Détection des ellipses dans des corpus de sous-titres en anglais

Anissa Hamza[1] Delphine Bernhard[1]
(1) LiLPa - EA 1339, Université de Strasbourg, France
hamzaa@unistra.fr, dbernhard@unistra.fr

RÉSUMÉ

Cet article présente une méthodologie de détection des ellipses en anglais qui repose sur des patrons combinant des informations sur les tokens, leur étiquette morphosyntaxique et leur lemme. Les patrons sont évalués sur deux corpus de sous-titres. Ces travaux constituent une étape préalable à une étude contrastive et multi-genres de l'ellipse.

ABSTRACT

Ellipsis Detection in English Subtitles Corpora

This article presents a methodology for detecting ellipses in English based on patterns combining information on tokens, their part-of-speech tags and their lemma. The patterns are evaluated on two subtitles corpora. This work is a preliminary step towards a contrastive and multi-genre study of the ellipsis phenomenon.

MOTS-CLÉS : ellipse, anglais, corpus, sous-titres, détection automatique.

KEYWORDS: ellipsis, English, corpus, subtitles, automated detection.

1 Introduction

L'ellipse renvoie à une incomplétude syntaxique de la phrase qui ne présente pas d'incidence sur la transmission de son contenu sémantique. En effet, grâce à la présence d'un antécédent linguistique ou extralinguistique, le co-locuteur parvient à interpréter la forme invisible d'une séquence donnée et à établir la relation avec le reste des éléments dans le discours. De cette définition, découlent deux principes fondamentaux pour que l'ellipse soit fonctionnelle : (i) une structure syntaxique incomplète `[s]` créant ainsi un vide syntaxique dans la structure de la phrase (en d'autres termes, ce vide ne contient aucune forme linguistique) ; (ii) un antécédent `[a]` qui permet la récupération et la transmission du contenu sémantique des éléments effacés. Une relation de dépendance entre le site elliptique et l'antécédent est ainsi apparente. L'exemple 1 ci-dessous [1] illustre un cas de *pseudogapping* représenté par l'ensemble vide ∅, où le syntagme verbal *put the first one through you* est ellipsé pour éviter la répétition et alléger la forme de la phrase. Ainsi, le syntagme effacé *put the first one through you* est nécessaire uniquement pour la construction syntaxique de la phrase puisque son sens est déduit du contexte, ici, linguistique [2].

(1). Get down ! Get down off it, you old cuckold, I don't care who you are. I'll **put the first one through you !**$_{[a]}$ I swear it, I **will** ∅$_{[s]}$ now ! One ! Two !

(2). Descends ! Descends de là, vieux connard ! Je me fous que ce soit toi. Je tire, je te tire dessus ! Je te jure que **je vais tirer** maintenant. Un ! Deux !

1. Les sites elliptiques sont marqués par ∅ dans les exemples.
2. Sauf mention contraire, les exemples sont extraits du corpus de développement décrit dans la section 3.1.

D'un point de vue contrastif [3], cette ellipse n'est pas autorisée en français puisque les auxiliaires des temps composés ne peuvent apparaître seuls dans la phrase. Comme le montre l'exemple ci-dessus, le site elliptique est traduit par la périphrase *aller+inf* qui exprime le futur proche *je vais tirer*. L'ellipse post-auxiliaire est ainsi rendue impossible **je vais* ∅ puisque *aller* est, en quelque sorte un pseudo-auxiliaire qui ne peut apparaître seul dans cette configuration.

Si certaines catégories de l'ellipse semblent simples à repérer et à caractériser, d'autres présentent davantage de défi à relever tant au niveau théorique qu'au niveau appliqué. L'objectif de cette contribution est ainsi de décrire une méthodologie de détection globale du phénomène elliptique en anglais, mettant en avant les différents enjeux qu'il pose aux outils du TAL, notamment aux étiqueteurs morphosyntaxiques. L'identification automatique des ellipses constitue la première étape d'un travail visant une analyse contrastive (anglais-français) et multi-genres du phénomène elliptique, afin de comprendre son fonctionnement en discours et les problèmes posés pour la traduction, et notamment la traduction automatique.

Nous présentons dans un premier temps une typologie des ellipses en anglais, en vue de leur détection automatique (section 2). Nous décrivons ensuite les corpus utilisés et les patrons d'identification définis par rapport à cette typologie (section 3). Enfin, nous détaillons une évaluation de ces patrons (section 4).

2 Typologie des ellipses à détecter en anglais

Les études théoriques de l'ellipse ont permis d'établir un éventail de classifications. Ces classifications varient d'une approche à l'autre : classification établie selon la composition syntaxique et l'agencement grammatical, selon la situation pragmatique, ou alors selon le contexte proprement dit. Nous suivons ici la taxonomie des ellipses établie par les approches syntaxiques contemporaines de l'étude de l'ellipse, notamment celle de van Craenenbroeck & Merchant (2013). Sachant que certaines ellipses ne figurent dans aucune classification (ellipse du sujet et de l'auxiliaire) et compte tenu des problèmes qu'elles posent lors d'une traduction automatique, nous avons néanmoins décidé de procéder à leur détection. van Craenenbroeck & Merchant retiennent dans leur classification trois types : les ellipses du syntagme verbal dans lesquelles sont classées les ellipses verbales et les ellipses post-auxiliaires, les ellipses propositionnelles dans lesquelles sont classés le *sluicing* et ses sous-catégories, et enfin les ellipses nominales. Afin d'atteindre notre objectif de détection automatique, nous avons simplifié la catégorisation de ces ellipses en les classant uniquement par élément déclencheur et en n'établissant pas de sous-catégorie. Ainsi, nous avons élaboré notre classification en la fondant entièrement sur les critères morphosyntaxiques qu'il est possible de formaliser dans les outils dont nous disposons. À titre indicatif, on trouvera ci-dessous des exemples illustrant les catégories qui seront détectées automatiquement. Les noms utilisés pour les patrons de détection présentés dans la section 3 sont indiqués entre parenthèses.

Ellipse du syntagme verbal

— L'ellipse verbale est l'omission du syntagme verbal et de ses compléments, laissant visible uniquement l'auxiliaire ou l'opérateur. Dans l'exemple ci-dessous, l'ellipse est déclenchée par l'opérateur *do* (`post-do`).
(3). John plays the piano but Maria doesn't ∅.

3. Approcher l'ellipse d'un point de vue contrastif sert à révéler davantage d'irrégularités pouvant contribuer à comprendre sa nature et à la définir. En effet, si ce phénomène passe généralement inaperçu au sein d'une même langue, sa complexité est vite mise en avant, grâce à un effet miroir, lors du passage à une autre langue, puisque certaines langues ne l'autorisent que rarement.

— L'ellipse post-auxiliaire renvoie à l'omission du groupe verbal déclenchée soit par un modal soit par un auxiliaire :
  — Déclenchée par un modal (`post-mod`)
    (4). Lauren can play the guitar and Mike can ∅, too. (Merchant, 2019)
  — Déclenchée par une inversion sujet-verbe ou une *question tag* (`vs-tag`)
    (5). Sit down. Should I ∅ ?
  — Déclenchée par *have* ou *be* (`post-be/have`)
    (6). Are you leaving tomorrow ? No, I'm not ∅".
  — *Pseudogapping*. En anglais, le *pseudogapping* est une ellipse déclenchée par un auxiliaire (`post-aux`), un opérateur (`post-do`) ou un modal (`post-mod`) : la forme non finie du verbe (cas des modaux et de *do*), ou le participe (cas des auxiliaires), sont omis, laissant après l'auxiliaire ou le modal une partie du prédicat.
    (7). John invited Sarah, and Mary did ∅ Jane. (Gengel, 2013)
  — Déclenchée par le marqueur de l'infinitif *to* (`post-to`)
    (8). Do you want me to ∅ ?

Ellipse propositionnelle

— *Gapping* : ellipse où le verbe fini est effacé dans une ou plusieurs constructions parallèles ou propositions coordonnées :
  (9). Mary carries a suitcase, and John ∅ a bag.
— *Sluice* : omission de la proposition entière à l'exception du pronom *wh-* comme dans l'exemple 10 (`post-wh`). Selon le nombre et le type des éléments restant après le pronom *wh-*, une sous-catégorie du *sluice* peut être identifiée [4].
  (10). Someone is knocking at the door, but I don't know who ∅.
— Ellipse dans les questions fragmentaires (`qs-frag`) : cette ellipse a été identifiée lors de l'analyse de certains types de discours (dialogue informels notamment). Elle renvoie à l'omission de l'auxiliaire (dans les temps composés) ou de l'opérateur (dans les temps simples) accompagné du sujet dans les interrogatives, laissant visibles dans la phrase seulement le verbe et ses compléments [5].
  (11). ∅ Going somewhere ?
  (12). ∅ Eat something ? No.

Ellipse nominale Compte tenu de sa rareté dans plusieurs langues, cette ellipse est marginalement étudiée. La liste ci-dessous n'est pas exhaustive mais représente les ellipses que nous avons sélectionnées pour la détection automatique, notamment en raison des problèmes qu'elles posent à la traduction automatique (par exemple de l'anglais vers l'arabe) :

— Ellipse déclenchée par le *'s* du génitif (`post-geni`)
  (13). He took John's car but not Mary's ∅.
— Ellipse déclenchée par un quantifieur (`post-quant`)
  (14). Thank you, but I already have some ∅.
— Ellipse déclenchée par un nombre (`post-card` et `post-ord`)
  (15). If they have eggs, bring me six ∅.
  (16). I have two interesting books, the first ∅ is in French while the second ∅ is in English.

4. Dans la présente contribution, aucune distinction n'est faite entre les sous-catégories du *sluice*.
5. On peut aussi trouver des questions fragmentaires où l'auxiliaire est ellipsé comme dans *You married ?* Nous n'envisageons pas le traitement de ces occurrences ici.

3 Corpus et méthodologie

3.1 Corpus de développement et d'évaluation

Les catégories d'ellipses décrites dans la section précédente sont détectées à l'aide de patrons. Les patrons de détection ont été mis au point manuellement à partir d'un corpus de développement de 5 362 tokens regroupant 331 exemples d'ellipses. Ces occurrences, repérées manuellement dans leur contexte, et mêlées à 120 phrases non-elliptiques, sont toutes extraites de documents authentiques publiés entre 1960 et 2014 et qui n'ont aucun lien avec le corpus d'évaluation : pièces de théâtre (H. Pinter), nouvelles (G. Green, F. Forsyth, J. Arden, ...), articles de presse (The Guardian, Mail), dialogues et romans (M. Barbery, J. Coe). Pour compléter ce corpus, nous avons également utilisé des exemples issus de (McShane & Babkin, 2016) et (Rønning *et al.*, 2018b) afin d'augmenter le nombre d'occurences elliptiques et couvrir ainsi le plus de variation possible.

Afin d'évaluer la performance des patrons, deux sous-corpus d'évaluation de tailles différentes ont été utilisés. Ils appartiennent tous deux au registre conversationnel de sous-titres. En effet, l'ellipse, en tant que propriété de discours spontané, est plus fréquente dans ce type de discours (Baird *et al.*, 2018). Par ailleurs, le corpus des sous-titres sélectionné propose également une version française, ce qui permettra à l'avenir de réaliser une étude contrastive. Le premier (Corpus 1) est extrait de séries répertoriées dans Opus *OpenSubtitles* [6] (Tiedemann, 2012), et compte 197 302 tokens et 1 270 occurrences d'ellipses. Le deuxième (Corpus 2) contenant 36 676 tokens et 396 occurrences d'ellipses, est constitué des sous-titres de séries Broadchuch et Downton Abbey, récupérés des DVD (Strong & Lyn, 2018; Percival *et al.*, 2018) et compilés par nous-même. Pour repérer les types d'ellipses à l'aide des patrons établis, les deux corpus, de développement et d'évaluation, ont été étiquetés morphosyntaxiquement à l'aide de l'étiqueteur morphosyntaxique de Stanford (Manning *et al.*, 2014). Par ailleurs, ces corpus ont été annotés manuellement par l'une des auteurs en ajoutant un code selon l'élément déclencheur de l'ellipse. Une vérification a ensuite été effectuée par les deux auteurs, à l'aide d'expressions régulières simples afin de détecter les occurrences éventuellement manquantes non annotées. En effet, l'ellipse reste un phénomène peu fréquent et il est donc nécessaire de parcourir une grande quantité de texte pour obtenir un nombre satisfaisant d'occurrences, ce qui augmente la possibilité d'oublier des occurrences lors de l'annotation.

3.2 Définition de patrons de détection

La détection des ellipses repose sur des patrons définis à l'aide de l'outil *TokensRegex*, inclus dans Stanford CoreNLP (Chang & Manning, 2014), qui permet de rechercher des séquences de tokens en combinant différentes informations (forme, lemme, partie du discours, etc.). Nous avons préféré des patrons opérant sur les tokens plutôt que sur l'analyse syntaxique en dépendance car ces derniers sont difficiles à mettre en oeuvre dans certains cas, qui nécessitent de prendre en compte l'ordre des mots dans la phrase (ellipses déclenchées dans les *question tags* ou par l'inversion du sujet et du verbe). Nous avons aussi observé de nombreuses erreurs dans les analyses syntaxiques obtenues.

La Table 1 donne quelques exemples de patrons (la totalité des patrons définis pour chaque type d'ellipse est présentée dans la Table 3). Un ou plusieurs patrons ont été établis pour identifier une même catégorie d'ellipse, en prenant en compte les conditions morpho-syntaxiques de chaque catégorie présentée en section 2. Les patrons sont relativement longs et détaillent toutes les conditions possibles de façon à atteindre le plus grand degré de précision possible. Par exemple, certains comportements syntaxiques récurrents sont observables dans le cas de l'ellipse du sujet et de l'auxiliaire dans la

6. http://opus.nlpl.eu/OpenSubtitles-v2016.php

question fragmentaire (`qs-frag`, voir exemples 11 et 12). En effet, si la phrase commence par un verbe principal (étiqueté *VB*) et se termine par un point d'interrogation, on peut supposer qu'il y a deux vides syntaxiques : un vide sujet et un vide auxiliaire. Le type de l'auxiliaire omis est ensuite interprété selon la flexion du verbe :

— S'il est réduit à un participe passé *ed* ou présent *ing*, l'auxiliaire absent peut être soit *have* soit *be*,

— S'il est une forme non-finie, l'opérateur *do* ou un modal manque,

L'un des patrons définis pour les questions fragmentaires est présenté dans la Table 1. Il repère une phrase contenant un verbe non-conjugué et se terminant par un point d'interrogation. Seuls trois types de tokens peuvent précéder ce verbe, de manière optionnelle : un tiret ou un point (marqueurs typographiques d'un dialogue dans les sous-titres), éventuellement suivis d'un adverbe ou d'un adjectif, puis d'une virgule. Par ailleurs, on ne devra pas trouver un pronom personnel après le verbe.

Type	**Patron**
`qs-frag`	`^/[-.]+/? [{pos:/RB|JJ/}]? /,/? [{pos:/VB|VB[^PZD].*/}] [!{pos:/PRP/} & ! /,/] []* /[?]/`
`post-do`	`[!{pos:/W.*/}][{pos:/(PRP.?|NN.?)/}] []{0,2} [{lemma:do} & !{pos:/VB[GN]/}] [{lemma:not}]? [{pos:PRP}]? /[.;!:?,]+/`
`post-geni`	`[!{lemma:of}] [!{pos:PRP}] [pos:POS] /[.;!?:,]/`

TABLE 1: Exemples de patrons de détection.

Pour l'ellipse `post-do`, plusieurs critères syntaxiques sont à considérer pour que *do* soit déclencheur d'une ellipse. Il est déclencheur lorsque dans la phrase il est un :

— verbe de suppléance :
(17). They sang a song or mother **did** ∅.

— auxiliaire de négation :
(18). Do you believe in miracles ? **I don't** ∅.

— auxiliaire d'inversion contrainte par certains adverbes :
(19). He didn't allow him to speak. Neither **did** I ∅.

— auxiliaire emphatique dans les assertives affirmatives qui ne contiennent pas d'auxiliaire (*have/be*) ou de modal :
(20). You just missed a toast. Oh yes, you **did** ∅.

Ces quatre cas autorisent le vide syntaxique laissé par l'omission du verbe lexical qui, s'il était restitué, n'impacterait pas la grammaticalité de la phrase, à condition que l'ensemble du syntagme verbal (VB et compléments) soit restitué. De plus, *do* est supplétif dans la mesure où il ne se charge pas *entièrement* du sens du verbe de la phrase, mais le *complète*. Il est ainsi considéré comme déclencheur d'une ellipse post-auxiliaire. Le patron `post-do` présenté dans la Table 1 récupérera donc toute phrase contenant le lemme *do* précédé par un pronom personnel ou un nom et non précédé par un pronom *wh-*. Le lemme *do* peut être suivi par *not* marqueur de négation ou d'un pronom personnel mais ne devra pas être suivi par une forme verbale.

Un dernier exemple que nous pouvons illustrer ici est le cas du génitif `post-geni`. Dans les trois possibilités qui existent pour exprimer la possession en anglais (la préposition *of*, les pronoms possessifs *his, hers, yours, etc.* et le morphème *–'s/–s'*), seul le morphème *–s* est considéré comme déclencheur d'une ellipse nominale comme dans :

(21). But Rosetti was writing about her own death, not her grandparents' ∅.

En effet, l'utilisation des pronoms possessifs est strictement anaphorique et n'engendre aucun vide syntaxique dans la structure de la phrase. Le patron récupérera ainsi tous les tokens étiquetés POS (possession), non précédés par un pronom personnel ou par le lemme *of*[7], suivis immédiatement d'un signe de ponctuation.

3.3 Limites de la méthodologie

De la liste présentée dans la section 2, seul le *gapping* ne fera pas l'objet d'une détection automatique. En effet, en l'absence d'un élément déclencheur, il est difficile d'établir avec notre méthodologie un patron qui prendrait en compte toutes les variations et les propriétés syntaxiques d'une occurrence de *gapping* et ce quel que soit le niveau de l'analyse, syntaxique ou morphosyntaxique.

Pour ce qui est de l'analyse morphosyntaxique, l'établissement d'un patron pour détecter ce type d'ellipse est particulièrement complexe, compte tenu du nombre variable des éléments résiduels dans la phrase et de la multiplicité de leurs étiquettes morphosyntaxiques. L'absence du verbe entre le sujet et son complément d'objet, par exemple, peut difficilement être formalisée dans un patron TokensRegex, en raison des faux positifs que le patron pourrait repérer. En effet, la figure 1 ci-dessous illustre deux occurrences étiquetées : la première est une phrase non-elliptique et présente le cas de deux éléments (*a candy* et *a cake*) coordonnés avec *and* et étiquetés `NN`. Ces deux éléments n'entretiennent pas de relation sujet-objet comme c'est le cas de la deuxième phrase où *and* coordonne deux propositions, dans une configuration très similaire du point de vue de la séquence des étiquettes morphosyntaxiques (`DT NN CC DT NN`). Le verbe de la deuxième proposition (*the chief ∅ a bag*) est omis présentant de ce fait un *gapping*.

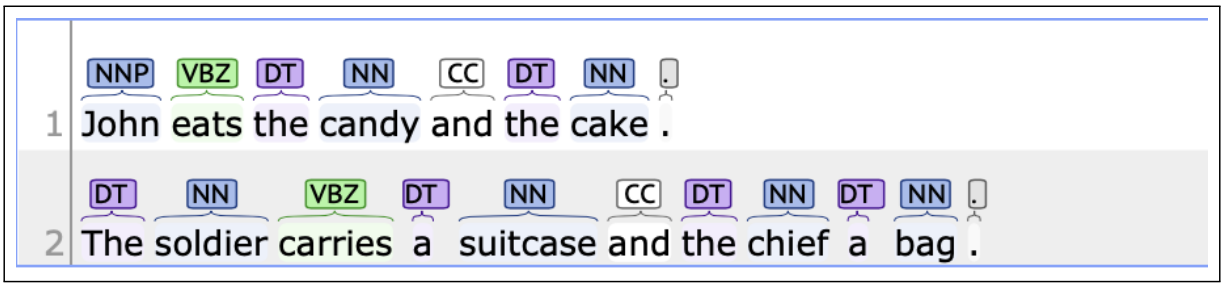

FIGURE 1: Exemple d'étiquetage d'une phrase non-elliptique et du *gapping*.

La détection du *stripping*, sous-catégorie du gapping, aurait pu être envisagée grâce aux marqueurs *too, as well*, et *also* toujours présents après le site elliptique, ou aux conjonctions de coordination *and* et *or* qui le précèdent :

(22). Jane Likes apples and Maria ∅ too.

Cependant, ce type de cas est très marginal et n'est pas représentatif du *gapping*. De la même manière les autres configurations, notamment les constructions parallèles déclenchant ce phénomène de *gapping*, ne pourront pas être détectées.

De ce fait, compte tenu des limites de TokensRegex d'une part, et de la complexité à fixer des conditions stables du phénomène elliptique dans le *gapping* d'autre part, nous avons dû écarter sa détection automatique[8].

7. Les constructions *friend of, part of* ne sont pas traitées comme elliptiques ici.
8. D'autres types d'ellipse que nous n'avons pas mentionnés dans la classification sont également difficiles à détecter automatiquement tels que *Bare ellipsis Argument* et les réponses fragmentaires.

4 Évaluation et discussion des résultats

Les résultats obtenus pour chaque type d'ellipse dans les trois corpus utilisés sont détaillés dans la Table 2, selon l'élément déclencheur.

Le discours conversationnel dont les sous-titres font partie est représentatif d'une communication caractérisée par une parole spontanée [9] et interactionnelle, susceptible d'offrir un usage de l'ellipse très varié. On remarque par exemple que les ellipses déclenchées par les modaux, *be/have, to, do, wh*, et les *questions tag* `vs-tag`, toutes apparentées à l'ellipse du syntagme verbal et propositionnelle sont beaucoup plus fréquentes que l'ellipse nominale. L'échange ci-dessous contenant deux ellipses, une déclenchée par *can* et l'autre par *to*, partageant le même antécédent *speak to Mark*, est repéré par deux patrons `post-mod` et `post-to` :

(23). - Beth, have you spoken to Mark ? - I can't ∅. - You've got to ∅.

Les patrons sont bien adaptés aux types d'ellipses qui sont présents dans le corpus de développement (sauf pour `qs-frag`). Il reste difficile d'y inclure toutes les variations syntaxiques qui peuvent se trouver dans un corpus de grande taille, comme le montrent les performances moindres obtenues sur les corpus d'évaluation 1 et 2. À l'exception de 4 catégories (`post-card, post-quant, post-be/have, vs-tag`), les résultats montrent une F-mesure plus élevée dans le corpus 2 que dans le corpus 1. Le rappel est globalement élevé, ce qui répond à nos besoins pour une étude ultérieure des ellipses dans des corpus de grande taille. Il est en effet plus simple de filtrer des faux positifs manuellement que de parcourir exhaustivement des corpus de grande taille pour retrouver les faux négatifs. Par ailleurs, même si le corpus de développement intègre des exemples issus de différents genres textuels, il faudra encore tester les patrons sur d'autres genres (discours parlementaire, journalistique, littérature, textes scientifiques, ...). Enfin, il restera à améliorer certains patrons : `qs-frag`, `post-be/have`, `post-geni`. Les résultats plus faibles observés pour ces types d'ellipse peuvent être justifiés par leur rareté et leur complexité qui se manifeste par des configurations non-observées dans un corpus de développement de petite taille.

Type d'ellipse	**Corpus de dév.**				**Corpus 1**				**Corpus 2**			
	#	P	R	F1	#	P	R	F1	#	P	R	F1
`post-do`	34	0,97	1,00	0,99	188	0,57	0,88	0,69	45	0,69	0,93	0,79
`post-mod`	72	0,96	0,96	0,96	147	0,71	0,96	0,81	44	0,79	0,86	0,83
`vs-tag`	31	0,96	0,87	0,92	305	0,58	0,97	0,72	72	0,58	0,93	0,72
`post-be/have`	31	0,84	0,84	0,84	185	0,47	0,70	0,57	52	0,50	0,67	0,57
`post-to`	30	1,00	0,93	0,97	39	0,62	0,87	0,72	36	0,79	0,83	0,81
`post-wh`	43	0,93	1,00	0,97	223	0,44	0,98	0,61	82	0,62	0,99	0,76
`qs-frag`	32	0,94	0,53	0,68	90	0,40	0,74	0,52	41	0,73	0,54	0,62
`post-card`	6	1,00	1,00	1,00	67	0,28	0,94	0,43	10	0,21	0,70	0,33
`post-geni`	19	1,00	0,89	0,94	5	0,25	0,60	0,35	12	0,83	0,83	0,83
`post-quant`	29	1,00	0,93	0,96	17	1,00	0,94	0,97	1	0,50	1,00	0,67
`post-ord`	4	1,00	0,75	0,86	4	0,50	1,00	0,67	1	1,00	1,00	1,00

TABLE 2: Résultats de l'évaluation. La colonne # correspond au nombre d'instances.

Les patrons n'ont pas toujours atteint une précision parfaite lors de leur application sur le corpus d'évaluation, notamment la catégorie `post-card` avec une précision de 0,28 dans le corpus 1 et

9. Les acteurs suivent un script imitant les conversations spontanées.

de 0,21 dans le corpus 2, ou encore la catégorie de post-geni qui a atteint seulement 0,25 dans le corpus 1. Après avoir parcouru les ellipses détectées ou non par chaque patron, les problèmes semblent être liés à deux types d'erreurs :

— Erreurs dues à la précision insuffisante des patrons qui renvoient :
  — au manque d'étiquettes suffisamment représentatives et précises pour annoter les déclencheurs cruciaux de l'ellipse par exemple la non-distinction entre *to* préposition *went to+nom* et *to* marqueur d'infinitif *want to+verb* (24) ou entre *do* comme auxiliaire supplétif/opérateur et *do* comme verbe plein (25) :
  (24). a. All_DT my_PRP$ life_NN I_PRP 've_VBP wanted_VBD somebody_NN to_TO talk_VB to_TO._. (Non elliptique)
  b. You_PRP can_MD call_VB me_PRP a_DT sap_VBP if_IN you_PRP want_VBP to_TO ∅._.
  (25). a. What_WP are_VBP you_PRP trying_VBG to_TO do_VB,—, Babe_NNP ?_. (Non elliptique)
  b. I_PRP know_VBP him_PRP better_JJR than_IN they_PRP do_VBP ∅._.
  — à la difficulté d'affiner le patron pour prendre en compte davantage de variations des structures elliptiques. On trouve par exemple des cas d'ellipse qs-frag dans les déclaratives, qui ont été annotées manuellement comme elliptiques, mais non détectés car aucun patron n'a été dédié aux phrases déclaratives. On pourrait en effet obtenir beaucoup de faux-positifs détectés dans les phrases à l'impératif :
  (26). Read_VB that_DT book_NN ?_. (qs-frag ellipse d'un pronom et de *have*)
  (27). Read_VB that_DT book_NN now_RB._. (impératif non elliptique)
  (28). Read_VB that_DT book_NN ._. (ellipse d'un pronom personnel et de *have* dans la réponse fragmentaire)
— Erreurs engendrées par l'étiqueteur : ces erreurs sont liées au mauvais étiquetage de certaines catégories, en raison de l'ambiguïté qu'un seul mot peut présenter. En effet, le *'s* comme marqueur du génitif *customer's* et le *'s* de la forme contractée de *be* ou *have* à la 3e personne *father's been there* (29) sont tous les deux étiquetés comme POS ci-dessous :
  (29). a. Whose car is this ? The_DT costumer_NN 's_POS ∅._.
  b. Father_NN 's_POS been_VBN there_RB._. (Non elliptique)
  Par ailleurs, les deux phrases ci-dessous sont détectées comme ellipses qs-frag pourtant la première ne l'est pas.
  (30). Can_MD you_PRP open_VB up_RP,_. please_ VB ?_.
  (31). We_PRP was_VBD hired_VBN as_IN his_PRP$ bodyguard_NN,_. see_VB ?_.
  Dédié à repérer toute occurrence contenant un verbe non précédé d'un nom et suivi d'un ? le patron a détecté *please*, ici incorrectement étiqueté VB au lieu de UH (interjection) comme elliptique. Cette ambiguïté morphosyntaxique est également à l'origine de la non détection des ellipses par le patron comme dans l'échange *Flattering her ? Oh yes he is* où *flattering* est étiqueté NN au lieu de VBG.

Par le biais de ces erreurs nous pouvons pointer les difficultés et les limites des patrons à base de tokens dues à un étiquetage pas suffisamment détaillé, voire erroné, ou encore en raison de l'instabilité du phénomène elliptique. Malgré ces restrictions, l'utilisation de patrons à l'aide de tokens reste toutefois, pour l'instant, une opération avantageuse dans la mesure où elle offre une méthodologie simple de détection. Elle pourrait être améliorée si d'autres paramètres étaient inclus dans l'élaboration des patrons. On pourrait par exemple enrichir le corpus de développement. Nous nous sommes limitées dans cet article à 331 occurrences elliptiques pour développer nos patrons mais une collecte manuelle d'occurrences diverses et variées et en aussi grand nombre que possible, pour fastidieuse qu'elle soit,

pourrait contribuer au perfectionnement des patrons. De plus, dans la mesure où *do, have, be* et *to* sont des déclencheurs essentiels de l'ellipse et compte tenu d'absence d'étiquettes les distinguant des autres catégories, il serait utile de parvenir à les étiqueter de manière plus fine en distinguant leur utilisation en tant d'auxiliaires et verbes pleins.

5 État de l'art

Les études de l'ellipse en TAL notamment pour l'anglais, restent peu nombreuses par rapport aux investigations théoriques. Le nombre limité de ces recherches est indubitablement lié aux défis que présente le phénomène elliptique. Comment la détecter quand on la remarque à peine ? Bos & Spenader (2011) par exemple expliquent le manque de recherches effectuées sur la détection automatique des ellipses, notamment des ellipses verbales, par la présence de deux difficultés principales. La première tient à l'absence d'outils ou de procédures servant à localiser l'ellipse et son antécédent, et la seconde, aux lacunes des recherches théoriques qui, une fois l'ellipse et son antécédent approximativement repérés, se concentrent sur la tâche de résolution, négligeant alors le problème de l'antécédent. À la lumière des travaux théoriques existants, il apparaît que les trois éléments-clefs à prendre en compte dans une étude de l'ellipse sont le site elliptique, l'antécédent et le contexte (syntaxique et/ou sémantique).

Pour ce qui est du site elliptique, il est important et nécessaire de définir ce qui manque (la nature et la fonction des éléments ellipsés). Le site elliptique semble poser des problèmes au TAL dans la mesure où le nombre de tokens ellipsés varie d'une occurrence à une autre. Pour ce qui est de l'antécédent, il est important de le localiser si la résolution de l'ellipse est envisagée au même titre que sa détection. Ceci peut indéniablement aider à délimiter le site elliptique. En revanche, le problème qui se pose, pour les outils du TAL en particulier, concerne la nature même de l'antécédent. Que faire dans le cas d'un antécédent extralinguistique ? Enfin le contexte : toutes les ellipses ne peuvent pas être détectées via la syntaxe. De plus, même dans le cas où la syntaxe semble suffire, le même type d'ellipse peut révéler des variations dépendantes de paramètres externes à la syntaxe. En effet, le genre et le type de discours peuvent affecter les comportements syntaxiques de l'ellipse.

Compte tenu de ces éléments clefs, les travaux sur l'analyse automatique de l'ellipse distinguent plusieurs étapes : (1) la détection des sites elliptiques [10] (2), la détection et la délimitation de l'antécédent, (3) la résolution de l'ellipse (comblement du site elliptique pour reconstituer une phrase non-elliptique). Pour ce qui est des types d'ellipses traitées, les ellipses du syntagme verbal restent les plus étudiées en traitement automatique des langues. Il existe également quelques travaux traitant du *sluicing* ou *sluice*, de l'ellipse du sujet ou du *gapping*.

McShane & Babkin (2016) décrivent un système, appelé ViPER (*VP Ellipsis Resolver*) qui vise à détecter et résoudre certains types particuliers d'ellipses du syntagme verbal avec un haut niveau de précision. La détection des ellipses traitées se fait en 3 étapes, combinant des patrons de surface et une analyse syntaxique en dépendances avec Stanford CoreNLP. Liu *et al.* (2016) proposent un système par apprentissage pour détecter les ellipses du syntagme verbal, entraîné à partir de deux corpus annotés (celui de Bos & Spenader (2011) et celui de Nielsen (2005)). L'objectif du système est de prendre une décision binaire concernant les modaux et les auxiliaires (site elliptique ou non). Les descripteurs utilisés pour la classification reposent sur l'étiquetage morphosyntaxique, la lemmatisation et l'analyse syntaxique en dépendance. La F-mesure obtenue est 69,52 pour le corpus de Bos & Spenader (2011) et 75,39 pour celui de Nielsen (2005).

10. Appelée détection de la cible (*target detection*) par Liu *et al.* (2016).

De nombreux travaux se focalisent de fait plutôt sur la détection de l'antécédent ou la résolution de l'ellipse, considérant ainsi la première étape de détection des ellipses comme résolue. Hardt (1992) propose un algorithme à base de règles pour déterminer les antécédents des ellipses du syntagme verbal (*Verb Phrase ellipsis*). L'algorithme prend pour entrée un syntagme verbal elliptique et une liste de syntagmes verbaux apparaissant à proximité (même phrase ou deux phrases précédentes). L'algorithme élimine alors les antécédents impossibles et trie les autres syntagmes verbaux par ordre de préférence. Pour les besoins de l'évaluation, les exemples d'ellipses ont été collectés dans le corpus Brown (étiqueté en parties du discours) à l'aide d'expressions régulières détectant les auxiliaires qui n'ont pas de verbe dans le contexte proche. Pour ce qui est du *sluice*, on pourra retenir les travaux récents de Baird *et al.* (2018) et Rønning *et al.* (2018a,b) : le premier article concerne la classification manuelle du type de *sluice* dans un corpus de sous-titres (à partir de *sluices* détectés automatiquement à l'aide d'expressions régulières), tandis que les deux derniers se focalisent sur la résolution des *sluices*. Le *gapping* est traité par Schuster *et al.* (2018) afin de produire des représentations de l'analyse syntaxique en dépendance qui encodent explicitement le matériel élidé. Les expériences présentées reposent notamment sur des phrases sélectionnées à partir d'une relation de dépendance spécifique (`orphan`) dans un des corpus *Universal Dependencies* (UD) pour l'anglais et des phrases collectées manuellement à partir de diverses ressources. Le *gapping* est également au cœur des travaux de Droganova *et al.* (2018a) qui mettent en avant la représentation choisie pour ce phénomène dans les corpus UD, consistant à promouvoir un des dépendants orphelins à la position du parent manquant et conduisant les analyseurs à prédire des relations entre des mots qui ne sont généralement pas reliés par une relation de dépendance. Par ailleurs, le phénomène est rare et donc peu représenté dans les corpus d'entraînement, ce qui complique encore la tâche pour les analyseurs syntaxiques. Pour combler le manque d'exemples dans les corpus d'entraînement, Droganova *et al.* (2018b) ont produit semi-automatiquement des phrases artificielles qui sont similaires à des constructions elliptiques du point de vue de leur structure. Enfin, Droganova & Zeman (2017) mettent en avant les nombreuses erreurs d'annotation manuelle relevées pour les constructions elliptiques dans les corpus UD, ce qui rend compte de la complexité du phénomène.

Les travaux présentés ici se concentrent généralement sur un type particulier d'ellipse, alors que nous proposons de détecter une grande variété de phénomènes elliptiques. Par ailleurs, les objectifs de l'étude de l'ellipse en TAL concernent avant tous les problèmes posés à l'analyse syntaxique (annotation manuelle, analyse automatique). Notre objectif est différent, dans la mesure où nous envisageons une analyse linguistique contrastive (comparaison anglais-français) et multi-genres afin d'étudier l'impact des phénomènes elliptiques sur la traduction humaine et automatique.

6 Conclusion et perspectives

Nous avons présenté une méthodologie de repérage automatique des ellipses en anglais évaluée sur des corpus de sous-titres. Il est important d'ajouter qu'à ce stade, établir des patrons à base de séquences de tokens associés à leur étiquette morphosyntaxique et leur lemme peut paraître rédhibitoire dans le sens où les conditions sont limitées aux tokens dont la moindre variation (même une erreur de tokénisation) entrave la reconnaissance de l'ellipse. Cette étape est cependant un préalable nécessaire à un premier repérage des occurrences elliptiques. L'utilisation de ces patrons peut également constituer une réelle amorce en vue d'une détection et d'une classification des ellipses à base d'un apprentissage automatique. De plus, une détection automatique de l'ellipse et de son antécédent pourrait, à notre sens, contribuer à une classification ainsi qu'à une reconnaissance automatique des genres de discours. En effet, si l'ellipse se révèle comme une particularité d'un genre spécifique, sa détection automatique permettrait certainement de définir le genre de texte analysé. De la même façon, sachant que l'ellipse

reste l'une des erreurs les plus fréquentes et cruciales rencontrées dans la traduction automatique aujourd'hui, aux côtés de la métaphore et certaines figures de style, sa détection et sa résolution informatisées sont ainsi les enjeux inévitables dans la recherche dans ce domaine.

7 Annexe

Type	**Nb de patrons**	**Patrons**
post-do	2	^[{lemma:no}]? [/[.;!:?,]/]? [{pos:/(PRP.?|NN.?)/}] []* [{lemma:do} & !{pos:/VB[GN]/}] [{lemma:not}]? [{pos:PRP}]? /[.;!:?,]+/
		[!{pos:/W.*/}][{pos:/(PRP.?|NN.?)/}] []{0,2} [{lemma:do}& !{pos:/VB[GN]/}] [{lemma:not}]? [{pos:PRP}]? /[.;!:?,]+/
vs-tag	4	[]* [{lemma:/(be|do|have)/} | {pos:MD}] [{lemma:not}]? [{pos:PRP}] /[.!?:]+/
		[]* [{lemma:/(be|do|have)/} | {pos:MD}] [{lemma:not}]? [{pos:DT}]? [{pos:PRP}] /[.]+/$
		^ [{lemma:/(be|do|have)/} | {pos:MD}] [{lemma:not}]? [{pos:/PRP|NN.*/}]? /,/? [{pos:/PRP|NN.*/}]? /[?]+/$
		[]* /,/ [{lemma:/(be|do|have)/} | {pos:MD}] [{lemma:not}]? [{pos:/PRP|NN.*/}]? /,/
post-mod	2	[{pos:/PRP|NN.*|WP|CC/} | {lemma:/,/}] [{pos:RB}]? [{pos:MD}] [{pos:/[^V].*/}]* /[.,:!;?-]+/
		[{pos:/PRP|NN.*|WP/}] [{pos:RB}]? [{pos:MD}] [{pos:CC} | {pos:"IN";lemma:/(if|unless)/}] []* /[.,:!;?-]+/
post-be/have	2	[!{pos:/V.*/} & !{pos:/WP.*/}] [{pos:/PRP.?|NN.?/}] [{pos:RB}]? [{lemma:/be|have/}] [{lemma:not}]? /[.;!?:,-]+/
		^[{pos:/PRP.?|NN.?/}] [{pos:RB}]? [{lemma:/be|have/}] [{lemma:not}]? /[.;!?:,-]+/

TABLE 3: Totalité des patrons utilisés pour la détection. Les patrons sur fond grisé sont expliqués de manière détaillée dans la section 3.2.

Type	Nb de patrons	Patrons
post-to	2	[{pos:/VB.*/ }] [{pos:PRP}]? [{lemma:not}]? [{pos:to}] /[.?,:!;]+/
		[{pos:/VB.*/ }] [{pos:RB}]? [{pos:JJ}]? [{lemma:not}]? [{pos:to}] /[.?,:!;]+/
post-wh	1	[{pos:/^W.*/} & {lemma:/(wh(at|y|o|en|ere|ich|ose)|how)/}] [!{pos:/V.*/} & !{pos:MD}]* [{lemma:/[,.!;:?-]+/} | {pos:CC}]
qs-frag	3	^/[-.]+/? [{pos:/RB|JJ/}]? /,/? [{pos:/VB|VB[^PZD].*/}] [!{pos:/PRP/} & ! /,/] []* /[?]/
		^/[-.]+/? [{pos:/RB|JJ/}]? /,/? [{pos:/VB|VB[^PZD].*/} !{lemma:thank}] [{pos:/PRP|NN/}]? /[?]/
		/,/ [{pos:/VB|VB[^PZD].*/} & !{lemma:thank}] [{pos:/PRP| NN/}]? /[?]/
post-geni	1	[!{lemma:of}] [!{pos:PRP}] [pos:POS] /[.;!?:,]/
post-quant	1	"/[sS]ome|[aA]ny/ /[.;!?:,]+/
post-card	2	[!{pos:/N.*/}] [{pos:CD} & !{lemma:one}] /[.;!?:,]+/
		[!{pos:/N.*/} & !{pos:JJ} & !{lemma:/th(at|ese|is|ose)/}] [{pos:CD} & {lemma:one}] /[.;!?:,]+/
post-ord	1	/the/ [{pos:/JJ.*/;ner:ORDINAL}] /[.;!?:,]+/

TABLE 3: Totalité des patrons utilisés pour la détection (suite).

Références

BAIRD A., HAMZA A. & HARDT D. (2018). Classifying Sluice Occurrences in Dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, p. 1580–1583.

BOS J. & SPENADER J. (2011). An annotated corpus for the analysis of VP ellipsis. *Language Resources and Evaluation*, **45**(4), 463–494.

CHANG A. X. & MANNING C. D. (2014). *TokensRegex : Defining cascaded regular expressions over tokens*. Rapport interne CSTR 2014-02, Department of Computer Science, Stanford University.

DROGANOVA K., GINTER F., KANERVA J. & ZEMAN D. (2018a). Mind the Gap : Data Enrichment in Dependency Parsing of Elliptical Constructions. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, p. 47–54.

DROGANOVA K. & ZEMAN D. (2017). Elliptic Constructions : Spotting Patterns in UD Treebanks. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, p. 48–57.

DROGANOVA K., ZEMAN D., KANERVA J. & GINTER F. (2018b). Parse Me if You Can : Artificial Treebanks for Parsing Experiments on Elliptical Constructions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, p. 1845–1852.

GENGEL K. (2013). *Pseudogapping and ellipsis*. Oxford, Royaume-Uni : Oxford University Press.

HARDT D. (1992). An algorithm for VP ellipsis. In *Proceedings of the 30th annual meeting of the Association for Computational Linguistics*, p. 9–14.

LIU Z., PELLICER E. G. & GILLICK D. (2016). Exploring the steps of verb phrase ellipsis. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, p. 32–40.

MANNING C. D., SURDEANU M., BAUER J., FINKEL J., BETHARD S. J. & MCCLOSKY D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, p. 55–60.

MCSHANE M. & BABKIN P. (2016). Detection and Resolution of Verb Phrase Ellipsis. *LiLT (Linguistic Issues in Language Technology)*, **13**.

MERCHANT J. (2019). Ellipsis : a survey of analytical approaches. In J. VAN CRAENENBROECK & T. TEMMERMAN, Eds., *The Oxford Handbook of Ellipsis*, p. 19–45. Oxford, Royaume-Uni : Oxford University Press.

NIELSEN L. A. (2005). *A corpus-based study of Verb Phrase Ellipsis Identification and Resolution*. Thèse de doctorat, King's College London.

PERCIVAL B., BOLT B., HALL E., SPIRO M. & ENGLER M. (2018). Downton Abbey - Saisons 1 à 6 - L'intégrale de la série. DVD.

RØNNING O., HARDT D. & SØGAARD A. (2018a). Linguistic Representations in Multi-task Neural Networks for Ellipsis Resolution. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, p. 66–73.

RØNNING O., HARDT D. & SØGAARD A. (2018b). Sluice Resolution without Hand-Crafted Features over Brittle Syntax Trees. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 2 (Short Papers)*, p. 236–241, New Orleans, Louisiana.

SCHUSTER S., NIVRE J. & MANNING C. D. (2018). Sentences with Gapping : Parsing and Reconstructing Elided Predicates. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL 2018)*.

STRONG J. & LYN E. (2018). Broadchurch Saisons 1 + 2. DVD.

TIEDEMANN J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, p. 2214–2218, Istanbul, Turkey.

VAN CRAENENBROECK J. & MERCHANT J. (2013). *Ellipsis phenomena*, In M. DEN DIKKEN, Ed., *The Cambridge Handbook of Generative Syntax*, p. 701–745. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.