensuite un grand blanc.] De cinq ans.

Classe PAS La majorité des exemples de cette classe sont ici aussi des pseudo-phrases (8). On trouve cependant un certain nombre de phrases averbales (7). Certaines d'entre elles sont relativement figées, ou débutent par une locution figée (7b).

(7) a. Pas de problème.
b. Pas question que nos chauffeurs fassent les malins.

(8) a. [(...) dit le réalisateur Chilien, qui tient à garder de son pays la nationalité et l'accent.] Pas les mauvais souvenirs.
b. [Et suscite déjà des commentaires.] Pas franchement chaleureux.
c. [(...) rien ne manquait.] Pas même le bruit des vagues, ni les cris des mouettes et des cormorans.

Classe RESTE Certaines phrases de cette classe sont clairement des pseudo-phrases, par exemple celles qui commencent par une forme fléchie de *celui* (9a), ou encore celles composées d'un adverbe (voir *laborieusement* dans (1)). Nous n'avons pas créé de classes pour ces exemples parce ils sont peu représentés dans notre corpus. D'autres phrases de cette classe auraient pu être identifiées comme des pseudo-phrases si l'étiqueteur MElt indiquait le genre et le nombre (ce qui n'est pas le cas à l'heure actuelle) : il s'agit de celles composées d'un GN pluriel (9b) ou d'un adjectif (éventuellement précédé ou suivi d'un adverbe) féminin et/ou pluriel (9c).

(9) a. [Mais là, j'ai fait un dessin et je lui lirai une histoire.] Celle du lapin tout bleu !
b. [Daniel Kuntz a expliqué la nature aux écoliers.] Des enfants très attentifs.
c. [(...) les choses sont parfois plus difficiles en zones rurales.] Plus délicates.

Pour les phrases composées d'un GN singulier (préfixé d'un déterminant) — ou d'un GAdj masculin singulier —, il est délicat de savoir s'il s'agit d'une pseudo-phrase ou d'une phrase averbale, car, rappelons-le, ce cas conduit à des ambiguïtés telles que celle observée en (2c) lorsque la tête du GN est un nom abstrait[9]. Il semble toutefois que l'usage permet de classer avec quasi-certitude certains de ces GN comme des phrases averbales. Citons *une excellente initiative*, *un grand moment*, *un comble*, *un bide*, *un rêve*, *une grande/sale soirée*, *un jour inoubliable*, *une vraie/véritable galère*, *un petit régal*, ou *du grand art*, *la tuile*.

[9]Lorsque la tête du GN ne peut désigner qu'un objet concret ou un animé, il s'agit d'une pseudo-phrase.

La classe RESTE contient toutefois des éléments qu'il semble difficile de classer en pseudo-phrases ou phrases averbales en s'appuyant uniquement sur un étiquetage morpho-syntaxique et sur des expressions régulières.

4 Conclusion et perspectives

L'outil rudimentaire d'extraction des pseudo-phrases que nous avons mis au point pour cette étude nous a permis d'identifier plusieurs classes de pseudo-phrases (CONJ, COORD, PRÉP, PAS) qu'il semble possible de détecter automatiquement. En effet, les faux positifs identifiés dans ces classes semblent pouvoir être écartés automatiquement au moyen de listes (*et pour cause*, *pas de problème*), ou à condition de reconnaître certaines structures (par exemple *GP, GN*), ce qui semble possible à partir d'un simple étiquetage morphosyntaxique. Par conséquent, le développement d'un outil de détection des pseudo-phrases (et donc des ponctuations fortes abusives) qui aurait une bonne précision semble accessible. Toutefois, le rappel d'un tel outil ne serait probablement pas aussi bon (cf. Classe RESTE).

Une autre piste à explorer consiste à repérer les pseudo-phrases après l'analyse syntaxique des phrases graphiques et à calculer les liens de dépendances syntaxico-sémantiques qui les relient à la phrase précédente. L'évaluation d'une telle approche demande d'enrichir des corpus de référence tels que le Corpus arboré de Paris 7 (Abeillé *et al.*, 2003) par des annotations sur les pseudo-phrases. Ces annotations manquent à l'heure actuelle, le phénomène de ponctuation forte abusive ayant été ignoré dans la communauté TAL.

Références

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In A. ABEILLÉ, Ed., *Treebanks*. Kluwer, Dordrecht.

BAILLY C. (1944). *Linguistique Générale et linguistique française (2ème édition)*. Berne : Francke.

BENZITOUN C., DISTER A., K K. G., KAHANE S., PIETRANDREA P. & F F. S. (2010). tu veux couper là faut dire pourquoi propositions pour une segmentation syntaxique du français parlé. In *Actes de CMLF 2010)*, La Nouvelle Orléans, USA.

BERRENDONNER A. (2002). Les deux syntaxes. In M. CHAROLLES, P. L. GOFFIC & M.-A. MOREL, Eds., *Y a-t-il une syntaxe au- delà de la phrase ?*, p. 23–36. Verbum, 24 (1-2).

CANDITO M., CRABBÉ B. & SEDDAH D. (2009). On statistical parsing of french with supervised and semi-supervised strategies. In *Proceedings of the EACL 2009 workshop : Grammatical Inference for computational linguistics*, Athens, Greece.

DENIS P. & SAGOT B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort. In *Proceedings of PACLIC 2009*, Hong Kong.

GREVISSE M. (2007). *Le Bon Usage (14ème édition par André Goose)*. Paris-Louvain la Neuve : Duculot.

LAURENS F. (2007). Analyse et formalisation des types de phrases averbales du français. Mémoire de Master, Université Paris 7.

ROZE C., DANLOS L. & MULLER P. (2010). LEXCONN : a french lexicon of discourse connectives. In *Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010)*, Moissac, France.

SAGOT B. & BOULLIER P. (2008). SXPipe 2 : architecture pour le traitement présyntaxique de corpus bruts. *Traitement Automatique des Langues (T.A.L.)*, **49**(2), 155–188.