La constitution d'un corpus d'ellipses erronées pour la représentation et la transmission des connaissances

Laura Noreskal[1]

[1]Université Paris Nanterre, MoDyCo UMR 7114, 200 Avenue de la République, 92001 Nanterre, France
laura.noreskal@parisnanterre.fr

Résumé. Cette recherche fait partie d'un projet national, écri+[1], visant la mise en place d'un dispositif d'évaluation, de formation et de certification de la maîtrise du français. Dans cet article, nous proposons de présenter nos choix méthodologiques pour la constitution d'un grand corpus de constructions elliptiques erronées dans le but de leur détection automatique. Le corpus constitué permettra non seulement de tester différentes méthodes de traitement automatique du langage mais également d'avoir une meilleure compréhension des facteurs générateurs d'ellipses erronées chez les étudiants.

Mots-clés. Ellipse, corpus, français, traitement automatique du langage, rédaction d'étudiants

Abstract. This research is part of a national project aimed at setting up a system for evaluating, training and certifying French language proficiency, ecri+. In this article, we propose to present our methodological choices for the constitution of a large corpus of faulty elliptic constructions in order to detect them automatically. The corpus will allow not only to test different methods of automatic language processing but also to have a better understanding of the factors generating erroneous ellipsis in student writings.

Keywords. Ellipsis, corpus, French, natural language processing, student writing

[1]Numéro ANR : ANR17NCUN0015

1 Introduction

À l'ère du numérique, les pratiques pédagogiques sont modifiées par l'évolution des technologies et des usages. Face aux difficultés que rencontrent de nombreux francophones dans la maîtrise du français écrit à l'université, il semble nécessaire de tirer profit de l'innovation numérique pour déployer de nouvelles méthodes d'apprentissage et de représentations des connaissances. Dans ce but, un collectif d'universités développe actuellement un dispositif national d'évaluation, de formation et de certification des compétences d'expression et de compréhension écrites en français : écri+ (ANR17NCUN0015). Il se compose de nombreux modules médiatisés, parmi lesquels un module de rédaction de textes brefs avec détection et correction automatique d'erreurs.

Dans ce cadre, notre contribution porte plus spécifiquement sur la constitution d'un large corpus représentatif des ellipses sujettes aux erreurs dans les rédactions des étudiants. L'ellipse mobilise beaucoup de contraintes linguistiques, comme par exemple des contraintes de parallélisme entre la séquence source et la séquence ellipsée (Abeillé 2008, Desmets 2008, Bîlbîie 2011). Leur non-respect provoque immédiatement une instabilité de l'énoncé (ex : **Les enfants sont allés en Allemagne et Italie*). Parmi les zones de difficultés rédactionnelles identifiées, l'ellipse est donc particulièrement révélatrice des compétences des étudiants. Plusieurs modules ont été élaborés dans le cadre des ressources de l'UOH (Université Numérique des Humanités)[2] afin de dresser un premier repérage des contextes syntaxiques et rédactionnels sensibles favorisant l'apparition d'ellipses (coordination, juxtaposition, comparatives, etc.) et de proposer un premier classement d'occurrences reposant sur la formulation d'hypothèses sur la source des erreurs (défaut du lien logique entre propositions, défaut de la valence du verbe ellipsé, etc.). En complément à ces modules, nous avons décidé de mener une étude linguistique sur les productions correctes et erronées de constructions elliptiques. Cette recherche sur les ellipses sera la première expérimentation pour identifier les erreurs syntaxiques dans les textes brefs d'étudiants.

2 http://www.uoh.fr/front; voir en particulier la plateforme http://uoh.univ-montp3.fr/j_ameliore_ma_maitrise_du_francais/portail/.

2 État de l'art

2.1 Les constructions elliptiques

L'ellipse est un phénomène très fréquent qui a été beaucoup étudié en français (Zribi-Hertz 1986 ; Busquets & Denis 2001 ; Mouret 2007 ; Abeillé 2008 ; Desmets 2008, Bîlbîie 2011). Dans un énoncé constitué de deux propositions A et B liées par une relation sémantique, une ellipse est l'omission dans B d'éléments récupérables grâce à A qui fournit un contexte direct. La proposition B est donc syntaxiquement incomplète mais reste compréhensible grâce aux informations données par les éléments réalisés dans A. La phrase elliptique est donc régie par des conditions sémantiques, pragmatiques et syntaxiques (Desmets 2008 ; Abeillé & Mouret 2008). De nombreuses recherches ont permis de distinguer plusieurs types d'ellipses dont le gapping (Ross 1970) *Pierre aime les fraises et Julie les abricots*, le sluicing (Ross 1969) *Il a réussi et j'ignore comment*, les ellipses "polaires" *Jean est parti et Pierre aussi*, les phrases à factorisation droite *Marie aime mais Pierre déteste le football* et les VP-Ellipsis ou ellipses modales *Pierre a mangé autant de bonbons qu'il a pu*.

Les recherches ont également montré que les ellipses mobilisaient beaucoup de contraintes linguistiques, notamment des contraintes de parallélisme syntaxique, sémantique et pragmatique entre la séquence source et la séquence ellipsée (Abeillé 2008, Desmets 2008, Bîlbîie 2011). Chaque élément ellipsé doit avoir un élément parallèle qui partage les mêmes caractéristiques linguistiques que lui.

(1) *Il grêle et danse.
(2) Il chante et danse.
(3) *Je vais en Espagne et Marie Allemagne.
(4) Je vais en Espagne et Marie en Allemagne.

Ainsi, dans l'exemple (1), l'élément ellipsé (pronom personnel) et l'élément réalisé (pronom impersonnel) ne partagent pas les mêmes caractéristiques linguistiques, ce qui provoque une instabilité de l'énoncé. Dans l'exemple (3), la contrainte non respectée est syntaxique. En effet, le verbe *aller* étant bivalent, il attend un sujet agentif (syntagme nominal) et un syntagme prépositionnel complément locatif. Or, en (3), les deux éléments réalisés dans la séquence ellipsée sont des syntagmes nominaux (*Marie* et *Allemagne*). Ainsi, les exemples (1) et (3) ne respectent pas la contrainte de parallélisme, ce qui empêche la compréhension de l'énoncé. Les exemples (2) et (4) sont des contextes corrects où les contraintes s'appliquent de façon satisfaisante.

2.2 La détection et la résolution automatique d'ellipses

Depuis quelques années, plusieurs recherches ont été menées sur l'annotation de corpus pour la détection et la résolution automatique d'ellipses (entre autres Nielsen

2005 ; Bos & Spenader 2011 ; Gandón-Chapela 2017). Elles ont principalement cherché à détecter et à résoudre des VP-Ellipsis dans des corpus anglophones tels que le British National Corpus (BNC) et le Wall Street Journal (WSJ). Mais aucune n'a porté sur la correction d'ellipses erronées ou sur les productions écrites des étudiants francophones, et c'est sur cette question spécifique que porte notre travail. Cet article expose donc l'étape primordiale de constitution d'un corpus spécifique qui permettra de tester différentes méthodes d'apprentissage automatique pour automatiser la détection et la correction des ellipses.

3 Constitution du corpus

3.1 Méthodologie de constitution du corpus

Afin de constituer notre corpus d'ellipses erronées, nous devons répondre à deux problématiques. La première concerne les types de rédaction privilégiant l'apparition d'ellipses erronées. Nous cherchons à constituer un corpus d'erreurs d'une taille considérable afin de pouvoir comprendre le phénomène et l'analyser. De ce fait, il est nécessaire de savoir dans quel type de productions écrites des étudiants les erreurs sont le plus présentes. Pour répondre à cette question, nous avons collecté plusieurs rédactions d'étudiants de différents niveaux et domaines d'études tels que Science du Langage, Histoire, LLCER Anglais, Droit et Économie. Nous avons ensuite classé les rédactions en deux catégories : les rédactions spontanées (partiels et exercices faits en classe) et les rédactions préparées (devoirs maison, rapports de stage et mémoires). En catégorisant les rédactions, nous pensons pouvoir déterminer la probabilité d'apparition d'ellipses erronées selon la nature de la production écrite. Les rédactions spontanées n'étant pas soumises aux relectures, nous pensons qu'elles contiennent plus d'erreurs que les rédactions préparées. Afin de valider notre hypothèse, nous avons recueilli les deux catégories de rédaction.

Néanmoins, déterminer quel type de rédaction est le plus susceptible de contenir des erreurs ne suffit pas. Il est nécessaire de savoir également quel type d'ellipse est le plus sujet aux erreurs. Notre travail implique de comprendre au mieux le phénomène afin de produire, au terme de notre étude, un outil capable de détecter les ellipses erronées. Il nous faut donc pouvoir établir une typologie des erreurs d'ellipses dans les rédactions. Nous avons alors décidé d'extraire les ellipses erronées avec un contexte conséquent : un contexte gauche de 2 à 3 phrases et un contexte droit d'une phrase. L'ellipse est un phénomène qui ne se produit généralement pas au-delà de deux phrases mais nous avons considéré que le contexte ne devait pas être négligé car il pourrait nous donner des informations sur les causes de l'erreur. Par ailleurs, chaque rédaction contenant une ellipse erronée a été gardée dans son intégralité dans le but de garder des informations comme le placement de l'ellipse dans le texte de rédaction.

Ainsi, depuis le début de notre recherche, nous avons collecté 250 rédactions dont 82 devoirs maison, 1 mémoire, 7 rapports de stage, 119 partiels et 41 exercices faits en classe.

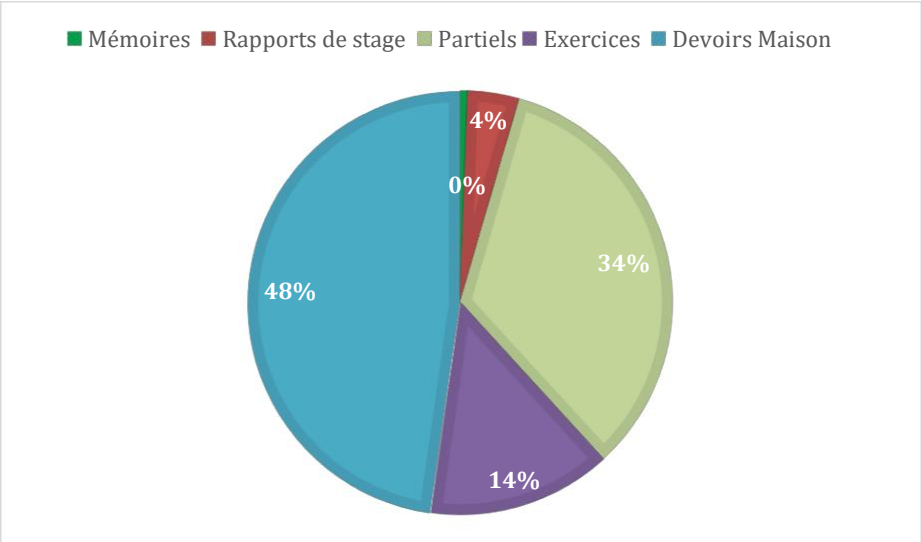
Table 1. Exemples d'ellipses erronées du corpus

Rédaction	Type de rédaction	Niveau	Domaine
Il s'agit soit d'un couple hétérosexuel stérile ou ayant des problèmes de fertilité, soit **<err>** un couple homosexuel ne pouvant, par nature, pas avoir d'enfant.	Exercice	Licence 1	LLCER Anglais
[...] La référence est générique car on parle de toutes les salades et pas **<err>** une en particulier […]	Partiel	Licence 1	Sciences du langage
[…] Toutes ces dispositions permettent à l'élève de réussir et donc **<err>** essayer d'augmenter l'égalité des chances […]	Devoir maison	Master 1	MEEF

3.2 Premières observations

Après avoir récupéré ces 250 rédactions d'étudiants, nous avons remarqué que les ellipses de type gapping posent le plus de problèmes. Nous avons pu extraire 178 constructions elliptiques erronées dont 25 des exercices faits en classe (14%), 1 du mémoire (0%), 7 des rapports de stage (4%), 60 des partiels (34%) et 85 des devoirs maison (48%) (voir Table 2).

Table 2. Répartition en pourcentage des erreurs d'ellipses



Au début de notre recherche, nous avions émis l'hypothèse que les rédactions spontanées (partiels et exercices) contiendraient plus d'erreurs que les rédactions préparées. Or, nous avons pu remarquer que les devoirs maison, que nous avons considérés comme des rédactions préparées, sont les productions qui contiennent paradoxalement le plus d'erreurs dans le corpus constitué à ce jour. Il s'agit d'un point qui nécessite de plus amples réflexions. Les devoirs maison, bien que préparés, ne sont peut-être pas sujets à la même considération de la part des étudiants que les rapports de stage ou les mémoires. Est-ce parce que l'enjeu du travail n'est pas le même pour les devoirs maison que la qualité de l'écrit est moins surveillée ? De fait, les erreurs sont donc beaucoup plus présentes dans ces devoirs que dans les autres rédactions préparées.

Toutefois, comme attendu, les résultats obtenus sur cet échantillon de montrent que le nombre d'erreurs dans les rédactions spontanées est également important (48%). Les résultats confirment donc que les rédactions les plus sujettes aux erreurs sont celles qui ne sont pas obligatoirement soumises à l'étape de relecture. Ainsi, les mémoires (0,6%) et les rapports de stage (4%) sont les rédactions qui contiennent le moins d'erreurs malgré leurs tailles, ce qui rend le relevé d'erreurs peu rentable sur ce type de données.

4 Conclusion et perspectives

Dans cet article, nous avons exposé les différentes études menées sur les constructions elliptiques puis nous avons détaillé les choix méthodologiques pour la constitution de notre corpus dans le but de créer un outil de détection automatique des ellipses dans les copies des étudiants. Nous avons donc choisi de collecter tous les types de rédactions d'étudiants afin d'observer quelles étaient les rédactions les plus sujettes aux erreurs. De plus, nous avons également cherché à savoir quel était le type d'ellipse

le plus susceptible d'être erroné. Les premières observations montrent que les constructions elliptiques erronées sont majoritairement présentes dans les rédactions sans relecture et que le gapping est le principal phénomène touché par une mauvaise construction.

En perspective, nous comptons continuer à collecter davantage de données afin d'atteindre un volume plus important (plus de 200 erreurs) dans le but de tester plusieurs méthodes de traitement automatique du langage pour la détection automatique. Un large corpus permettra également d'analyser plus en détails les caractéristiques propres aux rédactions des étudiants et ainsi permettre une meilleure compréhension des problèmes d'ellipses et de constructions qu'ils rencontrent lors des productions écrites. Notre étude donnera lieu alors à de nouvelles méthodes de travail et de formation pour aider les étudiants à maîtriser leur langue maternelle lors de leurs études universitaires.

Références

1. Abeillé A., Bilbiie G., & Mouret F. : Gapping in Romance : a fragment analysis, *International Conference on Elliptical constructions*, LLF & Université Paris Diderot, Chicago Center, Paris (2008)
2. Abeillé A., Mouret F. : Quelques contraintes sur les coordinations elliptiques en français. *Revue de sémantique et de pragmatique* (2008)
3. Bédard M., Tissot I. : L'Ellipse du complément du nom dans une coordination à deux ellipses. Actes du *Xe Colloque des étudiants en sciences du langage* (2006) 109-134
4. Bilbiie G. : Grammaire des constructions elliptiques. Une étude comparative des phrases sans verbe en roumain et en français, *Thèse de Doctorat*, Université Paris 7 (2011)
5. Bos J., Spenader J.: An annotated corpus for the analysis of VP Ellipsis. *Language Resources and Evaluation* 45 (2011) 463-494
6. Bourreau P. : Traitements d'ellipses : deux approches par les grammaires catégorielles abstraites, Actes de la *20e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013*) pp.215-228
7. Busquets J., Denis P. : L'Ellipse modale en français. *Cahiers de Grammaire* 26 (2001) 55-74
8. Dagnac A. : L'Ellipse modale en français : argument pour une ellipse du TP. *Congrès Mondial de Linguistique Française* - CMLF'08 (2008)
9. Desmets M. : L'ellipse dans les constructions en comme, *LINX* (2008)
10. Gandón-Chapela1 E.: Hunting for Post-Auxiliary Ellipsis in a parsed corpus of English. *Research in Corpus Linguistics 4* (2016) 33-38
11. Hardt D.: An empirical approach to VP Ellipsis. *Computational Linguistics* 23(4) (1997) 525–541
12. Ingham R., Roger G. : L'Ellipse verbale en français : point de vue diachronique. Actes du *XXVIIe Congrès international de linguistique et de philologie romanes*. (2013)

13. Kertz L., Kehler A., Elman J. L.: Evaluating a Coherence-Based Model of Pronoun Interpretation. *Ambiguity in Anaphora Workshop Proceedings* (2006) 49-56
14. Kiefer M., Schuler S., Mayer C., Trumpp N. M., Hille K., & Sachse S.: Handwriting or Typewriting? The Influence of Pen- or Keyboard-Based Writing Training on Reading and Writing Performance in Preschool Children. *Advances in cognitive psychology, 11(4)* (2015) 136-46. doi:10.5709/acp-0178-7
15. Mouret F. : Grammaire des constructions coordonnées en français, *Thèse de Doctorat*, Université Paris 7 (2007)
16. Nielsen L.: A corpus-based study of verb phrase Ellipsis identification and resolution. *Thèse de Doctorat*. King's College London (2005)
17. Troia G., Harbaugh, A., Shankland R., Wolbers K., & Lawrence, A.: Relationships between writing motivation, writing activity, and writing performance: Effects of grade, sex, and ability. *Reading and Writing.* 26(1) (2013) 17-44
18. Ross J. R.: Gapping and the order of constituents. In Manfred Bierwisch & Karl Erich Heidolph (eds.), *Progress in linguistics* (1970) 249-259
19. Ross, J. R.: Guess who? In Robert Binnick, Alice Davison, Georgia Green, and Jerry Morgan (eds.), Papers from the *5th regional meeting of the Chicago Linguistic Society*, 252-286. Chicago Linguistic Society : Chicago, Ill (1969)
20. Zribi-Hertz A. : Relations anaphoriques en français. Thèse de Doctorat d'État, Université Paris 8 (1986)