# UNIVERSITÉ DE STRASBOURG

*ÉCOLE DOCTORALE 520* **« Humanités »**

**UR 1339 LiLPa (Linguistique, Langues, Parole)**

# THÈSE présentée par :


## Anissa HAMZA


soutenue le : **20 septembre 2019**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : **Sciences du langage**

**Linguistique générale et Linguistique anglaise**

# *La détection et la traduction automatiques de l'ellipse*

# *Enjeux théoriques et pratiques*

**THÈSE dirigée par :**

| | |
|---|---|
| **Madame BOISSEAU Maryvonne** | Professeur, Université de Strasbourg |
| **Madame BERNHARD Delphine** | Maître de conférences, Université de Strasbourg |

**RAPPORTEURS :**

| | |
|---|---|
| **Madame BOUILLON Pierrette** | Professeur, Université de Genève |
| **Monsieur MILLER Philip** | Professeur, Université de Paris-Diderot |

---

**AUTRES MEMBRES DU JURY :**

| | |
|---|---|
| **Monsieur GRASS Thierry** | Professeur, Université de Strasbourg |
| **Monsieur LOOCK Rudy** | Professeur, Université de Lille |
| **Monsieur YVON François** | Professeur, Université Paris Sud |

*À mes parents qui, de leur lointain village, m'ont toujours soutenue dans mon long parcours.*

*« Parce qu'une thèse, ça ne s'écrit pas tout seul »*
*DoXtra*

Qu'il me soit permis tout d'abord d'exprimer ma gratitude à Maryvonne Boisseau qui m'a accueillie dès mon inscription en master et m'a guidée dans le choix de mon projet de thèse, pour enfin la diriger et la conduire grâce à des échanges réguliers très fructueux, grâce à ses lectures critiques, mais aussi grâce à la confiance qu'elle m'a toujours témoignée. Ce travail est ainsi le fruit des nombreuses discussions que nous avons eues tout au long de ces années. Merci de m'avoir amenée jusqu'au terme de cette recherche.

Je voudrais ensuite exprimer de chaleureux remerciements ainsi que ma reconnaissance à Delphine Bernhard, co-directrice de recherche qui m'a initiée au Traitement Automatique des Langues, et qui est à l'origine de mon intérêt pour ce domaine et des rencontres enrichissantes que j'ai pu faire tout au long de ce travail. Sa bienveillance, sa patience, son expertise et sa disponibilité m'ont permis d'explorer de nouvelles idées et de nouvelles manières de mener mes recherches. Merci de m'avoir ouvert le champ des possibles.

Je souhaite également remercier très sincèrement les professeurs, membres de mon jury : Pierrette Bouillon et Philip Miller pour avoir accepté d'être rapporteurs de cette thèse, les professeurs Thierry Grass, Rudy Loock et François Yvon pour avoir accepté de faire partie de ce jury. Je tiens également à remercier François Yvon de m'avoir accueillie au laboratoire LIMSI lors du stage que j'y ai effectué en 2017.

Tous mes remerciements vont également à

– Pierre Gançarski et Franck Burlot, membres de mon comité de suivi de thèse, pour nos rencontres et l'intérêt qu'ils ont manifesté pour mon travail.

– Daniel Hardt, Pranav Anand, Jim McCloskey, Adrian Brasoveanu et Austin Baird pour leur accueil chaleureux et leur accompagnement lors de mon séjour de recherche à l'université de Californie Santa Cruz.

– Catherine Schnedecker, directrice de l'école doctorale 520, qui s'est toujours intéressée à l'avancement de mes travaux, et qui a toujours été présente pour m'écouter et me conseiller.

– Rudolph Sock, directeur de l'EA 1339 LiLPa de l'université de Strasbourg, pour son amitié et sa sympathie, et pour l'ambiance de travail stimulante au sein du laboratoire.

– Julia Pustch, Directrice de DLADL, et Anne Bandry, Doyenne de la faculté des langues pour leur soutien sans faille.

Je tiens à remercier tout particulièrement Chantal pour sa présence de tous les instants et pour les relectures de ma thèse. En m'accompagnant par son écoute attentive mois après mois, elle m'a permis de travailler dans un cadre agréable et serein.

Je remercie également Julie pour les responsabilités partagées au sein de l'EA LiLPa et l'ER FDT, pour nos communications et pour nos longs échanges scientifiques stimulants. Je la remercie aussi pour ses relectures et son amitié. J'espère que notre collaboration se poursuivra !

Je voudrais aussi remercier Rodrigo pour ses conseils et son aide sur les questions statistiques et pour nos discussions fructueuses au sein du LiLPa.

Un grand merci à :
– mes amis Aurélie, Béatrice, Sarra, Martine, Marguerite, Jean-Luc, Nicolas, Jean-Yves, Othmen, Khaoula, et surtout à Marine,
– mes collègues Gilles, Camille, Laurence, Didier,
– mes camarades doctorants et docteurs Bruno, Hasna, Seto, Yuliya, qui m'ont tous aidée, chacun à leur manière, à garder bon sens et joie de vivre.

*Last but not least*, merci à Christophe pour sa grande patience et son soutien indéfectible. Cette thèse te doit beaucoup, et moi aussi.

*En juillet 2018, une éclipse de lune annoncée comme la plus longue du siècle, accompagnée de conditions météorologiques permettant une observation aisée, a suscité un réel intérêt populaire. Une petite fille de mon entourage, se réjouissant de l'attraction dont parlaient les adultes, a lancé joyeusement à la ronde : « Ce soir, nous allons voir une ellipse de lune ». Mot d'enfant sans doute, mais quelle jolie confusion ! La proximité de la sonorité de ces mots ne reflétait-elle pas également un cousinage sémantique ? Car semblable au soleil qui, caché par la lune, n'est jamais totalement invisible grâce à l'étrange lumière ocre dont il la nimbe lors d'une éclipse, le mot ou le syntagme dans l'ellipse, remarquable par son silence, n'est jamais entièrement absent, ni totalement présent. L'éclipse de lune me ramenait donc, de façon inattendue, à mon sujet d'étude.*

# Table des matières

# Liste des tableaux

# Liste des figures

## Listes des abréviations

GN Groupe Nominal
GV Groupe Verbal
SN Syntagme Nominal
SV Syntagme Verbal
PP Syntagme prépositionnel
AP Syntagme adjectival
Ø Site elliptique
TA Traduction Automatique
TR Traduction de Référence
TV Verbe Transitif
PAE (Post-auxiliary ellipsis) Ellipse post-auxiliaire
VPE (Verbal Phrase Ellipsis) Ellipse du groupe verbal

Abréviations de corpus

| | | |
|---|---|---|
| <CEx> | Corpus d'exemples utilisés | |
| <CDEV> | Corpus de développement | |
| <CP> | Corpus d'évaluation EUROPARL | Genre politique |
| <CL> | Corpus d'évaluation PLECI Littéraire | Genre littéraire |
| <CJ> | Corpus d'évaluation PLECI Presse | Genre journalistique |
| <CPr> | Corpus d'évaluation TED | Genre promotionnel |

# Étiquettes morphosyntaxiques de TokensRegex

**fondées sur les abréviations Penn Treebank**[1]

| POS | Description Tag | Exemple |
|---|---|---|
| CC | coordinating conjunction | and |
| CD | cardinal number | 1 |
| DT | determiner | the |
| EX | existential there | there is |
| FW | foreign word | les |
| IN | preposition, subordinating conjunction | in, of, like |
| IN/that | that as subordinator | that |
| JJ | adjective | green |
| JJR | adjective, comparative | greener |
| JJS | adjective, superlative | greenest |
| MD | modal | could, will |
| NN | noun, singular or mass | table |
| NNS | noun plural | tables |
| NP | proper noun, singular | John |
| NPS | proper noun, plural | Vikings |
| POS | possessive ending | friend's |
| PP | personal pronoun | I, he, it |
| PP$ | possessive pronoun | my, his |
| RB | adverb | however, usually, naturally, here, good |
| RBR | adverb, comparative | better |
| RBS | adverb, superlative | best |
| RP | particle | give up |
| SENT | Sentence-break punctuation | . ! ? |
| TO | infinitive 'to' | to go |
| UH | interjection | uhhuhhuhh |
| VB | verb be, base form | be |
| VBD | verb be, past tense | was, were |

[1] https://www.sketchengine.eu/penn-treebank-tagset/ (accès vérifié le 30 juillet 2018 à 14:10)

| | | |
|---|---|---|
| VBG | verb be, gerund/present participle | being |
| VBN | verb be, past participle | been |
| VBP | verb be, sing. present, non-3d | am, are |
| VBZ | verb be, 3rd person sing. present | is |
| VH | verb have, base form | have |
| VHD | verb have, past tense | had |
| VHG | verb have, gerund/present participle | having |
| VHN | verb have, past participle | had |
| VHP | verb have, sing. present, non-3d | have |
| VHZ | verb have, 3rd person sing. present | has |
| VV | verb, base form | take |
| VVD | verb, past tense | took |
| VVG | verb, gerund/present participle | taking |
| VVN | verb, past participle | taken |
| VVP | verb, sing. present, non-3d | take |
| VVZ | verb, 3rd person sing. present | takes |
| WDT | wh-determiner | which |
| WP | wh-pronoun | who, what |
| WP$ | possessive wh-pronoun | whose |
| WRB | wh-abverb | where, when |

## Conventions

Nous avons utilisé les guillemets chevrons « » pour citer et le **gras** et l'*italique* pour le soulignement d'insistance et les mots anglais.

Dans les exemples analysés, nous avons utilisé le **gras** pour indiquer l'antécédent et l'élément déclencheur de l'ellipse, l'*italique* pour les citer dans le corps du texte, et, occasionnellement, le soulignement pour attirer l'attention sur d'autres éléments.

Nous utilisons le signe **Ø** pour signaler le site elliptique et les crochets [ ] pour son interprétation.

*Cette représentation relève de choix arbitraires, retenus pour des raisons pratiques et ne relève d'aucune théorie particulière.*

Les citations en anglais ne sont pas traduites.

# Introduction

*Eclipsis est defectus dictionis, in quo necessaria verba desunt.*

*St Isidore*[2]

Notre travail de recherche, intitulé *La détection et la traduction automatiques de l'ellipse : enjeux théoriques et pratiques*, poursuit deux objectifs essentiels. Il s'agit tout d'abord de vérifier la possibilité de détecter automatiquement le phénomène elliptique pour, ensuite, explorer les procédures facilitant une traduction automatique acceptable. En effet, la traduction automatique de l'ellipse, pour qu'elle soit juste et prenne en compte le vide que l'ellipse introduit dans la chaîne linéaire, suppose que cette dernière ait été préalablement détectée.

Par conséquent ce double objectif complexifie l'objet de cette thèse qui s'inscrit dans un paradigme interdisciplinaire faisant intervenir la linguistique contemporaine sur l'ellipse, les « nouvelles » linguistiques, outillée et de corpus, et, enfin, le Traitement Automatique des Langues (TAL) dont le développement s'articule à celui de la linguistique et de l'informatique et des recherches actuelles dans le domaine de l'Intelligence Artificielle.

Devant l'impossibilité d'être en position d'expertise pointue dans chacun de ces domaines, notre travail emprunte à chacun d'eux des outils ponctuels nécessaires à l'avancée de nos travaux, et, selon les mots de Zribi-Hertz (1996, présentation), il s'inscrit dans une « démarche qui sous-tend une recherche en mouvement », du fait

[2] *Etymologiarum, Liber I* 'De Grammatica', ch. XXXIV 'De Vitiis', sec.10 (van Craenenbroeck & Merchant 2013, 740).

de son évolution constante. Il s'agit en effet de mener une analyse exploratoire et expérimentale en procédant par étapes afin de cerner les enjeux théoriques et appliqués qu'implique le traitement automatique de l'ellipse, comme nous l'énoncions plus haut, tant au niveau de sa détection qu'à celui de sa traduction.

Qualifier notre approche d'expérimentale signifie que l'étude que nous présentons est tout à la fois une expérience de pensée, de recherche et d'écriture sur un fait de langue qui, en dépit d'une littérature scientifique abondante, ne va pas de soi. En nous focalisant entièrement sur ce phénomène linguistique très précis qu'est l'ellipse, du grec *élleipsis*, lui-même portant la racine indo-européenne *leik* signifiant « laisser », et qui renvoie dès son origine à la notion de manque, de disparition, de suppression ou d'imperfection, nous avons immédiatement rencontré la complexité apparente et immédiate du phénomène et très vite orienté nos observations vers l'intérêt de le soumettre aux performances des nouveaux systèmes de traduction automatique comparées à celles des traducteurs humains.

Notre ambition n'est ainsi nullement de proposer une nouvelle théorie de l'ellipse ou d'arbitrer les nombreuses querelles disciplinaires à son sujet. Il s'agit plutôt de montrer, après avoir délimité le cadre théorique de notre étude, qu'il est possible d'élaborer au sujet de l'ellipse, des procédures automatiques de détection et de traduction, sans pour autant être assurée à ce stade de l'absolue justesse des résultats qui ne peuvent être que provisoires, compte tenu de l'évolution très rapide des recherches en Traduction Automatique (TA).

De cette manière, identifier et observer un fait de langue, tel l'ellipse, dans son passage d'une langue à une autre revient à relever deux défis simultanés. Le premier est d'ordre identitaire : qu'est-ce qu'une ellipse ? Quand et comment la repère-t-on ? Sous quelles conditions est-elle possible ? Le second, d'ordre fonctionnel, plus particulièrement au niveau de la traduction, vise à considérer son usage dans la langue source et l'éventuel effet miroir que lui tend la langue d'accueil. En effet, l'ellipse, par sa nature à la fois discrète et complexe, pose nombre de questions, concernant autant la recherche théorique que la recherche appliquée. En adoptant une définition simple, nous parlerons d'ellipse lorsqu'un ou plusieurs segments (au

niveau syntaxique) sont absents du discours, sans que le sens général de la séquence en soit modifié, restant ainsi compris par les co-locuteurs dans la plupart des cas. Comme l'écrit Fontanier : « l'ellipse consiste dans la suppression de mots qui seraient nécessaires à la plénitude de la construction, mais que ceux qui sont exprimés font assez entendre pour qu'il ne reste ni obscurité ni incertitude » (1968, 35). L'exemple le plus parlant d'une ellipse est la fameuse phrase extraite du roman de Herman Melville, *Bartleby the Scrivener* (1853), *I would prefer not to*. En effet, l'une des occurrences de cette réplique reproduite ci-dessous (1), présente une ellipse signalée avec l'ensemble vide Ø après le marqueur de l'infinitif *to* qui est ici son déclencheur. Pour comprendre ce qui est omis, on se réfère à la phrase précédente où le segment *help me compare this sheet here* est considéré comme antécédent qui aide à l'interprétation.

(1) I want you to **help me compare this sheet here**—take it,"
and I thrust it towards him.
"I would prefer not **to** Ø," said he. (Melville [1853],
2003, 40)

Ce premier exemple permet donc de noter, d'ores et déjà, la présence d'un antécédent, nécessaire à l'interprétation, et d'un déclencheur. Toutefois, l'antécédent des ellipses n'est pas toujours aussi facile à repérer et la catégorisation des ellipses est rendue difficile du fait du nombre de déclencheurs possibles.

L'exemple (2) ci-dessous présente une ellipse dans la seconde partie de la phrase :

(2) <CEx> I told her to leave me alone, but she **would**n't Ø.

On voit qu'il s'agit d'une ellipse déclenchée par un modal *would* et l'on peut rétablir le syntagme absent en allant le récupérer dans la première proposition. Cette forme elliptique sera catégorisée comme une ellipse post-auxiliaire (PAE). Notre travail montrera qu'il existe bien d'autres types d'ellipses – déjà répertoriées et catégorisées par des études antérieures – mais nous pouvons dès maintenant percevoir, grâce à ces deux exemples, que les problèmes de détection sont liés, entre autres, à l'antécédent (présent explicitement ou non), et à la présence ou non d'un élément déclencheur.

L'ellipse a toujours suscité de nombreux débats et ce depuis les premières études faites sur le langage et les langues (avant même de nommer la linguistique en tant que telle), s'appuyant au fil du temps sur les apports rhétoriques, philologiques, stylistiques, théoriques et enfin appliqués pour la compréhension du phénomène. Si l'importance et la constance des recherches sur le sujet peuvent amener à croire que l'ellipse ne réserve plus rien de nouveau aux chercheurs, il n'en reste pas moins que certains questionnements, théoriques, comme appliqués, sont loin d'être résolus. En réalité, incertitude sémantique, ambiguïté, et complexité caractérisent le phénomène, ce qui s'observe d'emblée lorsque l'on s'intéresse à son histoire dans l'évolution de la réflexion sur le langage et les langues. À seulement parcourir en effet les acceptions de l'ellipse à travers la période antique, la période médiévale, celle de la Renaissance jusqu'aux périodes modernes et contemporaines, on ne manquera pas de remarquer qu'à chacune d'elles correspond une tendance singulière, qu'elle soit d'ordre rhétorique, grammatical, ou stylistique.

En rhétorique, par exemple, l'ellipse a été répertoriée comme une figure de style parmi d'autres. Les grammairiens, par contre, en relation avec la conception d'une « norme », la caractérisent par l'absence d'un mot dans une phrase. Il en est de même dans le domaine du cinéma et de la littérature où, dans la réduction du contenu d'une scène, l'ellipse renvoie à la suppression de parties de la narration au profit de l'avancée de l'intrigue. Enfin, la notion de manque se teinte d'une connotation d'imperfection qui s'étend métaphoriquement jusqu'au vocable géométrique, où l'ellipse est un cercle aplati, donc imparfait. Cette transversalité ainsi que l'utilisation répandue de son acception lui ont conféré une sorte de notoriété, qui, paradoxalement, n'a contribué qu'à lui léguer un statut théorique instable à partir d'une identité, et partant d'une définition, aux contours flous, notamment dans les sciences du langage. Ainsi, dès la période antique, les variations des considérations à son sujet l'ont fait passer, au cours du temps, d'une notion simple dont les principes semblaient élémentaires, à celle d'un phénomène complexe, véritable pierre d'achoppement divisant grammairiens, rhétoriciens et philosophes, notamment en raison des hésitations et des confusions suscitées par les

exemples fournis, de telle sorte qu'en tant que phénomène langagier, elle a échappé à toute définition claire et précise. Ainsi, dès la tradition grammaticale antique et médiévale, l'ellipse a brouillé les frontières de la grammaire et de la rhétorique, ce qui lui a conféré ce statut marginal, simplificateur et ambigu puisque chacune des disciplines, grammaire ou rhétorique, estimait que l'étude du phénomène relevait de la compétence de l'autre. On a donc tenté de remédier à ce déficit définitionnel par une classification appartenant soit à la grammaire, soit à la rhétorique.

Par ailleurs, pour comprendre ce qu'est l'ellipse, les rhéteurs comme les grammairiens se sont beaucoup servis d'œuvres littéraires et de textes poétiques cependant restreints aux deux langues anciennes que sont le grec et le latin. Ce n'est qu'à partir des recherches de Sanctius, grammairien espagnol du XVI$^{e}$ siècle (1523-1601), que leur intérêt s'est étendu aux langues vernaculaires comme l'espagnol. Les exemples furent alors extraits d'œuvres littéraires dont le style était travaillé, mais dont le choix pouvait apparaitre souvent arbitraire car simplement fondé sur la popularité des écrivains et de leurs écrits ; ainsi en-est-il, par exemple, de l'utilisation exhaustive des textes de Virgile où les extraits n'ont étayé ni théorie solide, ni aucune description détaillée du phénomène puisqu'ils ne représentaient aucunement la langue parlée spontanée alors en usage et n'illustraient de ce fait qu'une petite partie des phénomènes linguistiques (Chanet 1983, Lallot 1983, Pitavy & Bigot 2008). Ce n'est que plus tard, au Moyen Âge, que l'abandon progressif de ces exemples parfaitement écrits au regard de la norme en usage, se fera au profit d'exemples extraits de discours d'orateurs.

La première explication syntaxique et exhaustive de l'ellipse revient donc à Sanctius qui introduit le principe d'*économie* pour illustrer la notion de manque, propre au phénomène, marquant par là une rupture dans l'histoire du concept. L'innovation apportée par ce grammairien a, comme on vient de le mentionner, consisté, d'une part, à s'intéresser à l'espagnol, langue vernaculaire, là où tous ses prédécesseurs s'étaient restreints aux langues classiques latine et grecque, et, d'autre part, à cibler ses recherches sur la totalité du discours et non sur une seule de ses parties.

À partir de cet apport, de la Renaissance aux périodes contemporaines, l'intérêt porté à l'ellipse ne cessera de croître, lui faisant acquérir un statut à part entière. Elle n'est plus désormais un phénomène que l'on étudie parmi d'autres. Son usage n'est plus considéré comme simple procédé stylistique, mais comme révélateur de problématiques inhérentes à l'usage de la langue elle-même.

Ainsi la linguistique moderne a-t-elle vu émerger au siècle dernier tout un éventail d'approches proposant une variété d'analyses sur le même phénomène, à commencer par les contributions de Harris et de Chomsky. Leurs travaux marquent le début d'une conception nouvelle de l'ellipse qui se maintiendra dans les investigations ultérieures incluant le recensement du phénomène et de ses structures. Chez Harris, le terme d'ellipse n'apparait pas à proprement parler, mais il développe ses questionnements autour de ce qu'il appelle *effacement*, terme qui se rapproche du signe zéro de Saussure et qui continue à être utilisé aujourd'hui pour renvoyer à l'ellipse et la définir. Chez Chomsky, c'est la place octroyée aux catégories vides qui est utilisée pour parler de l'ellipse (Chomsky 1993, 1995).

Comme on le voit plusieurs notions se mêlent encore pour définir l'ellipse : celles de manque et complétude (chez Bally), d'effacement (chez Harris), de catégories vides (chez Chomsky). On pense à Saussure quand on lit que l'ellipse marque une rupture entre langue et pensée, à Lamy lorsque l'on rencontre la notion d'abréviation, à Halliday & Hasan si l'on croise celle de présupposition.

Mais à présent, la recherche d'une définition de l'ellipse n'est toutefois plus une préoccupation essentielle des études qui s'attachent plutôt à définir les conditions de son apparition dans le discours ainsi que ses fonctions, à établir son impact syntaxique et sémantique dans la chaîne de la communication et à la distinguer d'autres phénomènes linguistiques avec lesquels elle peut se confondre, comme l'anaphore par exemple.

L'intérêt que l'ellipse a suscité – et continue de susciter – chez les linguistes permet d'ores et déjà de faire ressortir quatre pans fondamentaux sur lesquels reposerait sa définition actuelle la plus complète. Ces quatre versants visent tous un point d'équilibre théorique entre la forme et le sens véhiculé par le phénomène

elliptique : le versant syntaxique privilégie la structure visible et les éléments omis, le versant sémantique s'attache au sens et aux valeurs de cette structure, le versant pragmatique situe l'ellipse dans son contexte de production, enfin, le versant énonciatif se focalise sur la place de la relation entre les sujets (locuteur et co-locuteur) dans le processus de récupérabilité (interprétation des éléments manquants ou *retrievability* en anglais).

De cette manière, dans la littérature scientifique actuelle traitant de l'ellipse, sa caractérisation est établie à partir de critères appartenant à la syntaxe, la sémantique, la pragmatique, et l'énonciation. Mais ce sont les syntacticiens qui ont apporté jusqu'à présent le plus d'éléments de réponses aux interrogations portant sur le phénomène. L'approche syntaxique vise trois objectifs principaux : (i) définir les caractéristiques du phénomène elliptique au sein de la linguistique, (ii) recenser les conditions sous lesquelles il apparaît, et enfin (iii) identifier les éléments contribuant à sa résolution.

Ces axes de recherche ont pour résultat l'émergence de plusieurs approches, structurales et non structurales, affirmant ou niant la présence d'une structure syntaxique dans le site elliptique. Parmi ces travaux, nous citons plus particulièrement ceux de van Craenenbroeck & Merchant (2013), Johnson (2011), Culicover & Jackendoff (2005), Hardt (1993), Miller (2014) dont les nombreuses recherches pour parvenir à une définition de l'ellipse, se rejoignent sur certaines caractéristiques, telles son universalité et sa complexité. Ces positionnements ont bien évidemment induit des taxonomies sujettes à variation au sein d'un même courant.

Par ailleurs, bien que le statut théorique de l'ellipse ait été traité d'une manière large et diversifiée par les linguistes théoriciens, il reste peu exploré dans le domaine du TAL où pourtant les études en cours envisagent notamment le repérage automatique de différents types d'ellipses. Nous notons particulièrement les efforts entrepris, allant des approches fondées sur corpus annotés de l'ellipse, jusqu'aux tentatives de sa détection et de sa résolution automatisées, par Hardt (1993), Nielsen

(2004), Bos & Spenader (2011), McShane & Babkin (2016), et Rello (2010), entre autres.

Il nous est apparu cependant que si ces études ont bien cherché à détecter automatiquement le phénomène elliptique, elles n'ont jamais énoncé le réel objectif de cette détection, comme si la seule détection était en soi un objectif suffisant. Or, toute opération de traitement automatique a une finalité ; dans le cas de l'ellipse on peut penser que la détection peut permettre une meilleure visualisation de ce qui la déclenche, ou, et c'est ce qui motive notre travail, on peut faire l'hypothèse qu'une détection automatique fine de tout type d'ellipse facilitera son traitement automatique lors de sa reconnaissance en vue de sa traduction. S'agirait-il là d'une finalité pratique permettant une traduction automatique sans erreurs ?

En fait, depuis 2016, la traduction automatique suit, pour ainsi dire, les progrès dans la recherche sur l'Intelligence Artificielle, discipline fondée sur un ensemble de techniques informatiques s'inspirant des mécanismes cognitifs humains. Le passage de la traduction statistique (à base de segments mémorisés par l'ordinateur) à la traduction neuronale (en référence au réseau des neurones biologiques), offre chaque jour davantage d'amélioration de la qualité des opérations effectuées tant il est vrai que ce véritable changement de paradigme, induisant des performances accrues de nouveaux outils, livre désormais des traductions de plus en plus pertinentes et précises. Cette évolution ne manque pas d'intéresser tant les chercheurs du secteur académique que les grands organismes, les développeurs, les professionnels ou les médias. Cependant, l'analyse d'une sélection d'ellipses a pu nous montrer que certaines d'entre elles résistaient *encore* au développement de la traduction automatique. L'extrait ci-dessous est un exemple d'ellipse déclenchée par le modal *will* introduisant l'effacement du segment *ask her*. *Will* est ainsi traduit automatiquement par le verbe *aller*, une sorte de pseudo-auxiliaire qui ne convient pas ici.

(3) <CEx> - Chloe was there, so you can **ask her**.
- We **will** Ø.

- Chloé était là, alors tu peux lui demander.
- Nous **allons**.
(Google Traduction, juillet 2018)

Cette difficulté réside vraisemblablement dans l'impossibilité pour les systèmes de traduction à repérer l'antécédent de l'ellipse qui dépend directement du contexte d'insertion de l'occurrence. C'est ainsi que nous défendons l'idée qu'une détection automatique d'une ellipse pourrait contribuer à la compréhension des erreurs engendrées lors de sa traduction (en l'occurrence de l'anglais au français). C'est ce que notre étude vise précisément à appréhender, tant il semble évident qu'avec l'évolution des outils de traduction à l'heure actuelle, cette détection automatisée qui embrasse un vaste champ d'analyse des composants de la phrase, sera amenée à jouer un rôle important dans l'évaluation et l'amélioration de la traduction de l'ellipse.

Ainsi, en lien avec les considérations théoriques et conceptuelles nécessaires aux fondements de tout travail de recherche, en nous appuyant par ailleurs sur les définitions et les clarifications des notions-clefs évoquées précédemment, nos premières observations nous ont orientée vers la formulation des hypothèses de travail suivantes :

i) La définition de l'ellipse menant à une classification morphosyntaxique de ses manifestations, jointe à l'établissement des critères nécessaires à son repérage, devraient faciliter sa détection automatique.

ii) Les erreurs observées lors de détection pourraient contribuer à la compréhension des sources d'erreurs (de l'ellipse) dans la traduction automatique du phénomène.

Pour vérifier nos hypothèses, nous nous appuyons sur les analyses morphosyntaxiques qui, dans un premier temps, paraissent suffisantes à la détection automatique de *certaines* catégories d'ellipse puisqu'elles facilitent la décomposition du phénomène afin de mieux le situer parmi d'autres. Ceci constitue le point de

départ de notre méthodologie, sachant toutefois que cette position n'implique aucunement un rejet du versant sémantique. Cependant, dans la mesure où nous établissons une classification morphosyntaxique à des fins de repérage du phénomène, l'objet de notre recherche vise donc de prime abord à appréhender son identité syntaxique à l'aide de critères, qui, comme nous le verrons, explicitent les conditions d'apparition de l'ellipse et apparaissent comme les premiers indices d'une présence elliptique.

Par ailleurs, notre méthodologie adopte une démarche semi-contrastive qui offre la possibilité de dégager les spécificités de l'ellipse en anglais à partir de l'analyse de sa traduction en français. Pour ce faire, nous aurons recours à l'utilisation d'un corpus parallèle et multi-genres permettant de repérer les variations inter et intra-langues manifestées par le phénomène elliptique dans la langue et le discours qui l'ont généré et d'analyser sa distribution.

Nous présenterons donc les enjeux de cette détection et les problèmes soulevés par l'évolution de la traduction automatique. Nous évoquerons également les difficultés méthodologiques inhérentes à notre démarche interdisciplinaire et les limites d'ores et déjà perçues de notre entreprise, avant d'énoncer les différentes étapes de notre démarche fondée sur des observations de nature empirique et pratique.

La description de nos travaux est ainsi présentée selon l'avancée de notre réflexion se déroulant à travers les étapes présentées dans les cinq chapitres suivants :

– le premier chapitre, intitulé « Ancrages théoriques et pratiques », revient sur des notions clefs directement impliquées dans le processus elliptique et décrit l'une des classifications qui a émergé des approches syntaxiques contemporaines. Il présente également les nouvelles orientations du TAL, croisant recherches théoriques et pratiques qui, en visant sa reconnaissance automatique, ne manquent pas d'initier un regain d'intérêt à l'égard du phénomène elliptique. L'objectif de ce chapitre est ainsi de dégager, à partir des travaux existants, une définition opérationnelle de

l'ellipse menant à une classification de ses différentes catégories en vue de sa détection automatique ;

– pour ce faire, le deuxième chapitre, intitulé « Méthodologie », est consacré entièrement à l'élaboration d'une méthode de détection. Il s'agit tout d'abord de présenter la notion de corpus, allant de sa constitution (à partir de différents genres) à son exploitation, pour détailler ensuite les outils utilisés (dont la plupart est issue de l'ensemble CoreNLP). L'ancrage de cette notion dans notre recherche s'effectue à partir de critères de collecte qui ont permis de sélectionner un corpus adapté à la vérification de nos hypothèses. Enfin, pour expliquer certains choix visant à pallier les difficultés rencontrées, les limites et les contraintes de la méthodologie et des outils sont examinées, comme le sont celles du corpus à exploiter. Le résultat de la détection automatique menée en suivant la méthode exposée, est présenté dans les deux chapitres qui suivent ;

– le troisième chapitre intitulé « Évaluation de la détection automatique de l'ellipse dans un corpus de sous-titres », détaille l'établissement de patrons à partir des analyses morphosyntaxiques des exemples sous formes de conditions et de requêtes pour repérer chaque catégorie d'ellipse identifiée. Sont présentés ici les résultats de l'application de ces patrons sur le corpus de sous-titres à l'aide de TokensRegex (utilitaire de Stanford CoreNLP permettant de définir des expressions régulières sur les tokens). Leur performance sera alors évaluée à travers le recensement des erreurs rencontrées et classées par typologie. Le quatrième chapitre, quant à lui, retrace la performance des patrons évaluée cette fois-ci sur un corpus multi-genres, et analyse la distribution des catégories d'ellipses dans les genres analysés, l'objectif étant d'observer dans quelle mesure ces patrons peuvent être utilisés sur des corpus autres que celui exploité dans le chapitre précédent ;

– consacré à la traduction de l'ellipse, humaine et automatique, dans lesquelles l'étendue de la complexité du phénomène devient patente, le dernier chapitre œuvre à la vérification de notre deuxième hypothèse concernant les éventuelles retombées de la détection automatique de l'ellipse. Notre objectif est alors de vérifier s'il est possible d'établir une passerelle entre les erreurs de la détection et celles de la

traduction automatique des ellipses, afin d'ouvrir un champ d'investigation susceptible d'initier de nouvelles procédures aboutissant à une traduction acceptable du phénomène. En nous appuyant sur la méthodologie de la linguistique contrastive, nous proposerons à ce stade, une évaluation qualitative d'une sélection d'occurrences elliptiques pour relever tant les difficultés que les avancées de la traduction automatique dans le cas précis de l'ellipse post-auxiliaire.

Il convient toutefois de signaler que certains résultats de cette recherche risquent probablement d'apparaître obsolètes au moment où notre travail en cours sera déposé, et ceci en raison du développement technologique duquel il est dépendant. Comme on a pu le souligner précédemment, la dynamique de ce secteur ne cesse d'étonner. Lors de notre première inscription en doctorat en 2014, par exemple, nous avions traduit à l'aide de Google une série d'occurrences elliptiques que nous avons à nouveau vérifiées à la fin de notre parcours actuel. Le résultat montre que certains problèmes repérés précédemment sont d'ores et déjà résolus, produisant ainsi une traduction tout à fait acceptable, comme le montre l'exemple (4) ci-dessous.

(4) <CEx> I said I **will** Ø.

- J'ai dit que je **vais.** (Google Traduction Juin 2014)
- J'ai dit que je **le ferais**. (Google Traduction Juillet 2018)

De plus, avant d'engager l'exposé de notre recherche dans les pages qui vont suivre, nous aimerions (re)poser ou (re)formuler certaines questions restées à l'arrière-plan de notre progression : si l'ellipse est ce silence que l'on entend, quelles sont donc ses limites ? Comment délimiter ses frontières ? un regard morphosyntaxique est-il suffisant pour approcher le phénomène ? Quelle classification pourrait aider à sa reconnaissance automatique ? Quelle(s) retombée(s) réalisable(s) pouvons-nous imaginer pour sa détection ?

Nous souhaitons par ce travail, apporter sinon des réponses complètes, du moins des amorces de réponses à ces questions et nous espérons surtout que les éléments significatifs que nous mettrons au jour à travers notre approche, contribueront à

améliorer les procédures de détection automatique ainsi que les performances des traductions du phénomène elliptique.

# Chapitre 1

# Ancrages théoriques et pratiques

*McCloskey notes that speakers and writers often leave out informationally redundant grammatical material—such as when the verb "call" is omitted in "Jay Z called, but Beyoncé didn't." This process, known as ellipsis, is widespread across the languages of the world, and is particularly common in informal language and dialogue*[3].

Il n'apparaît pas utile à notre sujet de recherche de décrire plus avant le développement historique des débats évoqués dans notre introduction qui ont jalonné l'histoire de l'ellipse, ni même de discuter les mérites des différentes théories la concernant. Reconnue désormais comme un fait de langue établi, l'ellipse trouve pleinement sa place au cœur des recherches contemporaines, non seulement en linguistique théorique mais aussi en linguistique outillée combinée au Traitement Automatique des Langues (TAL). Pour conduire nos travaux relevant d'une démarche appliquée et pratique, il nous a donc semblé plus opportun de mettre en lumière des mises au point significatives relevées parmi la « masse » théorique de données disponibles.

Dans ce chapitre, nous clarifierons tout d'abord la définition de l'ellipse afin de dégager la plus opérante. Nous présenterons la classification sélectionnée en fonction de nos choix théoriques. Enfin, nous décrirons les études envisageant le traitement informatisé de certains types d'ellipses. Ces études ont en effet directement influencé notre propre questionnement et notre méthodologie et, à ce titre, constituent l'un des socles fondateurs de notre démarche.

[3] https://reports.news.ucsc.edu/linguistics/ (consulté le 13 juillet 2018 à 15:15).

## 1. Identité de l'ellipse : vers une définition opérationnelle

### 1.1. Complétude et incomplétude : ellipse, phrase et discours

Considérée comme propriété de la phrase et/ou du discours, l'ellipse renvoie donc à une partie absente de la structure syntaxique, pourtant parfaitement interprétée en se référant (consciemment ou inconsciemment) au contexte linguistique ou extralinguistique. En d'autres termes, l'ellipse se définit globalement comme une forme invisible dont paradoxalement la compréhension est immédiate. Elle contredit d'une certaine manière la vision saussurienne du signe, avec son signifiant et son signifié, puisqu'en l'occurrence, les destinataires ou récepteurs du message elliptique ne repèrent sans doute qu'un « signifié » grâce à la présence tangible d'un antécédent. Ce processus est décrit par le linguiste américain Jason Merchant comme une interruption dans la relation entre la forme invisible et son sens exprimé. Merchant s'inscrit ainsi dans une démarche syntaxique[4] qui vise justement à interpréter la dichotomie signifié/signifiant dans le cas précis de l'ellipse. Selon lui (2019, 19)[5],

> It [ellipsis] represents a situation where the usual form/meaning mappings, the algorithms, structures, rules, and constraints that in non-elliptical sentences allow us to map sounds and gestures onto their corresponding meanings, break down.

La présence d'un antécédent est donc obligatoire pour que l'ellipse soit d'une part autorisée, et d'autre part fonctionnelle, indiquant par-là l'existence d'une structure non-elliptique sous-jacente, équivalant à la phrase elliptique. Or, l'antécédent peut être linguistique ou extralinguistique. Dans ce dernier cas, il est nécessaire de se référer à la situation d'énonciation pour interpréter l'ellipse.

[4] Nous avons remarqué une certaine instabilité dans la délimitation entre syntaxe et sémantique chez certains linguistes américains dont Merchant. Cette instabilité relève peut-être d'une variation dans la définition des notions de phrase, énoncé, discours.

[5] Au cours de notre recherche, nous avons eu accès aux versions de travail des articles publiés entre 2014 et 2018. Ces articles, dans leur version définitive, sont désormais réunis dans *The Handbook of Ellipsis* (voir la bibliographie), auquel nous faisons référence ici.

Pour s'en approcher, il convient de ce fait, de s'intéresser aux différentes relations établies entre la structure de la phrase elliptique, le sens qu'elle véhicule, et son contexte. Abeillé & Mouret résument le processus comme suit :

> On parle d'ellipse, de manière générale, lorsque l'interprétation d'une forme syntaxique requiert plus que ce qui est fourni par les éléments qui la composent et que le matériel nécessaire pour obtenir cette interprétation est récupérable dans le contexte immédiat. On appelle éléments résiduels les éléments réalisés dans la forme elliptique. (Abeillé & Mouret 2010, 1)

Ces aspects que nous venons de mettre en lumière révèlent les diverses identités linguistiques de l'ellipse et la raison pour laquelle elle continue de susciter l'intérêt des linguistes en devenant objet d'études à part entière de plusieurs approches modernes et contemporaines. En effet, si l'ellipse continue à fasciner, c'est bien parce que, comme le souligne Merchant, les analyses qui lui sont consacrées débouchent directement sur la finalité de la syntaxe, à savoir, « to discern the nature of the form/meaning correspondence » (2019, 45).

Il ne saurait bien sûr être question ici de résoudre définitivement les divergences conceptuelles qui rendent si complexe l'appréhension du phénomène. Mais afin de répondre concrètement aux objectifs de notre étude, il est en revanche nécessaire de considérer l'acception opérationnelle de la notion : nous parlons donc d'ellipse lorsque un ou plusieurs segments du discours manquent dans une phrase dont le sens, dans la plupart des cas, est compris par le destinataire de façon aussi claire que le serait son équivalent non-elliptique. Ginzburg & Miller (2019, 75) abondent dans ce sens :

> The characterizing feature of ellipsis is that elements of semantic content are obtained in the absence of any corresponding syntactic form. The syntax thus appears to be incomplete. More specifically, the implicit semantic content is recovered from elements of the linguistic and non-linguistic context.

Pour le locuteur, il est évident que le procédé elliptique n'altèrera en rien la réception correcte de son message. Au contraire, s'il omet certains segments, c'est parce qu'il éprouve, inconsciemment (le plus souvent), le fait que, selon le principe

d'économie linguistique[6] à l'œuvre dans toutes les langues, leur apparition dans la phrase gênerait la fluidité de son discours et en impacterait la cohésion. Dans l'ellipse (5) ci-dessous, exemple qui illustre ce qui précède, l'interlocuteur n'a aucune difficulté à comprendre le non-dit *believe in miracles*. Le site elliptique est représenté par le signe de l'ensemble vide Ø et son antécédent est mis en gras :

(5) <CDEV> A: Do you **believe in miracles**$_{(1)}$ Lieutenant Veechfsky?
B: I do **Ø**$_{(2)}$.
Ø = [believe in miracles]

Ainsi, de la définition posée à ce stade et l'exemple ci-dessus, nous retiendrons trois éléments importants :

– un segment essentiel à la complétude de la structure syntaxique manque. Ce segment constitue le site elliptique $_{(2)}$.

– l'interprétation du segment manquant est guidée par la présence d'un antécédent (linguistique ou extralinguistique). Cet antécédent $_{(1)}$ est généralement assez proche du site elliptique et peut soit le précéder soit le suivre. Le repérage de l'antécédent devient alors un paramètre crucial dans l'interprétation de l'ellipse.

– la taille du site elliptique est délimitée par les éléments présents et par la catégorie et la fonction de l'antécédent $_{(1)}$ et $_{(2)}$.

Ces éléments sont étroitement liés aux cinq critères identifiés par Quirk *et al.* (1985, 884-888), énoncés ci-dessous :

– l'élément ellipsé est récupérable,

– la construction elliptique est grammaticalement *défectueuse*,

– l'insertion de l'élément ellipsé rend la phrase grammaticale (le sens est le même que celui de la phrase elliptique),

– les éléments omis sont récupérables dans le texte et,

– ils ont la même forme que l'antécédent.

Comme les termes *défectueuse, insertion, récupérable* le laissent entendre, l'ellipse présente un réel défi à la définition canonique de la phrase puisqu'elle

[6] Le principe d'économie linguistique a été introduit par A. Martinet dans *Économie des changements phonétiques* (Martinet, 1955).

touche à sa complétude. Chacun peut se souvenir des consignes de l'enseignant qui demande à ses élèves de répondre à une question en formulant une phrase *complète*, ou encore d'écrire une phrase commençant par une majuscule et se terminant par un point. C'est cette notion de complétude (et incomplétude) qui rend l'ellipse énigmatique : relève-t-elle de la syntaxe dans la mesure où la phrase complète est une construction gouvernée par des règles syntaxiques qui organisent les relations entre le sujet, un prédicat et ses compléments ? Ou relève-t-elle de la sémantique, dans la mesure où la compréhension du sens par l'interlocuteur est globale ? En réalité, l'idée de complétude semble être davantage liée à la notion de *norme*, généralement fixée selon la fréquence d'usage.

S'appuyant sur la tradition syntaxique générative, Depraetere & Langford (2012, 11) rappellent qu'une simple juxtaposition de mots ne suffit pas pour qu'apparaisse une phrase dont l'identité est liée à l'assemblage de constituants ayant une fonction en son sein :

> Sentences do not just consist of a series of juxtaposed words that each belong to a particular part of speech. […]
> Different parts of speech cluster together and form constituents, which have certain functions in the sentence.

Ceci vaut également pour le français. Dans leur *Grammaire Méthodique du Français*, Riegel *et al.* (2014, 203) relient l'idée de complétude évoquée plus haut au fait que le contenu a été « encodé » dans le « cadre formel » constitué par une phrase et qui est susceptible de s'élargir dans certaines limites :

> […] tout contenu acquiert un caractère de complétude pour le fait même qu'il a été encodé dans le cadre formel d'une phrase. Le cadre phrastique, et avec lui, le contenu véhiculé, peut s'élargir au gré du locuteur et en fonction de l'information à transmettre, pourvu que ce soit dans les limites des possibilités architecturales de la langue et des capacités mémorielles de l'interprétant pour les traiter.

Nous leur emprunterons, à ce stade, la définition de la phrase qu'ils proposent (2014, 203) :

> […] un assemblage de mots à la fois *significatif* et *grammatical*, c'est-à-dire conforme à des règles de construction. Plus précisément, une phrase est une construction que l'on peut définir et identifier par la conjonction de trois caractéristiques, les deux premières syntaxiques et la dernière interprétative. (Notre soulignement)

Riegel *et al.* (*id*.) montrent ainsi que se « déploient » dans la phrase « le réseau des relations (les fonctions grammaticales) et les classes d'unités simples (les parties du discours) et complexes (les groupes de mots) qui constituent l'architecture syntaxique des énoncés ». Il s'agit de ce que l'on peut nommer propriété « syntaxique ». La phrase peut ainsi être décrite à l'aide d'un système de règles de dépendance entre les constituants. Par « propriété syntaxico-sémantique », Riegel et *al.* entendent la fixation du sens des mots et le rapport qui existe entre eux.

La question se pose alors de savoir si, dans le cas d'une phrase elliptique, les relations entre les constituants, présents ou non, sont affectées. L'exemple (6) est un échange contenant une question suivant un ordre canonique *what have you bought ?* et une réponse fragmentaire *Trousers*, donc elliptique. La compréhension immédiate de l'échange suffit à montrer que les relations sont effectivement établies, non seulement entre les constituants de chacune des phrases A et B, mais aussi entre les phrases elles-mêmes, en ne laissant aucune ambiguïté.

(6) <CEx> A: What have you bought? B: ∅ Trousers.

Ces relations sont possibles parce que la phrase est elle-même une unité au sein du discours :

> La phrase appartient bien au discours. C'est même par là qu'on peut la définir : la phrase est l'unité de discours. (Benveniste 1966, 130)

La conséquence évidente de ce fait désormais établi est la nécessité de prendre en compte dans nos analyses, le contexte *au-delà* de la phrase pour une meilleure compréhension et un meilleur traitement de l'ellipse. Nous aurons l'occasion de vérifier d'ailleurs, que l'antécédent d'une ellipse ne se trouve pas forcément dans la

même phrase que le site elliptique et qu'il peut soit le précéder, soit le suivre immédiatement ou même en être éloigné[7].

L'enjeu est alors de comprendre, au sein de la phrase, du discours et hors du discours lui-même, quels éléments favorisent l'apparition d'une ellipse. Autrement dit, qu'est-ce qui permet l'effacement de tel ou tel constituant et comment peut-on recouvrer l'élément ellipsé[8] ? Aux interrogations sur l'identité de l'ellipse et la structure du site elliptique, s'ajoutent donc celles portant sur les conditions permettant de la recouvrer. En anglais, les termes les plus fréquemment utilisés sont ceux de *recoverability*[9] et *licensing*, qui désignent précisément les conditions locales permettant l'ellipse. Il s'agit donc ici de travailler véritablement sur le fonctionnement de l'ellipse, ce qui constitue le cœur même de notre recherche puisque de la manière dont nous comprendrons son mécanisme, dépendra la mise au point des procédures outillées en vue de sa détection automatique.

### 1.2. Ellipse et anaphore

Sous les termes d'*effacement*, d'*anaphore zéro*, de *pronominalisation nulle*, *de représentation zéro*, les définitions de l'ellipse ont varié pour être à présent ramenées à celle d'une construction tronquée, incomplète ou fragmentaire, cependant parfois confondue avec d'autres phénomènes linguistiques telle que l'anaphore. La relation entre les deux phénomènes a donc fait l'objet d'une attention particulière et des questions importantes semblent demeurer en suspens : quelle est la nature de leur relation ? Convient-il d'envisager, ou non, une analyse des ellipses au sein même du phénomène anaphorique où elle serait alors perçue comme lui étant apparentée ? Car si l'on propose de définir l'ellipse comme relation impliquant une dépendance entre l'élément omis et son antécédent, on rejoint en effet la définition de l'anaphore

[7] Ce qui constitue un discours est un ensemble de phrases réalisées dans une situation donnée. Pourtant, devant un ensemble de phrases, il est plus courant de parler de *texte* que de discours y compris en linguistique textuelle.

[8] Dans la présente recherche, nous avons choisi d'employer le verbe « ellipser » qu'on ne trouve pas dans le dictionnaire mais que Lallot, à titre d'exemple, a utilisé au participe passé dans sa traduction d'Appolonius Dyscole. Le verbe « élider », qui aurait parfois pu convenir, renvoie à d'autres phénomènes d'effacement comme l'élision.

[9] souvent traduit par « récupérabilité » alors que « recouvrement » serait sans doute aussi juste.

dans la mesure où cette dernière est également fondée sur une relation de dépendance existant entre un élément et un référent dans la phrase ou le contexte linguistique voisin.

Il existe cependant une différence de taille entre les deux phénomènes. L'un, l'ellipse, produit un effacement, donc un vide, dans la structure syntaxique, et le sens de ce vide est compris et interprété par le co-locuteur grâce à l'interface syntaxico-sémantique ; l'autre, l'anaphore, est une reprise d'un élément du discours sous une autre forme, ce qui n'implique nullement un vide syntaxique. Pour avoir un sens, cette reprise est tributaire d'un référent[10], accessible dans le contexte. L'exemple suivant illustre ainsi la reprise anaphorique de *some vacations* par le pronom *ones*.

(7) <CDEV> - Do you have some vacations this year?
- Short *ones*.

L'ajout du référent *vacations* à la suite de *ones* rendrait la phrase non acceptable. En revanche, l'apparition de l'élément omis dans la structure syntaxique d'une phrase elliptique n'impacterait pas sa grammaticalité, tel est le cas dans l'exemple (8).

(8) <CEx> They were ordered to **leave** but they **did**n't Ø.

La distinction entre anaphore et ellipse n'est toutefois pas toujours aussi évidente, notamment lorsque des auxiliaires ou des modaux servent à les déclencher comme dans l'exemple (9).

(9) <CDEV> It didn't occur to him to **disobey the harsh note**. He never **had**. No one he knew ever **had.**

Nous nous posons alors la question suivante : *had* est-il anaphorique ou déclencheur d'ellipse ? En d'autres termes *had* remplace-t-il *entièrement* le verbe *disobey* ou vient-il *seulement* le suppléer ? Avec la simple définition extraite du

[10] Le « référent » de l'anaphore n'est ni plus ni moins ce que certains appellent son « antécédent », où l'on retrouve le même terme que pour l'ellipse (Voir Francis Cornish 2010, par exemple).

CNRTL[11], nous constatons qu'il est difficile de trancher entre remplacer entièrement (sous le principe de l'anaphore) et compléter ce qui existe déjà (sous le principe de l'ellipse). Ginzburg & Miller (2019, 75) apportent une réponse claire à ce questionnement puisque selon eux, l'ellipse peut toujours être analysée comme similaire à l'anaphore :

> Ellipsis is similar to anaphora, except that there is no overt anaphoric element involved. The elements present in an elliptical clause are predicates, arguments or adjuncts of what is omitted. It is the presence of these elements that makes it possible to recognize the ellipsis. It is in principle always possible to reanalyse any elliptical phenomenon as a case of anaphora, by hypothesizing an unpronounced pro-form in the elliptical site (zero-anaphora) and/or by considering the licenser to be anaphoric[12].

À ce stade, nous présentons ci-dessous les critères sur lesquels nous nous fondons pour adopter une définition de l'ellipse à visée pratique :

i) la complétude/incomplétude, c'est à dire la présence dans une phrase d'un constituant absent permettant néanmoins la compréhension ;
ii) le recouvrement/récupérabilité du segment omis grâce à la présence d'un antécédent identifiable dans la phrase elle-même, dans les phrases adjacentes, dans le discours ou dans la situation (en anglais : *Context*) ;
iii) l'existence d'une dépendance syntaxique entre l'élément ellipsé et au moins l'antécédent.

### 1.3. Synthèse des questionnements portant sur l'identité de l'ellipse

En guise de synthèse provisoire, nous empruntons à Phillips & Parker (2014, 80) la figure (1) ci-dessous, parce qu'elle offre une vue d'ensemble des différentes interrogations portant sur le phénomène elliptique. Incomplet à certains égards, ce schéma fait néanmoins apparaître les questions essentielles précédemment

[11] « Suppléer : mettre à la place d'une autre qui fait défaut, ou qui est insuffisante ou incomplète, une chose qui en tient lieu ; remplacer. » http://www.cnrtl.fr/definition/suppl%C3%A9er (consulté le 19 juin 2018 à 15:39)
[12] Cette prise de position nous intéresse particulièrement dans la mesure où elle facilitera l'élaboration des requêtes menant d'abord à une reconnaissance du phénomène elliptique.

soulevées : l'interaction entre syntaxe et sémantique, la distinction problématique entre ellipse et anaphore, la reconstitution de la structure effacée dans le site elliptique.

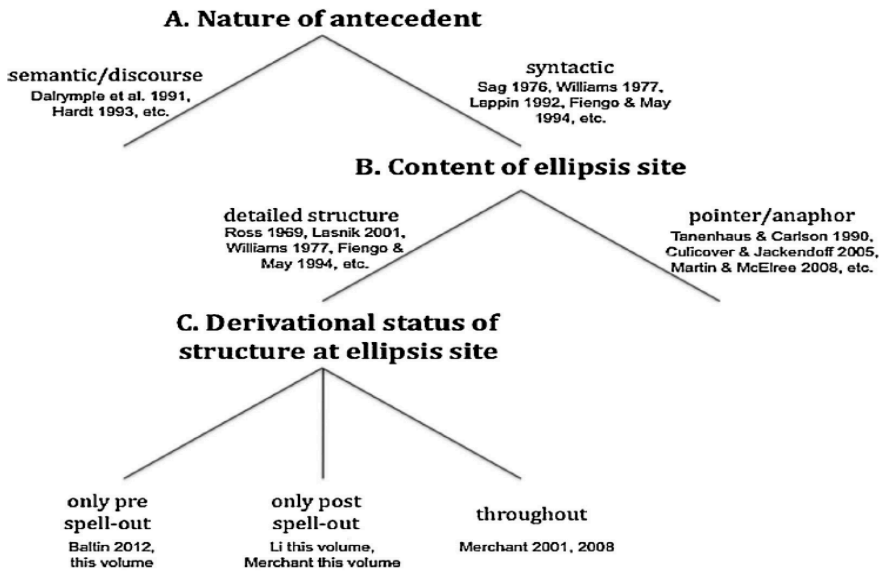



**Figure 1 : Questionnements au cœur des recherches sur l'ellipse synthétisés par Phillips & Parker (2014)**

Enfin, comme on le voit, plaider en faveur d'une nature syntaxique de l'antécédent soulève principalement des questions relatives au contenu du site elliptique et à sa structure. On comprend aussi qu'à partir des interrogations mises en avant par ce schéma, trois approches théoriques majeures sont aujourd'hui déterminantes :

i) une approche syntaxique devant permettre la résolution de l'ellipse par la reconstruction de la phrase ;
ii) une approche sémantique et discursive privilégiant le sens ;
iii) une approche hybride réunissant l'aspect syntaxique et l'aspect sémantique[13].

[13] Les auteurs de ces approches sont cités plus loin.

## 2. Éléments théoriques : vers une classification opérationnelle

### 2.1. Brève présentation de quelques éléments théoriques utiles

La répartition tripartite que nous venons de présenter n'est toutefois pas aussi tranchée qu'il y paraît. Comme on a pu l'énoncer précédemment, les interrogations portant sur l'ellipse se sont le plus souvent focalisées sur la relation entre le segment manquant (l'élément dont on fait l'économie), le vide engendré par cette économie dans la structure de la phrase, et l'antécédent. La question de savoir s'il existe une structure syntaxique silencieuse, non prononcée, à l'endroit du site elliptique, résume bien le questionnement, tant de Bîlbîie (2013), de Phillips & Parker (2014) que de Merchant (2019) (tel que présenté dans la figure 2 ci-dessous). Une réponse par l'affirmative implique une théorie énonçant une syntaxe abstraite des phrases. Si, au contraire, la réponse est négative, cela sous-entend qu'il n'y a ni élément silencieux, ni structure syntaxique abstraite, mais que seul existe ce qui est prononcé, ce que Merchant (2019, 21) énonce dans le principe suivant : « ce qu'on entend, c'est ce qu'on a »[14]. Opter pour l'une ou l'autre réponse conduit à emprunter une approche structurale pour l'une ou non structurale pour l'autre [15] (illustrées dans la figure 2).

En effet, les approches non structurales rejettent l'existence d'une structure syntaxique non-prononcée. Elles dépendent d'une théorie du sens qui permet de le générer sans avoir recours à une structure syntaxique dans le site elliptique (Ginzburg & Sag 2000, Culicover & Jackendoff 2005). À l'inverse, les approches structurales affirment qu'il existe une structure syntaxique dans le site elliptique et que son sens découle des mécanismes syntaxiques du contexte voisin. Cette vision ouvre deux axes de recherche dont l'un, *PF-deletion*[16]*,* présume un effacement phonologique (Ross 1969, Sag 1976, Hankamer 1979, Merchant 2001, Johnson 2004), et l'autre, le

[14] (angl. Wyhiwyg = *What you hear is what you get*).

[15] Merchant les nomme en anglais *structural* vs *non-structural approches*. Deux traductions sont possibles : structurelles ou structurales : nous avons opté pour cette dernière.
Les informations succinctes présentées dans cette partie sont extraites de quatre ressources bibliographiques essentielles : Merchant (2019), Aelbrecht (2010), Bîlbîie (2013) et Gandon-Chapela (2016).

[16] *Phonetic form deletion*

*LF-copy/null-anaphora*[17] préconise un remplacement par un élément lexical nul pour la prononciation (Hardt 1993, Lobeck 1995, Fiego & May 1994, Lappin 1999, Chung *et al*. 1995).

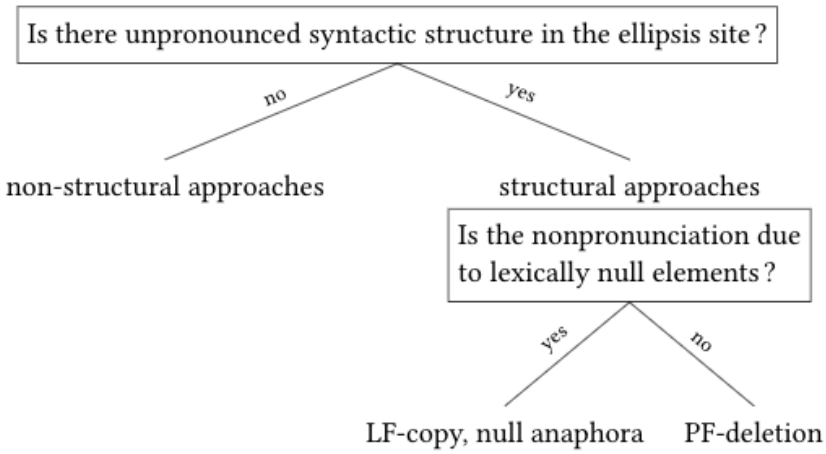



**Figure 2 : Les approches contemporaines de l'ellipse (Merchant 2019)**

Merchant (2019, 25) compare cette recherche à celle des astrophysiciens en quête du trou noir ; la réponse aux questions posées ne pouvant être qu'indirecte :

> How does one decide whether some piece of syntactic structure is or isn't there, particularly when that structure in any case does not lead to any pronounced difference? *Indirectly, of course*. *Detecting and arguing for such 'missing' structures is analogous to searching for a black hole: one can tell it's there only by its effects on surrounding material.* The logic of the hunt for elided structure is similar. If one finds effects that seem to be due to missing material, there is an argument that such structure exists. (Notre soulignement)

Il existe également une troisième approche avancée et discutée d'une manière approfondie par Schwabe & Winkler (2003), qui défendent l'hypothèse selon laquelle certains types d'ellipse peuvent être expliqués et analysés sans avoir recours aux mécanismes de l'ellipse. Les partisans de cette théorie estiment que pour expliquer

[17] Proforme nulle

une phrase *dite* elliptique, il est possible d'avoir recours à d'autres mécanismes, notamment sémantiques et discursifs, indépendants de ceux de l'ellipse[18].

L'intérêt de ces différentes approches réside pour nous dans les classifications (dans le sens de catégorisation) qu'elles ont établies et qui varient d'une approche à l'autre. Les facteurs de variation relèvent de l'agencement grammatical et de l'organisation syntaxique, ou du contexte, ou de la situation pragmatique, facteurs qui recoupent les critères que nous avons retenus pour la définition de l'ellipse (voir plus haut, 1.2). En vue d'une exploration et exploitation automatique, il est nécessaire en effet de regrouper les ellipses selon des critères clairs et opérationnels afin de les traiter automatiquement de manière optimale[19].

## 2.2. Taxonomie de l'ellipse de van Craenenbroeck & Merchant

Parmi les catégorisations existantes, nous retenons, en dépit de ses problèmes (signalés par le symbole # dans le tableau), celle de van Craenenbroeck & Merchant (2013) puisqu'il s'agit d'une reformulation qui englobe d'une manière élaborée les classifications syntaxiques déjà existantes.

Van Craenenbroeck & Merchant classent ainsi l'ellipse en trois catégories qu'ils nomment *predicate, clausal* et *nominal ellipses* au sein desquelles des sous-catégories sont identifiées. Les tableaux ci-dessous synthétisent les différents types d'ellipses tels qu'ils les présentent[20] :

[18] En dehors des mécanismes de mouvement et d'extraction expliqués par Johnson (2009) dans le cas de *gapping*.

[19] Nous n'entrons pas dans le détail des distinctions établies entre toutes les approches théoriques ayant envisagé le traitement de l'ellipse ni les nuances résultant des débats existant entre elles. Ces débats ont généré des classifications différentes du phénomène et qui varient d'une approche à l'autre, selon la composition syntaxique et l'agencement grammatical ou selon la situation pragmatique, ou encore le contexte proprement dit. Notons au passage que la grammaire de Quirk *et al.* (1985, 888), par exemple, classe les ellipses selon le type de récupérabilité. Ils nomment alors *textual ellipses* celles où la récupérabilité des éléments dépend du co-texte, *structural ellipses* celles où la récupérabilité est sous-jacente à la structure grammaticale, *situational ellipses* lorsque la récupérabilité est exophorique (en-dehors du contexte extralinguistique). Cette classification, qui prend en compte à la fois l'organisation syntaxique et l'interprétation, nous parait à l'origine des classifications syntaxiques existantes à l'heure actuelle dans la mesure où la catégorisation même repose sur les conditions de récupérabilité.

[20] Le tableau inclut toutes les ellipses illustrées dans Merchant & van Craenenbroeck y compris celles qui ne sont pas attestées en anglais britannique ou américain. Sauf mention contraire, les exemples sont extraits de van Craenenbroeck & Merchant (2013).

| ***Predicate ellipsis*** | | |
|---|---|---|
| Sous-catégorie | Définition | Exemple |
| *VPE* | Omission du groupe verbal.[#] | John likes candy, but Bill doesn't Ø. |
| *Pseudogapping* | Le *pseudogapping* est une ellipse dans laquelle un verbe lexical (accompagné ou non de ses arguments compléments), après un auxiliaire est supprimé laissant derrière lui des éléments résiduels. | She'll read something to Sam, but she won't Ø to Bill.<br>Deux caractéristiques majeures du *pseudogappin*g :<br>La présence d'un auxiliaire est un critère nécessaire pour le déclencher.<br>Le cas *pseudogapping* avec plusieurs résidus est très rare. |
| *British English do* | « The former is a British English construction that is on the surface identical to VPE, but for the presence of a non-finite form of the verb do next to the ellipsis site » [#] (p 703) | John will eat candy and Bill will do Ø, too. |
| *Modal complement ellipsis* | Cette ellipse diffère de la VPE dans la mesure où le déclencheur de l'ellipse est nécessairement un modal déontique. | Jan wil niet meedoen, maar hij moet Ø. (Exemple néerlandais)<br>Jan ne veut pas participer mais il doit le faire. |
| *Predicate Phrase ellipsis* | *Predicate phrase ellipsis* est une omission des compléments (locatifs ou prédicatifs) après *be* et *have.* | Ben will be in the garden, though he'd rather not be Ø. |

| **Clausal ellipsis** | | |
|---|---|---|
| *Sluicing* | Le *sluicing* est l'économie de la proposition entière à l'exception du pronom *wh—* | Someone knocked at the door, but we don't know who **Ø**. |
| *Sprouting* | Le *sprouting* est un sous-type du sluicing dans lequel le pronom *wh—* n'a aucun lien avec l'antécédent exprimé. | Ed is eating, but I don't know what **Ø**. |

| | | |
|---|---|---|
| *Swiping* | « In swiping constructions, a wh-PP has been sluiced, but the canonical order of preposition and wh-phrase (the former preceding the latter: about what) has been inverted. » (p 718) | Ed gave a lecture, but I don't know what about Ø. |
| *Spading* | Le *spading* renvoie à l'économie de la proposition qui est remplacée par un pronom démonstratif. | Jef eid iemand gezien, mo ik weet nie wou **da**.<br>Jef has someone seen but I know not who that<br>(Jef saw someone, but I don't know who.) |
| *Fragment answers/*Réponses fragmentaires | « Fragment answers are subsentential XPs with the same propositional content and assertoric force as utterances of fully sentential syntactic structures. » (p 719) | What did you buy? B: A boat |
| *Gapping* | Le *gapping* renvoie à l'omission du verbe fini dans une ou plusieurs constructions (parallèles ou propositions coordonnées), laissant apparent des résidus. | Mary carries a suitcase, and John Ø a bag. |
| *Stripping* | Le *stripping* est considéré comme une catégorie du *gapping*. Les deux catégories diffèrent dans le nombre des résidus : le *stripping* en autorise un seul, tandis que le *gapping* en autorise plusieurs. | Ed likes stiletto heels and Maggy Ø too. |
| *Null complement anaphora* | Décrit comme the « odd man out » de la liste, ce type implique l'effacement d'une proposition complément sans qu'il reste de « survivors » (p 719) | Ed wanted Bill to help Mary, but he refused. (p 718) |

| Nominal ellipsis | | |
|---|---|---|
| L'ellipse nominale désigne l'économie du nom, du pronom, ou d'un groupe nominal dont la fonction grammaticale est sujet ou complément (ou une partie) du verbe | Merchant & van Craenenbroeck proposent quatre questions-test pour repérer les ellipses nominales :<br>A : X a-t-il besoin d'un antécédent nominal ou est-il utilisé seul ? L'ellipse requiert un antécédent.<br>B : X forme-t-il un comparatif ou un superlatif ? Si oui, c'est une ellipse.<br>C : X peut-il signifier tout ce que peuvent normalement signifier les adjectifs ?<br>D : X peut-il former un pluriel ayant une morphologie nominale de pluriel (qui diffère des désinences adjectivales)? Si la réponse à cette question est non, il s'agit d'une ellipse. | van Craenenbroeck & Merchant (2013, 732), illustrent ces questions avec le cas de *poor*<br><br>"For example, *poor* in English is a nominalization by most of the above tests: it needs no antecedent, it does not form a comparative, and it does not have the full range of meanings found when used as a modifier of nouns. (The last test does not give a meaningful result in this case, as the nominalization is a collective, which triggers plural agreement on predicates and cannot be pluralized or used as a predicate itself: *He is a poor.)<br>a. The poor deserve our help.<br>b. If you have money, you should help the poorer (than you).<br>c. A: Look at the poor kitty stuck in the tree!<br>B: *That's no poor – he lives there.<br>d. *The poors are everywhere in this town!" |

**Tableau 1 : Les catégories d'ellipses par van Craenenbroeck & Merchant (2013)**

L'analyse de cette classification soulève un certain nombre de questionnements d'ordre terminologique et classificatoire. D'abord, certaines notions-clefs nécessaires à la compréhension de la classification ne sont pas définies : qu'est-ce qu'un prédicat selon van Craenenbroeck & Merchant ? Faut-il ou non y inclure le modal/auxiliaire ? Dans les exemples illustrés, l'omission s'effectue seulement sur une partie du prédicat et non sur le prédicat dans son entièreté. Dans la mesure où l'auxiliaire et le modal restent toujours apparents et que ce sont eux qui déclenchent l'ellipse post-auxiliaire, ils sont situés en dehors du site elliptique, et ils ne peuvent matériellement être omis. De ce fait, il apparait difficile de parler de l'ellipse du prédicat entier puisque seul le groupe verbal ou une partie de ce dernier

est omis. Nous nous appuierons sur ce constat (tout au long du présent travail), pour déterminer, comme nous le verrons plus loin, l'une de nos conditions classificatoires.

Vient ensuite le questionnement lié à la sous-catégorisation établie : est-elle toujours pertinente ? Les flottements existant dans l'établissement de la classification de certaines catégories sont parfois sources d'incohérences, et le *gapping* en est l'exemple. Cette ellipse est classée dans les ellipses propositionnelles alors que seul le verbe est omis (et non l'entièreté de la proposition). De plus, les ellipses nommées *predicate phrase ellipses* ne requièrent pas une catégorisation distincte et peuvent être associées à la VPE. C'est d'ailleurs ce que la plupart des syntacticiens s'accordent à défendre.

Tous ces questionnements, loin d'être exhaustifs, montrent bien la nécessité d'établir une classification aussi fine que possible pour pouvoir repérer les ellipses automatiquement. La classification proposée par van Craenenbroeck & Merchant est-elle alors pertinente dans notre cas ? Partiellement sans doute. Aussi proposons-nous, afin d'atteindre notre objectif de détection automatique, une simplification dans la catégorisation des ellipses en les classant uniquement par élément déclencheur et en n'établissant pas de sous-catégorie. Les détails de cette classification sont présentés à la fin de ce chapitre. Dans la partie suivante, nous présentons des travaux menés sur certains types d'ellipse à partir d'un corpus d'étude, ensemble de données authentiques exploitables informatiquement, dont l'analyse constitue aujourd'hui l'une des opérations courantes.

## 3. Approches outillées de l'ellipse : de la linguistique de corpus au TAL

Nous distinguons les études à visée théorique où domine la linguistique dite de corpus, des travaux où se combinent les apports de la linguistique de corpus aux méthodes du TAL. Nous examinerons en premier lieu l'approche de Miller (2011) et de Miller & Pullum (2013) pour l'ellipse post-auxiliaire. Ces études, dans le cadre de la linguistique de corpus, s'attachent aux conditions impactant les occurrences elliptiques d'un point de vue strictement linguistique. Nous proposerons ensuite un état des lieux concernant la recherche plus ou moins automatisée appliquée à la VPE, et au *sluicing*. Ces études visent la détection de l'ellipse, le repérage de l'antécédent

et tentent de résoudre l'ellipse. Seront ainsi examinés les travaux de Hardt (1992), Hardt & Rambow (2001), Bos & Spenader (2011) McShane & Babkin (2016).

On peut remarquer que dans les corpus majoritairement exploités, la langue anglaise est celle qui est la plus représentée. Nous notons particulièrement l'utilisation des corpus suivants : le *Brown Corpus*, le *British National Corpus (BNC)*, *Corpus of Contemporary American English* (*COCA)*, *Wall Street Journal (WSJ)*, le *New York Times* et *Opensubtitles*[21]. Il s'agit seulement dans cette partie de signaler des études pertinentes pour notre propre travail. Nous ne rentrerons donc pas dans le détail dans ce qui est développé dans ces articles mais nous mettrons en évidence ce qui relève des procédures méthodologiques et des résultats obtenus.

### 3.1. Approche sur corpus illustrées par Miller & Pullum

Les deux articles de Miller (2011) et Miller & Pullum (2013) constituent un ensemble dans lesquels sont analysés les critères discursifs d'apparition des expressions anaphoriques *do so, do it/this/that* pour mettre en lumière, a contrario, ce qui permet, dans le discours l'apparition de la PAE (*post-auxiliary ellipsis*) (Miller 2011). L'analyse de Miller est fondée sur l'exploitation extensive des corpus COCA et BNC sur la base d'un échantillon de 450 exemples. Il retient principalement le critère de fréquence, puis au moyen d'une analyse des exemples, il dégage les conditions qui permettent l'une ou l'autre expression à l'exclusion des autres (observations, constats, manipulations). Il formule alors des hypothèses sur la base de certains critères[22] :

(i) Le registre[23] : oral, fiction, journalistique et académique (ou plus généralement scientifique). Par exemple, à l'inverse de la PAE, *do so* est plus fréquent dans le registre soutenu :

[21] Toutes les informations concernant les corpus exploités figurent dans l'annexe I p. 263.
[22] Nous renvoyons le lecteur à la série d'exemples analysés dans (Miller 2011, 9).
[23] Nous reviendrons dans le deuxième chapitre sur les différents concepts de genre, type de discours, registre, style. À l'instar de Miller (2011) et comme nous l'avons déjà précisé dans l'introduction, l'un des objectifs de ce travail est d'analyser l'influence du registre/genre des textes sur les différents types d'ellipse. Nous y présenterons en détail les discours étudiés.

> It appears that *do so* is much more frequent in Academic and Newspaper than in Spoken or Fiction whereas PAE with finite auxiliary *do* is more frequent in the latter than in Academic and Newspaper. (Miller 2011, 83)

(ii) L'alternative entre les polarités positive et négative favorisant l'occurrence des PAE.

(iii) Le degré de saillance de l'antécédent.

(iv) La distance entre la PAE et son antécédent.

Observant la distance entre la PAE et son antécédent, Miller & Pullum remarquent que l'antécédent peut-être exophorique, ce qui a été nié par Hankamer & Sag (1976) dans une étude précédente. Miller & Pullum (2013, 7) prennent clairement position contre Hankamer & Sag et affirment ainsi :

> Our claim is that Hankamer was wrong about exophoric PAE in that it can indeed be freely deployed in exophoric uses in all the situations where it satisfies the general discourse conditions on its use that apply in anaphoric contexts, too.
> It is not even quite correct to say that exophoric PAE is rare. At the very least, it is misleading to say that. Many circumstances prevent the exophoric use of PAE, but it seems to occur as often as the demands of the non-linguistic context happen to motivate it – it is free to occur exophorically within *the range of the circumstances that allow it to occur at all*.

Dans l'exemple (8) trouvé dans le COCA, la PAE est exophorique :

> (10) The aisles at the Lakewood Wal-Mart are surprisingly packed at 11 p.m. 'Can we? Can we?' Vanessa tugs at her mother, pointing to a rack of 'Lady and the Tramp" DVDs. Diaz shrugs. OK. (Miller & Pullum, 2013, 15)

Ainsi, Miller & Pullum (2013) estiment que la PAE peut être exophorique mais devra répondre à certaines conditions du discours pour être permise.

On retiendra de ces articles :

– qu'ils utilisent les méthodes de la linguistique de corpus ;

– que leur finalité est de découvrir le fonctionnement d'un phénomène linguistique afin de le théoriser ;

– que les conclusions sont à confirmer par un travail sur une quantité plus importante de données.

Pour ce qui concerne l'ellipse, en l'occurrence la PAE, ces articles mettent en avant certaines conditions favorisant son apparition (registre, saillance de l'antécédent, coréférence) et confirment le rôle de l'antécédent et du contexte. Si la linguistique de corpus permet de travailler sur des données considérables, l'analyse exploite un échantillon (ici 450 exemples) à partir duquel il est possible d'extrapoler et de formuler des hypothèses à vérifier ensuite.

D'autres études du même type existent. Nous citerons celles de Beecher (2008) et Nykiel (2015) pour le *sluicing* qui n'apportent toutefois pas d'éléments nouveaux à la méthodologie. Cependant, l'utilisation d'un corpus d'analyse a permis à ces travaux de confronter le phénomène elliptique aux différents usages de la langue y compris d'un point de vue diachronique (Nykiel 2015) et de faire ressortir des occurrences inhabituelles. Ces dernières présentent un défi à la détection automatique. Le relever constitue l'un des intérêts des approches fondées sur corpus visant à repérer automatiquement les ellipses.

### 3.2. Approches fondées sur corpus pour une détection automatique de l'ellipse

La finalité des études que nous venons de présenter est de proposer un éclairage argumenté et quantitativement documenté sur les phénomènes linguistiques étudiés afin d'apporter des éléments nouveaux à la connaissance de tel ou tel phénomène. Allant au-delà de ces objectifs, le Traitement Automatique des Langues (TAL), qui ne saurait faire l'économie de certaines procédures propres à la linguistique de corpus, offre la possibilité d'appliquer ces connaissances à des fins pratiques, telle la détection de l'ellipse qui fait l'objet de notre travail. Grâce aux travaux théoriques précédemment énumérés, il apparaît en résumé que le site elliptique, l'antécédent et le contexte sont les éléments-clefs nécessaires à prendre en compte afin de mener une étude sur le phénomène. De même, lorsqu'un traitement automatique est envisagé, il est impératif de prendre en compte ces éléments en tenant compte des problèmes particuliers qu'ils posent au TAL :

**(i) Le site elliptique** : les problèmes qu'il pose au TAL tiennent principalement à sa taille variable d'une occurrence à une autre, brouillant parfois l'identification des frontières de ce site.

**(ii) L'antécédent** : le problème à résoudre, pour les outils du TAL en particulier, concerne sa nature même et la distance qui le sépare du site elliptique. Que faire dans le cas d'un antécédent extralinguistique ?

**(iii) Le contexte** : toutes les ellipses ne peuvent pas être détectées via le contexte syntaxique. Par ailleurs, le genre et le type de discours peuvent affecter les structures syntaxiques d'une phrase. La complication surgit lorsqu'il s'agit d'intégrer ces variations et les particularités inhérentes au genre, dans les outils sélectionnés pour son traitement automatique.

Les procédures élaborées pour traiter ces problèmes identifiés en vue d'un traitement automatique sont propres au TAL et sont constamment ajustées à l'objectif fixé. Elles peuvent être automatisées ou semi-automatisées. Dans le cas de l'ellipse, les quelques études conduites se sont attachées à appréhender soit l'ellipse seule, soit la reconnaissance de son antécédent, soit sa résolution. L'objectif de cette section est de montrer l'évolution apparente des procédures méthodologiques envisagées dans chaque démarche. En effet, deux méthodes ont été utilisées pour analyser l'ellipse :

– les méthodes à base de règles où des règles sont construites par un expert du domaine, le linguiste par exemple, pour classer ou identifier un phénomène.

– les méthodes par apprentissage supervisé où le système est entrainé à déterminer les règles lui-même à partir des données fournies qui ont préalablement été annotées manuellement.

Nous nous pencherons ainsi sur les études pionnières en la matière celles de Hard (1992, 1993) et Hardt & Rambow (2001) pour ensuite présenter les approches reposant sur l'annotation[24] manuelle de corpus et l'apprentissage automatique dans

[24] L'une des opérations utilisées en TAL qui consiste à ajouter des informations au corpus. (Voir chapitre 2).

le cas de la VPE et du *sluicing*. Notre examen est limité aux apports méthodologiques et techniques.

### 3.2.1. Études pionnières de Hardt et Hardt & Rambow

C'est à Daniel Hardt (1992)[25] que nous devons la toute première approche informatisée de l'ellipse verbale en anglais, fondée sur corpus, mais introduisant des procédures automatisées relevant du TAL. Dans cette première étude, l'objectif était de détecter les VPE (dans l'acception de Merchant & van Craenenbroeck) en utilisant la commande `grep`[26] et de récupérer leurs antécédents automatiquement. D'autres études discutant les résultats obtenus et approfondissant les questions abordées ont suivi (Bos & Spenader 2011, McShane & Babkin 2016).

Lors de ses premières explorations, Hardt a travaillé sur 304 exemples extraits du *Brown Corpus.* L'algorithme à base de règles qu'il a élaboré contient trois fonctions importantes :

– La fonction *remove-impossible* consiste à éliminer tous les antécédents impossibles et improbables pour une VPE, par exemple les structures où les VPE sont contenues dans les antécédents (ACD).

– La fonction *assign-levels* renvoie à l'attribution d'un degré de préférence aux antécédents plausibles.

– Enfin la fonction *select-highest* sélectionne les antécédents qui ont le niveau de préférence le plus élevé. Cette fonction s'applique aux cas dans lesquels plus d'une règle de préférence sont attribuées à l'antécédent. Si plusieurs éléments ont le même niveau de préférence, l'antécédent sélectionné est l'élément le plus proche de la VPE dans la périphérie gauche de la phrase.

[25] Plusieurs versions améliorées ont été présentées depuis. En effet, les premiers travaux de Hardt sur le traitement automatique des ellipses remontent à 1992 et sont toujours en cours.

[26] grep = *global regular expression print.*

Hardt (1992) a obtenu 94% de précision[27] dans son analyse, ce qui correspond ainsi à une détection de 285 antécédents corrects (sur 304). Il a ainsi réussi à mettre en œuvre un programme informatique capable de résoudre certaines catégories de la VPE en copiant l'antécédent dans le site elliptique.

Hardt a poursuivi ses investigations en collaboration avec Rambow pour repérer les VPE et les interpréter. Ainsi, dans leur article (2001), les corpus d'étude contiennent des extraits du *Wall Street Journal (WSJ)* annotés dans le *Penn Treebank*. Différents types de traits ont été pris en compte afin d'identifier ceux qui sont en corrélation avec la présence ou l'absence de VPE, à savoir des traits de surface, des traits morphologiques et syntaxiques, et des traits sémantiques (Hardt & Rambow 2001, 291-92).

Les traits de surface concernent non seulement le nombre et la distance entre les mots et entre les phrases mais aussi la longueur de l'antécédent. Les traits morphologiques et syntaxiques sont liés à la présence des auxiliaires dans l'antécédent et dans le site elliptique, à la voix, à la structure syntaxique et à la sous-catégorisation des verbes. Les traits sémantiques et discursifs sont liés quant à eux aux compléments, à la polarité, et à la structure du discours.

Hardt & Rambow ont ensuite entraîné un système pour apprendre des règles, afin de décider s'il y lieu de réaliser une VPE ou non, et à quel moment dans un système de génération de texte. Les règles sont apprises à l'aide du système d'apprentissage RIPPER (Cohen 1996). Ils obtiennent ainsi un taux de 7,5% d'erreurs. Leurs résultats ont montré que les traits qui affectent le plus la VPE sont la distance et la relation syntaxique et discursive entre l'antécédent et le site elliptique.

### 3.2.2. Annotation manuelle et apprentissage supervisé de la VPE

L'étude de Bos & Spenader (2011) s'inscrit ainsi dans la continuité de ces travaux. Il s'agit d'une démarche caractérisée par une annotation semi-automatique des 25

[27] Précision et rappel, de l'anglais, *precision and recall,* sont des critères de mesure de performance. La précision correspond au nombre d'éléments correctement détectés par rapport à tous les éléments détectés. Le rappel correspond au nombre d'éléments correctement détectés par rapport à tous les éléments à détecter présents dans le corpus. Une mesure qui combine la précision et le rappel est la F-mesure.

sections du *WSJ*. L'objectif principal de cette étude est d'annoter les éléments liés à la VPE et comparer les exemples du corpus sous étude aux exemples standard trouvés dans les recherches théoriques. Les éléments annotés sont notamment les propriétés syntaxiques du contexte elliptique et de son voisinage, à savoir, les auxiliaires qui déclenchent les VPE (voir le tableau 2 ci-dessous), le début et la fin de l'antécédent, le type de la configuration syntaxique de l'antécédent (VP, TV, NP, PP ou AP), ainsi que le type de configuration syntaxique de la séquence entre l'antécédent et le site elliptique (2011, 468).

| Type | Instances |
|---|---|
| aux | do don't does doesn't did didn't done doing |
| aux | am 'm are aren't ain't is 's isn't was wasn't were 're weren't be been |
| aux | have 've haven't has 's hasn't had 'd hadn't |
| modal | can cannot can't |
| modal | could couldn't |
| modal | may mayn't |
| modal | must mustn't |
| modal | might mightn't |
| modal | will won't |
| modal | would, wouldn't |
| modal | shall shan't |
| modal | should shouldn't |
| semi-modal | need needn't |
| semi-modal | dare daren't |
| semi-modal | ought oughtn't |
| infinitival | to |

**Tableau 2 : Auxiliaires déclencheurs de la VPE dans l'étude de Bos & Spenader (2011, 468)**

Le tableau (2) présente les variations morphologiques des déclencheurs des VPE (auxiliaires, modaux et semi-modaux), pris en compte par Bos & Spenader lors de la première phase d'annotation. Comme nous le verrons à la fin de ce chapitre, nous avons établi notre propre classification du phénomène elliptique sur la base de ces mêmes déclencheurs afin d'envisager sa reconnaissance automatique. Cette étude a présenté des résultats quantitatifs en termes de :

> frequency of the various VPE triggers, the distribution of the syntactic category of VPE antecedents with respect to VPE triggers and source–target patterns, and the distribution of source–target patterns across VPE triggers. (Bos & Spenader 2011)

Les auteurs ont montré une couverture de l'annotation des VPE plus importante que les méthodes précédentes. En effet, 487 cas de VPE ont été détectés par rapport à 260 seulement par Hardt (1997) dans le même corpus.

Inspirée par les annotations faites par Bos & Spenader, l'une des dernières approches informatisées des VPE revient à Liu *et al*. (2016). Contrairement à la majorité des approches existantes, qui se sont focalisées sur la détection du verbe ellipsé d'une part et l'identification de l'antécédent d'autre part, Liu *et al.* proposent une décomposition de la démarche en trois étapes importantes durant lesquelles des comparaisons sont faites entre divers systèmes d'apprentissage :

> 1. Target Detection (T),
> where the subset of VPE targets is identified.
> 2. Antecedent Head Resolution (H),
> where each target is linked to the head of its antecedent.
> 3. Antecedent Boundary Determination (B),
> where the exact boundaries of the antecedent are determined from its head. (Liu *et al*. 2016, 33)

En effet, à partir d'un corpus annoté en parties du discours, la première étape consiste à identifier automatiquement les candidats qui semblent déclencher la VPE, à savoir, les modaux ou les auxiliaires[28] (*be, do, have*) en entrainant un classifieur par régression logistique $Log^T$ à l'aide de diverses informations : étiquetage morphosyntaxique, lemme, analyse en dépendances syntaxiques. Dans la deuxième étape, les candidats susceptibles d'être des antécédents sont repérés automatiquement selon la position-tête qu'ils occupent dans la phrase. Pour ce faire, Liu *et al.* ont pris en compte les trois phrases qui précèdent immédiatement le site elliptique et la phrase même du site elliptique. De ce fait, les antécédents cataphoriques, qui représentent 1% des occurrences dans l'étude de Bos & Spenader (2011), ont été ignorés. Pour mener à bien cette deuxième étape, deux modèles ont été entrainés : $Log^H$ un classifieur qui permet de distinguer les antécédents possibles et $Rank^H$ un modèle qui permet de spécifier les préférences entre ces antécédents.

[28] Appelés dans leur recherche *Light verbs* : ces verbes sont décrits comme « légers » et sémantiquement faibles puisque lorsqu'ils sont utilisés dans la phrase, ils dépendent d'autres mots pour avoir un sens complet.

Enfin, vient la troisième et dernière étape qui consiste à délimiter les frontières de l'antécédent potentiel dans la phrase. Différents candidats, avec des frontières différentes, sont générés automatiquement. Plusieurs éléments-clefs ont été considérés pour cerner les caractéristiques des antécédents, notamment, le parallélisme entre la tête et le vide dans le site elliptique[29].

L'annotation qu'ont faite Liu *et al*. (2016) prend en compte des caractéristiques syntaxiques et lexicales simples. Les trois étapes s'inscrivent dans une méthode d'apprentissage automatique à partir de classifications établies et des analyses syntaxiques et morphosyntaxiques effectuées en plus de l'annotation. La combinaison de toutes ces opérations semble présenter plusieurs avantages selon les auteurs :

> We have explored a decomposition of Verb Phrase Ellipsis resolution into subtasks, which splits antecedent selection in two distinct steps. By modeling these two subtasks separately with two different learning paradigms, we can achieve better performance then doing them jointly, suggesting they are indeed of different underlying nature. (Liu *et al*. 2016, 39)

Dans une recherche concurrente menée au cours de la même année, Kenyon-Dean *et al.* (2016) ont reproché à cette étude la non-inclusion des candidats de *to-VPE*, déclencheurs des ellipses verbales, et qui représentent environ 5% des données analysées dans le corpus de *WSJ* comme le montre le tableau ci-dessous[30] :

| Auxiliary Type | Example | Frequency |
|---|---|---|
| Do | does, done | 214 (39%) |
| Be | is, were | 108 (19%) |
| Have | has, had | 44 (8%) |
| Modal | will, can | 93 (17%) |
| To | to | 29 (5%) |
| So | do so/same[3] | 67 (12%) |
| TOTAL | | 554 |

**Tableau 3 : La distribution des auxiliaires déclencheurs de la VPE dans les 25 sections de WSJ (Kenyon-Dean *et al.* 2016, 1735))**

[29] D'autres caractéristiques sont détaillées dans Liu *et al.* (2016).
[30] Le total correspond en réalité à 555 occurrences.

L'objectif de cette recherche est d'améliorer les moyens d'identification des meilleurs antécédents et la résolution de l'ellipse en utilisant les techniques d'apprentissage automatique entrainées sur un corpus annoté. Deux étapes ont caractérisé cette démarche, d'abord la détection de la VPE puis l'identification de l'antécédent. Cette méthode s'inscrit dans un apprentissage supervisé entrainé à partir de données analysées syntaxiquement :

> a novel approach to detecting and resolving VPE by using supervised discriminative machine learning techniques trained on features extracted from an automatically parsed, publicly available dataset. (Kenyon-Dean *et al.* 2016, 1734)

La figure (3) illustre les deux étapes suivies pour résoudre une VPE :

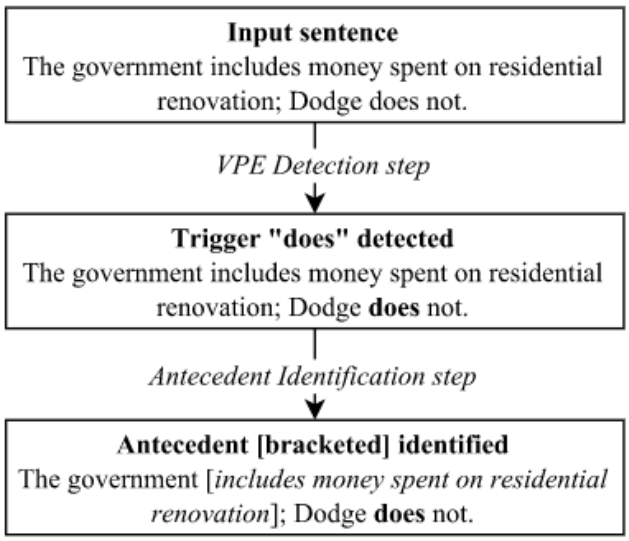



**Figure 3 : VPE trouvée dans le WSJ (démarche suivie par Kenyon-Dean *et al.* (2016, 1735))**

Le texte a été analysé syntaxiquement en utilisant l'analyseur syntaxique de Stanford disponible dans le *CoreNLP package* (Manning *et al.* 2014) pour une extraction des caractéristiques de la VPE. Les traits utilisés sont les suivants :

— l'auxiliaire déclencheur de la VPE : son type et son identité,

— le contexte : trois mots avant et trois mots après l'auxiliaire déclencheur ainsi que leurs fonctions dans le discours,

— la syntaxe : relation entre les auxiliaires et leur contexte syntaxique.

Cette approche a révélé l'importance d'identifier les auxiliaires et les différents traits cités ci-dessus puisque l'exactitude calculée pour la détection de la VPE a atteint le taux de 80.78% contre 69.52% pour l'étude précédente sur le corpus de Bos & Spenader (2011).

De leur côté, McShane & Babkin (2016) décrivent et évaluent ViPER (*VP Ellipsis Resolver*), un système dont l'objectif est de détecter et de reconstruire certains types particuliers d'ellipses du syntagme verbal avec un haut niveau de précision. Ce système est basé sur des principes linguistiques tels le parallélisme syntaxique simple (comme dans la coordination)[31] et la corrélation entre les modalités[32], ainsi que sur l'évaluation de la présence nécessaire ou non des constituants de la phrase.

ViPER est ainsi programmé pour identifier les structures simples et les analyser, mais ignore les structures complexes. La distinction entre les structures d'ellipses simples et complexes est majoritairement faite selon un critère clef qui est le *sponsor* (les éléments contribuant à l'interprétation de l'ellipse, que nous traduisons par antécédent) (McShane & Babkin 2016, 4). Une fois que la proposition contenant l'antécédent est identifiée, McShane & Babkin proposent de distinguer les éléments à inclure de ceux à exclure de l'antécédent pour permettre une reconstruction de l'ellipse. Pour offrir ensuite à ViPER la possibilité de gérer la modalité, McShane & Babkin (2016) retiennent (sur les 10 répertoriées dans la théorie de Nienburg & Raskin, 2004) neuf types de modalités susceptibles d'être déclencheurs d'une VPE. Ces modaux et auxiliaires sont d'abord détectés (grâce aux marques de ponctuation qui montrent une interruption dans le discours), puis analysés syntaxiquement avec le reste des éléments du *sponsor* à l'aide des outils de *Stanford CoreNLP*.

[31] Défini comme « Structural Parallelism. Each structurally parallel context contains an ellipsis clause directly preceded by a conjunct that is syntactically connected to in one of three ways that can be loosely described as exhibiting syntactic parallelism: coordination (be it of clauses or verb phrases), parataxis, and certain types of main/adverbial clause pairs » (McShane & Babkin 2016, 8).
Le parallélisme syntaxique a été pris en compte par de nombreuses recherches antérieures pour expliquer l'ellipse, notamment dans (Goodall 1987, Hobbs & Kehler 1997, McShane 2005).

[32] Une version plus approfondie est fondée sur la microthéorie de la modalité introduite par (Nirenburg & Raskin 2004) dans la théorie de la sémantique ontologique.

Enfin, pour la résolution de la VPE, la démarche à base de règles de McShane & Babkin s'appuie sur les approches *LF-copying* puisque ViPER est censé copier l'antécédent identifié dans le site elliptique. Contrairement aux démarches suivies dans les autres méthodes, ViPER ne ne procède pas par apprentissage supervisé. Cet outil a été utilisé uniquement pour vérifier la résolution des structures elliptiques qu'il a classées comme simples.

Les auteurs n'ont pas utilisé de corpus préalablement annotés pour l'entrainement du système car ils estiment que le travail d'annotation requiert beaucoup de temps. De plus, tout comme Bos & Spenader (2011), ils considèrent que la plupart des annotations faites sur les corpus ne sont pas facilement réutilisables et nécessitent des réajustements pour pouvoir tester toutes leurs hypothèses.

Les résultats de ce travail de détection et de résolution sont prometteurs et ont ainsi atteint un nombre significatif de VPE détectées et résolues. Sur les 393 VPE analysées, McShane & Babkin rapportent 239 cas correctement résolus. Alors que 75 exemples ont été partiellement résolus, 80 (20%) exemples sont complètement incorrects et nécessitent une redéfinition de certaines caractéristiques.

### 3.2.3. Annotation manuelle et apprentissage supervisé du sluicing

Parmi les autres catégories analysées, plusieurs recherches fondées sur corpus ont été menées pour un repérage automatique des *sluicing*. Nous citons par exemple l'étude de Fernandez *et al.* (2005) sur la détection automatique des '*root*' *sluices* qui sont des phrases isolées (extraites du BNC) apparaissant généralement dans un dialogue comme *Who? What ? Why?*

Une recherche plus récente, similaire à cette dernière a été menée par Baird, *et al.* (2018)[33]. L'objectif est d'examiner la distribution des *sluices* en redéfinissant la classification faite par Fernandez *et al.* sur le *BNC*. En effet, Baird *et al.* (2018) ont développé un système par apprentissage afin de détecter et de classer

[33] Cette recherche s'inscrit dans le cadre du *Santa Cruz Ellipsis Project* coordonné et mené par Pranav Anand & Jim McCloskey. L'objectif de ce projet est d'annoter le contenu implicite des *sluices* et de mettre à disposition de la communauté scientifique un large corpus avec des annotations aussi riches que possible. http://ohlone.ucsc.edu/SCEC/ (consulté le 14 avril 2017 à 17:20).

automatiquement des occurrences du *sluicing* à partir d'un corpus de sous-titres[34]. La méthodologie suivie par Baird *et al.* se déroule en trois étapes déterminantes. La première consiste à repérer toutes les occurrences *wh*– à l'aide d'expressions régulières réunies dans un programme Python. La deuxième repose sur une annotation manuelle (classification) d'un échantillon représentatif du nombre d'occurrences détectées dans l'étape précédente : cet échantillon sera utilisé comme corpus d'entraînement. Enfin, un système a été élaboré dans la troisième étape pour classer automatiquement les occurrences de *sluices* dans le reste du corpus n'ayant pas été utilisé comme corpus d'entrainement.

Lors du repérage automatique des *sluices* dans le corpus, Baird *et al.* ont identifié les deux catégories suivantes :

– *Embedded sluicing* qui renvoie à une proposition contenant un verbe comme *know, ask, say, understand*, suivi d'un mot *wh*–, et,

– *Root Sluicing* qui désigne toute occurrence contenant le mot *wh*– sans aucun verbe.

À partir d'un échantillon de 4 500 exemples, les occurrences *root et embedded sluicing* précédemment détectées sont ensuite annotées manuellement en prenant en compte l'usage qu'en font les locuteurs dans une situation d'énonciation donnée. Quatre usages sont alors dégagés :

- **Direct** : le *sluice* interroge une partie indéfinie de l'antécédent qui n'est pas nécessairement connue par le locuteur, et qui peut être implicitement ou explicitement exprimée.

[34] Le corpus *Opensubtitles* accessible via http://opus.nlpl.eu/ (consulté le 13 août 2018 à 13:18).

(11) A: He didn't come.
B: Why?
A: Break up[35].

– **Clarification :** le *sluice* interroge tout l'antécédent, exprimant notamment la surprise ou la confusion. Cette catégorie peut inclure également les usages illocutoires des mots *wh*– et les *sluices* sans antécédent linguistique.

(12)A: Captain ! It 's the Tomb of Heroes !
B: What?
B: How can it be?[36]

– **Reprise :** le *sluice* interroge la partie définie et explicite de l'antécédent. L'élément interrogé est nécessairement connu par le locuteur.

(13) A: They made her mad.
B: Who ?
A: The devils

– **Aucun**[37] : cette catégorie renvoie aux occurrences qui ne sont pas des *sluices*. Elles ont été détectées par erreur en raison d'un mauvais étiquetage (POS)[38], ou de questions figées.

(14) A: I saw you.
B: Saw what ?

Les trois auteurs de cette recherche ont annoté individuellement 100 occurrences et ont atteint un taux d'accord de 84%. Cette étape était nécessaire avant d'aborder l'élaboration du modèle qui a été établi pour deux classifieurs : *NaiveBayes* et *Decision tree*[39]*.* Ce dernier a atteint de bons résultats dans la détection et la classification des *sluices* et a obtenu un taux de 80% de *sluices* correctement identifiés et de 67% correctement classés comme le montre le tableau (4) ci-dessous.

[35] Tous ces exemples sont extraits de Baird *et al.* (2018).
[36] Malgré la présence d'un point d'interrogation, on remarque l'expression de l'incrédulité et de la surprise. Ce qui fait que le statut de « what » est ambigu, même si la « surprise » concerne en effet l'assertion « it's the Tomb of Heroes ! ».
[37] *None* dans la version anglaise
[38] L'étiquetage consiste à donner à chaque mot du corpus une étiquette présentant des informations morphosyntaxiques.
[39] Méthode de classifications à base d'arbre de décision.

| Predicted Class | True Sluices | Correctly Categorized |
|---|---|---|
| clar | 0.81 | 0.76 |
| dir | 0.69 | 0.61 |
| rep | 0.57 | 0.57 |
| Total | 0.80 | 0.67 |

**Tableau 4 : Résultats de l'arbre de décision des sluices dans Baird *et al.* (2018, 1583)**

Le travail récent de Rønning *et al*. (2018a,b) se focalise sur la résolution des sluices et leur traitement entièrement automatique à l'aide des réseaux neuronaux récurrents milti-tâches. Cette étude présente plusieurs avantages dans la mesure où, par rapport aux travaux antérieurs, elle ne s'appuie ni sur une sélection manuelle des traits utilisés pour l'apprentissage, ni sur une annotation syntaxique préalable des données.

> Syntactic information is arguably important for sluice resolution, but we show that multi-task learning with partial parsing as auxiliary tasks effectively closes the gap and buys us an additional 9% error reduction over previous work. Since we are not directly relying on features from partial parsers, our system is more robust to domain shifts, giving a 26% error reduction on embedded sluices in dialogue.

### 3.2.4. Bilan des approches sur corpus

Grâce aux travaux menés sur corpus, nous retiendrons que chaque méthode a emprunté des indices et des critères de détection appartenant à différents éléments-clefs présents dans le contexte (déclencheur, antécédent, co-référence), sélectionnés sur une base d'options théoriques qui privilégie, selon les objectifs à réaliser, soit une syntaxe simple, soit l'annotation. Ces critères sont bien évidemment établis à partir de la définition même de l'ellipse, c'est-à-dire à partir de la relation entre sens et forme. La plupart de ces démarches ont nécessité un travail d'annotation manuelle plus au moins important, ce qui est généralement le plus coûteux. Pour certaines de ces approches, la détection et la résolution de l'ellipse

constituaient les objectifs primordiaux, pour d'autres, seule la détection du site elliptique a été envisagée. Les structures complexes ont été mises de côté (en particulier dans l'approche de McShane & Babkin) puisque ces dernières présentent encore des difficultés jusque dans le champ même des recherches linguistiques fondamentales. Lorsqu'aucun tri n'a été fait entre les structures simples et complexes, un taux d'erreur important a été rapporté.

En bref, ces études sont issues soit d'une approche à base d'un apprentissage supervisé et développées à partir de données d'entrainements conséquentes, soit d'une approche à base de règles en mettant en avant les configurations syntaxiques et morphosyntaxiques des constructions elliptiques observées dans un corpus de développement. C'est en nous inspirant de ces travaux que nous avons opté pour une approche à base de patrons développés à partir de critères morphosyntaxiques, comme nous le verrons dans les chapitres suivants.

D'autres études ont également été menée sur la VPE (par exemple Nielsen 2004). La VPE reste sans aucun doute l'ellipse la plus étudiée à l'aide de procédures de traitement automatique. Comme en linguistique théorique, la VPE présente un défi de taille à la linguistique outillée. Ces recherches ont principalement visé l'anglais. À notre connaissance, aucune étude n'a été envisagée sur la VPE en français. L'une des études que nous retenons ici sur la VPE a été menée par Pilevar *et al.* (2011), qui ont aligné automatiquement un corpus parallèle de sous-titres[40] anglais-persan. Les méthodes utilisées s'avèrent similaires à celle de Bos & Spenader (2011). Les auteurs ont rapporté que plus de 10 515 occurrences de VPE sont présentes dans le corpus anglais. Ils ont ensuite montré que les VPE sont beaucoup moins nombreuses en persan où elles sont traduites, donc disparaissent.

D'autres catégories ont également fait l'objet d'un traitement informatisé. Rello (2010) par exemple a envisagé une détection automatique de l'ellipse du sujet et sa distribution en espagnol dans le corpus ESZIC[41] en exploitant deux genres de discours, juridique et médical. Le *gapping* est traité par Schuster *et al.* (2018) afin de

[40] Extraits du corpus OPUS (*open-subtitles*) qui contient plus de 3,7 millions de mots dans chaque langue.
[41] Explicit Subjects Zero-pronouns and Impersonal Constructions.

produire des représentations de l'analyse syntaxique en dépendance qui encodent explicitement le matériel ellipsé. 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)[42] 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 se trouve donc peu représenté dans les corpus d'entraînement, ce qui complique encore la tâche des analyseurs syntaxiques. Pour combler le manque d'exemples dans les corpus d'entraînement, Droganova *et al*. (2018b) ont produit semi-automatiquement des phrases artificielles, 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.

## 4. Conclusion intermédiaire : notre classification

La quantité de travaux et de données sur la question de l'ellipse est immense et nous n'avons présenté ici brièvement que les approches contemporaines pertinentes à la démarche appliquée retenue pour ce travail. Les débats générés autour de (i) la nature de l'élément manquant dans le site elliptique (ii) de l'antécédent, et (iii) de la relation que ces deux derniers entretiennent entre eux, sont à l'origine de la division des approches contemporaines de l'ellipse entre approches structurales, non structurales ou hybrides. Nous nous sommes alors limitée à la présentation de la taxonomie établie par van Craenenbroeck & Merchant (2013) parce qu'elle nous

[42] Nivre *et al.* (2016).

semble la plus étoffée parmi les classifications existantes et qu'elle englobe les ellipses traitées dans les études modernes.

Pour conclure le tour d'horizon préalable à notre propre investigation, après avoir examiné les orientations nouvelles des approches contemporaines sur l'ellipse par l'intégration de la méthodologie sur corpus, nous avons montré que parmi les catégories d'ellipse, la VPE est celle qui a suscité le plus d'intérêt dans le domaine du TAL. Depuis les premiers travaux menés par Hardt (1992), les procédures mises en place pour traiter le phénomène ne cessent d'évoluer vers des méthodes de plus en plus sophistiquées, et reposant sur l'apprentissage supervisé.

Cet aperçu non exhaustif des études de l'ellipse sur corpus permet en particulier de montrer la difficulté de mettre en place une démarche purement automatique pour le traitement des différents types d'ellipse. Le chercheur intervient toujours tout au long de la procédure méthodologique pour élaborer certains paramètres nécessaires au bon déroulement de l'apprentissage. Il apparaît ainsi que sur le plan des procédures, la principale différence entre linguistique de corpus et TAL, est le mode d'exploitation des corpus : en linguistique de corpus, il s'agit d'*interroger*[43] un corpus à l'aide de requêtes pour faire émerger des phénomènes et les quantifier, tandis qu'en TAL, on élabore des programmes informatiques pour effectuer des opérations comme celles consistant, par exemple à reconnaître l'antécédent d'une ellipse, à repérer le site elliptique même ou encore à restituer l'ellipse.

L'intérêt de ces systèmes automatiques réside pour certains (à base de règles par exemple) dans la possibilité d'effectuer le choix de se dispenser d'annotations manuelles des occurrences elliptiques (McShane & Babkin 2016, 2), ce que nous pouvons considérer comme une étape préliminaire d'un chemin vers un apprentissage automatique. Pour pouvoir envisager un traitement opérationnel, les procédures automatiques nécessitent dans la plupart des cas une simplification des catégories identifiées dans la littérature sur l'ellipse. Nous notons par exemple le cas de McShane & Babkin (2016, 14) qui n'établissent pas de distinction entre la VPE et la PAE puisque les modaux sont considérés au même niveau que les *matrix*

[43] Nous empruntons ce terme à Loock (2016).

*verbs* comme *agree, ask, avoid, beg, challenge, choose*. Ce choix de simplification peut parfois être imposé à la fois par les outils utilisés et par les objectifs à réaliser, ce qui constitue un point de rencontre avec notre recherche.

Ainsi, au terme de ce chapitre et partant des travaux menés, nous cherchons pour notre part à proposer un traitement automatique de l'ellipse sur corpus dont l'objectif est de cerner les enjeux théoriques et pratiques de sa détection et traduction automatiques. La méthodologie que nous exposerons plus loin permet de tester nos hypothèses de recherche, à savoir :

i) La définition de l'ellipse menant à une classification morphosyntaxique de ses manifestations, jointe à l'établissement des critères nécessaires à son repérage, devraient faciliter sa détection automatique.

ii) Les erreurs observées lors de détection pourraient contribuer à la compréhension des sources d'erreurs (de l'ellipse) dans la traduction automatique du phénomène.

À cette fin, nous avons élaboré notre classification en la fondant entièrement sur les critères morphosyntaxiques qu'il est possible de formaliser à l'aide des outils dont nous disposons. Par cette classification à base d'éléments déclencheurs, nous ne visons pas à théoriser le phénomène elliptique, mais à mettre en avant les conditions morphosyntaxiques de son apparition comme étape préliminaire à une analyse approfondie de sa détection d'une part et de sa traduction humaine et automatique d'autre part. De ce fait, cette classification n'implique en aucun cas le rejet de telle ou telle sous-catégorie identifiée dans les recherches contemporaines, mais tient simplement compte des contraintes techniques nécessaires pour éviter des erreurs de repérage effectuées par le système de détection. Sont présentés ci-dessous les différents types d'ellipse de notre classification et des exemples les illustrant. Une référence est faite aux catégories identifiées dans van Craenenbroeck & Merchant (2013) pour chaque déclencheur. L'ellipse peut être :

— déclenchée par un modal : un modal peut en effet déclencher des ellipses du syntagme verbal tels que la VPE et le *pseudogapping.*

(15) Lauren can **play the guitar** and Mike **can** Ø, too. (van Craenenbroeck & Merchant, 2019)

— déclenchée par *be ou have* : ces deux auxiliaires[44] peuvent déclencher des ellipses du syntagme verbal comme la VPE, le *pseudogapping,* et l'ellipse des compléments[45].

(16) <CDEV> If you're falling for him, I'**m** not Ø.

— déclenchée par *do* : les ellipses du syntagme verbal que *do* déclenche sont la VPE, et le *pseudogapping*.

(17) <CDEV> They **kept attacking** and we **did**n't Ø.

— déclenchée par le marqueur de l'infinitif *to* : cas de la VPE et du *pseudogapping*

(18) <CDEV> I didn't **beat her** and I didn't try **to**. Ø?

— déclenchée par une inversion sujet-verbe (modal ou auxiliaire) ou une *question tag*

(19) <CDEV> You don't **believe me**. - **Should I** Ø?

— déclenchée par un pronom *wh–* : cas du *sluicing* et ses sous-catégories, et d'autres ellipses propositionnelles :

(20) <CDEV> Someone **is knocking at the door**, but I don't know **who** Ø.

— déclenchée dans une *question fragmentaire* : ellipse identifiée lors de l'analyse de certains genres de discours (dialogue informels et recettes de cuisine). Elle renvoie à

[44] Pour ce qui est des auxiliaires, nous adoptons dans la présente recherche la catégorisation simplifiée des syntacticiens et identifions *be, have* et *do* comme des auxiliaires ; nous les distinguons des verbes pleins.
« The set of auxiliary verbs in English is a closed class […]. This closed class includes the verbs *be, have* and *do* (often referred to as the primary auxiliaries) […] » (Depraetere & Langford, 2012, 22).

[45] Nous rappelons que dans la classification de van Craenenbroeck & Merchant, l'ellipse des compléments est classée dans la catégorie des ellipses du syntagme verbal.

l'omission de l'auxiliaire accompagné du sujet dans les interrogatives, laissant apparents dans la phrase le verbe lexical et les compléments[46].

(21) <CDEV> Ø Going somewhere?
(22) <CDEV> Understand that?

Les catégories que nous établissons ci-dessous répondent aux quatre questions-tests que van Craenenbroeck & Merchant ont formulées pour repérer des ellipses nominales. La liste que nous dressons n'est pas exhaustive mais recense parmi les ellipses nominales celles que nous avons sélectionnées pour leur détection automatique, essentiellement en raison des problèmes qu'elles posent à la traduction automatique (en particulier jusqu'à une période récente de l'anglais vers le français et, encore aujourd'hui, vers l'arabe[47]).

— Ellipse déclenchée par le *'s* du génitif

(23) <CDEV> He tooks John's **bag** but not Mary**'s** Ø.

— Ellipse déclenchée par un *quantifieur*

(24) <CDEV> Thank you, but I already have **some** Ø.

— Ellipse déclenchée par un nombre cardinal

(25) <CEx> If they have **eggs**, bring me **six** Ø.

— Ellipse déclenchée par un nombre ordinal

(26) <CDEV> I will do my best to finish **the first activity**, but keep in mind that you have to deal with **the second** Ø.

[46] Il est à noter dans l'exemple (21), que l'absence de l'antécédent ne pose pas de réel problème à la détection de l'ellipse dans la mesure où l'économie de la structure grammaticale de la phrase suffit à remarquer qu'il manque un auxiliaire et son sujet. Cependant, pour restituer l'ellipse, comme le montre l'exemple (22), un antécédent est nécessaire puisqu'aucune flexion (en tant qu'indice) n'est présente. En effet, ce qui est ellipsé peut aller du simple modal et sujet *Will you eat something?* à une proposition *Would you like to eat something ?*

[47] Nous tenons à conserver ces exemples qui sont à présent résolus grâce aux progrès du TAL, mais qui jusqu'en 2016 ont fait l'objet d'une analyse qui a constitué une étape dans la conduite de la présente recherche. Nous revenons sur ce point dans le dernier chapitre.

Le code (ou appellation) que nous attribuons à chaque catégorie, lors de l'analyse et de l'annotation, est présenté entre accolades { }, la colonne de gauche reprend les grandes catégories de la classification van Craenenbroeck & Merchant :

| | | |
|---|---|---|
| Ellipse du syntagme verbal | Déclenchée par un modal | {post-mod} |
| | Déclenchée par *be* et *have* | {post-have/be} |
| | Déclenchée par *do* | {post-do} |
| | Déclenchée par *to* (marqueur d'infinitif) | {post-to} |
| | Déclenchée une inversion sujet-verbe ou une *question tag* | {vs-tag} |
| Ellipse propositionnelle | Déclenchée par un pronom *wh*– | {post-wh} |
| | Question fragmentaire | {qs-frag} |
| Ellipse nominale | Déclenchée par le *'s* du génitif | {post-geni} |
| | Déclenchée par un quantifieur | {post-quant} |
| | Déclenchée par un nombre cardinal | {post-card} |
| | Déclenchée par un nombre ordinal | {post-ord} |

**Tableau 5 : Appellations attribuées aux catégories d'ellipse**

Nous avons ainsi fait le choix de détecter automatiquement ces ellipses pour deux raisons essentielles :

— La première tient à leur fréquence dans les langues que nous traitons (l'anglais et le français),

— La deuxième est liée au repérage des erreurs de traduction provoquées par les occurrences elliptiques.

Ce dernier aspect s'inscrit dans un cadre technologique en pleine expansion et impose une adaptation continue des objectifs de la recherche à une réalité mouvante. Certaines ellipses encore sources d'erreurs pour la traduction

automatique il y a encore quelques mois, voire les quelques semaines précédant le moment où nous écrivons, ne le sont plus désormais. D'autres, de même, ne le seront peut-être plus lors de l'impression de ce travail, du moins pour ce qui concerne la paire de langues que nous avons choisie. Cependant, bien que l'acceptabilité des traductions de ces ellipses s'améliorent continuellement aujourd'hui, nous avons considéré utile de poursuivre leur détection, en quête de caractéristiques peut-être universelles, transférables d'une langue à une autre et afin d'explorer, dans de futures recherches, l'impact du phénomène elliptique sur la traduction de langues encore peu dotées et moins étudiées que le français, telles que l'arabe, le berbère ou le chinois, entre autres.

Pour revenir à notre objectif immédiat, il convient donc de relever deux éléments nécessaires à la compréhension de la méthodologie établie et, plus tard, des analyses effectuées. En premier lieu, bien que nous n'empruntions pas la classification de van Craenenbroeck & Merchant pour la détection automatique des ellipses, il peut arriver que certains exemples s'y réfèrent naturellement lors de leur analyse qualitative, notamment lorsqu'une explication syntaxique est requise. Par ailleurs, certaines ellipses identifiées dans la classification syntaxique ne sont pas incluses dans notre projet de détection en raison de l'absence d'un élément déclencheur d'ellipse, ce qui rend la détection impossible (cas du *gapping*, par exemple), et d'une ambiguïté syntaxique engendrée par les éléments ellipsés impossibles à modéliser[48].

Enfin, le tableau ci-dessous met en lumière les trois éléments qui différencient en surface les classifications proposées dans ce chapitre.

[48] Ces éléments sont impossibles à formaliser à l'échelle d'une détection à base de *tokens* compte tenu du taux d'erreurs important qui en résulte (cas des réponses fragmentaires). Nous reviendrons en détail sur ce point dans le chapitre 2.

| La classification de van Craenenbroeck & Merchant (2013) | Notre classification |
|---|---|
| Classification syntaxique : prend en compte la fonction syntaxique de la catégorie ellipsée dans la phrase. | Classification morphosyntaxique : prend en compte l'élément déclencheur et l'ordre des étiquettes morphosyntaxiques afin de faciliter l'établissement des requêtes de détection et réduire le taux d'erreurs. Les ellipses sont classées selon le déclencheur. |
| Établissement d'une sous-catégorisation pour un seul type d'ellipse :<br>Par exemple, selon le nombre de résidus présents après le pronom *wh–* et la fonction du pronom même, une sous-catégorie du *sluicing* a été nommée. | Seul l'élément déclencheur est la source d'une sous-catégorisation. Ainsi, aucune distinction n'est faite entre les catégories du *sluicing* dans la mesure où le même patron[49] à base de tokens[50] reconnaît les sous-catégories citées dans Merchant & van Craenenbroeck. |
| Identification des réponses fragmentaires comme elliptiques. | Catégorie que nous reconnaissons mais que nous ne détectons pas dans le présent travail. Nous avons ajouté les questions fragmentaires compte tenu de leur fréquence dans les exemples collectés. |

Nous ajouterons que si le critère morphosyntaxique de l'ellipse est au cœur de notre recherche, il apparait alors évident qu'un travail ultérieur, à base d'apprentissage automatique s'appuyant sur une analyse approfondie des conditions et des critères syntaxiques, pourrait être envisagé afin de classer automatiquement le phénomène selon les classifications syntaxiques existantes. Dans ce cas, une classification syntaxique rigoureuse sera pertinente.

Le chapitre à venir, qui décrira les phases de notre méthodologie d'analyse de l'ellipse, s'attachera aussi à expliciter le choix d'un corpus d'étude et l'adoption d'une méthode de détection outillée.

[49] Combinaison de critères linguistiques pour repérer les catégories d'ellipse. Voir le chapitre 3 pour plus de détails.

[50] Un anglicisme utilisé en informatique pour signifier un mot, une donnée, une entité, ou une variable analysée. Par exemple la phrase *L'hiver est dur.* contient les token*s* suivants : *L' hiver est dur.* (avec une espace entre le *L'* et *hiver* et entre *dur* et *« . »*).

# Chapitre 2

# Méthodologie

> *Ellipsis creates challenging scientific and engineering problems. Although research over the past 50 years has shown that the principles permitting ellipsis involve many different types of information (grammatical structure, context, real-world knowledge), the precise mix of these principles and their interaction is still an open question*[51].

D'un point de vue théorique et méthodologique, le recensement des ellipses que nous venons de proposer pour permettre ultérieurement leur détection automatique, peut sembler trop large à appréhender dans le cadre d'un travail de thèse. Or, il s'agit bien, dans un premier temps, de cerner l'ellipse *en général*, plutôt que l'une ou l'autre en particulier et bien qu'il ne soit pas possible, pragmatiquement, de proposer des procédures de détection pour *toutes* les ellipses dans le cadre de notre étude, nous souhaitons donner toute sa place audit phénomène dans la révolution actuelle de la traduction automatique, tout en sachant qu'à l'inverse du processus de détection, nous nous concentrerons dans le chapitre 5, « Étude de la traduction automatique de l'ellipse post-auxiliaire », sur la traduction d'un seul type d'ellipse. Pour ce faire, notre démarche est le résultat combiné de l'ensemble des choix disciplinaires, théoriques et appliqués que nous avons faits au fur et à mesure que nous défrichions le sujet, en privilégiant le travail sur corpus, la linguistique

[51] https://reports.news.ucsc.edu/linguistics/ (accès le 20 octobre 2017 à 20:33).

outillée et le traitement automatique des langues, mais également les analyses manuelles. Ce chapitre présente ainsi les décisions que nous avons prises pour répondre aux questions suscitées par la constitution d'un corpus et le choix des outils informatiques. Les limites et les contraintes du modèle mis en place, seront présentées en fin de chapitre.

Cependant, avant de présenter les étapes nécessaires à la détection automatique de l'ellipse, il nous appartient d'apporter quelques clarifications sur ce que nous entendons par *traitement automatique* du phénomène elliptique. Comme suggéré précédemment, les objectifs de notre recherche ne visent pas à traiter informatiquement de façon exhaustive les identités du phénomène elliptique qu'elles soient d'ordres syntaxique, pragmatique ou sémantique, mais à étudier essentiellement, en premier lieu, leurs caractéristiques morphosyntaxiques, étape préliminaire indispensable au repérage et à la compréhension du phénomène afin de permettre sa détection, puis sa traduction.

Comme nous l'évoquions, le traitement automatique que nous proposons est le résultat d'un ajustement continu de contraintes théoriques et techniques rencontrées lors de l'exploitation du corpus d'étude. Il vise principalement la compréhension des caractéristiques des ellipses recensées afin de dresser une typologie d'erreurs qui se manifestent lors de sa reconnaissance et sa traduction automatiques.

La compréhension du phénomène elliptique en vue d'une formalisation[52] à des fins pratiques repose donc sur des méthodes nécessitant une simplification dans la représentation théorique de ses catégories (voir la classification présentée dans le chapitre précédent). Pour répondre à un nombre de contraintes ne pouvant être formalisées dans le cas précis de l'ellipse (antécédent extralinguistique, absence d'un déclencheur, annotation automatique, etc.), il nous a fallu faire des choix spécifiques, parfois imposés par les objectifs à atteindre et/ou les outils disponibles, et, de ce fait, la détection automatique de certains types d'ellipse n'a pu être envisagée.

[52] dans le sens de modéliser ou d'établir des algorithmes pour représenter telle ou telle occurrence.

La méthode de détection présentée au cours de ce chapitre repose sur une description morphosyntaxique de l'ensemble des catégories et sur l'identification des conditions et des critères déclenchant une ellipse (qui seront expliqués dans le chapitre 3). À travers cette méthode, nous n'avons nullement la prétention d'apporter une compréhension *complète* du phénomène, ce qui d'ailleurs semble impossible en raison de son caractère instable et des limites des outils disponibles à l'heure actuelle. En d'autres termes, il ne s'agit pas d'opposer notre travail de détection aux classifications existantes de l'ellipse (que nous n'empruntons pas pour la détection mais pour l'analyse des occurrences), mais seulement de montrer comment, en dépit de ses limites, notre démarche appliquée permet de souligner *la complexité* du phénomène elliptique. Les étapes constituant notre approche s'inscrivent plus largement dans une méthodologie hybride faisant appel à la linguistique théorique et à la linguistique outillée. Sont exposées ci-après les différentes étapes suivies : i) la constitution du corpus, ii) la présentation des outils informatiques utilisés iii) l'étiquetage du corpus et son annotation manuelle, iv) l'identification des conditions morphosyntaxiques et l'établissement des patrons, et v) l'interprétation des résultats. Ces deux derniers points feront l'objet des chapitres 3 et 4.

## 1. Constitution des corpus

Nous avons montré dans le précédent chapitre comment les recherches menées sur l'ellipse durant les dernières décennies ont tiré parti du développement de la linguistique de corpus. L'étude d'ellipses sur corpus, en effet, tirant sa singularité d'une démarche qui s'appuie sur des données authentiques, bénéficie d'avancées technologiques significatives participant au diagnostic d'apparition des occurrences elliptiques et au repérage de leurs distributions typologiques. L'usage exclusif des exemples forgés initiés par les approches de la grammaire générative n'offre pas les mêmes perspectives. Comment en effet, à partir d'exemples forgés, normés et décontextualisés, approcher le phénomène, dont la nature même est instable ? En d'autres termes, comment l'aborder à partir de ces exemples formatés, alors que l'une de ses apparitions les plus fréquentes se trouve dans le discours oral, sujet à

des variations continues ? De plus, en partant du caractère évolutif des langues naturelles, comment peut-on rendre compte des différentes formes elliptiques influencées par le genre du discours dans lequel elles apparaissent si l'on travaille sur ces exemples formatés, stables et invariants ?

Par l'ajustement constant des critères de collecte des données aux objectifs à atteindre, la démarche menant à la constitution d'un corpus est nécessairement analytique. Remarquons ici que le traitement du phénomène elliptique dans la grammaire générative, malgré les restrictions émises plus haut, ne semble toutefois pas incompatible avec la méthodologie de la linguistique de corpus, les deux approches présentant chacune un intérêt particulier : le positionnement théorique de la grammaire générative permet la description linguistique approfondie des catégories de l'ellipse en vue de sa détection automatique tandis que la méthodologie sur corpus permet l'étude de la variation à travers les occurrences attestées et le traitement informatisé du corpus. De ce fait, la complémentarité des deux approches apparaît évidente, l'une n'excluant pas l'autre. Bien entendu, cette orientation suscite d'autres interrogations qui portent plus spécifiquement sur la notion même de corpus : quel choix de corpus pour quel type d'ellipse ? Quel corpus pertinent pour établir et évaluer des patrons de détection ? Dans quelle mesure tel ou tel corpus est-il représentatif ? Et, par conséquent, quel est le degré de fiabilité des résultats obtenus ? Il est à cet égard assez surprenant de constater que malgré le recours actuel aux corpus, la notion même recouvre des réalités très diverses, en fonction de la discipline et des objectifs. En définitive, qu'est-ce qu'un corpus ? Quels sont les types de données le constituant ? Comment évaluer sa pertinence ?

Dans son acception large, comme indiqué précédemment, un corpus est constitué d'un ensemble de données *authentiques* réunies à des fins spécifiques. Certains corpus, par exemple, servent à l'observation de phénomènes linguistiques, tandis que d'autres sont utilisés afin de dégager leurs caractéristiques, et par là, les définir. Ils peuvent être des textes écrits ou oraux (voir le critère d'authenticité ci-après) et soumis à un ou plusieurs critères. Tous ces paramètres sont pris en compte par le chercheur lui-même en fonction des objectifs qu'il vise.

Notre définition du corpus est tirée des travaux de Sinclair (1996) et McEnery *et al.* (2006)[53] selon lesquels le corpus est défini par des critères linguistiques préalables à sa constitution et à ses objectifs, et en lien avec les outils informatiques.

> A computer corpus is a corpus which is encoded in a standardized and homogenous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance. (Sinclair 1996, 4)
> A collection of sampled texts, written or spoken in machine-readable form which may be annotated with various forms of linguistic information. (McEnery *et al.* 2006, 4)

Ainsi un corpus est-il, du moins pour la présente étude, un ensemble de textes écrits, sélectionnés selon certains critères, pour répondre aux objectifs fixés de traitement et d'analyse des ellipses. Il existe plusieurs manières de compiler des corpus : selon la nature des données (écrites ou orales), selon le type ou le genre de discours (journalistique, politique, institutionnel, …), ou encore selon les objectifs de la recherche. En effet, un corpus qui a été conçu pour analyser un phénomène syntaxique, par exemple, n'est pas interrogé de la même manière qu'un corpus destiné à l'étude d'une thématique de discours. Ceci s'applique également à sa constitution.

### 1.1. Critères de constitution

Les critères de constitution, d'ordre quantitatif et/ou qualitatif, sont établis par le chercheur lui-même. Pour notre part, nous avons adopté les critères résumés dans Loock (2016, 20-21) : échantillonnage, représentativité, authenticité, exploitation automatisée, et intégration d'informations linguistiques (critères en linguistique et en traductologie notamment).

**– Échantillonnage :** le corpus constitué n'est qu'une partie d'un ensemble qu'il représente en n'en donnant qu'un aperçu (par exemple, un genre de texte, un phénomène linguistique, …). Cet échantillon doit être constitué de textes choisis

[53] En effet, cette définition trouve ses origines chez Leech selon qui le corpus est :
« They [Computer corpora] are generally assembled with particular purposes in mind and are often assembled to be (informally speaking) representative of some language or text type » (Leech in Svartvik 1992, 116).

« avec suffisamment de soin pour qu'ils puissent être considérés comme typiques du genre » (Loock 2016, 20).

**– Représentativité** : un corpus est représentatif s'il est élaboré en vue de répondre aux objectifs d'une recherche donnée. Il est généralement admis qu'un corpus ne peut être représentatif que si les résultats de l'étude menée à son endroit peuvent être appliqués et généralisés à l'ensemble de la langue. Or, il ne s'agit pas toujours d'une représentativité quantitative[54]. Le corpus ne se définit pas par sa taille mais par la scientificité des données le constituant. Ceci sous-entend qu'un corpus de petite taille ne dévalorise aucunement l'étude menée, et ne remet pas en cause sa représentativité[55].

**– Authenticité :** les données qui constituent un corpus sont des utilisations réelles de la langue. Elles peuvent être simplement orales ou écrites, mais peuvent être également une transcription de l'oral (transcription d'interview) ou une adaptation orale de l'écrit (bibliothèque sonore). Elles peuvent relever d'un domaine ou d'un contexte particulier et correspondre à une époque spécifique.

**– Aptitude à l'exploitation automatisée :** l'une des caractéristiques présentes dans les définitions modernes de corpus est ainsi sa « forme électronique exploitable par une machine » (Loock 2016, 21).

**– Intégration d'informations linguistiques** : selon les objectifs d'analyse du corpus, plusieurs informations peuvent être ajoutées en utilisant des logiciels adaptés : annotation manuelle de corpus, étiquetage morphosyntaxique ou analyse syntaxique des phrases (voir Loock 2016, 21).

[54] Selon Sinclair (2008, 30), quelle que soit la taille du corpus analysé, ce dernier n'est pas destiné à représenter toutes les modalités de la langue. En fait, le critère de représentativité se manifeste à deux reprises : d'abord lors de la constitution du corpus (comme moyen de rassembler les énoncés qui partagent les mêmes caractéristiques) et lors de l'interprétation des résultats (comme moyen de vérifier l'adaptabilité du corpus à la nature du phénomène étudié, et par conséquent, à la possibilité de généraliser les résultats).

[55] Cependant, une phrase peut-elle constituer à elle seule un corpus ? S'il est évident que dans certains cas (langue peu dotée, ancienne ou rare par exemple) un corpus de taille moyenne peut être utile, il apparaît, dans d'autres cas, que ce corpus modeste ne peut être justifié si des données sont disponibles par ailleurs en quantité suffisante, tant il est vrai que, pour ne pas retomber dans les travers de la linguistique à partir d'exemples forgés, l'étude d'un fait de langue, pour qu'elle soit représentative, doit se fonder sur un certain nombre d'occurrences.

Cet ensemble de critères semble suffisant à l'établissement d'un corpus d'étude exploitable suivant les pratiques de recherche adoptées par la majorité des linguistes aujourd'hui et qui incluent notamment l'observation, l'analyse et la discussion des résultats, opérations amorcées à partir d'un corpus défini en lien avec les objectifs de la recherche. En réalité, pour nous, fonder l'analyse des ellipses sur corpus, en particulier sur un corpus parallèle, juxtaposition de textes originaux avec leurs traductions[56], apparaît efficace dans la mesure où croiser leurs occurrences permet de mettre en relief les différentes formes de l'ellipse et amène à prendre en compte non seulement les vides syntaxiques mais aussi les liens entre ces derniers et les unités visibles du discours dans deux langues différentes. C'est d'ailleurs ce lien qu'il permet d'établir entre la linguistique et la traduction qui nous a incitée, entre autres motivations, à choisir un corpus parallèle.

### 1.2. Présentation de nos corpus

Deux corpus sont utilisés dans le présent travail : un corpus que nous appelons corpus de développement (nous utilisons l'abréviation <CDEV> pour le désigner) constitué pour établir les patrons de détection automatique, et un corpus d'évaluation (<CE>), constitué pour mesurer la performance des patrons et évaluer la possibilité de leur application sur un autre corpus. Le corpus de développement est lui-même issu d'un corpus d'exemples <CEx> que nous avons constitué préalablement.

Nous entendons par *patrons*, appliqués à la détection de l'ellipse, des critères formulés à partir des conditions implémentables dégagées pour chaque type d'ellipse. Suivant la structure canonique de la phrase grammaticale en anglais (SVO), ces conditions seront préalablement définies pour plusieurs sortes d'ellipses et traduites, comme nous le verrons plus tard, dans le langage de requête propre à l'outil utilisé.

[56] Ce type de corpus est à distinguer des corpus comparables où les textes sont écrits indépendamment les uns des autres dans des registres de langues ou dans des langues différentes mais traitent du même thème et datent de la même époque.

### 1.2.1. Corpus de développement

Un corpus de développement est un ensemble de textes ou de phrases considérés comme pertinents pour constituer et établir, dans le cadre de ce travail, des patrons de détection. La constitution d'un corpus de développement, comme tout corpus, doit être fondée sur les critères précis de pertinence, de fiabilité et d'échantillonnage des données.

Les patrons de détection ont été mis au point manuellement à partir d'un corpus de développement (voir figure 4) de 5 362 tokens regroupant 331 exemples d'ellipses. Ces occurrences, repérées manuellement[57] dans leur contexte, sont toutes extraites de documents authentiques publiés entre 1960 et 2014 et n'ont aucun lien avec le corpus d'évaluation. Ce corpus de développement est constitué de pièces de théâtre (H. Pinter), nouvelles (G. Green, F. Forsyth, J. Arden, ...), articles de presse (*The Guardian*, The Daily Mail), dialogues de roman (J. Coe)[58]. Pour compléter ce corpus de développement, nous avons également utilisé des exemples issus de (McShane & Babkin, 2016) et (Rønning *et al.*, 2018b) afin d'augmenter le nombre d'occurrences elliptiques et couvrir ainsi le plus de variations possible.

```
26
27  I don't wish to kill her, he thought, insistently, looking over at her bed. Or do I?
28  It's not working, can you show me how?
29
30  I want to ask you a question. Will you allow me to?
31
32  I sometimes think that you're quite strong enough to do the things you want to.
33  I can't stand the disgrace. I can't.
34
35  So you told him that?
36  Maybe I have.
37
38  It wasn't mine. It was the club's.
39
40  I didn't want to,I must own up to that.
41
```

**Figure 4 : Extrait du corpus de développement**

[57] Au départ, ces ellipses sont sélectionnées selon l'ordre de leur occurrence dans les œuvres. Un tri, qui a consisté à éliminer les ellipses très fréquentes, a ensuite été fait pour équilibrer l'échantillon et varier les occurrences elliptiques.

[58] Voir la bibliographie pour les références des documents utilisés pour la constitution des corpus CDEV et CEx.

Nous avons ajouté aux phrases elliptiques de ce corpus 120 phrases non-elliptiques qui présentent soit le contexte qui précède ou qui suit les occurrences elliptiques, soit des occurrences indépendantes de ces dernières, ce qui nous a semblé nécessaire pour obtenir des séquences susceptibles d'être détectées comme étant elliptiques alors qu'elles ne le sont pas (faux positifs[59]) et par là, envisager l'amélioration des patrons.

Le dernier point restant à aborder est le statut des exemples que nous avons sélectionnés pour constituer le corpus de développement destiné à établir nos patrons de détection. Les exemples apparaissent tels qu'ils sont utilisés dans leur contexte, « attestés » et contextualisés. Conformément au choix théorique que nous avons fait, nous avons décidé de ne pas développer des patrons sur la base de phrases *bien formées*, ou d'exemples fabriqués, utilisés précédemment à des fins pédagogiques, puisqu'établir des patrons à partir d'exemples fabriqués permettrait difficilement la détection d'ellipses « non conformes ». En effet, les variations induites par le type de discours ne seraient alors jamais prises en compte. L'exemple qui illustre ce cas est l'ellipse de la paire sujet-auxiliaire dans la question nommée « fragmentaire » dans notre classification : *Leaving tomorrow ? Yes, I am,* ellipse qui ne répond à aucun cas illustré dans les théories traditionnelles.

### 1.2.2. Corpus d'évaluation

Pour vérifier la performance de nos patrons, nous avons constitué un corpus d'évaluation à partir de cinq sous-corpus présentant différents genres de discours (voir le tableau 6 ci-dessous) et avons extrait des échantillons. Les genres sélectionnés appartiennent aux domaines suivants : littéraire, promotionnel, politique, conversationnel et journalistique. Nous reviendrons en détail sur les caractéristiques de ces genres dans le chapitre 4.

[59] détectés comme positifs, mais en réalité négatifs, voir note de bas de page 91 *infra.*

| Corpus | Genre de discours représenté | Taille de l'échantillon utilisé (N Tokens)[60] | Informations |
|---|---|---|---|
| EUROPARL[61] | Politique | 283 205 | Actes du parlement européen des années 1996 à 2011<br>Conçu au départ pour servir aux recherches menées dans le domaine de la traduction automatique, il a été utilisé depuis dans de nombreuses autres études visant le traitement automatique des langues. |
| Sous-titres de séries TV[62] | Conversationnel | 233 978 | Pour constituer ce corpus, nous avons sélectionné des extraits des deux premières saisons de la série *Broadchurch* ainsi que des extraits des cinq premières saisons de la série *Downton Abbey*, récupérés des DVD (Strong & Lyn, 2018 ; Percival *et al*., 2018) et compilés par nous-même. Pour compléter cette collection, nous avons sélectionné au hasard un fichier de sous-titres *Opus subtitles* que nous avons vérifié. |
| PLECI[63] | Journalistique | 41 522 | PLECI est un corpus constitué par les universités de Poitiers et de Louvain, regroupant des articles de presse et quelques textes littéraires.<br>Dans ce corpus, nous avons sélectionné uniquement les articles et les romans publiés entre 1990 et 2014, traduits de l'anglais vers le français. |
| | Littéraire | 55 286 | Dans le cadre du corpus PLECI, le corpus littéraire que nous avons choisi est constitué de quatre romans écrits en anglais et traduits en français. Il s'agit de *Night Over Water* (1991) de Ken Follet, *Strawberry Tree* (2011) de Ruth Rendell, *A Widow for One Year* (1998) de John Irving et enfin *Harry Potter and the Order of the Phoenix* (2003) de J.K Rowling. |

[60] Nous ne donnons ici que la taille des échantillons exploités.
[61] http://www.statmt.org/europarl/ (accès vérifié le 10 juin 2017 à 10:49).
[62] Voir annexe II p.265.
[63] Nous avons accès à ce corpus depuis le 15 mars 2015. Nous remercions Raluca Nita (Université de Poitiers) d'avoir mis ce corpus à notre disposition.

| | | | |
|---|---|---|---|
| TED[64] | Promotionnel | 46 172 | Ce corpus est extrait d'un ensemble de transcriptions des conférences TED qui ont été alignées automatiquement pour être utilisées dans les recherches en TAL. Les conférences TED sont des conférences destinées à rencontrer l'intérêt de tous les genres de public en fonction de leur contenu. |

**Tableau 6 : Corpus utilisés et genres de discours les représentant**

En plus des critères précédemment exposés, c'est le libre accès aux textes collectés dans ces corpus qui a majoritairement favorisé leur choix. À l'exception du PLECI, tous les autres corpus présentaient l'avantage d'être directement téléchargeables en ligne et, malgré les droits qui leur restent associés, les licences permettent leur usage dans le cadre d'activités de recherche. Un autre critère de sélection de ces corpus tient à leur année de production sachant que notre étude porte sur l'anglais contemporain et sa traduction en français (de 1990 à 2014). Enfin, le dernier critère s'appuie sur la qualité des traductions proposées par des traducteurs expérimentés[65] et non sur une production « non professionnelle »[66]. On ajoutera, pour ce qui concerne l'aspect théorique de l'ellipse, que le recours à un corpus échantillonné par genre permet de vérifier dans quelle mesure le phénomène est spécifique à tel ou tel type de discours.

D'un point de vue pratique, il apparaissait plus économique de recourir à un ensemble de corpus déjà compilés (à l'exception du corpus conversationnel que nous avons compilé nous-même) afin d'utiliser le gain de temps obtenu à l'approfondissement des différentes étapes de la recherche.

Notre apport à ce niveau de réalisation a donc consisté à repérer, réunir et organiser tous ces sous-corpus afin de constituer un sous-corpus d'étude suffisamment représentatif permettant l'analyse des occurrences d'ellipses. La décision de rassembler un nombre conséquent de corpus variés est donc fondée sur l'hypothèse que la fréquence et la forme des occurrences elliptiques varient d'un

[64] https://wit3.fbk.eu/ (accès vérifié le 12 juin 2017 à 11:15).

[65] À l'exception du corpus TED, traduit par des traducteurs bénévoles (on ignore s'il s'agit de traducteurs professionnels expérimentés).

[66] Il ne s'agit pas ici d'un jugement de valeur. Par « non professionnelle », nous renvoyons aux volontaires qui proposent leurs services en dehors des circuits académiques et professionnels.

genre à l'autre et d'une langue à l'autre et que, en conséquence, les variations de leur comportement syntaxique, examinées soigneusement, pourraient permettre l'enrichissement des patrons de détection automatique.

Pour conclure cette étape, nous soulignerons que notre acception de l'ellipse comme fait de langue, requiert tout d'abord la confrontation du phénomène elliptique aux différentes constructions *canoniques* du système langagier. Notre objectif dans l'utilisation d'un corpus parallèle est ainsi d'évaluer les différentes variations du phénomène, tant au niveau intra-langue qu'au niveau inter-langue. Inclure les genres dans une telle étude semble compatible avec l'évaluation de la traduction du phénomène elliptique, puisqu'avant de traduire le texte, une compréhension du contexte, du type de discours et de ses conventions est nécessaire. Viennent ensuite les variations internes aux systèmes des langues que l'on traduit et qu'un corpus parallèle permet d'observer au plus près. En effet, ces discours sont issus de communications sociales ayant eu lieu dans des situations d'énonciation différentes. La différence entre ces sous-corpus discursifs est déterminée par les objectifs de communication communs aux participants. Ces objectifs peuvent engendrer des variations de structures syntaxiques dans les phrases utilisées, elliptiques ou non elliptiques, que les locuteurs contrôlent, consciemment ou non.

Fonder une analyse sur un corpus parallèle implique également un travail sur des textes et sur leur(s) traduction(s). Notre objectif de traiter l'ellipse dans la traduction humaine et automatique nous impose d'opter pour ce genre de corpus. L'approche comparative de deux langues permet en effet de mettre en relief leurs traits communs pour identifier par la suite les caractéristiques particulières à chacune d'elles et observer un fait de langue spécifique. Afin de cerner les éventuelles convergences et divergences, une telle comparaison offre une analyse fine des fonctionnements des langues en question ainsi qu'une compréhension de leurs mécanismes respectifs. De plus, ce sont les différences rencontrées dans le corpus,

qui, exploitées lors de l'analyse des processus de traduction, pourront permettre de prédire les erreurs engendrées par l'ellipse.

L'étude de ces deux variations (inter-langue et intra-langue) peut ainsi apporter une compréhension nouvelle du phénomène et contribuer à sa compréhension et à sa traduction, comme nous le verrons dans le chapitre 5 de ce travail.

## 2. Présentation des outils utilisés

La linguistique de corpus propose aujourd'hui un panel d'outils aptes à répondre aux besoins du linguiste en lui permettant la réalisation de ses objectifs dans les meilleures conditions possibles. Dans notre étude, les outils permettant d'intervenir sur un corpus dans une étude des phénomènes elliptiques, sont notamment les outils d'étiquetage. Nous inscrivons notre analyse de l'ellipse dans une approche hybride combinant l'aspect qualitatif et quantitatif afin de cerner les caractéristiques de ses occurrences.

Pour l'ensemble de nos travaux, nous utilisons les outils spécifiques réunis dans Stanford **CoreNLP**, développés à l'université de Stanford (Manning *et al.*, 2014). Notre choix s'est porté rapidement sur ce logiciel puisqu'il fournit des modèles d'analyse pour l'anglais. D'autres modèles pour les langues telles que l'arabe, le chinois, le français, l'allemand et l'espagnol sont aussi disponibles. Par ailleurs, la praticité et la simplicité de son utilisation ont également motivé notre choix. Enfin, cet outil offre également plusieurs niveaux d'analyses :

> It can give the base forms of words, their parts of speech, whether they are names of companies, people, etc., normalize dates, times, and numeric quantities, mark up the structure of sentences in terms of phrases and syntactic dependencies, indicate which noun phrases refer to the same entities, indicate sentiment, extract particular or open-class relations between entity mentions, get the quotes people said, etc[67].

Nous avons utilisé les deux versions de cet ensemble (en ligne et téléchargeable), tout d'abord, la version téléchargeable pour exploiter le corpus, puis la version en ligne qui permet de visualiser facilement les annotations réalisées pour une phrase.

[67] https://stanfordnlp.github.io/CoreNLP/index.html (accès vérifié le 6 juin 2019 à 10:05).

Pour exploiter le corpus et élaborer nos patrons, nous avons eu recours à l'analyse morphosyntaxique, la lemmatisation et l'étiquetage des entités nommées.

Le **Stanford Part-of Speech (POS) Tagger**[68] est un outil qui attribue à chaque token une étiquette[69] présentant des informations morphosyntaxiques (nom, verbe, adjectif, …). L'étiqueteur anglais utilise le jeu d'étiquettes *Penn Treebank*[70] qui est aujourd'hui le plus fréquemment utilisé et qui sert de référence à la majorité des outils d'étiquetage. La section 3.1 décrit l'utilisation de cet outil dans notre approche.

Le **Stanford Named Entity Recognizer (NER)** permet la reconnaissance des entités nommées : noms de personnes ou de sociétés, ou encore des noms de gènes et de protéines, etc.[71] Trois classes d'entités sont particulièrement performantes en anglais : les personnes, les organisations et les lieux. La figure (5) montre un exemple de la reconnaissance des entités nommées :

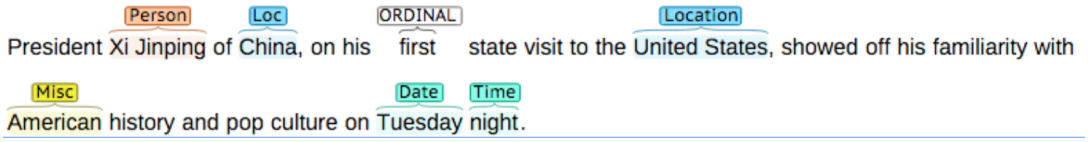



**Figure 5 : Reconnaissance des entités nommées**[72]

Lors de l'établissement de patrons, nous avons utilisé la reconnaissance des entités nommées lorsque les étiquettes morphosyntaxiques n'étaient pas suffisantes pour affiner les patrons (voir le patron {post-card}). De la même manière, nous avons eu recours à l'utilisation des lemmes (forme canonique du mot : singulier pour les noms, infinitif pour les verbes, etc.), que CoreNLP permet également de générer pour chaque token présent dans le corpus (figure 6).

[68] https://nlp.stanford.edu/software/tagger.html (accès vérifié le 06 juin 2019 à 10:10).
[69] Les étiquettes utilisées figurent dans la liste des abréviations.
[70] L'expression *Penn Treebank* a été employée pour la première fois par Geoffrey Leech dans le cadre du projet *Penn Treebank* utilisé pour annoter des gros corpus comme le *Brown Corpus, Wall Street Journal Corpus,* et le *Switchboard Corpus* dans le domaine du TAL. Ces annotations concernent les différentes propriétés des phrases et des textes.
[71] https://nlp.stanford.edu/software/CRF-NER.html (accès vérifié le 06 juin 2019 à 10:12).
[72] Exemple figurant sur le site.

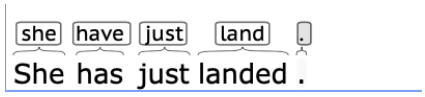



**Figure 6 : Lemmatisation**

**TokensRegex** est un outil proposé dans CoreNLP qui permet de créer des requêtes sous forme de patrons, sur la base d'une syntaxe d'expressions régulières définies sur le token :

> TokensRegex is a framework for defining cascaded patterns over token sequences. It extends the traditional regular expression language defined over strings to allow working with tokens. In other words, it generalizes from matching over sequences of characters (strings) to matching over sequences of tokens. Furthermore, it uses a multi-stage extraction pipeline that can match multiple regular expressions—a practically more useful scenario than single pattern matching. (Manning *et al.* 2014, 1)

Les patrons sont testés un par un, d'abord sur l'interface en ligne que propose CoreNLP. Ils sont ensuite lancés automatiquement sur les corpus via Eclipse, un environnement de développement permettant de développer et d'exécuter des programmes Java[73].

**InterText** est un outil qui permet d'aligner et gérer des textes parallèles (original et traduction) au niveau de la phrase[74]. Aligner un corpus parallèle, c'est assembler les mots, les phrases ou les paragraphes du texte original avec leurs équivalents dans le texte traduit, ce qui permet de porter un regard contrastif sur un phénomène linguistique. L'utilisateur d'InterText peut apporter des modifications et des corrections en cas d'erreurs, ce qui, de plus, justifie le choix de cet outil.

[73] https://www.eclipse.org/downloads/packages/release/mars/r/eclipse-ide-java-developers (accès vérifié le 06 juin 2019 à 14:07).
[74] https://wanthalf.saga.cz/intertext (accès vérifié le 20 juillet 2018 à 16:30).

## 3. Étiquetage morphosyntaxique et annotation des corpus

### 3.1. Étiquetage morphosyntaxique

L'étiquetage, comme nous l'avons déjà énoncé, est l'une des opérations fondamentales et incontournables à effectuer sur un corpus donné en vue d'une analyse complexe des phénomènes linguistiques. À l'aide du contexte, la tâche de l'étiquetage consiste à attribuer à chaque mot du corpus une étiquette, comme illustré ci-dessous :

> La phrase *the cat ate the mouse* étiquetée à l'aide du *Stanford tagger*
> the/DT cat/NN ate/VBD the/DT mouse/NN

Étiqueter morphosyntaxiquement un corpus, c'est ainsi donner à chacun des tokens de la séquence des informations nécessaires à l'analyse linguistique. L'un des intérêts de cet étiquetage est d'écarter les ambiguïtés de la langue notamment lorsqu'une forme peut correspondre à plusieurs catégories grammaticales, ce qui pourrait fausser les analyses automatiques des phénomènes (cas auxquels nous serons confrontée dans le prochain chapitre). La difficulté est alors d'entraîner l'étiqueteur à attribuer la catégorie spécifique de chaque mot, surtout quand ce dernier peut être catégorisé de différentes manières (exemple de *walk* qui peut-être un verbe ou un nom).

S'il n'existe pas d'étiqueteur pouvant agir en même temps sur les deux langues dans un corpus parallèle, il est néanmoins possible de procéder de manière indépendante à une analyse pour chaque langue. Par contre, les patrons ne seront pas semblables dans les deux langues, d'une part parce que les critères morphosyntaxiques ne sont pas les mêmes dans chacune d'elles, et d'autre part, parce que les jeux d'étiquettes utilisés sont différents[75].

Dans notre travail, l'étiquetage est entièrement automatique (Stanford CoreNLP, exemple donné dans la figure 7). Seule la version anglaise de notre corpus a été

[75] À l'exception du jeu d'étiquette universel (par exemple, *Universal Pos Tags* du projet *Universal Dependencies*)

étiquetée morphosyntaxiquement puisque les requêtes de détection sont uniquement définies pour le corpus anglais.

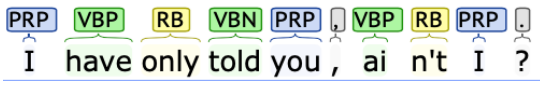



**Figure 7 : Étiquetage morphosyntaxique établi par les outils CoreNLP**

### 3.2. Annotation manuelle des ellipses

L'annotation peut être effectuée sur un corpus écrit, oral ou oral transcrit, et peut être manuelle (aucun outil n'est utilisé), automatique (entièrement faite par une machine) ou semi-automatique (intervention de l'humain). Deux méthodes sont utilisées dans l'annotation automatique : l'annotation à base de règles définies par un linguiste expert et l'annotation par apprentissage (à l'aide d'un corpus déjà annoté manuellement, et servant de référence pour entraîner un outil).

Les erreurs constatées dans le processus d'annotation relèvent pour la plupart du choix même du type d'annotation qui peut être entièrement ou partiellement automatique. Lorsque l'annotation est entièrement automatique, les erreurs sont (le plus souvent) engendrées par des outils parfois mal adaptés aux données, ou par des erreurs présentes dans le corpus (erreurs liées à la sauvegarde, saisie, ...). Lorsque l'annotation est manuelle ou semi-automatique, les erreurs sont dues à l'interprétation subjective des annotateurs ou à un manque de précision du manuel d'annotation. En effet, le caractère subjectif de l'annotation se manifeste dans les résultats obtenus lors des questionnements fondamentaux autour du quoi, du comment et du pourquoi de l'annotation. Afin de parer au mieux à cette subjectivité, Mélanie-Becquet & Landragin (2014, 117) suggèrent l'utilisation d'un « manuel d'annotation » :

> Pour que les annotations ne soient pas trop subjectives, un manuel d'annotation strict et directif s'avère nécessaire. Il faut cependant que le schéma d'annotation tienne compte des ambiguïtés et flous possibles, et autorise une certaine souplesse dans l'affectation des valeurs.

Il existe évidemment plusieurs moyens d'évaluer la fiabilité des annotations. Celui probablement le plus répandu consiste à vérifier le degré d'accord entre les annotateurs, c'est-à-dire, par exemple, à faire annoter un corpus par deux ou trois annotateurs différents pour observer ensuite ce qui leur est commun.

Par ailleurs, le type d'annotation effectuée sur un corpus dépend de plusieurs facteurs : du cadre théorique et des objectifs de l'annotateur lui-même, de la nature des données (ce que l'on pourrait annoter), et de la nature du phénomène analysé (syntaxique, sémantique, pragmatique). Toute la méthodologie établie à l'aide d'outils spécifiques dans le processus d'annotation s'inscrit d'abord dans un cadre théorique permettant d'initier des recherches appliquées. Ainsi à titre d'exemple, PRAAT[76] pour la transcription phonétique et ANALEC[77] pour les chaines de références sont-ils, parmi d'autres, des outils fréquemment utilisés pour annoter un corpus, tous deux issus d'investigations phonétiques, sémantiques et syntaxiques.

Dans le cas de l'ellipse, il est nécessaire de s'interroger sur l'intérêt du processus d'annotation dans la mesure où, par définition, une partie du discours est rendue invisible alors que les outils nécessaires à l'annotation sont conçus précisément pour annoter les catégories visibles. Pour dépasser la réflexion aporétique, l'annotation du contexte syntaxique, la plupart du temps chargé d'indices en lien avec la présence d'une ellipse, a été l'un des recours envisagé par les linguistes afin de contourner l'impossibilité initiale. En effet, une fois les informations ajoutées au corpus, il est possible de lancer des requêtes fines et précises à partir des outils choisis pour repérer ces indices, voire détecter l'ellipse.

Dans la présente recherche, nous avons annoté manuellement, en ajoutant son code à chaque type d'ellipse trouvée, 3 échantillons :

– le premier échantillon contient 1000 phrases extraites au hasard de chaque genre de discours. Chaque phrase extraite est à la fois précédée et suivie de 2 lignes pour avoir accès au contexte. L'objectif ici était d'analyser les

[76] http://www.fon.hum.uva.nl/praat/ (accès vérifié le 17 avril 2018 à 4:49).
[77] http://lattice.cnrs.fr/Telecharger-Analec (accès vérifié le 17 avril 2018 à 4:50).

différences entre les genres en relevant la distribution du phénomène. Le résultat de ce travail est présenté dans le chapitre 4.

– les deux autres échantillons de tailles différentes 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). Le premier échantillon compte 197 302 tokens et 1 270 occurrences d'ellipses. Le deuxième contient 36 676 tokens et 396 occurrences d'ellipses. L'objectif ici est d'évaluer la performance des patrons et de calculer le rappel. Les résultats de cette détection sont présentés dans le chapitre 3.

L'annotation manuelle a requis une lecture de l'intégralité des échantillons sélectionnés. Une vérification a ensuite été effectuée à l'aide d'expressions régulières simples[78] afin de détecter les occurrences éventuellement manquantes non annotées. En effet, l'ellipse reste un phénomène relativement peu fréquent et il est donc nécessaire de parcourir une grande quantité de texte pour obtenir un nombre satisfaisant d'occurrences.

## 4. Bilan : apports et contraintes de la méthodologie adoptée

Quelle que soit l'approche méthodologique adoptée, le chercheur est conduit à ajuster constamment l'ensemble des objectifs à atteindre et les outils dont il dispose pour mener à bien le déroulement de sa recherche. Pour ce qui nous concerne, l'ellipse, par l'absence *remarquable* de la forme la représentant, pose, comme on l'a vu, nombre de défis à relever, nous amenant à situer son analyse aux frontières de plusieurs choix et ajustements théoriques, méthodologiques et outillés. Nous ne reviendrons pas en détail sur les contraintes habituelles à laquelle toute recherche fondée sur corpus est soumise, à savoir : la collecte de données aléatoire qui peut remettre en question la pertinence de la recherche, la taille du corpus analysé qui peut parfois influencer la représentativité des résultats, et la quantité suffisante d'instances pour une analyse statistique des données, difficile à atteindre en raison de la rareté du phénomène observé.

[78] Voir annexe III p.267.

En pratique, selon les critères évoqués précédemment, le corpus est présenté comme un échantillon représentatif des textes sélectionnés indiquant par là l'une de ses limites. En effet, ce n'est pas parce qu'un échantillon du corpus représente son genre de textes, qu'il est représentatif de la langue entière. Une question supplémentaire peut être soulevée concernant la validité de l'échantillon : peut-on en tirer des conclusions généralisables ? En fait, sachant que la raison d'être du corpus est de permettre l'analyse d'un échantillon parmi d'autres, il se révèle obligatoirement incomplet et il rend illusoire l'idée d'envisager son extension à un travail mené sur l'ensemble de la langue. Pourtant, le chercheur tente en permanence d'atteindre cette représentativité en enrichissant le corpus avec davantage de données, de textes et de ressources. Ce travail, aussi bénéfique soit-il, n'élimine pas pour autant les écueils. On pourrait par exemple reprocher au corpus journalistique son instabilité apparente et de ce fait sa non représentativité, puisque le type de discours qu'il représente est soumis à une évolution rapide due à l'utilisation fréquente de néologismes et d'expressions propres à ce registre. Dans notre cas, nous avons voulu, par ce rassemblement de genres, couvrir le plus grand nombre de variations et d'irrégularités qu'une ellipse peut manifester à travers la différence des genres. Ceci, évidemment, ne nous aide pas à éviter le problème de la représentativité mais nous amène néanmoins à relever quelques pistes relatives à la nature du phénomène qui peuvent être généralisables et déboucher sur d'autres perspectives. Dans cette ligne de pensée, la taille du corpus pèse aussi sur la représentativité des résultats dans la mesure où l'investigation d'un corpus de grande taille, notamment lors de l'annotation de l'ellipse (comme nous le verrons), prend beaucoup de temps. La représentativité de notre corpus est ainsi soumise au comment et au pourquoi de la constitution du corpus. Suivant Sinclair (1991, 13), nous pouvons dire « the results are only as good as the corpus is ».

Il nous reste à vérifier la pertinence du corpus des sous-titres pour analyser l'ellipse. Nous avons voulu au départ exploiter un corpus de transcriptions orales où l'ellipse apparait plus fréquente. Cette idée a été écartée compte tenu des difficultés que ce type de corpus présente. Il s'agit en effet de difficultés techniques qui se

situent notamment au niveau des problèmes de la collecte et de l'enregistrement des conversations spontanées, tâches à la fois compliquées et fastidieuses pour de nombreuses raisons[79]. Par ailleurs, la tâche de transcription des données s'avère également chronophage. S'ajoute à cela la difficulté de trouver un corpus de transcriptions de l'oral traduites. C'est précisément pour tenter de contourner ces difficultés que nous avons opté pour un corpus de sous-titres de séries télévisées qui, s'il ne présente pas la caractéristique de spontanéité immédiate, offre cependant les caractéristiques de l'oral propre à contenir des ellipses.

Ces conversations sont en effet caractérisées par des interactions courtes, des omissions grammaticales, des structures simples, dues nécessairement à l'espace et au nombre de signes imposés par le format audio-visuel. Il serait par exemple totalement aberrant de traduire de manière exhaustive un dialogue en couvrant totalement l'image de texte. Sont à prendre en compte dans ce corpus les exigences techniques particulières liées à la synchronisation image-texte et aux stratégies d'adaptation qui justifient parfois le recours à l'ellipse. Ces paramètres spatio-temporels constituent ainsi les premières nécessités à considérer lors de l'analyse du corpus et lors de l'interprétation des résultats. Par ailleurs, une fois ces éléments d'ordre technique mis à part, se pose la question *existentielle* de l'ellipse : a-t-on recours à l'ellipse parce que le dialogue l'exige et comme le requiert la langue source (cas de la non-traduction des *question tags* comme nous le verrons dans le chapitre 5), ou faut-il ellipser même ce qui ne l'est pas dans la langue source pour économiser de l'espace lors du sous-titrage ? La réponse à cette question dépend très clairement du type d'ellipse qu'on analyse, et de la situation dans laquelle il apparait, éléments que nous essayerons de comprendre plus finement dans les chapitres à venir.

[79] Bích (2011) a décrit le changement chez un même locuteur dans le cas d'une conversation spontanée. Il explique : « Par exemple, lorsqu'un locuteur parle d'un travail dans lequel il est fortement impliqué, il arrive qu'il passe alternativement par des phases de langage qu'on pourrait caractériser comme très "spontanées", surtout s'il s'adresse à un interlocuteur qu'il connaît bien, et, presque simultanément par d'autres phases qu'on pourrait dire au contraire "très soutenues", surtout lorsqu'il semble parler en tant que représentant de sa profession ».

Nous exposons maintenant les difficultés rencontrées lors de l'application de notre méthode, les que nous avons mises en œuvre ainsi que les éléments que nous n'avons pas résolus.

### 4.1. Traitement manuel du corpus en raison des limites des outils

Lors de l'exploitation du corpus, l'un des problèmes rencontrés tient à l'absence d'outils capables de traiter facilement un corpus parallèle et à celle d'outils capables de traiter le vide causé par l'ellipse au niveau monolingue. Un traitement manuel et une combinaison de plusieurs outils deviennent alors nécessaires.

#### 4.1.1. Un recours : l'annotation manuelle

Comme nous l'avons signalé dans les étapes méthodologiques de notre travail, nous avons eu recours à l'annotation manuelle du corpus pour identifier les catégories d'ellipse et analyser leur distribution. Lorsque le travail de recherche est placé au cœur du domaine du TAL, nombreuses sont les opérations qui requièrent une annotation préalable des données. Dans notre cas, il s'agit d'une annotation servant à créer une *référence* afin de mettre au point les patrons et d'évaluer leur performance. La qualité de cette référence est soumise aux critères d'identification des ellipses, puisqu'il n'y a pas eu de mesure d'accord inter-annotateur. L'annotation manuelle présente ainsi des limites qui peuvent remettre en cause la validité des résultats obtenus. À l'échelle d'une thèse de doctorat en sciences humaines, en effet, la vérification du taux d'accord entre plusieurs annotateurs n'est généralement pas possible, et de ce fait, la fiabilité de l'annotation dépend du seul auteur de la recherche. De plus, le risque d'oubli d'une ellipse ou d'une autre est presque inévitable lorsque le fichier annoté est d'une taille considérable. Cet oubli peut être dû à la rareté du phénomène (discours politique par exemple), ou à l'intensité de sa fréquence (corpus des sous-titres). Pour contourner cet écueil, trois options ont été choisies :

1- élargir l'étude approfondie des patrons à tous les corpus exploités ;

2- évaluer la performance des patrons sur des échantillons de taille raisonnable (corpus d'évaluation) ;
3- annoter manuellement 1 000 phrases de chaque corpus et mesurer la significativité de la différence entre les occurrences d'ellipses observées dans différents genres. Bien sûr, cela est indépendant de la performance des patrons, et se fait sur des annotations validées manuellement. Cela ne permet pas d'améliorer ou de faciliter l'annotation manuelle mais permet en revanche de valider ou d'invalider les hypothèses formulées dans le chapitre 4 sur la distribution des ellipses dans les genres ;
4- utiliser des expressions régulières simples et effectuer une vérification à la fin de l'annotation manuelle pour diminuer le risque d'oubli. Ces expressions représentent une structure simplifiée des patrons d'identification[80].

### 4.1.2. Alignement : pour une analyse de la traduction de l'ellipse

Même si cette étape ne présente pas d'intérêt pour notre détection automatique, nous avons choisi d'aligner les corpus pour mieux observer simultanément l'ellipse dans les deux langues et, ainsi, analyser sa traduction.

Lors de notre travail de Master, nous avions aligné manuellement notre corpus. Cependant, l'alignement manuel devient vite une activité chronophage quand il s'agit d'un corpus plus vaste. Pour notre corpus actuel, l'alignement automatique des phrases pour les corpus qui n'étaient pas déjà alignés (sous-titres) s'est fait à l'aide de l'outil InterText.

L'intérêt de cette étape d'alignement dans le cas précis de l'ellipse est notamment de pouvoir cerner et extraire les caractéristiques linguistiques propres à chaque langue et de repérer les éventuelles similitudes ou différences immédiatement visibles à l'œil nu. À titre d'exemple, l'extrait ci-dessous met en avant les différences de fonctionnement des auxiliaires entre le français et l'anglais, impliquant la non-traduction de la *question tag* en français.

[80] Il est important de noter que ce travail est effectué pour vérifier et non pour annoter. Une lecture totale du corpus a été faite au préalable pour annoter manuellement.

(27) <tuv xml:lang="en"><seg> I've only told you, ain't I Ø?
<tuv xml:lang="fr"><seg> J'ai déjà tout dit Ø !

En effet, il s'agit d'une ellipse déclenchée par *have* dans la *question tag* où le verbe *tell* en anglais est ellipsé. En français, *ain't I ?* n'est pas du tout traduit puisque le système de la langue n'utilise pas ce genre d'interlocution dans le discours. Pourtant, si le traducteur avait ajouté *n'est-ce pas* ? ou *non ?* éventuels équivalents dans la langue française, il aurait introduit une autre catégorie d'ellipse qui ne toucherait pas uniquement le verbe mais la totalité de la proposition[81].

Nous ajoutons aussi que cette étape d'alignement permet de visualiser les segments et leurs correspondances sans avoir à les rechercher manuellement. Elle représente alors un gain de temps non-négligeable au profit des étapes qui suivront cette phase préliminaire.

### 4.2. Limites et contraintes liées à la méthode appliquée au phénomène elliptique

Les contraintes que nous présentons ci-dessous sont apparues lorsque nous avons essayé d'adapter les différents outils présents dans CoreNLP à l'analyse de l'ellipse. Le problème qui se pose est toujours lié à la capacité de la machine à rendre compte du vide engendré par l'ellipse. Ce que nous devons alors réaliser est de repérer non le vide mais les indices laissés par ce vide.

#### 4.2.1. Difficulté à établir des patrons à base d'une analyse syntaxique en dépendances

Généralement considérée comme plus performante et précise que les autres analyses, l'analyse syntaxique en dépendances identifie les relations syntaxiques entre les mots de la phrase (sujet, objet direct ou indirect, etc.). Son objectif principal est de décrire le corpus en vue d'une exploitation linguistique dans le domaine du TAL. La question qui pourrait se poser alors concerne la raison pour laquelle nous n'avons pas développé nos patrons à partir de ce type d'analyses.

[81] Il ne s'agit pas dans ce chapitre d'entrer dans le détail de l'analyse des exemples donnés à titre d'aperçu au lecteur qui retrouvera une discussion approfondie dans le chapitre 5. Nous rappelons que l'objectif de ce chapitre est d'exposer la méthodologie de travail.

Pour rappel, nous avons constitué nos patrons à base de tokens combinant des expressions régulières et des étiquettes morphosyntaxiques. Cette méthode a ses limites que nous détaillerons au fur et à mesure de l'analyse des résultats. Cependant, du fait de ces limites, il nous a semblé pertinent au départ d'avoir recours à l'analyse syntaxique pour dégager les relations de dépendance entre les éléments, ce qui devait permettre de formuler des requêtes encore plus précises que celles établies à l'aide de tokens. L'idée initiale était donc d'établir nos patrons de détection à base de dépendances entre les différents éléments de la phrase. L'outil sélectionné pour réaliser cette tâche a été Semgrex (Chambers *et al.,* 2007), un utilitaire proposé en complément des analyseurs syntaxiques Stanford. En fait, l'intérêt d'une analyse en dépendances syntaxiques se manifeste dans la mise en relief des relations entre les segments. Compte tenu de la nature du phénomène elliptique qui peut aussi se définir par rapport à son voisinage, c'est-à-dire en lien avec les autres éléments de la phrase, le linguiste n'éprouve généralement pas de difficulté à analyser une séquence elliptique. L'analyseur syntaxique de CoreNLP, comme le montre l'exemple ci-dessous (figure 8), a parfaitement établi le lien, avec la conjonction *and,* entre le modal *will* et le verbe antécédent *help*. Pourtant, ceci ne semble pas être le cas dans toutes les ellipses que nous avons analysées.

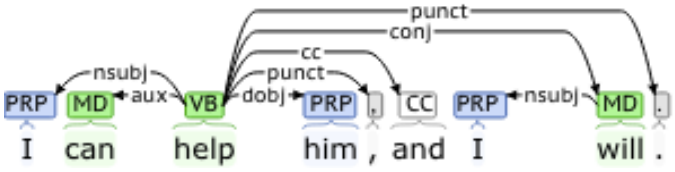



**Figure 8 : Analyse en dépendance de la phrase *I can help him, and I will.***

L'analyse en dépendances dans le cas de l'ellipse a en effet été écartée rapidement et ceci pour deux raisons. La première tient aux relations interrompues entre la catégorie visible de la phrase et le site elliptique. Lorsqu'un segment est omis, l'analyseur syntaxique commet des erreurs et n'exprime pas la relation sujet-verbe correcte, comme ici dans l'inversion entre le *you* et *were* (figure 9).

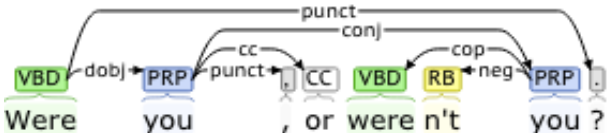



**Figure 9 : Analyse en dépendance de la phrase *Were you, or weren't you ?***

La deuxième s'explique par la difficulté à spécifier l'ordre des tokens dans une analyse en dépendance. L'exemple ci-dessous (figure 10) montre que l'analyseur syntaxique est malheureusement incapable, en l'absence d'un connecteur, d'établir des dépendances dans une même chaîne de conversation, entre un élément d'une phrase A et un élément d'une phrase B[82].

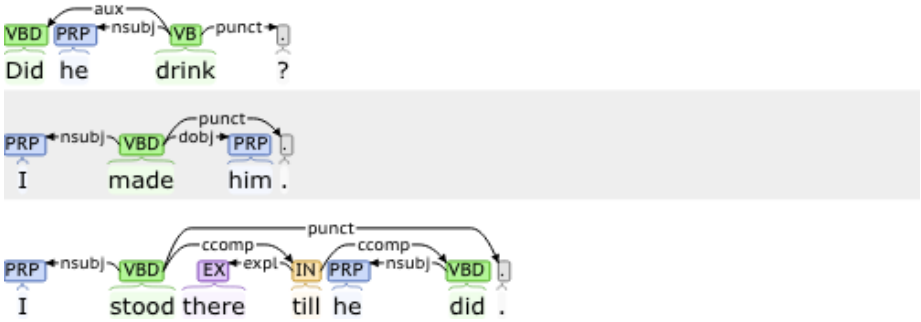



**Figure 10 : Analyse syntaxique du dialogue Did he drink ? I made him. I stood there till he did.**

En effet, c'est bien la difficulté à indiquer l'ordre des relations, qui diffère d'un type d'ellipse à un autre, et de les inclure toutes dans un patron qui nous a incitée à écarter l'analyse syntaxique en dépendances.

### 4.2.2. Ellipses écartées

Ayant rejeté l'analyse en dépendances syntaxiques de notre étude, il est important de souligner que la détection à base de tokens a elle aussi ses limites qui, lorsqu'elles sont additionnées à la difficulté d'aborder le phénomène analysé, rendent la détection de certaines catégories irréalisable. En effet, dans l'impossibilité

[82] Par ailleurs, il s'agit également ici d'une limite des patrons établis à base TokensRegex qui sont limités à la phrase.

de formaliser des règles très fines dont le phénomène a besoin, la détection des ellipses présentées ci-après n'a pas été envisagée.

#### 4.2.2.1. Le gapping

Il nous est très compliqué d'établir un patron qui prendrait en compte toutes les variations et les propriétés syntaxiques d'une occurrence de *gapping*. L'exemple (28) illustre un cas de *gapping* :

(28) <CEx> In the photograph, Aunt Sadie's face, always beautiful, **appears** strangely round, her hair **Ø** strangely fluffy, and her clothes **Ø** strangely dowdy.

Pour résumer, le *gapping* pourrait par exemple se rencontrer lorsque dans la structure syntaxique de la proposition, le sujet (*her hair* et *her clothes*) est précédé d'une conjonction de coordination et suivi d'un ou plusieurs tokens qui ne sont ni des auxiliaires ni des verbes. Ils sont toujours suivis d'un point.

L'établissement d'un patron pour détecter ce type d'ellipse est particulièrement complexe, compte tenu des variations du nombre 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 peut repérer. La figure (11) 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 Ø a bag*) est omis présentant de ce fait un *gapping.*

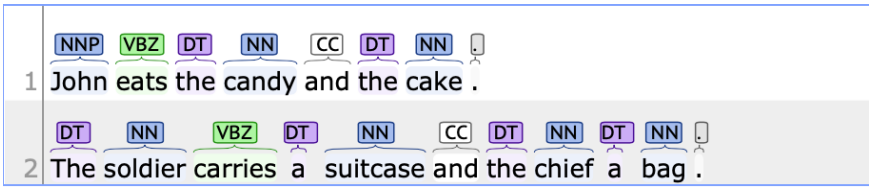



**Figure 11 : Étiquetage d'une phrase non-elliptique et du gapping**

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 :

(29) <CEx> Jane likes apples and Maria Ø too.

Cependant, ce type de cas n'est pas représentatif du *gapping* et les autres configurations, notamment les constructions parallèles déclenchant ce phénomène de *gapping*, ne pourront pas être détectées[83].

Par conséquent, la précision du patron reste très restreinte dans TokensRegex compte tenu du fait que les éléments résiduels n'ont pas toujours la même étiquette dans tous les exemples, qu'ils n'entretiennent pas toujours la même relation entre eux, et que leur nombre varie d'une séquence à une autre.

De ce fait, en raison des limites de l'outil d'une part, et de la difficulté à fixer des conditions stables du phénomène elliptique dans le *gapping* d'autre part, nous avons dû écarter sa détection automatique dans le cadre de cette recherche.

#### 4.2.2.2. Réponses fragmentaires

Nous avons détecté automatiquement les questions fragmentaires (où le sujet et l'auxiliaire manquent) mais nous avons écarté les réponses fragmentaires (- *Something to drink ? – Water, please*) identifiées dans la classification de van Craenenbroeck & Merchant en raison du taux d'erreurs observé, entravant ainsi les autres détections. En effet, si le patron autorise la détection de ces ellipses dans les

[83] Le *gapping* peut également se trouver dans une forme interrogative : *Did mom come first, or daddy ?* Plusieurs vides syntaxiques sous le principe du *gapping* peuvent également se suivre comme : *Jane loves to study rocks, and geography, too.* (Lobeck 1995: 27)

réponses (c'est-à-dire dans des phrases affirmatives), le taux d'erreurs est important car il est impossible de cerner dans un nombre limité de patrons toutes les variations que ces constructions engendrent. Cette catégorie rejoint donc le problème du *gapping* identifié ci-dessus[84].

Nous dirons cependant qu'en raison du développement permanent des outils informatiques, ces contraintes ne doivent pas être considérées comme définitives, mais se présentent comme des défis ultérieurs à relever en vue d'une meilleure reconnaissance automatique de l'ellipse.

### 4.3. Contraintes fondamentales

Le dernier obstacle que nous abordons dans cette partie et auquel un travail sur corpus parallèle nous expose est celui de l'acceptabilité des exemples constitués. Dans le cadre de notre recherche la notion d'*acceptabilité* est particulièrement importante tout d'abord en amont, lorsqu'il s'agit de constituer des corpus, et, en aval, dans l'évaluation de la traduction des occurrences elliptiques à l'intérieur des corpus mêmes. Nous nous limiterons dans ce chapitre à exposer les limites de l'acceptabilité concernant la constitution de corpus, et développerons celles liées à la qualité de la traduction dans le chapitre 5, consacré aux traductions de l'ellipse.

La notion d'acceptabilité appliquée aux extraits constitués dans un corpus est souvent opposée à celle de norme, grammaticale ou contextuelle. Qu'est-ce qu'un exemple, une phrase ou un énoncé acceptable ? Depuis la réflexion menée sur cette notion par Harris, l'acceptabilité est liée au jugement qu'un locuteur d'une langue ou d'une société donnée, prononce à l'égard d'un énoncé pour le valider. En effet, il ne suffit pas qu'une suite de segments soit grammaticalement et syntaxiquement agencés selon un ordre canonique ou une règle établie pour qu'elle ait un sens et soit acceptable. Il serait nécessaire de vérifier, tester et valider l'acceptabilité de ce qui est véhiculé auprès d'une population pour la *mesurer*. En prenant en compte plusieurs facteurs qui entrent en jeu chez les locuteurs d'une même langue (milieu

[84] Une explication plus détaillée est donnée dans le chapitre 3.

social, âge, espace géographique, profession, la liste n'étant pas exhaustive), l'acceptabilité apparait donc très relative.

Or, ces remarques ne sont pas toutes à prendre en considération dans le cas précis de la constitution de nos corpus de développement et d'évaluation. En effet, le corpus de développement a été constitué à partir d'exemples issus de documents authentiques, sans aucun souci de conformité à une norme quelconque. De cette manière, notre démarche « est donc basée sur l'usage et non sur les règles de grammaire dont on vérifierait la bonne/mauvaise application » (Loock 2016, 23). Nous avons conscience du fait que certains exemples puissent dérouter par leur non-acceptabilité sémantique, mais l'objectif premier poursuivi dans cette phase a été de repérer et de vérifier manuellement, toujours pour ce qui est du corpus de développement, l'apparition du phénomène elliptique. Notre corpus d'évaluation, servant à tester l'efficacité des patrons dans leur détection automatique des ellipses, répond au même positionnement.

Pourtant, nous n'ignorons pas que, hors des exemples forgés, se pose effectivement la question de l'*acceptabilité* des énoncés extraits d'un corpus, car une fois extraits de leur contexte (qui, dans le cas d'un corpus électronique, peut être relativement limité), ces extraits devenus des exemples pourraient, eux aussi, si l'on n'y prend pas garde, *fonctionner* comme des exemples fabriqués. Autrement dit, il ne suffit pas d'avoir un corpus pour que les exemples soient *sûrs* ou *imparables*, mais il importe que ces exemples soient repérés comme appartenant à des variétés de langue, des registres et des contextes particuliers.

Pour cette raison, tous nos exemples, appartenant à des registres identifiés à l'intérieur du corpus, sont, par nature, *acceptables,* dans le sens où ils correspondent à un usage de la langue tel qu'il a été porté par ces registres.

Pour aller plus loin dans la prise en compte de registres dans l'analyse de l'ellipse, nous citerons en particulier les travaux de Miller (présentés dans le premier chapitre) qui porte une attention particulière aux variations induites par le registre et fonde son analyse sur une sélection *aléatoire* d'occurrences dans le COCA. En adoptant cette démarche, Miller (2011) a pu remarquer que le type de registres, ce que nous

avons nommé *genres* (oral, fictionnel, journalistique et académique), influence la fréquence des ellipses post-auxiliaires avec *do* et *do so*.

Bien entendu, le fait de considérer le corpus *en l'état*, n'induit nullement un désintérêt pour ses caractéristiques de quelque ordre qu'elles soient, mais correspond à une préoccupation majeure d'étudier un fait de langue (ici l'ellipse) au plus près de sa réalité.

En résumé, comme nous l'avons déjà précisé, notre méthodologie de travail a été établie grâce aux études et aux conclusions menées sur l'ellipse par les linguistes théoriciens et informaticiens qui s'intéressent particulièrement à la place du phénomène elliptique dans la structure grammaticale de la phrase et à son traitement informatisé. Cependant, on se souvient qu'à ce jour, très peu d'études globales et approfondies envisageant une détection automatique de *tous* les types d'ellipse à l'intérieur d'un corpus ont été réalisées. De plus, jusqu'à présent, cette détection, quand elle a été réalisée, ne l'a été que partiellement. Par exemple, chaque étude s'est focalisée sur *un type* d'ellipse, mettant de côté les autres catégories. Par ailleurs, ces études ont considéré la détection comme une finalité en soi et se sont limitées à la description du phénomène. Or, notre recherche conçoit la détection non plus comme une finalité immédiate, mais comme un préalable conditionnant la réalisation d'un autre objectif, à savoir la résolution des problèmes posés par l'ellipse soumise à sa traduction automatique. De ce fait, elle tente d'ouvrir une voie nouvelle applicable aux langues sous étude, dans le traitement automatique et global de l'ellipse. En effet, si l'on parvient à détecter l'ellipse automatiquement, les problèmes de sa traduction automatique pourront être plus facilement répertoriés.

Nous avons montré dans ce chapitre que la complexité de l'ellipse se révèle à chacune des étapes menant à l'élaboration de patrons de détection automatique, allant du recensement des critères nécessaires à leur réalisation, jusqu'à la prise en compte des difficultés rencontrées dans l'exploitation du corpus, prouvant ainsi, si cela était encore nécessaire, la grande plasticité du phénomène étudié, voire de l'activité langagière.

Un établissement de patrons de détection fondé sur les relations de dépendance entre les constituants de la phrase, nous a semblé pertinent au départ, mais les problèmes rencontrés lors de l'analyse de dépendance entre le site elliptique et les autres éléments de la phrase nous ont amenée à écarter cette option d'analyse pour établir nos patrons sur la base d'une analyse morphosyntaxique et des tokens.

Nous examinerons dans les chapitres à venir les résultats obtenus après avoir suivi la méthodologie précédemment exposée, appliquée à la détection automatique des ellipses. Le chapitre 3 présente ainsi un bilan de cette démarche de détection concernant le genre conversationnel tandis que le chapitre 4 est consacré à la distribution des ellipses et à leur détection dans l'ensemble des genres du corpus d'évaluation. En nous focalisant alors spécifiquement sur l'ellipse post-auxiliaire, à partir des erreurs relevées dans cette phase de détection automatique, nous consacrerons le chapitre 5 à l'exposition des obstacles rencontrés dans la traduction (humaine et automatique) mais aussi aux nombreuses perspectives qu'elle ouvre.

# Chapitre 3

# Évaluation de la détection automatique sur un corpus de sous-titres

*Ellipsis creates challenging scientific and engineering problems. Although research over the past 50 years has shown that the principles permitting ellipsis involve many different types of information (grammatical structure, context, real-world knowledge), the precise mix of these principles and their interaction is still an open question*[85].

Comme nous l'avons annoncé, l'objectif de ce chapitre est de présenter les résultats obtenus suite à l'application de notre démarche à la reconnaissance automatique des ellipses pour les confronter ensuite à nos hypothèses de recherche. Il s'agira tout d'abord d'exposer les conditions morphosyntaxiques d'apparition des ellipses recensées à partir du corpus de développement, puis de décrire les patrons utilisés pour aboutir à la reconnaissance de chaque catégorie du phénomène étudié. La performance de ces patrons sera ensuite évaluée, conduisant au relevé d'une typologie des erreurs repérées dans les deux corpus de développement et d'évaluation.

Avant d'aller plus loin, au risque de nous répéter, nous tenons à souligner que l'objectif principal de notre approche est bien la détection de l'ellipse et non sa résolution. Par détection, nous entendons repérage. La résolution, quant à elle, renvoie à l'interprétation (dans le sens de reconstruction) des éléments effacés. Ce rappel est important puisque les deux opérations requièrent des démarches entièrement différentes. Prenons à titre d'exemple la reconnaissance de

[85] https://reports.news.ucsc.edu/linguistics/ (accès le 20 octobre 2017 à 20:33).

l'antécédent du site elliptique qui est nécessaire à l'interprétation du segment ellipsé, mais n'est pas forcément indispensable à la détection du phénomène, du moins pour les ellipses que nous envisageons dans ce travail. Une hiérarchie peut alors être établie entre deux types de critères : des critères morphosyntaxiques, généralement suffisants pour l'identification des ellipses, et des critères syntaxiques, sémantiques, pragmatiques et énonciatifs, nécessaires à leur interprétation.

## 1. Conditions morphosyntaxiques des ellipses et lecture des patrons utilisés

À partir des ellipses récurrentes que nous avions précédemment répertoriées manuellement dans le corpus de développement, nous avons élaboré des patrons à l'aide de l'outil TokensRegex inclus dans Stanford CoreNLP. Ces patrons, comme nous l'avons dit, sont établis en combinant une syntaxe à base d'expressions régulières à des informations obtenues par analyse automatique des corpus (étiquettes morphosyntaxiques, lemmes et entités nommées) afin d'effectuer des requêtes sur les corpus. Ces requêtes présentent les structures des occurrences elliptiques et sont lancées sur les corpus préalablement étiquetés.

Par souci de clarté, souvenons-nous, contrairement aux travaux présentés dans le premier chapitre qui se concentrent généralement sur un type particulier d'ellipse, nous nous proposons de détecter une grande variété d'ellipses. Ce choix a été fait afin de mieux appréhender ce phénomène et de pouvoir sélectionner l'ellipse posant le plus de problèmes aux outils informatiques, pour enfin analyser sa traduction automatique, à laquelle, malgré son évolution actuelle, elle semble résister.

Nous détaillerons ci-après, pour chaque déclencheur (éléments-indices de la présence elliptique), les conditions morphosyntaxiques qu'il est possible de formaliser. Nous présenterons en même temps la lecture de chaque patron, sachant que nous en utiliserons 20 (en tout) afin de mener la détection des différentes catégories d'ellipse et qu'un ou plusieurs d'entre eux ont parfois été établis pour identifier une même catégorie d'ellipse. Leur nombre par déclencheur varie, dépendant des conditions que l'on peut ou non combiner à l'intérieur d'un seul patron.

Rappelons tout d'abord que le processus elliptique est soumis à deux conditions : la nécessité d'une omission grammaticale et un contexte l'y autorisant. L'omission grammaticale, soumise aux règles d'une structure syntaxique non prononcée, répond le plus souvent au besoin d'économie et de fluidité du discours. Afin de cerner les caractéristiques du phénomène véhiculé par cette structure non prononcée, il est impératif de repérer, lors des analyses, les différentes configurations morphosyntaxiques dans les phrases elliptiques. C'est alors ce repérage qui permet d'identifier les éléments déclencheurs signalant le phénomène elliptique, éléments que nous proposons à présent de passer en revue à travers les principales ellipses relevées (sans prétendre à l'exhaustivité) et que nous présentons ci-dessous.

### 1.1. Ellipse {post-do}

*Do* est déclencheur d'ellipse verbale comme illustré ci-dessous :

(30) <CDEV> - you're a brave boy. You didn't **cry out**. Neither **did** I Ø, when I was your age.

De fait, nous considérerons *do* comme déclencheur d'ellipse lorsqu'il est :

– verbe de suppléance : *do* est déclencheur d'ellipse quand il apparaît comme suppléant du verbe lexical ellipsé, *do* n'ayant pas de sens en lui-même dans ce cas de suppléance,

(31) <CDEV> You just missed the toast my dear. Oh yes, you **did** Ø.

– auxiliaire nécessaire à la formulation des négations dans les temps simples en anglais,

(32) <CDEV> Did you miss me? You bet I did Ø. - I'll bet you **didn't Ø.**

– auxiliaire utilisé lors d'une inversion contrainte par certains adverbes.

(33) <CDEV> He didn't allow him to speak to her. Neither **did** I Ø.

– auxiliaire inséré dans les déclaratives qui ne contiennent pas d'auxiliaire (*have* et *be*) ou de modal pour créer un effet d'emphase. En effet, l'emploi dit emphatique de *do* sert généralement à confirmer (en cas de polémique explicite ou non), contredire, nier ou interroger[86].

(34) <CDEV> You just missed the toast my dear. Oh yes, you **did** Ø.

Cet auxiliaire est de première importance dans la langue anglaise puisque son absence peut engendrer des malentendus dans un dialogue, notamment dans les réponses aux *question tags*[87] (par exemple, il serait ambigu de répondre avec seulement *yes* à la question *she doesn't smoke, does she?)*.

Ces quatre cas autorisent le vide syntaxique laissé par l'omission du verbe lexical, qui, s'il était restitué, n'affecterait pas la grammaticalité de la phrase, à condition que l'ensemble du syntagme verbal (base verbale et compléments) soit récupérable. 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. Pour résumer le contexte grammatical, *do* peut déclencher une ellipse dans une phrase interrogative comme c'est généralement le cas dans une *question tag*, dans une phrase affirmative ou exclamative (parfois impliquant une inversion sujet verbe du fait de la présence de certains adverbes comme *neither*) et dans une phrase négative où *do* est combiné dans une forme contractée avec un opérateur de négation.

La détection de ces ellipses a nécessité l'élaboration de deux patrons, le premier récupérant toute phrase contenant le lemme *do* précédé par un pronom personnel ou possessif ou un nom. Le lemme *do* peut être suivi de *not* marqueur de négation ou d'un pronom personnel, mais il ne doit pas être étiqueté comme participe présent

[86] Notons également que l'emphase peut être attribuée à *oh yes,* sans oublier l'emphase prosodique que pourrait révéler *do* à l'oral (Herment 2011).
[87] Voir le chapitre 5 pour un examen plus détaillé des *question tags*.

ou participe passé. L'apparition du lemme *no* et d'une marque de ponctuation est optionnelle avant le pronom ou le nom.

Le deuxième patron, quant à lui, a été établi pour éviter la détection inutile des interrogatives non-elliptiques, comme *what do you do* ? Ainsi, en plus des conditions énumérées dans le premier patron pour les tokens suivant *do*, le deuxième devra récupérer toute phrase contenant le lemme *do* non précédé d'un pronom *wh*– avant le pronom ou le nom.

| | | |
|---|---|---|
| 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}]? /[.;!:?,]+/) |

### 1.2. Ellipse post-modale[88] `{post-mod}` et ellipse `{post-be/have}`

La simple observation permet de constater que lorsqu'ils déclenchent une ellipse, les modaux et les auxiliaires partagent plus au moins les mêmes conditions morphosyntaxiques. L'omission du verbe lexical dans les constructions parallèles, dans les constructions coordonnées ou les réponses courtes, et les *question tags*, entre autres, constitue une ellipse. Comme on le voit dans l'exemple (35) ci-dessous, la récupérabilité du complément ellipsé, c'est-à-dire du groupe verbal *get a bit of air*, n'a aucune incidence sur la structure syntaxique.

(35) <CDEV> You go and get a bit of air. - Yes, I **will** Ø.

De la même manière, les conditions syntaxiques des modaux comme déclencheurs de l'ellipse rejoignent celles de *do.* Ainsi l'ellipse post-modale apparaît-elle dans une phrase interrogative (ceci est généralement le cas des *question tags* : *should I?*), dans

[88] Nous ne reviendrons pas sur la distinction ellipse / anaphore pour ce qui est des modaux et des auxiliaires, puisque notre distinction est fondée sur le vide syntaxique qui pourrait ou non autoriser syntaxiquement l'ajout du verbe lexical à droite du modal ou de l'auxiliaire.

une phrase affirmative ou exclamative (réponses courtes *Yes, I **will** Ø*.), lors d'une inversion des positions sujet et verbe (*Neither will you*, entre autres exemples), et dans une phrase négative où le modal est suivi d'un marqueur de négation (*He won't.*)

Deux patrons sont utilisés pour détecter ces ellipses :

Le premier récupère tout modal précédé d'un pronom personnel, d'un nom, d'un pronom *wh–* ou d'une conjonction de coordination ou d'une virgule, sans qu'il soit suivi d'une forme verbale. Un adverbe peut optionnellement précéder le modal directement. Tout autre élément est autorisé avant la ponctuation finale de la phrase. Ce patron est notamment dédié aux propositions coordonnées (cas de *pseudogapping*, par exemple). Un deuxième patron est construit pour repérer les autres constructions et des éléments optionnels sont autorisés, tels qu'un adverbe pouvant précéder le modal ou une conjonction de coordination ou une conjonction de subordination, *if* ou *unless* pouvant le suivre.

| | | |
|---|---|---|
| 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)/}] []* /[.,:!;?-]+/` |

Quant aux auxiliaires *have* et *be*, et selon la classification syntaxique de van Craenenbroeck & Merchant, ils peuvent être soit déclencheurs d'ellipse de compléments (attributs, exemple 36) (dans la terminologie des auteurs cités ci-dessus *predicate phrase ellipsis*), soit d'une VPE (37), soit d'un *pseudogapping* (38).

(36) <CDEV> Because it's not a boarding house. It never **was** Ø.
(37) <CDEV> You haven't finished the room Petey. – I **have** Ø.
(38) <CDEV> Mary hasn't bought a jacket, but she **has** Ø a dress.

Techniquement, l'avantage de notre classification en termes d'économie de moyens et de temps, réside dans le fait qu'un même patron peut servir à repérer les

trois catégories ci-dessus, en prenant en compte l'élément déclencheur et la place de la ponctuation dans la phrase. En effet, le *pseudogapping* apparaissant le plus souvent dans les constructions comparatives et coordonnées dans lesquelles la forme non finie du verbe (cas des modaux et de *do*), ou le participe (cas de *have* et *be*), sont omis, laisse un résidu après l'auxiliaire ou le modal. C'est d'ailleurs ce *résidu* (*remnant* en anglais) qui le distingue de la VPE[89].

(39) <CEx> Mother is **working** tomorrow, but father isn't Ø. (VPE)
(40) <CDEV> Mother is **working** tomorrow, but she isn't Ø next week. (*Pseudogapping*)
(41) <CEx> He **eats apples** more than she **does** Ø. (VPE)
(42) <CEx> He **eats** apples more than she **does** Ø oranges. (*Pseudogapping*)

En fait, dans la VPE, l'auxiliaire ou le modal clôt la proposition du site elliptique, se passant du reste des éléments. En revanche, dans le *pseudogapping*, l'auxiliaire ou le modal est suivi d'une partie de la proposition (Miller, 2014). Cependant, le résidu de la proposition elliptique est soumis à la contrainte du parallélisme, dans la mesure où il forme un contraste avec le dernier élément de la proposition qui précède (*apples* est mis en opposition à *oranges* dans l'exemple 42). Si ce parallélisme n'est pas respecté, le *pseudogapping* n'est pas autorisé (43).

(43)*He eats apples more than she does **apples.**

Ainsi, un patron qui sera établi pour identifier automatiquement un déclencheur, auxiliaire ou modal, autorisera n'importe quel élément qui le suit à l'exception des formes conjuguées du verbe[90].

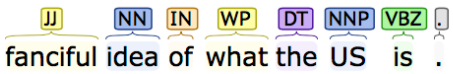
À ce stade, nous soulignons que nous aurions pu établir des patrons qui détecteraient à la fois des ellipses post-modales et post-have/be compte tenu du fait que ces ellipses partagent plus au moins les mêmes conditions morphosyntaxiques.

[89] La proposition contenant le site elliptique du *pseudogapping* est, le plus souvent, introduite par *but, as if, than* (lorsqu'il y a une comparaison). Le *pseudogapping* est envisagé dans la plupart des théories comme un type de VPE.
[90] Selon Levin (1986, vii), le *pseudogapping* peut également apparaître dans les questions réponses (formulées par le même locuteur) : *Does that make you mad ? It would me.* La détection de ces ellipses est également effectuée à l'aide du patron de la VPE déclenchée par un modal.

Nous avons néanmoins été freinée par une contrainte technique due aux étiquettes différentes caractérisant chacun des deux déclencheurs, ce qui a évidemment influencé la construction des restrictions. En effet, seul le modal détient une étiquette MD le représentant tandis que *have* et *be* sont étiquetés comme des verbes lexicaux, source d'erreurs dans la détection que nous analyserons plus en détail dans la section 3 du présent chapitre.

Nous avons alors constitué deux patrons en ayant recours aux lemmes pour repérer les ellipses déclenchées par *be* et *have*, le premier étant apte à reconnaitre toute ligne contenant le lemme *be* ou *have*, précédé par un nom ou un pronom, situés directement en début de phrase, et, optionnellement, un adverbe. Le lemme peut également être suivi de *not* avant le signe de ponctuation. Le deuxième patron contient davantage de restrictions : il ne détectera pas, par exemple, les lemmes *be* et *have* précédés par un pronom *wh–.* Cette dernière restriction permet d'éviter la détection des constructions non elliptiques contenant des mots en *wh–,* étiquetés comme pronoms *wh–* (dans la proposition qu'ils introduisent) et créent par là des faux positifs[91]. Ces propositions, comme le montrent l'exemple (44), suivent la structure grammaticale : pronom *wh– + sujet + be.*

(44)




Il est important de signaler que l'étiqueteur Stanford utilise trois étiquettes pour les mots *wh–* (disponibles dans la liste des abréviations) :

(WDT) pour un *déterminant wh–* (*which)*
(WP) pour un pronom *wh–* (*who, what*)
(WRB) pour un adverbe *wh–* (*when, where*)
(WP*) pour un pronom possessif.

Avec un tel étiquetage, aucune différence est établie entre *what* pronom relatif sans antécédent et *what* pronom interrogatif introduisant une interrogative

[91] En statistique :
– Faux positif : occurrences détectées comme elliptiques alors qu'elles ne le sont pas ;
– Faux négatif : occurrences elliptiques qui ne sont pas détectées alors qu'elles sont des ellipses ;
– Vrai négatif : occurrences qui ne sont pas détectées et qui ne sont effectivement pas elliptiques.

indirecte. Compte tenu de la difficulté à attribuer automatiquement[92] une étiquette à ces pronoms, la grammaire a simplifié cet étiquetage en faveur d'une identification en mot *wh– (wh— word)*, ici avec l'étiquette WP, ce que de nombreux outils d'étiquetage ont appliqué. L'une des conséquences de cette distinction est ainsi apparente dans l'établissement des patrons, puisqu'aucune précision ne semble être possible en lien avec cette catégorie de pronoms *wh–* (relatif ou interrogatif). De ce fait, nous avons écarté toutes les étiquettes des pronoms *Wh–* du patron.

| | | |
|---|---|---|
| Post-be/have | 2 | `^[{pos:/PRP.?\|NN.?/}] [{pos:RB}]? [{lemma:/be\|have/}] [{lemma:not}]? /[.;!?:,-]+/` |
| | | `[!{pos:/V.*/} & !{pos:/WP.*/}] [{pos:/PRP.?\|NN.?/}] [{pos:RB}]? [{lemma:/be\|have/}] [{lemma:not}]? /[.;!?:,-]+/` |

### 1.3. Ellipses {post-to}

Nous suivons Miller & Pullum (2013) dans leur positionnement à l'égard du déclencheur *to* et de leur classification de ce type d'ellipse dans la catégorie de l'ellipse post-auxiliaire :

> We assume (with e.g. Gazdar et al. 1985) that infinitival *to* is a defective non-finite auxiliary verb. This analysis is not endorsed in Huddleston et al. (2002), but is robustly and convincingly defended by Levine (2012). (Miller & Pullum 2013, 2)

Suivant notre classification, le marqueur de l'infinitif *to* est déclencheur de l'ellipse lorsqu'il est utilisé sans base verbale en clôturant généralement la phrase. Il est généralement précédé des verbes exprimant une volonté et/ou une cause comme *promise, threaten, cause, manage, fail, refuse, hesitate, fail,* etc.

[92] Afin d'affiner l'étiquetage, on pourrait par exemple, apprendre à l'étiqueteur que les interrogatives indirectes sont obligatoirement introduites par un certain type de verbes.

Dans notre corpus de développement, les ellipses déclenchées par le marqueur *to* apparaissent avec *have* ou avec des verbes exprimant l'action, l'assentiment ou le désaccord, le besoin ou le manque. Comme le montre l'exemple ci-dessous, le segment verbal composé de la base verbale et de ses compléments dans la première phrase, est ellipsé après *to*, dans la réponse :

(45) <CDEV> Keep yourself agreable to Mr MacClure. - I won't **fail** to Ø.

Pour repérer ces ellipses, la requête ne devra donc autoriser aucune forme verbale après *to*. Les deux patrons nécessaires à leur repérage sont alors constitués de telle sorte que le premier puisse reconnaitre *to* précédé de toute forme verbale, et éventuellement d'un pronom personnel et de *not*, tandis que le deuxième détecte *to* lorsqu'il est précédé d'un adverbe et/ou d'un adjectif (comme dans *Are you coming to see me ? I'm eager to.*).

| | | |
|---|---|---|
| Post-to | 2 | `[{pos:/VB.*/ }] [{pos:PRP}]? [{lemma:not}]? [{pos:to}] /[.?,:!;]+/` |
| | | `[{pos:/VB.*/ }] [{pos:RB}]? [{pos:JJ}]? [{lemma:not}]? [{pos:to}] /[.?,:!;]+/` |

### 1.4. Inversion sujet-verbe `{vs-tag}`

L'élément déclencheur de ce type d'effacement est soit une *question tag*[93], soit un adverbe. Ainsi le site elliptique peut-il contenir l'un des déclencheurs présentés précédemment (modal ou auxiliaire). Dans l'exemple (46), la *question tag* est déclenchée par un modal :

(46) <CDEV> You couldn't have hallucinations twice, **could you Ø**?

Toutefois, établir une catégorie distincte pour cet effacement pourrait sembler non pertinent dans la mesure où le déclencheur peut être l'un de ceux déjà identifiés précédemment, rendant donc inutile une nouvelle catégorisation. Comme nous le

[93] Nous reviendrons plus en détail sur le fonctionnement de la *question tag* dans le chapitre 5.

verrons, il s'agit cependant d'une catégorie à laquelle nous avons eu recours pour affiner les requêtes et diminuer le taux d'erreurs. En effet, le patron établi par exemple pour identifier les ellipses post-modales dans des séquences affirmatives ne prenant pas en compte celles des *question tags*, implique la nécessité de réaliser un patron spécifique à ce dernier cas qui tiendra davantage compte des contraintes morphosyntaxiques comme notamment le type de ponctuation et l'ordre des constituants. Grâce à cet affinement, le taux d'erreurs est diminué (ce qui n'aurait pas été le cas si les deux ellipses évoquées avaient été traitées par le même patron). De cette manière, cette *catégorie* n'en est pas une à part entière, mais elle nous est utile et consacrée aux ellipses post-modales et auxiliaires qui apparaissent dans des phrases interrogatives ou affirmatives (lorsque le sujet et les verbes sont inversés), contrainte encore une fois non prise en compte dans les autres catégories que nous avons établies. Elle constitue donc une réelle simplification pour envisager la détection d'un nombre non négligeable d'ellipses.

Pour ces cas particuliers, et par souci de précision, nous avons mis en place trois patrons ayant des restrictions différentes. Le premier détecte les cas où les modaux*, do, have et be* sont suivis d'un pronom personnel lui-même suivi d'un point (cas de sujet inversé) ou d'un point d'interrogation (*question tag*), d'exclamation ou d'un double-point ; *not* peut se trouver optionnellement avant le pronom personnel. Le deuxième patron est consacré aux ellipses en début de phrase tandis que le troisième autorise la virgule avant les déclencheurs. Ce dernier patron peut paraitre suffisant pour détecter une *question tag* et l'ellipse déclenchée par l'inversion du sujet et du verbe. Mais on remarque que lorsque ces conditions sont réunies dans un seul patron, le nombre d'erreurs est très élevé. De ce fait, certaines occurrences n'ont pas été détectées et d'autres, qui l'ont été, ne présentaient pas d'ellipses. C'est la raison pour laquelle nous avons partagé les conditions sur plusieurs patrons.

| | | |
|---|---|---|
| Vs-tag | 3 | `[]* [{lemma:/(be\|do\|have)/} \| {pos:MD}] [{lemma:not}]? [{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.*/}]? /,/` |

### 1.5. Ellipse {post-*wh*}

Les pronoms *wh–* déclenchent un *sluicing* permettant l'économie de la proposition entière à l'exception du pronom lui-même.

(47) <CDEV> Nobody can ride him for a long time. I'm going to train him on the long halter. Billy Buck is going to show me **how Ø**.

Cette ellipse est très fréquente dans les dialogues et les conversations quotidiennes. Dans la plupart des cas, les locuteurs ne répètent pas la proposition concernée par la demande d'information que le mot *wh—* suffit à exprimer (ici dans l'exemple 48 cité, la raison de l'injonction).

(48) <CDEV> Hold your fire, immediately! – **Why** Ø?

Pour prendre en compte les variations syntaxiques engendrées par les ellipses post-wh dans un dialogue, il est nécessaire de détecter tous les mots *wh–* (*who, what, how, which, why, where, how many, how much, when)* suivis de constituants (prépositions comprises) dépourvus de verbes finis précédant une marque de ponctuation (point d'interrogation, point d'exclamation, point). Ces conditions permettent une détection de syntagmes comme *why not? what else? how many +N?*

Grâce à la catégorisation de l'ellipse post-wh intégrant ces paramètres, nous pouvons envisager la réalisation d'un patron susceptible de détecter aussi bien le *sluicing* que ses sous-catégories comme le *swiping* ci-dessous :

(49) <CDEV> It's got a groove worked all the way down one side, which fits a similar tongue on another stone, though nobody knows quite **what for**.

Pour ce faire, le patron devra être suffisamment précis pour ne pas englober la détection des questions *figées* (en anglais *frozen questions*[94]) comme dans (50) qui ne présente pas de cas de *sluicing*[95].

(50) <CEx> **Guess what?** It's time to go.

Nous avons donc constitué un seul patron pour repérer ces ellipses, capable de détecter tous les mots *wh–* (représentés à l'aide des lemmes et des étiquettes morphosyntaxiques pour plus de précision) suivis de toutes étiquettes à l'exception des verbes (toutes formes) et des modaux avant la marque de ponctuation ou une conjonction.

| | | |
|---|---|---|
| Post-wh | 1 | `[{pos:/^W.*/} & {lemma:/(wh(at\|y\|o\|en\|ere\|ich\|ose)\|how)/}] [!{pos:/V.*/} & !{pos:MD}]* [{lemma:/[,.!;:?-]+/} \| {pos:CC}]` |

[94] Susceptibles d'être très fréquentes dans le corpus du discours conversationnel.
Nous suggérons « questions figées » comme traduction de *frozen questions* à cause de l'ordre des mots. Nous sommes consciente que ces questions ne sont pas figées au sens où les lexicologues l'entendent.

[95] Ce travail de détection de *sluicing* s'inscrit dans la continuité des travaux que nous avons menés lors de notre séjour de recherche à l'université de Santa Cruz aux États-Unis.

### 1.6. Post-génitif {post-geni}[96]

En anglais, la possession est exprimée avec la préposition *of* (51), les pronoms possessifs (*his, hers, yours*, etc.) comme dans (52) et le morphème[97] *–'s* ou *–'* lorsque le nom est au pluriel (53).

(51) <CEx> It's only my word against the word of **Mr. Khelada**.
(52) <CEx> It's only my **word** against **yours**.
(53) <CDEV> It's only my word **against Mr. Khelada's.**

Dans le présent travail, seul le morphème *–'s* (que l'apostrophe le précède ou le suive, dans les cas où ce morphème est aussi la marque du pluriel) est considéré comme déclencheur de l'ellipse nominale. En effet, l'utilisation des pronoms possessifs est strictement anaphorique (*yours* renvoie à *word*) et n'engendre aucun vide syntaxique dans la structure de la phrase. Ainsi, la répétition explicite de *word* dans l'exemple (52) rendrait la phrase agrammaticale. En revanche, ce n'est pas le cas dans les exemples (54) et (55) où les noms *fault* et *car* peuvent apparaître dans la structure suivant le morphème *–s* :

[96] À ce stade, dans les catégories qui suivent, les éléments présentés sont déclencheurs des ellipses nominales. Suivant les conditions que Merchant & van Craenenboeck ont établies afin de repérer une ellipse nominale, nous avons choisi de la traiter lorsqu'elle est déclenchée par des quantifieurs, un génitif, et des nombres ordinaux et cardinaux. Considérées comme marginales dans ces deux langues avec peu d'études qui leur sont consacrées, les ellipses nominales révèlent pourtant des différences systémiques qui peuvent être repérées entre les langues, notamment lorsqu'elles ne sont pas apparentées (langues sémitiques et langues romanes par exemple). Alors qu'elles posent problème aux étiqueteurs morphosyntaxiques, certaines ellipses nominales n'engendrent pourtant pas beaucoup d'erreurs dans la traduction automatique de l'anglais vers le français. C'est ce paradoxe qui nous incite à maintenir leur détection, car il s'agit d'explorer, d'une façon générale, les erreurs produites en amont par ces étiqueteurs pour aboutir à des pistes évaluatives de la traduction automatique potentiellement applicables à d'autres langues.
Comme nous envisageons dans un travail futur une exploration de la traduction automatique anglais-arabe-français, cette première investigation limitée sera sans doute utile.

[97] Huddleston (1984, 268) remet en cause la dénomination du génitif et propose *PossP* pour *Possessive Phrase('s)* et *of phrase.*

(54) <CDEV> So it's all Papa's **fault**? It isn't my mother**'s** Ø or brother**'s** Ø.
(55) <CDEV> The car? It's actually Carl**'s** Ø.

Par conséquent, nous avons établi un patron qui récupérera tous les tokens étiquetés comme POS (possession), non précédés par un pronom personnel ou par le lemme *of*, suivis immédiatement d'un signe de ponctuation.

| | | |
|---|---|---|
| Post-geni | 1 | `[!{lemma:of}][!{pos:PRP}] [pos:POS] /[.;!?:,]/` |

### 1.7. Cas des cardinaux et ordinaux (`{post-card}, {post-ord}`)

L'ellipse nominale peut également être déclenchée par les nombres cardinaux comme dans l'exemple (56)[98].

(56) <CDEV> The police are struggling. I don't think this is gonna be done in a day or **two Ø**.

Nous remarquons dans l'exemple (57) ci-dessous l'ellipse nominale du nom *reason* déclenchée par les nombres ordinaux *first* et *second*. Sur le plan syntaxique, l'exemple peut sembler ambigu, à tel point qu'il est possible de se demander si ellipse il y a ou non, dans la mesure où sans l'affirmation de l'ellipse, les adjectifs ordinaux pourraient, si l'on n'y prenait garde, être considérés comme des substantifs ou des pronoms. Or, s'ils étaient pronoms, ils n'accepteraient pas la restitution du nom

[98] C'est lors de notre prise de conscience du phénomène elliptique que nous avons repéré cette catégorie d'ellipse, lorsqu'en naviguant sur les ambiguïtés dans la programmation, nous avions enregistré une plaisanterie qui circulait sur un forum de programmeurs.
My mom said:
"Honey, please go to the market and buy one bottle of milk. If they have eggs, bring six."
I came back with six bottles of milk.
She said: "Why the hell did you buy six bottles of milk?"
I said: "Because they had eggs!"
http://forums.imore.com/off-topic-lounge/281999-programmer-jokes-just-another-off-topic-thread-imore.html (consulté le 25 juin 2014 à 15:10).

*reason* dans la structure de surface, puisque par nature le pronom se substitue au nom.

(57) <CEx> There are **two reasons** why he's being so nice to Ivy.
The **first** Ø is to make me angry and I dread to think about the **second** Ø [99].

Nous avons ainsi constitué deux patrons pour repérer les ellipses déclenchées par les nombres cardinaux et un seul pour celles déclenchées par les nombres ordinaux.

Le premier patron repèrera tous les tokens étiquetés comme nombre cardinal CD, non précédé par un nom, et suivi immédiatement d'une marque de ponctuation. Par ailleurs, le lemme du nombre cardinal ne devra pas être *one.* Cette dernière restriction a été formulée pour éviter de repérer les reprises anaphoriques comme dans *the right one*. En effet, les frontières entre le phénomène elliptique et le phénomène anaphorique s'étendent aux étiquetages morphosyntaxiques où l'outil utilisé n'est pas en mesure, par exemple, de faire la différence entre *one,* lorsqu'il est nombre cardinal, et *one,* lorsqu'il est reprise anaphorique. De ce fait*,* nous avons ajouté, dans un deuxième patron, les restrictions visant à écarter la détection du cardinal *one* lorsqu'il est précédé de noms, d'adjectifs ou de pronoms démonstratifs.

| | | |
|---|---|---|
| Post-card | 2 | `[!{pos:/N.*/}] [{pos:CD} & !{lemma:one}] /[.;!?:,]+/` |
| | | `[!{pos:/N.*/} & !{pos:JJ} & !{lemma:/th(at\|ese\|is\|ose)/}] [{pos:CD} & {lemma:one}] /[.;!?:,]+/` |

Un seul patron a été utilisé pour les ellipses déclenchées par les ordinaux. Il repérera toute entité nommée *Ordinal* qui est étiquetée comme adjectif et qui est

[99] Lorsque cette ellipse, qui parait simple et pour laquelle une traduction littérale aurait été suffisante, a été traduite dans Google Traduction, elle a engendré des problèmes dans l'accord du nombre ordinal puisque l'antécédent n'est pas facilement identifié.
« Il y a deux **raisons** pour lesquelles il est si gentil avec Ivy. **Le premier** est de me mettre en colère et je redoute de penser à **la seconde** » (17 juin 2018). Nous notons néanmoins l'accord correct de *la seconde*.

précédée par l'article *the*. Cette entité est immédiatement suivie d'une marque de ponctuation.

| | | |
|---|---|---|
| Post-ord | 1 | /the/ [{pos:/JJ.*/;ner:ORDINAL}] /[.;!?:,]+/ |

### 1.8. Cas du quantifieur {post-quant}

Un autre cas à aborder ici est l'ellipse nominale déclenchée par un quantifieur. On remarque que les structures syntaxiques des ellipses provoquées par *some* (58) sont identiques à celles déclenchées par les nombres cardinaux et ordinaux en anglais.

(58) <CDEV> Will she have the **option**? - Thank you, but I already have **some** Ø.

Comme l'ellipse nominale provoquée par les nombres (cardinaux et ordinaux) et les quantifieurs (*some, any*[100]) ne semble pas présenter de variations morphosyntaxiques en anglais, les mêmes conditions peuvent être utilisées pour la détecter, à savoir, la phrase devra contenir une marque de ponctuation (selon la forme de la phrase, affirmative, interrogative ou négative), précédée par un quantifieur (*some* ou *any*).

| | | |
|---|---|---|
| Post-quant | 1 | /[sS]ome\|[aA]ny/ /[.;!?:,]+/ |

Il convient de signaler que la ponctuation joue un rôle crucial dans toutes les conditions définies pour établir les patrons de détection des ellipses, qu'elles soient anaphoriques (lorsque l'antécédent précède le site elliptique) ou cataphoriques (où le site elliptique précède son antécédent, identifiées par Halliday & Hasan (1976)).

[100] Dans la présente recherche, nous n'avons élaboré de patrons que pour *some* et *any.*

Ces ellipses cataphoriques peuvent être nominales ou verbales et sont suivies de la proposition contenant l'antécédent :

(59) <CDEV> Even though my mother cooked the first two Ø, I have to do the last **dessert**. (Ellipse nominale)
(60) <CEx> I **didn't** Ø, but he **left**. (Ellipse Verbale {post-do})

Or, les ellipses verbales cataphoriques inexistantes dans notre corpus de développement se retrouvent rarement en anglais et en français pour deux raisons principales. La première tient au fait qu'il s'agit majoritairement d'une spécificité des ellipses nominales (l'exemple 60 de l'ellipse verbale est rare). En effet, toutes les ellipses n'autorisent pas un emploi cataphorique, notamment le *pseudogapping* et le *gapping* où le site elliptique doit immédiatement suivre l'antécédent. La deuxième, et toujours dans le cas du *pseudogapping* et du *gapping*, tient, elle, au fait qu'un emploi cataphorique provoque un déséquilibre dans le parallélisme des propositions, critère auquel le *gapping* est soumis. Ces deux raisons sont illustrées dans l'exemple (61) où le site elliptique précède l'antécédent *will make a statement blasting* et crée par là un déséquilibre puisque l'effacement n'est pas autorisé.

(61) *If George Ø the newspaper reporters, Al **will make a statement blasting** the press. (Kehler 2002, 91)
[Ø = makes a statement blasting] (notre soulignement et notre interprétation du site elliptique)

Pour cette raison, l'un des paramètres de variation à prendre en compte, lors de l'interprétation des patrons dans le langage de TokensRegex semblerait être le repérage de toute marque de ponctuation (point, virgule et point d'interrogation, entre autres) suivant le site elliptique. On peut alors penser que ce repérage permet la réduction du taux d'erreurs. Or, s'il offre la possibilité de repérer un nombre important d'occurrences susceptibles d'aider à l'amélioration de patrons, il ne permet pas de détecter *précisément*, dans l'immédiat, le phénomène elliptique car toute marque de ponctuation n'est pas forcément indice de la présence d'une ellipse. Dans les patrons, ce paramètre reste néanmoins pertinent pour obtenir un taux de rappel aussi élevé que possible.

Par ailleurs, il est utile de rappeler que considérer les variations développées précédemment confère aux patrons de détection une précision apte à réduire le nombre d'occurrences détectées incorrectement comme ellipses. Ensuite, comme nous l'avons précédemment expliqué, nous établissons nos patrons de détection sans retenir intégralement les classifications établies, en particulier celles des sous-catégories d'ellipses classées dans la rubrique de l'ellipse du syntagme verbal, de l'ellipse propositionnelle et de l'ellipse nominale. En d'autres termes, un même patron, par exemple, sera utilisé pour détecter la VPE et la PAE, deux catégories dissociées qui peuvent toutefois partager un même élément déclencheur (cas de *do* entre autres).

### 1.9. Ellipse dans les questions fragmentaires `{qs-frag}`

En dépit de l'absence d'un élément déclencheur de cette catégorie, nous avons maintenu sa détection parce qu'il est possible d'en décrire les conditions morphosyntaxiques dans un patron. En effet, en considérant les indices que ce type d'ellipse laisse dans la phrase, il est généralement observé que le repérage automatique de l'auxiliaire et du sujet manquants peut être envisagé.

Nous avons remarqué que dans certains types et registres de discours, notamment en anglais, tels les dialogues informels et spontanés, lorsque le sujet est ellipsé, son auxiliaire l'est également, rendant parfois la compréhension problématique sans accès au contexte linguistique et/ou extralinguistique. L'exemple (62) ci-dessous illustre justement une question fragmentaire dans laquelle le sujet et son auxiliaire sont omis, laissant visibles seulement le participe présent de la forme aspectuelle *be + ing*, *expecting.*

(62) <CDEV> Expecting somebody?

La particularité de ce type de constructions elliptiques tient au fait que leur antécédent est rarement explicite. L'élément ellipsé est cependant sous-jacent à la construction canonique de la phrase.

Ce type d'ellipse vient remettre en question les études antérieures ayant eu recours à une reconnaissance automatique de l'antécédent pour les repérer. En effet,

si le repérage automatique de l'antécédent est nécessaire à l'interprétation du site elliptique, il reste cependant insuffisant lorsqu'il est exophorique, ne se trouvant pas dans le contexte linguistique. S'ajoute à ceci, dans certaines constructions, la difficulté présentée par l'absence d'une marque morphologique d'accord du verbe, qui, si elle était là, pourrait servir d'indice au repérage de ce qui manque. Ces occurrences apparaissent notamment dans les énoncés interrogatifs requérant une réponse affirmative ou négative de la part du co-locuteur (*yes/no questions*). Il existe en grammaire anglaise deux façons d'exprimer ce type de questions :

– Soit l'auxiliaire est apparent (formes verbales aspectuelles) ; dans ce cas l'interrogation est formulée comme suit : *auxiliaire + sujet + verbe* ? comme dans *Are you leaving?*

– Soit la phrase ne contient pas d'auxiliaire (formes verbales simples), dans ce cas la structure implique l'ajout de l'opérateur *do (présent/passé) + sujet + verbe ?* comme dans *Did she come?*

En réalité, l'interaction du registre et de la structure syntaxique est contraignante dans la plupart des cas. Selon le type de discours et son registre, le locuteur prend parfois la liberté d'abréger son propos en ne formulant que la moitié de la question. Il omet ainsi l'auxiliaire et parfois le sujet en ne laissant dans la phrase que le verbe principal. Comme on le voit dans l'exemple (63) ci-dessous, seuls le verbe principal *give* et les compléments sont présents.

(63) <CDEV> Ø Ever **give** money to Gary Thorp?
- No, Ø never **heard** of him.

L'absence de flexion (ou désinence) sur le verbe *give* dans l'occurrence (63) montre qu'il s'agit ici d'une ellipse de *do* et du sujet. Se pose alors le problème du temps de *do*. Or, il s'avère parfois difficile, voire impossible, de le déterminer en particulier lorsque les situations syntaxiques ne contiennent aucun indice, et que seul le contexte en fournit. De la même façon, l'exemple présenté ci-dessous (64) est ambigu du fait de l'incertitude pesant sur le temps de la phrase : *do* peut être au présent ou au passé sans que la grammaticalité de l'échange en soit affectée.

(64) <CDEV>- Ø **Pinch** anything?

Par contre, la forme du verbe plein restant dans la phrase interrogative aide au repérage des auxiliaires *have* et *be* lorsqu'ils sont ellipsés. En effet, dans l'exemple (64), *have* et *be* ne pourraient convenir pour restituer l'ellipse tout comme *do* ne le pourrait dans l'exemple (65) :

(65) <CDEV> Ø Seeing someone now? Course you are Ø!

C'est également le cas de l'exemple (66) où le participe passé *done* et la réponse du co-locuteur plaident en faveur d'une ellipse de l'auxiliaire *have*. Dans ce dialogue, sont fragmentaires à la fois la question et la réponse. En effet, dans la réponse qu'apporte le co-locuteur, une partie du syntagme verbal est ellipsée. L'antécédent de cette ellipse est apparent dans l'occurrence précédente et s'il devait réapparaître dans la structure grammaticale de la réponse, un ajustement devrait être effectué pour garantir sa grammaticalité[101] (Hardt, 1991). De ce fait, le remplacement de *something* par *anything* (réponse à *something* dans le cas d'une déclaration négative) est nécessaire : *done anything like that*.

(66) <CDEV>- Ø Ever **done** something like that? - I haven't Ø.

Si les flexions et les formes *ing* et *en*/*ed* des verbes restants aident à déterminer l'auxiliaire ou l'opérateur manquant, le temps auquel ce dernier est conjugué n'est pas aisément identifiable par un simple repérage syntaxique. Il requiert en outre de porter une attention particulière à la situation extralinguistique de l'énoncé.

Il est en effet difficile dans certains cas d'identifier le sujet ellipsé dans la phrase. Par défaut, lorsqu'il s'agit d'un dialogue, *you* est le plus fréquent et convient le plus souvent aux situations. Néanmoins, il arrive parfois que le sujet soit à la troisième personne. Là encore, lorsque l'auxiliaire est ellipsé avec le sujet, il est impossible de déterminer si le pronom ellipsé est celui de la deuxième personne ou de la troisième personne (par exemple, *you* ou *he*/*she*) puisque la terminaison qui permet de faire la

[101] Nous rappelons que, comme l'affirment les partisans des approches syntaxiques de l'ellipse, l'élément ellipsé n'est pas toujours une copie exacte de l'antécédent, et des modifications syntaxiques devront être effectuées pour assurer la grammaticalité et l'acceptabilité de la phrase.

distinction en anglais est la marque des auxiliaires. L'exemple (66) accepterait un sujet à la troisième personne sans risque d'agrammaticalité, sous réserve toutefois d'une réponse congruente.

De cette manière, pour constituer le patron permettant la détection automatique de ce type d'ellipse, il apparait nécessaire de prendre en compte les deux conditions ci-dessous :

– Si la phrase commence par un verbe lexical et se termine par un point d'interrogation, nous supposerons qu'il y a deux vides syntaxiques : un vide sujet et un vide auxiliaire.

– Si le verbe est réduit à un participe passé ou un participe présent, nous supposerons que l'auxiliaire absent est soit *have,* soit *be*. En revanche, si le verbe est une base verbale (*stem*), nous supposerons que l'opérateur *do* ou éventuellement un modal est ellipsé.

Ces conditions sont réunies dans 3 patrons :

Le premier repère chaque 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 de pronom personnel ou de virgule après le verbe.

Le deuxième patron écarte les occurrences comme *thank you* confondues avec les ellipses, alors que le troisième est consacré aux questions fragmentaires qui n'apparaissent pas en début du dialogue et qui peuvent être précédées par une virgule.

| | | |
|---|---|---|
| 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/}]? /[?]/ |

Pour conclure, il n'échappera pas au lecteur que si les patrons sont relativement longs, c'est bien pour détailler toutes les conditions nécessaires à la présence d'occurrences elliptiques de façon à atteindre le plus grand degré de précision. En revanche, ils ne sont pas exhaustifs et ne détectent pas toutes les occurrences existantes. En effet, seules les structures syntaxiques des ellipses figurant dans le corpus de développement sont transférées dans les patrons. L'observation des résultats de la détection dans le corpus d'évaluation permettra de dégager les autres conditions et d'envisager leur détection dans un travail futur.

## 2. Évaluation des patrons

Pour repérer les types d'ellipses à l'aide des patrons établis, les deux corpus (de développement et les deux échantillons du corpus d'évaluation, voir chapitre 2, p. 81), ont été étiquetés morphosyntaxiquement à l'aide de l'étiqueteur de Stanford. Pour mémoire, afin d'évaluer la performance (précision, rappel et F-mesure) des patrons, deux échantillons du corpus d'évaluation de tailles différentes ont été utilisés. Ils appartiennent tous deux au registre conversationnel de sous-titres. En effet, dans la présente recherche, nous avons classé les sous-titres de séries télévisées dans le genre conversationnel. Ce choix nous permet de considérer l'ellipse comme un phénomène linguistique résultant de l'activité langagière du locuteur dans un contexte de communication orale spontanée. Nous avons donc supposé qu'un corpus entièrement oral, représentatif d'une communication caractérisée par une parole spontanée[102] et interactionnelle, était susceptible d'offrir un usage de l'ellipse très fréquent et varié. C'est la raison pour laquelle nous avons opté pour une

[102] Les acteurs suivent un script imitant les conversations spontanées.

annotation d'un plus grand nombre d'échantillons dans ce genre de discours malgré les restrictions qu'il présente d'un point de vue linguistique[103].

Lorsque les patrons sont établis, ils sont appliqués à l'ensemble des exemples collectés dans le corpus de développement. La figure (12) ci-dessous présente un exemple de résultat : sont donnés le numéro de la ligne et de la phrase contenant l'ellipse, le texte de la phrase et son étiquetage, le type d'ellipse détecté, et le segment reconnu par le patron. Le contexte précédant ou suivant la phrase elliptique peut également être visible dans le même fichier. *Matches* renvoie à la catégorisation que le patron attribue au type d'ellipse tandis que *Annotated types* renvoie à celle que nous avons attribuée manuellement. Lorsqu'une différence se manifeste entre ces éléments, une vérification est requise.

```
---
Line number: 247
sentence number: 276
sentence text: – I don't, no.
–   :   –
I   PRP I
do  VBP do
n't RB  not
,   ,   ,
no  DT  no
.   .   .
matched expression: – I don't,
matched expression value: STRING(Post-do)
matched expression char offsets: (0,10)
matched expression tokens:[--1, I-2, do-3, n't-4, ,-5]
Matches: [post-do]
Annotated types : [post-do]
---
```

**Figure 12 : Ellipse déclenchée par *have* détectée par le patron**

L'amélioration continue des patrons est envisagée dès lors que leur exécution est lancée sur le corpus de développement : il s'agit d'une action répétitive dans laquelle la présence de toute anomalie ou erreur d'identification ou de dysfonctionnement est prise en compte, ce qui aboutit à la validation ou non du patron. À titre d'exemple, lors de son exécution sur le corpus de développement, le patron dédié à la détection

[103] En effet, les conversations spontanées sont caractérisées par une absence de constructions grammaticales complexes, par des ruptures de constructions grammaticales, des suppressions de segments, des tics de langage. Ces constructions peuvent être considérées comme inexploitables, en TAL notamment, puisqu'elles forment des séquences de fragments requérant généralement un contexte extralinguistique pour être interprétées dans leur totalité. Cependant, un discours fragmentaire n'est pas forcément elliptique, et peut au contraire suivre la *norme grammaticale* et répondre ainsi à des règles de construction identifiables.

des ellipses post-*wh* ne doit pas repérer les occurrences où le mot *wh*– est suivi d'une séquence verbale. En cas d'erreur, une vérification manuelle devra être effectuée pour identifier les origines de l'erreur comme nous le verrons par la suite.

Après leur validation, les patrons sont lancés sur le corpus d'évaluation. Il s'agira alors d'évaluer leur performance et de dresser une typologie des erreurs survenues lors de la détection.

Les résultats[104] obtenus pour chaque type d'ellipse dans les trois corpus annotés sont détaillés dans le tableau (7) ci-dessous, selon l'élément déclencheur.

| Type d'ellipse | CDEV | | | | Échantillon 1 | | | | Échantillon 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 | 0,59 | 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 |

**Tableau 7 : Résultats de l'évaluation[105]**

[104] Les résultats présentés dans ce chapitre sont parus dans (Hamza & Bernhard, 2019). Voir la bibliographie pour plus de détails.

[105] # correspond au nombre d'ellipses annotées.
Afin de calculer la précision, le rappel et la F-mesure, nous avons procédé comme suit :

$$\text{Précision} = \frac{\text{Nombre d'ellipses correctement détectées}}{\text{Nombre total des ellipses détectées}}$$

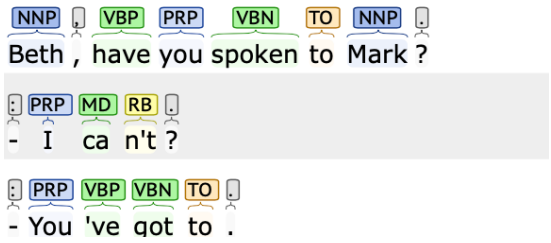
$$\text{Rappel} = \frac{\text{Nombre d'ellipses correctement détectées}}{\text{Nombre total des ellipses à détecter dans le corpus}}$$

$$\text{F-mesure} = 2 . \frac{(\text{Précision} . \text{Rappel})}{(\text{Précision} + \text{Rappel})}$$

Le tableau 7 montre que les patrons sont bien adaptés aux types d'ellipses qui sont présents dans ce corpus de développement, sauf pour celui de {qs-frag} dont le rappel est limité à 0,53. Ce taux peut être expliqué par la difficulté rencontrée à affiner davantage le patron pour détecter toutes les occurrences annotées (voir section 3.).

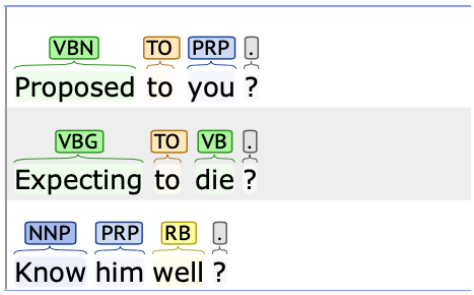
Les résultats de la détection dans les deux échantillons montrent un taux de rappel très encourageant pour certaines catégories mais très bas pour d'autres. Le discours conversationnel auquel appartiennent ces deux échantillons (et dont les sous-titres font partie) est représentatif de cette communication caractérisée par une parole spontanée et interactionnelle, susceptible d'offrir un usage de l'ellipse très varié, évoquée plus haut. On remarque par exemple que les ellipses déclenchées par les modaux, *be/have*, *to, do, wh–*, et les *question tags*, toutes apparentées à l'ellipse du syntagme verbal et à l'ellipse propositionnelle sont beaucoup plus fréquentes que l'ellipse nominale ({post-card, post-ord, post-quant, post-geni}). L'échange (67) ci-dessous contenant deux ellipses, l'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} :

(67)



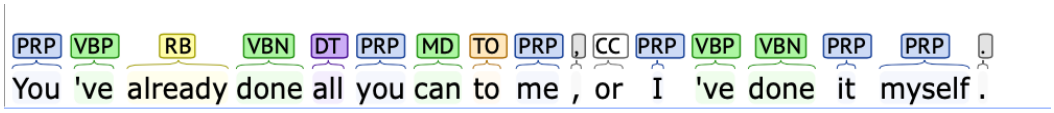

L'ordre des étiquettes morphosyntaxiques de cet exemple correspond aux conditions combinées dans les deux patrons qui reconnaissent comme elliptique. La détection est également correcte dans les questions fragmentaires (68) qui illustrent les trois cas répertoriés dans les patrons {qs-frag}.

(68)



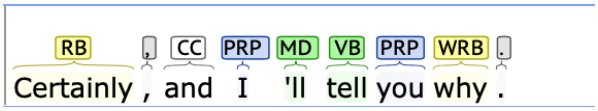

En effet, ces questions fragmentaires requièrent chacune un traitement différent. Les flexions de *proposed* et de expecting nécessitent l'auxiliaire *be* ou *have*, tandis que la flexion zéro de *know* indique que *do* manque en plus du sujet. Par ailleurs, le patron {post-mod} établi pour prendre en compte toutes les ellipses déclenchées par un modal a correctement détecté un cas de *pseudogapping* ci-dessous (69) déclenché par *can* nécessitant l'effacement de *do*. Le rappel atteint 0,96 dans l'échantillon 1 et 0,86 dans l'échantillon 2.

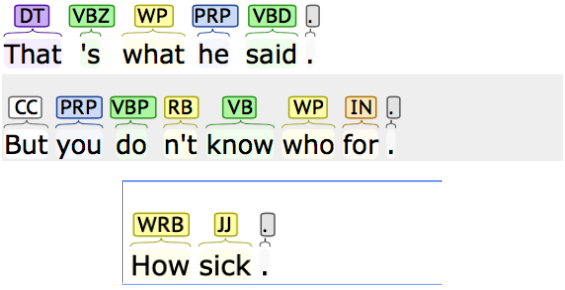
(69)




De la même manière, la particularité du patron {post-wh}, dont le rappel est particulièrement élevé (0,98 dans l'échantillon 1 et 0,99 dans le deuxième), réside dans sa capacité à récupérer tous les cas de *sluicing* notamment les catégories identifiées par Baird *et al*. (*root sluices* et *embedded sluices*), mais aussi le *swiping* et les *sluicings* exclamatifs qui répondent à la construction *wh– + adjectif* :

(70)

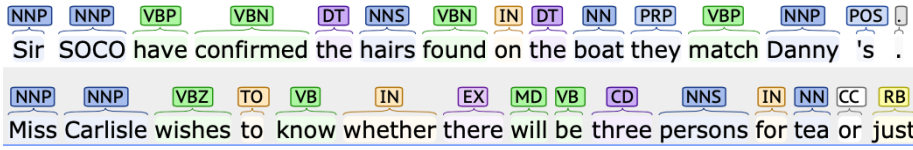




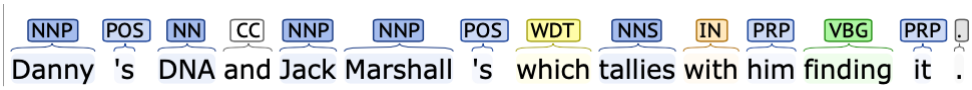

Ce rappel élevé répond à nos besoins pour une étude ultérieure des ellipses dans des corpus de grande taille. Il est en effet plus simple de filtrer des faux positifs manuellement que de parcourir exhaustivement des corpus de grande taille pour retrouver les faux négatifs.

Certains taux, en revanche, sont relativement bas. La précision du patron {post-geni} est seulement de 0,25 dans l'échantillon 1, et celle du {post-card} est de 0,21 dans l'échantillon 2. Ces patrons ont été constitués à partir des occurrences collectées dans le corpus de développement dont le nombre parait insuffisant pouvant souligner par-là la rareté du phénomène : 19 ellipses déclenchées par un marqueur génitif et 6 par un nombre cardinal. En effet, ces deux patrons prennent en compte uniquement les occurrences où le nombre cardinal et le génitif sont immédiatement suivis d'une marque de ponctuation comme dans l'exemple (71) mais ne prennent pas en compte ces déclencheurs au milieu de phrase (72), occurrence non détectée par conséquent.

(71)



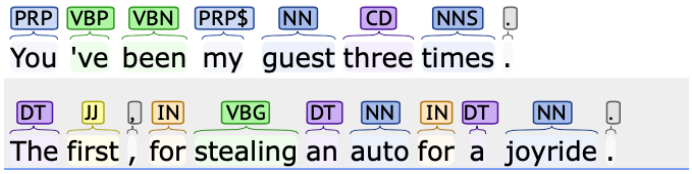

(72)




Pourtant, cette explication pourrait être écartée si l'on compare les résultats du patron {post-ord} dont le rappel a atteint 1,00 alors que seulement 4 occurrences ont

été incluses dans le corpus de développement, seules 4 détectées dans l'échantillon 1 et une seule dans l'échantillon 2. Cela tient au fait que les conditions d'apparition des ellipses déclenchées par les nombres ordinaux précédés de l'article *the* (exemple (73) ci-dessous) ont été plus faciles à formaliser dans un patron précis que celles des ellipses {post-geni} et {post-card}.

(73)




Par ailleurs, les résultats plus faibles observés pour ces types d'ellipse peuvent aussi être justifiés par leur rareté et la difficulté qui se manifeste par des variations syntaxiques non-observées dans un corpus de développement de petite taille. Il est inutile de rappeler que seules les conditions syntaxiques des occurrences figurant dans le corpus de développement sont interprétées dans les patrons. Il en existe d'autres que nous n'avons pas traitées mais qui seraient intéressantes à analyser.

En dépit de la précision imparfaite observée dans certaines catégories d'ellipse, ces patrons nous ont été particulièrement utiles lorsque nous les avons lancés sur le corpus d'évaluation juste après l'annotation manuelle et la vérification avec les expressions régulières. Ils avaient en effet détecté des ellipses que nous n'avions pas annotées en raison d'un oubli, imperfection inhérente à toute action humaine. Nous soulignerons par-là l'une des contributions de notre travail qui réside dans la possibilité d'utiliser ces patrons pour exploiter un large corpus et effectuer un premier repérage de tel ou tel type d'ellipse automatiquement. Ce premier repérage peut servir de base pour annoter et analyser les exemples afin de faire émerger davantage de variations et de configurations syntaxiques propres au phénomène elliptique qui peuvent être prises en compte dans les prochaines investigations. Dans le chapitre 4, nous allons évaluer la performance de ces patrons appliqués à un corpus multi-genre.

## 3. Typologie d'erreurs à partir d'une détection à base de tokens

Nous nous intéresserons ici aux résultats obtenus et visualiserons les différents types d'erreurs engendrées par notre procédure de détection, pour enfin proposer des pistes d'amélioration. Les patrons n'ont pas toujours atteint une précision parfaite lors de leur application sur le corpus d'évaluation, notamment en ce qui concerne les catégories précédemment mentionnées. Après avoir parcouru les ellipses détectées ou non par chaque patron, nous avons remarqué que certaines erreurs déjà identifiées dans le corpus de développement étaient impossibles à corriger. Toutefois, l'objectif de la détection est prioritairement qualitatif et rend compte des erreurs d'étiquetage lors de l'identification des ellipses. Comme on le remarque, les problèmes semblent être liés à la fois aux outils d'analyses et aux patrons eux-mêmes.

### 3.1. Erreurs dues à la précision insuffisante des patrons

Nous avons signalé à la fin du chapitre 2 la difficulté à établir des patrons pour repérer certaines catégories comme le *gapping* que nous avons notamment écarté. La détection des catégories que nous avons conservées n'en est pas pour autant simplifiée ni exempte de défauts puisque la précision de certains patrons est très basse en raison d'un manque d'étiquettes représentatives dans l'étiqueteur et de la difficulté à affiner un patron en y incluant toutes les irrégularités possibles des phénomènes elliptiques.

#### 3.1.1. Étiquettes insuffisamment représentatives

Les étiquettes[106] disponibles pour étiqueter les tokens d'une phrase donnée ne semblent pas assez représentatives pour annoter les déclencheurs cruciaux de l'ellipse.

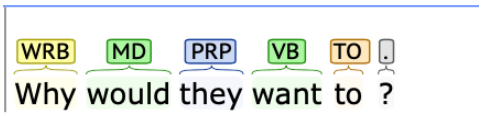
[106] Nous rappelons que l'étiqueteur Stanford utilise les étiquettes de Penn Treebank.

### 3.1.1.1. TO infinitif *vs* TO préposition

Dans ce jeu d'étiquettes, par exemple, aucune distinction n'est établie entre le *to* préposition et le *to* marqueur d'infinitif. Ces deux tokens portent tous les deux l'étiquette TO quelle que soit leur fonction dans la phrase, ce qui engendre alors beaucoup de faux positifs. Comme le montre le tableau (7), la précision des patrons {post-to} a atteint 1,00 dans le corpus de développement mais a baissé dans les échantillons plus larges : 0,62 dans l'échantillon 1 et 0,79 dans l'échantillon 2.

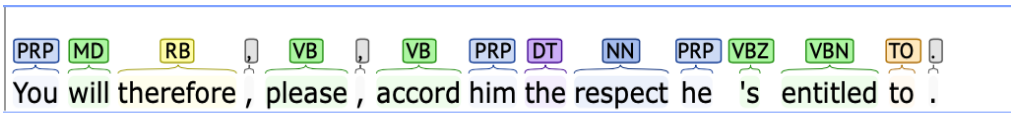
*To* en tant que marqueur d'infinitif déclenche une ellipse lorsque le segment verbal qui le suit, composé d'une base verbale et de ses compléments, est effacé, comme le montre l'exemple (74) :

(74)



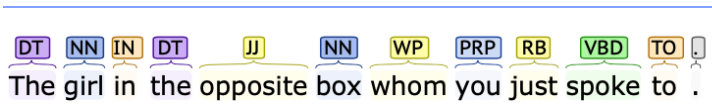

À l'inverse, la préposition *to* dans l'exemple (75) ci-dessous n'est pas elliptique puisque le syntagme nominal *the respect*, antécédent de la proposition relative *he's entitled to* (avec omission de *that*) est intégrée au syntagme prépositionnel complexe complément du verbe *accord*. Cet étiquetage imprécis est à l'origine du repérage de cette phrase comme elliptique alors qu'elle ne l'est pas.

(75)



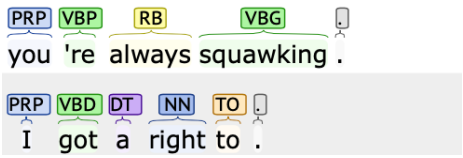

Le même cas est observé dans l'exemple avec le pronom *whom* ci-dessous :

(76)



Ces deux erreurs apparaissent aussi dans les relatives dont la syntaxe se rapproche des phrases interrogatives introduites par un pronom interrogatif, dans la mesure où le syntagme omis n'est pas effacé mais simplement déplacé. Pour améliorer le patron {post-to}, il aurait été possible de le restreindre aux seules occurrences ne contenant pas de pronom *wh–* précédant le *to*, mais la précision du patron peut baisser puisqu'il ne détectera pas les vraies ellipses déclenchées par *to* dans les interrogatives avec le pronom *wh–* comme dans *What if he tries to?* De plus, lors de l'annotation du corpus d'évaluation, d'autres cas non pris en compte dans l'établissement de patrons (en raison de leur absence dans le corpus de développement) ont été rencontrés. En effet les deux patrons {post-to} appliquent la nécessité que *to* soit *immédiatement* précédé par un verbe, un adverbe, un pronom, un adjectif ou *not*. De ce fait lorsque *to* est précédé par un nom, l'ellipse n'a pas été détectée comme c'est le cas de l'exemple (77) ci-dessous.

(77)

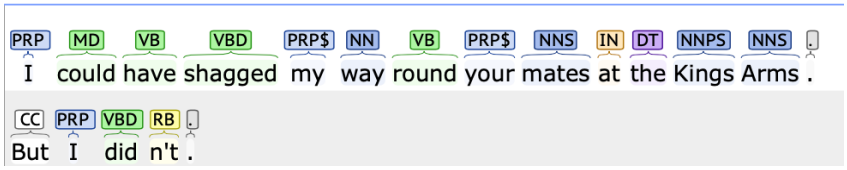



En réalité, il est difficile dans ce cas d'imaginer une amélioration puisque restreindre davantage le patron aux constructions non relatives ou autoriser les noms avant *to* engendre une non détection des autres ellipses (plus fréquentes) ou une détection de faux positifs. En revanche, un critère qui pourrait faciliter une méthodologie de détection à base de patrons ou par apprentissage automatique réside dans une éventuelle sélection selon la construction sémantique des verbes qui précèdent *to.* En effet, ce sont les verbes orientés vers l'accomplissement ou non d'une action qui semblent le plus souvent déclencher des ellipses. Une classification selon les propriétés primitives (états *vs* action) de ces verbes pourrait alors être envisagée.

### 3.1.1.2. Auxiliaire *vs* verbe plein

L'une des limites de l'analyse morphosyntaxique tient également à la non-spécificité des étiquettes attribuées aux auxiliaires. En effet, *have, be* et *do* sont étiquetés de la même manière que le serait un verbe lexical dans une phrase. L'exemple (78) présente une ellipse déclenchée par *do* dans une configuration négative *I didn't* où le syntagme *shag my way round your mates at the Kings arms* est omis.

(78)

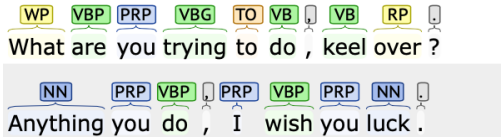



Les différentes étiquettes que *do,* tout comme *have* et *be,* peut prendre sont les suivantes :

VB (Verbe, forme de base)
VBD (Verbe au passé)
VBG (Verbe, gérondif ou participe présent)
VBN (Verbe, participe passé)
VBP (Verbe, qui n'est pas la 3ème personne du singulier au présent)
VBZ (Verbe, 3ème personne du singulier au présent)

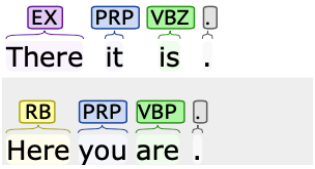
Pour couvrir toutes les apparitions possibles du déclencheur de l'ellipse, aucune de ces étiquettes ne devrait être exclue du patron. Par conséquent le nombre de faux positifs peut être très important en raison de la détection des verbes lexicaux, étiquetés de la même manière et qui ne sont pas déclencheurs d'ellipse. Pour pallier cette lacune et compléter l'analyse morphosyntaxique, nous avons opté pour l'utilisation de lemmes. Malgré cela, cette stratégie s'avère être d'un usage limité puisqu'on retrouve de nombreux faux positifs : la précision du patron {post-do} est de 0,57 dans l'échantillon 1 et de 0,69 dans l'échantillon 2, celle du patron {post-be/have} est encore plus basse 0,47 dans l'échantillon 1 et 0,50 dans l'échantillon 2. Les deux occurrences dans (79) présentent *do* comme verbe plein mais détecté par le patron comme elliptique.

(79)




Dans le premier exemple, *do* ne peut être déclencheur puisqu'il est dans sa forme infinitive *to+base verbale*. Il s'agit ici d'une construction que nous n'avons pas rencontrée dans le corpus de développement : cette restriction aurait pu être incluse dans le patron, pour qu'il ne détecte pas les occurrences de *do* lorsqu'il est précédé de *to*. Par contre, le deuxième exemple présente un réel défi puisqu'il s'agit encore une fois du déplacement du complément *Anything* avant le sujet et le verbe de la phrase et non de son effacement. Par ailleurs, c'est également ce problème que l'on rencontre pour les auxiliaires *be* et *have* (80) lorsque les compléments adverbiaux *there, here*, par exemple, les précèdent.

(80)




En réalité, les erreurs engendrées par le déplacement de ces compléments avant l'élément déclencheur ne sont pas seulement liées au manque d'une étiquette précise, mais elles tiennent également à la difficulté d'inclure toutes les variations dans les patrons.

### 3.1.2. Difficultés à affiner le patron

Comme nous l'avons signalé, par la longueur et la multitude des patrons, nous avons tenté de couvrir le plus de structures elliptiques possibles. Il apparait que les erreurs engendrées par les patrons insuffisamment affinés sont liées au fait que les ellipses ne sont pas assez représentées dans le corpus de développement et qu'en

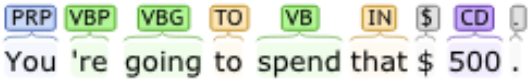
raison de la variation du phénomène, ce dernier devient impossible à formaliser dans son entièreté.

Les ellipses déclenchées par un cardinal, par exemple, bien qu'elles soient rarement étudiées dans la littérature scientifique, sont toutefois importantes et intéressantes dans la mesure où leur traduction d'une langue à une autre pose encore problème. Néanmoins, leur détection est très insuffisante (la précision n'a pas atteint 0,30 dans les deux échantillons) parce que le Stanford Tagger lui-même engendre beaucoup d'erreurs. En effet, le patron, étant constitué de telle sorte qu'il repère les cardinaux immédiatement suivis d'une marque de ponctuation, a détecté tous les chiffres (années, montants, etc.) qui, bien-sûr, ne déclenchent pas tous systématiquement un phénomène elliptique.

(81)


Par ailleurs, la convention d'écrire le signe *$* avant le nombre en anglais a influencé l'efficacité du patron qui a détecté les occurrences, comme (82) ci-dessous, non elliptiques avec la présence du $ :

(82)

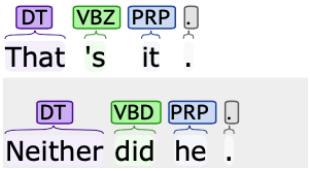



Une piste d'amélioration de ce patron, dans ce cas précis, pourrait sans doute être l'inclusion des entités nommées, dans une requête conditionnée à exclure toute entité nommée *Money* et *Date* telles que présentées dans l'exemple (83).

(83)

En outre, parmi les erreurs résultant d'une précision insuffisante du patron, nous relevons celles engendrées par une similitude de structures dans l'ordre linéaire de la séquence d'étiquetage. Cette similitude est observée dans le patron {vs-tag} dédié à repérer les ellipses dans les *question tags* et celles déclenchées par l'inversion du sujet et du verbe. C'est le cas de l'exemple (84) *That's it.* où l'ordre d'étiquettes qui correspond parfaitement à l'une des conditions exigées par le patron, fausse la détection. Cette phrase est alors détectée comme une ellipse dont l'étiquetage est semblable à celui d'une phrase comme *Neither did he,* elliptique.

(84)

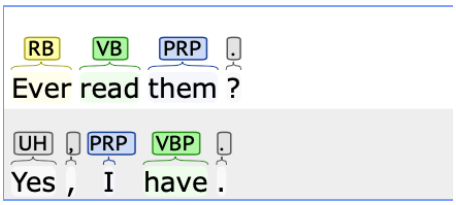



Cette confusion n'est pas seulement due à l'impossibilité d'affiner le patron ou à l'organisation des étiquettes dans la séquence (ordre linéaire des étiquettes) mais s'étend au repérage exact des pronoms (identification des catégories), puisque la distinction fondamentale entre *nature* et *fonction* ne semble pas être prise en compte par l'étiqueteur. En d'autres termes, l'étiquetage de ces exemples concerne deux éléments distincts, l'ordre linéaire et la catégorie de la séquence. Ces deux éléments sont interdépendants dans la mesure où des contraintes syntaxiques (tenant à l'ordre des éléments dans la phrase notamment) sont inhérentes aux catégories. Cette apparente complexité, élémentaire pour l'humain, n'est pas prise en compte par l'étiqueteur en dépit des indications couvrant ces deux aspects dans la formulation des conditions (la précision avec les lemmes notamment).

D'ailleurs, le patron {qs-frag} engendre le même type d'erreur et parait particulièrement intéressant à analyser. Le corpus de développement contient en plus des questions fragmentaires, des ellipses de l'ensemble du sujet-auxiliaire dans les phrases déclaratives pouvant être confondues avec les questions fragmentaires. Nous les avons annotées comme elliptiques mais nous avons dû restreindre le patron

à détecter seulement celles relevant des phrases interrogatives. En d'autres termes, les ellipses de la paire sujet-auxiliaire ont été annotées mais ne sont pas détectées puisqu'aucun patron ne leur a été dédié. On pourrait par exemple obtenir beaucoup de faux-positifs détectés dans les phrases à l'impératif. Ci-dessous, nous observons trois occurrences du même verbe *read* : la première (85) est elliptique car *read* est précédé par un auxiliaire *have* et un sujet *you* effacés voire enfouis dans une structure « profonde ». Le point d'interrogation et la réponse *I have* favorisent cette prise de position. Cette occurrence correspond donc parfaitement aux conditions réunies dans le patron et a été détectée.

(85)

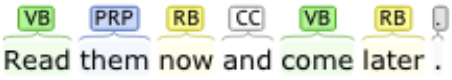



Les deux qui suivent sont des exemples donnés à titre d'illustration et ne figurent pas dans nos corpus. La deuxième occurrence de *read* (86) est également elliptique du fait de l'effacement du sujet *I* (et possiblement de l'auxiliaire). Cette ellipse apparait en revanche dans une phrase déclarative. Comme l'une des conditions qui lui ont été fixées est de repérer seulement les interrogatives, le patron ne la détectera pas.

(86)

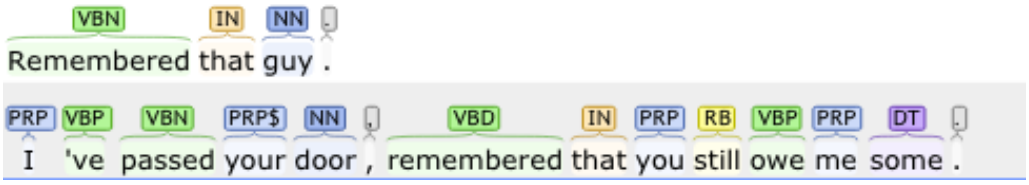

Cette restriction permet d'éviter le repérage des constructions impératives comme dans la troisième occurrence de *read* ci-dessous (87) où aucun effacement n'a eu lieu, et donc, aucune ellipse.

(87)




En poursuivant notre observation, nous notons que le verbe *read* est étiqueté dans les trois phrases comme VB et qu'aucune distinction entre ses formes n'a été faite. L'étiqueteur attribue l'étiquette VB (base verbale) à tout verbe (qui n'est pas –*ed/en* ou –*ing*) débutant une phrase, y compris lorsque ce verbe est en réalité un participe passé (ou un passé : l'exemple 87 peut être interprété comme une ellipse du sujet seul). En fait, l'étiqueteur ne repère pas qu'il y a une ellipse du sujet, il ne peut donc pas attribuer la bonne étiquette au verbe. L'étiquette VB ne peut être exclue du patron malgré le risque de détecter un nombre considérable de faux positifs et de laisser passer les occurrences elliptiques les plus fréquentes. Or, ce problème est particulièrement observé avec les verbes irréguliers. Dans certaines autres occurrences, l'étiqueteur parvient à différencier l'étiquette VBN (participe passé) de l'étiquette VBD (passé simple) comme l'illustre l'exemple (88) de *remembered* (bien sûr une interprétation comme *I've remembered that you still owe me some* est possible) :

(88)[107]

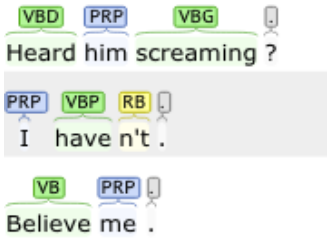



Pour aller plus loin, comme le verbe lexical qui débute la phrase peut porter, soit l'étiquette VB lorsqu'un modal ou *do* manque, soit VBN lorsque *be* ou *have* manque, nous avons exclu l'étiquette VBD, VBP et VBZ du patron {qs-frag}. Le résultat a été alors la non détection des occurrences elliptiques là où le verbe a été étiqueté comme

[107] Exemple donné à titre d'illustration et ne figurant pas dans les échantillons exploités.

VBD alors qu'il aurait dû être étiqueté comme VBN, comme *heard* dans l'exemple (89). Il apparait évident que l'auxiliaire *have* manque et la réponse le confirme.

(89)




En effet, même lorsque le patron est suffisamment affiné pour détecter les occurrences elliptiques, certaines erreurs sont apparues en raison d'un étiquetage erroné dû à plusieurs facteurs.

### 3.2. Erreurs engendrées par l'étiqueteur

Nous l'avons énoncé, entraîner un étiqueteur consiste à annoter d'abord manuellement une grande quantité de corpus variés pour ensuite effectuer un apprentissage, de manière à ce que l'étiqueteur produise automatiquement le même étiquetage sur d'autres corpus. Lorsque les corpus d'apprentissage et d'application n'appartiennent pas au même registre ou à la même langue de spécialité, l'étiqueteur peut rencontrer des difficultés susceptibles de provoquer des erreurs.

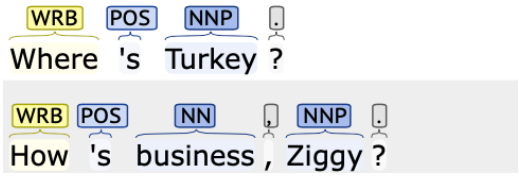
Les données ayant servi à l'apprentissage de l'étiqueteur morphosyntaxique Stanford sont extraites de plusieurs corpus dont le *WSJ* et présentent plusieurs genres de textes mais vraisemblablement pas de genres proches de l'oral[108]. L'exactitude de cet étiqueteur est de 96,86% (Toutanova & Manning, 2000). Il est difficile d'affirmer en l'état actuel de notre progression dans quelle mesure les différents genres de discours exploités dans la présente recherche, sont, ou non, à l'origine des ambiguïtés à résoudre lors de l'étiquetage. La question se pose néanmoins. Ce qui est certain, c'est qu'en dehors des problèmes liés strictement à l'étiquetage, le phénomène elliptique présente un défi supplémentaire à relever pour les étiqueteurs. De ce fait, en raison de l'ambiguïté qu'un seul mot peut présenter et

[108] https://nlp.stanford.edu/software/parser-faq.html#models (accès vérifié le 6 août 2019 à 19:12).

des incises[109] qui perturbent l'étiquetage correct des occurrences, on voit nettement que les erreurs relevées dans cette section sont directement liées au mauvais étiquetage de certaines catégories.

Nous avons ainsi remarqué que le 's comme marqueur du génitif et le 's de la forme contractée de *be* ou *have* à la 3ème personne sont tous deux souvent étiquetés comme POS (*possessive ending*). Dans les phrases présentées dans l'exemple (90) ci-dessous, le 's de *is* est étiqueté comme un possessif, ce qui a rendu la séquence tout à fait compatible avec les conditions de détection du patron {post-wh} qui exige une récupération des pronoms *wh*– lorsqu'ils sont suivis de n'importe quelle étiquette à l'exception du verbe et du modal. En effet, aucune des étiquettes suivant le pronom *wh*– (POS, PRP, RB, NNP) n'a été exclue du patron :

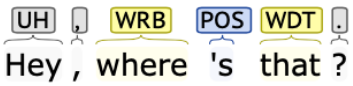
(90)




Une autre source d'erreur a été observée pour le même patron et apparaît avec *that* étiqueté comme WDT (*wh*– déterminant : la seule étiquette qui lui est dédiée) dans l'exemple (91) ci-dessous. Nous avons involontairement omis d'exclure *that* lors de l'élaboration du patron. En effet, *that* est source d'erreur lorsqu'il apparaît dans la phrase comme pronom démonstratif et qu'il est étiqueté comme un déterminant *wh*–. Ainsi sommes-nous toujours confrontée au problème d'identification des formes et cela malgré la désambiguïsation.

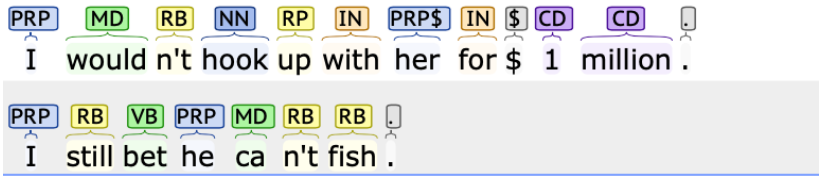
[109] Par simplification, nous appelons « incise » tout segment enchâssé (d'un simple mot à une proposition) à l'intérieur d'une phrase sans aucun mot de liaison (Arrivé *et al.,* 1986, 323). Tout au long de notre recherche, nous n'attachons pas d'importance au statut de l'incise qu'elle soit constituée d'un syntagme verbal, nominal, adverbial ou autre, dans la mesure où l'impact qu'elle engendre dans la détection est le même indépendamment de sa nature.

(91)



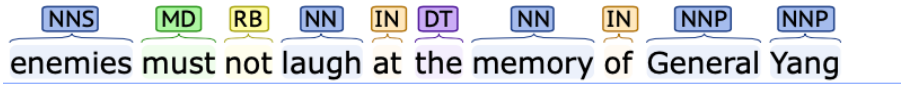

L'étiquetage erroné des éléments résiduels impacte également la précision des patrons. Si l'on considère le patron {post-mod} par exemple, nous remarquerons qu'il présente un taux de précision et de rappel très élevés (précision 0,71 et rappel de 0,96 dans l'échantillon 1 et précision de 0,79 et rappel de 0,86 dans l'échantillon 2). Les erreurs relevées dans ce cas renvoient notamment à l'étiquetage erroné des catégories suivant le modal. Dans l'exemple ci-dessous, *hook* et *fish* sont étiquetés comme noms NN alors qu'ils sont des verbes et auraient dû être étiquetés comme VB.

(92)



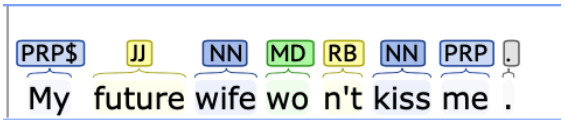

Dans les occurrences elliptiques détectées de manière erronée nombreux sont les exemples contenant des verbes étiquetés comme noms ayant induit les patrons en erreur, tels que *divorce, kiss, pardon, race, text*. Ces erreurs d'étiquetage portent notamment sur des mots qui peuvent être étiquetés selon deux manières (verbe et nom) où seules les contraintes d'ordre syntaxique permettent de les identifier. C'est pourquoi, même dans le cas où l'outil dispose d'une étiquette spécifique pour le déclencheur (ici MD), la précision du patron {post-mod} est limitée par les erreurs d'étiquetage morphosyntaxique et par conséquent par la détection de phrases non-elliptiques reconnues comme étant elliptiques. On remarque la même erreur dans les phrases négatives exemple (93) ci-dessous qui est également identifiée comme une ellipse {post-mod}.

(93)




De la même façon, la phrase (94) ci-dessous a été repérée comme *ellipse* post-modale. La présence de la conjonction *but* et les éléments résiduels *kiss me* qui, selon l'étiquetage, ne contiennent pas de verbe, remplissaient apparemment les conditions exigées pour la détection (à savoir de récupérer tout modal non suivi d'un verbe).

(94)

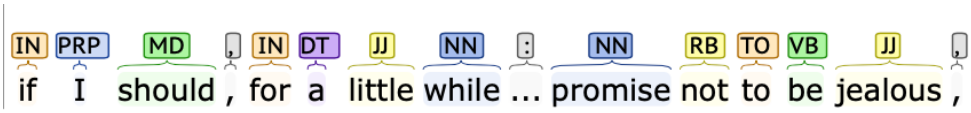



En effet, la confusion est créée par le mauvais étiquetage de *kiss* considéré comme un nom NN. Bien sûr, dans d'autres contextes, *kiss* peut être un nom. C'est la raison pour laquelle il est important de souligner que les erreurs de l'outil ne sont pas des erreurs *aléatoires* mais sont plutôt liées à des ambiguïtés hors contexte. Ceci peut être également lié aux corpus d'apprentissage. Par exemple, si le corpus ayant servi pour l'apprentissage de l'étiqueteur contient uniquement des occurrences *kiss* annoté manuellement comme nom (comme l'exigent les contraintes syntaxiques des occurrences), il est peu probable qu'il soit étiqueté comme verbe lorsqu'il apparaît comme tel dans un autre corpus.

D'autres erreurs relèvent particulièrement des marques de ponctuation, que l'on peut considérer comme un condensé de syntaxe, sémantique et prosodie et qui soulignent encore la complexité des problèmes à traiter par l'outil informatique. En effet, sachant que les patrons sont élaborés en tenant compte de la ponctuation pour suivre l'ordre des tokens, les modaux par exemple, sont parfois séparés du verbe par un segment enchâssé, provoquant de ce fait une erreur d'étiquetage du verbe suivant ce segment. Comme le montre l'exemple (95) ci-dessous, *promise* est

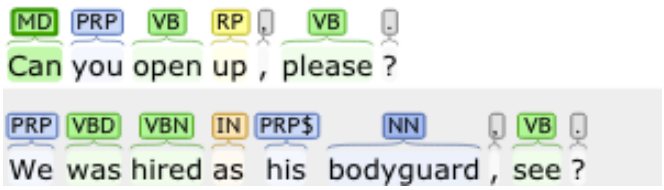
étiqueté comme un nom NN et le patron, suivant la condition récupérant tout modal suivi de n'importe quel élément sauf lorsqu'il est suivi d'un verbe, identifie la phrase comme elliptique[110]. Ainsi, cette erreur se trouve à la croisée des limites liées aux erreurs d'étiquetage et à la précision insuffisante des patrons.

(95)




Ce sont les incises qui ont généré cette erreur et de ce fait impacté la précision du patron {post-mod}. Cependant, il arrive parfois que même en l'absence de l'incise, l'étiquetage soit erroné. Rappelons par exemple le cas de l'omission du sujet et de l'auxiliaire dans la question fragmentaire {qs-frag}. Les deux phrases ci-dessous (96) sont détectées comme ellipses {qs-frag} pourtant la première ne l'est pas.

(96)




En effet, le patron {qs-frag} est dédié à repérer toute occurrence contenant un verbe non précédé d'un nom et suivi d'un point d'interrogation. Le patron a détecté comme elliptique *please*, ici incorrectement étiqueté VB au lieu de UH (interjection).

Par le biais de ces erreurs nous pouvons pointer les difficultés et les limites des patrons à base de tokens dans l'étude du phénomène elliptique. Pour résumer, nous retiendrons deux grands types d'erreurs :

[110] On pourrait imaginer une requête négative qui conditionne le non-repérage de trois noms qui se suivent par exemple, pour améliorer le patron. Ceci pourrait entrainer une baisse du taux de rappel, ignorant certaines occurrences potentiellement elliptiques.

– les erreurs résultant d'une précision insuffisante des patrons (choix d'annotation), ou d'un manque de couverture[111] : ces erreurs renvoient d'une part au manque d'étiquettes suffisamment représentatives et précises pour annoter les déclencheurs cruciaux de l'ellipse (leur sont alors substituées des étiquettes moins précises), et d'autre part, à l'impossibilité d'affiner le patron pour prendre en compte les variations des structures elliptiques[112].

– les erreurs engendrées par l'étiqueteur liées au mauvais étiquetage de certaines catégories, en raison de l'ambiguïté qu'un seul mot peut présenter.

Compte tenu de ces paramètres, issus d'un examen de données relativement peu nombreuses et de ce fait pouvant sembler peu représentatives, il est légitime de s'interroger sur l'intérêt de mener une étude quantitative automatisée pour appréhender l'ellipse en tant que phénomène instable et sujet à variations. En effet, nous avons conscience de la grande difficulté à intégrer ces nombreuses variations structurelles dans une analyse à base de tokens et dans l'élaboration de patrons limités à la phrase. La mise en œuvre d'outils complémentaires pour aborder le phénomène demeure encore un champ d'étude à explorer.

## 4. Vers une amélioration de patrons

### 4.1. Enrichissement du corpus de développement et affinement de l'évaluation

Nous nous sommes limitée dans le présent travail à 331 occurrences elliptiques pour développer nos patrons. Nous envisageons de poursuivre cette recherche, en particulier parce que les patrons ne sont pas assez efficaces, comme les erreurs que nous avons recensées le montrent. Une collecte manuelle d'occurrences diverses et variées et en aussi grand nombre que possible, pour fastidieuse qu'elle soit, enrichirait les corpus qui, d'un certain point de vue, manquent peut-être de *couverture*. Les patrons pourraient alors être perfectionnés et, lors de la phase d'évaluation, l'examen de leurs performances, croisées avec les résultats obtenus par

[111] Nous signifions par manque de couverture le nombre insuffisant de certains types d'ellipse dans le corpus de développement.

[112] Nous ne considérons pas cette impossibilité comme définitive et rédhibitoire compte tenu du développement des outils du TAL.

détection manuelle, permettrait de vérifier s'il existe ou non des ellipses non prises en compte par ces patrons.

### 4.2. Annotation des auxiliaires et des modaux : déclencheurs d'ellipse

Dans la mesure où les auxiliaires et les modaux sont des déclencheurs essentiels de la plupart des catégories d'ellipses, une combinaison de méthodes d'analyse entre les étiquetages morphosyntaxiques à d'autres annotations linguistiques d'ordre sémantique, discursif, voire même prosodique, peut aider à la précision des conditions contenues dans les patrons. Comme nous l'avons indiqué précédemment, à l'exception des modaux, les auxiliaires *have*, *be* et *do* n'ont pas d'étiquette qui les distingue des verbes lexicaux. Cela tient sans doute au fait qu'en syntaxe, une simplification a été effectuée, notamment pour *have* et *be* (ignorant par exemple le rôle de *be copule*). Il serait sans doute intéressant d'ajouter une étiquette (AUX)[113], à titre d'exemple, pour distinguer ces auxiliaires des verbes pleins, dans un premier temps. Une distinction entre auxiliaire et *copule* peut ensuite être envisagée pour peaufiner l'analyse[114]. Pour cette raison, et en prenant en compte ce que les génératives ont développé à travers les « théta-rôles » (rôles thématiques), autrement dit en réintroduisant un codage sémantique parallèlement à une analyse syntaxique des relations de dépendance, l'analyste pourrait alors compléter les conditions morphosyntaxiques indispensables.

Toutefois, l'amélioration de la méthode de détection visant une réduction du taux d'erreurs ne pourra exister sans le recours à l'exploitation d'autres corpus. Ainsi, dans le chapitre suivant, afin de dégager d'éventuelles pistes d'amélioration, nous essaierons de tirer profit du corpus *genré* pour évaluer la performance des patrons sur d'autres types de textes.

[113] Cette étiquette existe par exemple dans les Universal Pos tags du projet *Universal Dependencies* : il pourrait donc être possible, dans de futures recherches par exemple, d'utiliser un étiqueteur entraîné avec ce jeu d'étiquettes particulier.
[114] Une étiquette COP pourrait éventuellement convenir.

# Chapitre 4

# Distribution des ellipses dans un corpus *genré*

*Toute étude en corpus suppose en effet une caractérisation de l'ensemble de textes étudiés, de manière à se donner les moyens d'évaluer la portée des descriptions ou des propositions avancées ; a fortiori, toute étude qui vise la comparaison de corpus à partir de l'hypothèse d'une différence spécifique en rapport avec le phénomène étudié doit pouvoir justifier le choix des sous-corpus comparés.*

Marie-Paule Péry-Woodley[115]

Nous avons exposé dans le chapitre précédent le résultat d'une détection automatique des ellipses dans un corpus de genre conversationnel et dressé une typologie des erreurs y apparaissant. Nous suivrons dans le présent chapitre une démarche identique appliquée cette fois à des genres différents, c'est-à-dire, pour rappel, exécution de patrons sur des échantillons préalablement annotés et évaluation du taux de précision et du rappel des patrons dans les genres sélectionnés.

À travers la typologie d'erreurs recensées dans le travail effectué sur l'observation de dialogues, nous avons mis en avant la difficulté à établir une méthode de détection à base de tokens, ainsi que la précision parfois insuffisante des patrons due aux variations du phénomène étudié et aux erreurs d'étiquetage catégoriel. La question qui se pose est alors la suivante : dans quelle mesure les patrons que nous avons élaborés pour détecter les ellipses dans le genre conversationnel sont-ils transférables à d'autres genres ? Et, s'ils le sont, sont-ils adaptés aux contraintes discursives de ces autres genres ? Car en effet, au sein d'une même langue, des variations linguistiques d'un phénomène précis peuvent apparaitre lorsque celui-ci s'inscrit dans tel genre ou tel autre, politique, journalistique, littéraire, promotionnel,

[115] (Péry-Woodley 2005, 15), voir la bibliographie pour la référence complète.

par exemple, tels qu'ils sont exposés dans ce chapitre, variations parfois difficiles à inclure dans les patrons.

L'objectif de ce chapitre est ainsi de présenter les résultats d'une détection à base de patrons sur ces autres genres discursifs précédemment cités et choisis dans les corpus sélectionnés. Nous aborderons quelques particularités de ces genres qui peuvent à notre sens être à l'origine des différences existant entre les différences de performance. Nous signalerons également les nouvelles erreurs qui surviennent (lorsqu'elles ne figurent pas dans la typologie identifiée dans le chapitre précédent). En utilisant des tests statistiques, nous étudierons ensuite la distribution des types d'ellipses (association entre présence et absence d'ellipse par genre et différence entre les types d'ellipses dans chaque genre). Enfin, le bilan de notre méthodologie de détection conclura le chapitre.

## 1. Ellipse et genre : fréquence et variation

Sans craindre l'utilisation encore peu courante du terme *genré*, nous dirons que notre approche est ici une approche *genrée*, caractéristique sous-jacente de notre méthodologie d'analyse, déterminée, rappelons-le, par les objectifs à réaliser à partir du choix du corpus. Ce néologisme que nous employons à dessein pour désigner d'une manière générale les différents types de discours, est la traduction admise et couramment utilisée de l'expression anglaise *genre-based*. En effet, outre la répartition en littérature des textes en genres (fiction en prose, poésie, théâtre, etc.), le terme *genre* est souvent utilisé, et de façon indifférenciée, pour qualifier le style, le registre ou le type de discours. Ainsi peut-on parler, suivant Beacco (2004) par exemple, de *genre discursif*. Certaines recherches en linguistique textuelle font une distinction nette entre toutes ces notions (voir par exemple, Adam 1997). En ce qui nous concerne, lorsque nous utilisons le qualificatif *genré*, nous incluons dans ce terme unique, toutes les particularités des autres termes qui, selon les auteurs et les disciplines, identifient les textes étudiés en précisant, d'une certaine façon, un mode de fonctionnement.

Les genres discursifs soumis à notre observation sont issus de communications sociales ayant eu lieu dans des situations d'énonciation différentes. De ce fait, nous

partirons pour ce travail, des hypothèses énoncées par Biber & Conrad (2009, 9), selon qui l'étude d'un phénomène linguistique particulier appartenant à divers registres, peut aider à identifier le type de texte étudié. Ils remarquent, par exemple, que la distribution des noms et des pronoms est différente dans une conversation ou dans la presse écrite :

> However, conversation uses more pronouns and fewer nouns, while newspaper writing uses more nouns and fewer pronouns. In other words, the relative distribution of nouns and pronouns differs greatly between conversation and newspaper writing. The linguistic analysis of registers is based on such differences in the relative distribution of linguistic features, which are especially common and pervasive in some registers but comparatively rare in other registers. (Biber & Conrad 2009, 9)

De ce fait, l'approche de Biber & Conrad nous a amenée à faire l'hypothèse que la présence ou non d'ellipses dans tel ou tel discours pourrait participer à son identification en termes de genre discursif. Un travail incluant une étude statistique, serait alors propre à vérifier la concordance entre la fréquence des ellipses et le genre de discours, participant ainsi à la reconnaissance du genre en question. Cependant, sachant qu'étudier un phénomène linguistique comme l'ellipse sans prendre en compte son contexte d'occurrence en termes de *genre de texte* serait dépourvu de toute légitimité, la nécessité d'analyser les genres en lien avec l'ellipse s'est imposée d'elle-même, et ce, dès lors qu'un travail sur corpus a été envisagé.

Notre corpus présente cinq genres différents répartis en sous-corpus : journalistique, politique, littéraire, promotionnel et conversationnel (ce dernier a été traité de manière détaillée précédemment, voir le chapitre 3). Il est important de signaler que la catégorisation des genres telle qu'elle est présentée au cours de ce chapitre peut être sujette à débat, puisque des propriétés et des paramètres peuvent se croiser entre les différentes catégories, établissant ainsi une autre hiérarchie des genres et donc une autre manière de classifier. Le genre du discours politique[116], par

[116] Nous le verrons plus loin, nous faisons un raccourci en appelant « discours politique » les discours prononcés au parlement européen, bien que nous soyons consciente que le genre du discours politique englobe bien d'autres types de textes.

exemple, et le genre promotionnel[117] ont en commun le fait qu'ils sont tous deux d'abord écrits pour être produits à l'oral. Dans cette perspective, ils pourraient donc être classés dans la même catégorie. Cela pourrait également être le cas pour le genre journalistique (presse écrite) et le genre littéraire qui sont tous deux d'abord des discours écrits en vue d'une publication écrite, et donc destinés à la lecture silencieuse[118]. Notre objectif n'étant pas d'entrer dans le débat épistémologique sur les genres et notre travail se limitant aux exemples attestés de nos corpus dans leur version écrite, nous considérons que cette classification offre néanmoins, dans le cas de l'ellipse, une distinction opérationnelle pratique, puisqu'elle permet une répartition des variétés d'oral : spontané, non spontané, écrit oralisé.

Afin de vérifier la variation des phénomènes elliptiques et de compléter notre analyse du chapitre 3, nous rappelons que nous avons annoté manuellement 1 000 lignes extraites aléatoirement dans chacun des sous-corpus[119]. Pour l'annotation manuelle, les phrases sélectionnées aléatoirement sont présentées avec un contexte de deux phrases les précédant et les suivant. La phrase à annoter est présentée avec des étoiles comme l'illustre la figure (13). Le nombre (10476) correspond à la ligne d'occurrence dans le corpus initial, une ligne pouvant contenir plusieurs phrases. Dans cette même figure, nous repérons par exemple une ellipse déclenchée par un modal qui apparait dans la toute dernière ligne (10476) *We can't sanction them anymore. We can't.* ellipse que nous n'avons pas annotée puisqu'elle n'apparait pas dans une phrase à annoter, identifiée par des étoiles.

[117] Voir les caractéristiques plus loin.

[118] On peut considérer que la lecture silencieuse de ces textes écrits s'accompagne d'une phonation intériorisée, dans le cas des textes littéraires en particulier, mais on peut aussi défendre l'idée que ces textes ne prennent tout leur sens que lors d'une lecture à haute voix, d'une *performance*, ce dont le type de travail sur corpus que nous avons engagé ne saurait rendre compte.

[119] En raison de la taille importante des corpus et de l'aspect chronophage de leur vérification globale. Nous remercions Philip Miller pour ses conseils concernant les évaluations statistiques des ellipses dans les genres.

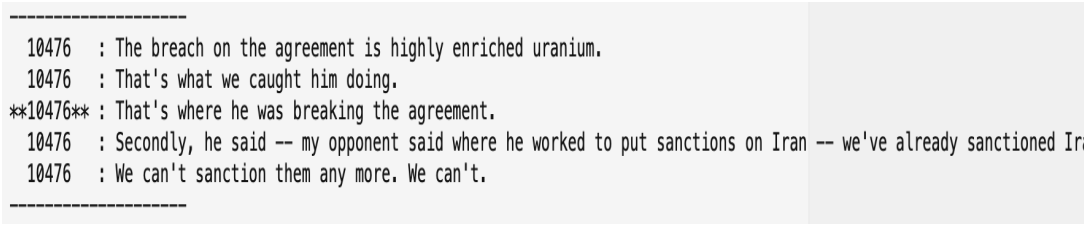



**Figure 13 : Exemple extrait d'un échantillon sélectionné au hasard du genre promotionnel**

Comme nous l'avons signalé, l'annotation des échantillons a demandé plus de temps pour certains corpus que pour d'autres, et ceci en raison de la quantité d'ellipses qui y étaient présentes (pour leur nombre élevé ou au contraire leur rareté) et de la longueur des phrases (temps de lecture plus long). De prime abord, on pourrait croire que plus un corpus contient d'ellipses, plus le temps consacré à l'annotation est long. Cette hypothèse est vite rejetée dans notre cas puisque le corpus conversationnel nous a pris moins de temps que tous les autres[120]. Le tableau (8) ci-dessous présente le nombre d'ellipses (pour chaque déclencheur) repérées dans chacun des échantillons.

[120] À titre d'information, ci-dessous le temps consacré à annoter (la colonne de droite correspond au nombre de tokens par échantillon) :

| | | |
|---|---|---|
| Genre conversationnel | 2h10 | 7102 |
| Genre promotionnel | 2h36 | 16404 |
| Genre littéraire | 2h49 | 17664 |
| Genre journalistique | 2h57 | 23121 |
| Genre politique | 3h06 | 26655 |

| Type/Genre | Conversationnel | Littéraire | Journalistique | Promotionnel | Politique |
|---|---|---|---|---|---|
| qs-frag | 15 | 3 | 2 | 4 | 0 |
| vs-tag | 14 | 13 | 0 | 8 | 0 |
| post-do | 11 | 6 | 7 | 10 | 2 |
| post-mod | 15 | 7 | 1 | 5 | 0 |
| post-be/have | 16 | 6 | 4 | 9 | 1 |
| post-to | 7 | 3 | 1 | 2 | 1 |
| post-wh | 16 | 5 | 2 | 10 | 1 |
| post-geni | 1 | 1 | 0 | 1 | 0 |
| post-quant | 0 | 0 | 0 | 1 | 0 |
| post-card | 2 | 1 | 1 | 4 | 0 |
| post-ord | 0 | 0 | 0 | 1 | 0 |
| Total | 97 | 45 | 18 | 55 | 5 |

**Tableau 8 : Nombre d'ellipses annotées par échantillon**

Arrêtons-nous un instant sur ce travail d'annotation qui nous permet quelques observations préliminaires. Les ellipses sont plus facilement détectées dans les phrases courtes car un contexte plus étoffé ne facilite pas l'annotation. Un facteur cognitif et psychologique pourrait également intervenir en jouant sur la capacité d'attention de l'annotateur. En effet, lorsqu'il annote l'un ou l'autre corpus avec un *a priori* sur la certitude de trouver moins d'ellipses dans tel ou tel genre, l'annotateur se concentre probablement de façon intense pour les repérer sans en oublier ; inversement, il peut tendre à relâcher son attention lorsque les ellipses sont plus fréquentes. Par ailleurs, la haute fréquence et/ou la rareté des ellipses pourrait simplement s'expliquer par les particularités-mêmes des genres retenus. Certains genres, notamment le genre conversationnel comme nous avons déjà pu le montrer, témoignent d'une plus grande fréquence d'usage de l'ellipse que d'autres.

Dans la suite, la performance des patrons sera présentée pour chacun des genres classés selon l'ordre croissant des ellipses annotées.

### 1.1. Genre politique / *European Parliamentary speeches*

Comme tout discours politique, les discours du Parlement Européen sont parfaitement codifiés et donc loin d'être spontanés. Motivés par des questions et des thèmes précis à traiter, ils s'inscrivent, de façon générale, dans la longue tradition de

la rhétorique telle qu'elle a été définie par Aristote selon trois genres : le discours délibératif (politique), le discours judiciaire et le discours épidictique (démonstratif). Le discours politique s'organise généralement autour d'une structure textuelle identique d'un discours à l'autre : salutations codifiées à l'égard des personnes présentes par ordre d'importance décroissant, introduction du propos, développement du propos et conclusion. Les phrases utilisées sont particulièrement longues et parfois complexes ; elles sont généralement déclaratives, parfois interrogatives ou impératives lorsque le locuteur qui prononce le discours recherche une interaction avec le public. L'une des figures très souvent retrouvée dans ce genre de discours est le sous-entendu. Nous utilisons le terme de *sous-entendu* pour faire référence à un phénomène d'ordre sémantique et contextuel qui relève à la fois de la connaissance partagée et du non-dit, comme si les choses allaient de soi et que le sens des termes n'avait pas besoin d'être précisé. Par exemple, le recours fréquent à l'expression « dans nos sociétés démocratiques » tient pour acquis le fait que toutes les sociétés européennes partagent un même idéal et une même pratique de la « démocratie », sorte d'utopie généralisée, alors que certains pays, pour des raisons historiques, peuvent s'en écarter.

Mais la particularité de ce corpus d'étude réside dans le fait qu'il présente un discours écrit à vocation orale. Par conséquent, en dépit de la rigidité de l'écrit codifié, peu propice à la construction elliptique, le passage à l'oral, nécessairement interactif, évoluant vers une plus grande spontanéité, voire vers l'improvisation, pourrait alors favoriser l'apparition d'occurrences elliptiques. Or, le genre politique n'est pas propice à l'apparition des phénomènes elliptiques, puisque, comme nous le constatons dans le tableau (8), les ellipses n'apparaissent que rarement dans ce type de discours. Contrairement aux ellipses rencontrées dans le genre conversationnel (présentées dans le chapitre précédent), celles du genre politique paraissent dans des phrases longues et sont difficilement repérables. Nous isolons ci-dessous, à titre d'exemple, une ellipse {post-wh} déclenchée par *how* dans une longue phrase où l'entièreté de la proposition *it concerns Portugal in a positive way* est ellipsée.

(97) <CP> Mr President, the Portuguese Presidency is extremely anxious to give a comprehensive answer to this question, particularly because it broaches an issue which concerns Portugal in a positive way, and I shall explain **how** Ø.

Les mesures de performance des patrons dans ce genre précis sont présentées dans le tableau (9) ci-dessous[121].

| Type | Nb | P | R | F1 | Vrais positifs | Faux positifs | Faux négatifs |
|---|---|---|---|---|---|---|---|
| qs-frag | 0 | / | / | / | 0 | 0 | 0 |
| vs-tag | 0 | / | / | / | 0 | 0 | 0 |
| post-do | 2 | 0,25 | 0,50 | 0,33 | 1 | 3 | 1 |
| post-mod | 0 | 0,00 | / | / | 0 | 10 | 0 |
| post-be/have | 1 | 0,00 | 0,00 | / | 0 | 8 | 1 |
| post-to | 1 | 0,00 | 0,00 | / | 0 | 1 | 1 |
| post-wh | 1 | 0,03 | 1,00 | 0,06 | 1 | 33 | 0 |
| post-geni | 0 | 0,00 | 0,00 | / | 0 | 2 | 0 |
| post-quant | 0 | / | / | / | 0 | 0 | 0 |
| post-card | 0 | 0,00 | / | / | 0 | 17 | 0 |
| post-ord | 0 | 0,00 | / | / | 0 | 1 | 0 |
| Total | 5 | 0,03 | 0,40 | 0,05 | 2 | 75 | 3 |

**Tableau 9 : Performance des patrons dans l'échantillon politique annoté[122]**

Le nombre de faux positifs détectés dans ce corpus est très élevé pour les patrons {post-mod}, {post-card} et {post-wh}. Les exemples ci-dessous sont incorrectement détectés comme elliptiques :

[121] Compte tenu du nombre limité des ellipses annotées dans certains corpus, nous avons décidé de donner le nombre des occurrences incorrectement détectées pour vérifier les sources de confusion du patron.

[122] / = impossible à calculer.

(98) <CP> We **should**, however, also bear in mind, Mr Papayannakis, that we cannot act in an area which will have such a big financial impact, armed only with the measures a presidency can propose in the space of six months.

(99) <CP> Since **1991**, the Community has provided significant financial support to the New Independent States including the countries of central Asia.

(100) <CP> It **must,** however, **be** stressed that the guidelines under discussion are related only to the Structural Funds, **whose** Objectives 1 and 2 specifically adopt the diversification of rural society as a priority.

Le problème n'est pas nouveau, pour ce qui est du patron {post-mod} : la phrase (98) par exemple est particulièrement longue et contient plusieurs segments enchâssés, induisant par-là l'étiquetage catégoriel erroné du verbe lexical, ici *bear*, comme NN alors qu'il aurait dû être étiqueté comme VB. Le deuxième patron {post-card} détecte tout nombre cardinal suivi immédiatement d'une ponctuation (exemple 99). L'exemple (100) a été détecté comme elliptique par deux patrons {post-mod} et {post-wh} malgré l'étiquetage correct de *be* et de *adopt* comme VB. En effet, le patron {post-mod} a détecté la séquence *It must, however,* et s'est arrêté avant *be.* Pour ce qui est de {post-wh}, l'exemple (100) montre que le patron n'a pas été suffisamment affiné puisque *whose* aurait dû être exclu. Ce dernier patron a néanmoins repéré la seule ellipse déclenchée par un pronom *wh–* annotée. Le taux de faux positifs détectés dans ce corpus met en lumière l'une des limites des patrons qui semblent plus adaptés aux phrases courtes ne présentant pas de structures complexes. Il est difficile d'imaginer une amélioration de ces patrons pour prendre en compte les segments enchâssés, puisque l'un des critères essentiels des patrons à base de tokens est la ponctuation. Toute proposition placée entre deux virgules n'est pas forcément une incise (cas de deux phrases parallèles par exemple).

### 1.2. Genre journalistique / Articles de presse écrite

Comme on le sait, les articles de presse sont soumis à des contraintes de format et de temps : un nombre de mots est généralement exigé et il existe une date butoir de remise de l'article. Le code de base de la presse écrite est de répondre aux questions essentielles telles que : qui a fait quoi, où, comment et pourquoi ? Il existe toutefois différents types d'articles de presse et les codes de chacun sont différents. La presse féminine, par exemple, s'écarte des conventions journalistiques habituelles dans ses emplois constants de néologismes[123], entre autres. Un journal comme *Le Monde Diplomatique* au contraire reste, dans la plupart des cas, fidèle à la norme langagière en vigueur. Dans le cas du corpus étudié, les médias (quotidiens et magazines) choisis sont : *The New York Times*, *The Economist*, *National Geographic*, *Le Monde Diplomatique*, *Time Magazine* et *Courrier International*, entre autres. Tous présentent des caractéristiques liées au respect des conventions journalistiques. La langue utilisée est généralement accessible et les structures grammaticales des phrases sont simples. L'ellipse, dans le cas précis de ce corpus, se voit plus fréquemment dans les titres qui, pour des raisons d'*accroche* du lecteur, sont brefs, ne comportant que le(s) mot(s) essentiels (l'effacement du segment *was found* dans *Maggie dead in bed at The Ritz*[124], par exemple)[125]. Dans son étude, Komur (2011, 260) décrit l'ellipse comme stratégie discursive utilisée pour « orienter l'attention du lecteur » :

> L'ellipse fait partie des stratégies discursives que le journaliste emploie consciemment pour orienter l'attention du lecteur. […] L'ellipse participe ainsi à la gestion de l'information et oriente l'attention des co-énonciateurs par isolement d'un constituant. […] En outre, l'ellipse rentabilise la communication grâce à la coopération journaliste-lecteur et apparaît comme un « phénomène dialogal ».

[123] Syntaxiques ou lexicaux.
[124] *The Sun*, 2013.
[125] Dans ce sous-corpus, l'antécédent de l'ellipse détectée dans les titres n'est pas toujours linguistique et requiert une connaissance de la situation d'énonciation afin d'interpréter les segments omis (contexte que l'article permet ensuite de reconstituer).

Par ailleurs, les omissions touchent parfois les verbes, parfois les prépositions. Dans l'exemple ci-dessous (101) extrait de notre corpus journalistique, on observe l'ellipse de *decipher that* déclenchée par le modal *can*.

(101) <CJ> I'm trying to **decipher that**, but I don't know if I **can** Ø.

Même si les patrons appliqués au discours journalistique ont permis de repérer une quantité d'ellipses plus élevée que dans le corpus politique, le nombre d'erreurs reste très important. On observe par exemple 21 faux positifs détectés par le patron {post-*wh*} et 25 par celui du {post-card}. Les erreurs identifiées sont liées aux erreurs d'étiquetage similaires à celles évoquées dans le genre politique (étiquetage du verbe après le pronom *wh*– comme NN). On remarque en revanche que si la précision est, elle, très basse à l'exception de {post-do} qui atteint 0,63, le rappel est élevé puisque les patrons ont récupéré la plupart des ellipses annotées.

| Type | Nb | P | R | F1 | Vrais positifs | Faux positifs | Faux négatifs |
|---|---|---|---|---|---|---|---|
| qs-frag | 2 | 1,00 | 1,00 | 1,00 | 2 | 0 | 0 |
| vs-tag | 0 | / | / | / | / | 1 | 0 |
| post-do | 7 | 0,63 | 0,71 | 0,67 | 5 | 3 | 2 |
| post-mod | 1 | 0,20 | 1,00 | 0,33 | 1 | 4 | 0 |
| post-be/have | 4 | 0,29 | 0,50 | 0,36 | 2 | 5 | 2 |
| post-to | 1 | 1,00 | 1,00 | 1,00 | 1 | 0 | 0 |
| post-wh | 2 | 0,09 | 1,00 | 0,16 | 2 | 21 | 0 |
| post-geni | 0 | / | / | / | 0 | 0 | 0 |
| post-quant | 0 | / | / | / | 0 | 0 | 0 |
| post-card | 1 | 0,04 | 1,00 | 0,07 | 1 | 25 | 0 |
| post-ord | 0 | / | / | / | 0 | 0 | 0 |
| Total | 18 | 0,19 | 0,78 | 0,31 | 14 | 59 | 4 |

**Tableau 10 : Performance des patrons dans l'échantillon journalistique annoté**

Les patrons sont, malgré les erreurs, plus adaptés à ce corpus qu'au corpus politique. Les phrases sont en effet plus courtes et contiennent moins d'incises.

### 1.3. Genre littéraire / Romans

Si nous avons choisi ce genre, c'est en raison de la fréquence de l'ellipse habituellement reconnue comme figure de style dans les textes littéraires. Il ne s'agit pas ici de passer en revue la totalité des genres littéraires, mais dans le cas précis de notre recherche, de nous pencher sur l'étude de textes représentatifs de ce que nous pourrions qualifier d'une langue littéraire « moyenne », au sens où, dans sa facture, elle reste une langue proche de l'usage standard. L'exemple ci-dessous contient une ellipse déclenchée dans la *question tag* repérée dans notre corpus :

(102) <CL> If you knew which was the haunted room you could simply avoid it, **couldn't you Ø?**

Dans ce corpus constitué de romans, la place dévolue à la narration et aux dialogues est d'importance. Les ellipses apparaissent plus généralement dans les dialogues et dans les descriptions. Le tableau (11) ci-dessous montre un taux de rappel relativement élevé pour certains patrons : 0,92 pour {vs-tag}, 0,83 pour {post-do} et 0,80 pour {post-wh}. Nous observons également que, en dépit des faux positifs détectés, les patrons {post-mod}, {post-be/have}, et {post-geni} ont réussi à repérer les ellipses annotées manuellement d'où le taux du rappel élevé.

| Type | Nb | P | R | F1 | Vrais positifs | Faux positifs | Faux négatifs |
|---|---|---|---|---|---|---|---|
| **qs-frag** | 3 | 0,50 | 0,33 | 0,40 | 1 | 1 | 2 |
| **vs-tag** | 13 | 0,71 | 0,92 | 0,80 | 12 | 5 | 1 |
| **post-do** | 6 | 0,45 | 0,83 | 0,59 | 5 | 6 | 1 |
| **post-mod** | 7 | 0,78 | 1,00 | 0,88 | 7 | 2 | 0 |
| **post-be/have** | 6 | 0,26 | 0,83 | 0,40 | 5 | 14 | 1 |
| **post-to** | 3 | 1,00 | 0,33 | 0,50 | 1 | 0 | 2 |
| **post-wh** | 5 | 0,29 | 0,80 | 0,42 | 4 | 10 | 1 |
| **post-geni** | 1 | 1,00 | 1,00 | 1,00 | 1 | 0 | 0 |
| **post-quant** | 0 | / | / | / | 0 | 0 | 0 |
| **post-card** | 1 | 0,00 | 0,00 | / | 0 | 18 | 1 |
| **post-ord** | 0 | / | / | / | 0 | 0 | 0 |
| **Total** | 45 | 0,39 | 0,80 | 0,52 | 36 | 56 | 9 |

**Tableau 11 : Performance des patrons dans l'échantillon littéraire annoté**

Lors de la vérification des résultats, nous avons remarqué que certaines ellipses n'ont pas été détectées dans les corpus PLECI (littéraire et journalistique) en raison

des erreurs dans le corpus d'origine[126]. En effet, l'exemple annoté comme {post-to} ci-dessous (103) contient des guillemets inutiles, étiquetés comme signe de ponctuation, suivis de l'auxiliaire *do* à la forme négative. De ce fait, l'exemple ne remplit pas les conditions combinées dans le patron qui exclut la détection de *to* lorsqu'il est suivi d'une forme verbale.

(103) He's going to " Don't you point that thing at me!

Nous signalons ainsi par là également l'une des limites de la détection à base de TokensRegex qui peut paraitre rédhibitoire dans le sens où les conditions sont restreintes aux tokens dont la moindre variation entrave la reconnaissance de l'ellipse. Dans un corpus de petite taille, cet impact aurait pu être atténué si un nettoyage de corpus avait été effectué, ce qui parait impossible dans un corpus de grande taille (sauf à envisager des procédures automatisées).

### 1.4. Genre promotionnel / *TED talks*

Le terme *promotionnel* peut induire en erreur puisqu'il peut évoquer le type de discours que l'on rencontre dans les textes publicitaires. Nous l'avons adopté pour caractériser le sous-corpus de *TED Talks* parce que les auteurs des conférences qui le constituent sont amenés à faire la promotion du thème et des idées qu'ils défendent, qu'elles soient vulgarisées ou non, et ce quel que soit le thème abordé pouvant relever d'un autre genre (discours scientifique par exemple). Ces conférences présentent la particularité de développer, dans un temps très court (15 à 20 min), un thème illustré par l'apport de chacun des participants. La contrainte de temps oblige à une maîtrise du discours qui perd son caractère spontané et doit être, au contraire, très structuré. Structure et contenu sont ainsi conditionnés par ce rapport au temps afin que la communication soit la plus efficace possible en direction des co-auditeurs et co-locuteurs. On retrouve ici toutes les caractéristiques propres au discours politique, celles d'un discours qui va de l'écrit très codifié et formaté vers la performance orale. S'ajoute à cela le recours à l'humour dont font parfois usage les

[126] Ces erreurs se produisent généralement lors de la saisie ou de la reconnaissance optique de caractères.

présentateurs pour garder l'attention des spectateurs. Dans les techniques employées dans les *TED talks* en vue d'une autopromotion, par exemple, on retrouvera les mêmes procédés que ceux utilisés couramment dans les discours publicitaires. Au regard de la contraction temporelle imposée, l'ellipse narrative est la plus fréquente. Nous passons d'une étape à une autre sans explicitation du processus. Cette ellipse narrative s'apparente à l'ellipse, en particulier temporelle, que l'on rencontre parfois dans les textes littéraires lorsque l'on passe, sans transition d'une période à une autre période éloignée de la précédente et qui a pour fonction de faire progresser la narration. On passe d'un discours écrit à des fins scientifiques au discours écrit, à vocation orale. Le discours que représente ce corpus s'inscrit dans le cadre de vulgarisation où tout style et toute forme de présentations sont autorisés.

Nous observons ci-dessous l'exemple d'une ellipse déclenchée par *do*.

(104) <CPr> And some of it **worked**, and some of it didn't Ø.

On constate dans le tableau ci-dessous (12) que si le rappel est bon pour toutes les ellipses (entre 0,50 et 1, à l'exception de {post-quant} et {post-ord} où le rappel est de 0), la précision des patrons, elle, varie de 0,12 à 1 selon le type d'ellipse recherché.

| Type | Nb | P | R | F1 | Vrais positifs | Faux positifs | Faux négatifs |
|---|---|---|---|---|---|---|---|
| **qs-frag** | 4 | 0,40 | 0,50 | 0,44 | 2 | 3 | 2 |
| **vs-tag** | 8 | 0,60 | 0,75 | 0,67 | 6 | 4 | 2 |
| **post-do** | 10 | 0,67 | 0,80 | 0,73 | 8 | 4 | 2 |
| **post-mod** | 5 | 0,57 | 0,80 | 0,67 | 4 | 3 | 1 |
| **post-be/have** | 9 | 0,33 | 0,56 | 0,42 | 5 | 10 | 4 |
| **post-to** | 2 | 0,67 | 1,00 | 0,80 | 2 | 1 | 0 |
| **post-wh** | 10 | 0,42 | 0,80 | 0,55 | 8 | 11 | 2 |
| **post-geni** | 1 | 0,50 | 1,00 | 0,67 | 1 | 1 | 0 |
| **post-quant** | 1 | / | 0,00 | / | 0 | 0 | 1 |
| **post-card** | 4 | 0,12 | 0,50 | 0,19 | 2 | 15 | 2 |
| **post-ord** | 1 | 0,00 | 0,00 | / | 0 | 0 | 1 |
| **Total** | 55 | 0,42 | 0,69 | 0,52 | 38 | 52 | 17 |

**Tableau 12 : Performance des patrons dans l'échantillon promotionnel annoté**

Le patron {post-card} par exemple a seulement 0,12 de précision en raison des faux positifs détectés et ce malgré la reconnaissance de 2 ellipses sur les 4 annotées. Pour tous les patrons, les erreurs observées appartiennent à la typologie établie précédemment. Par ailleurs, ce genre de discours confirme également que les variations induites par la présence de certains signes de ponctuation et autres caractères spéciaux, erreurs dans le corpus d'origine, entravent le bon fonctionnement des patrons puisque le patron {post-do} a omis de détecter l'ellipse déclenchée par *do* suivis d'un caractère parasite ó, qui remplace vraisemblablement un signe de ponctuation en raison de problèmes d'encodage.

(105) What ó I don't ó 'Petunia?'

### 1.5. Genre conversationnel/ Sous-titres de séries télévisées

Comme on l'a vu dans le chapitre précédent, nous avons choisi, pour représenter la variété standard de la langue et les conversations spontanées, un corpus de sous-titres appartenant à des séries télévisées, tout en restant consciente des limites de ce corpus (en raison notamment des stratégies d'adaptation audio-visuelle, la quantité limitée de signes dans l'espace dévolu aux sous-titres, etc.).

Le tableau (13) ci-dessous montre le résultat de la détection dans 1 000 autres lignes[127] annotées dans le genre conversationnel. Les erreurs identifiées restent les mêmes que celles présentées dans le chapitre 3. Contrairement aux autres genres, et sur la même quantité d'exemples annotés, ce corpus présente davantage d'ellipses. En effet, comme il n'y a que des dialogues et des échanges, et que la narration est rendue par l'image, l'ellipse y est ainsi favorisée.

[127] Nous avons annoté d'autres séries de 1000 lignes que nous n'avons pas exploitées dans le chapitre précédent pour mener une comparaison à d'autres genres sur le même nombre de tokens.

| Type | Nb | P | R | F1 | Vrais positifs | Faux positifs | Faux négatifs |
|---|---|---|---|---|---|---|---|
| qs-frag | 15 | 0,50 | 0,20 | 0,29 | 3 | 3 | 12 |
| vs-tag | 14 | 0,73 | 0,79 | 0,76 | 11 | 4 | 3 |
| post-do | 11 | 0,80 | 0,73 | 0,76 | 8 | 2 | 3 |
| post-mod | 15 | 0,88 | 0,93 | 0,90 | 14 | 2 | 1 |
| post-be/have | 16 | 0,65 | 0,69 | 0,67 | 11 | 6 | 5 |
| post-to | 7 | 0,63 | 0,71 | 0,67 | 5 | 3 | 2 |
| post-wh | 16 | 0,59 | 1,00 | 0,74 | 16 | 11 | 0 |
| post-geni | 1 | 0,50 | 1,00 | 0,67 | 1 | 1 | 0 |
| post-quant | 0 | / | / | / | 0 | 0 | 0 |
| post-card | 2 | 0,67 | 1,00 | 0,80 | 2 | 1 | 0 |
| post-ord | 0 | / | / | / | 0 | 0 | 0 |
| Total | 97 | 0,68 | 0,73 | 0,70 | 71 | 33 | 26 |

**Tableau 13 : Performance des patrons dans l'échantillon conversationnel annoté**

Si la quantité des occurrences détectées et celle des occurrences annotées manuellement semble limitée, il est intéressant de constater que les patrons sont, malgré un certain degré d'imprécision dont nous avons fait état, adaptés à ce genre de corpus. En effet, la majorité des ellipses annotées a été détectée (à l'exception du patron {qs-frag}, dont nous avons signalé la nécessité d'une amélioration, et qui présente seulement 0,2 de rappel).

### 1.6. Synthèse intermédiaire ellipse/genre

En fait, d'un point de vue quantitatif, les taux de précision, du rappel et de la F-mesure varient d'un genre à un autre et d'un type d'ellipse à un autre, comme illustré dans les figures (14), (15) et (16). Pour ce qui est du genre, les trois mesures augmentent lorsque l'on se rapproche de l'oral. Pour ce qui est du type d'ellipse, les ellipses post-auxiliaires sont clairement les plus rencontrées.

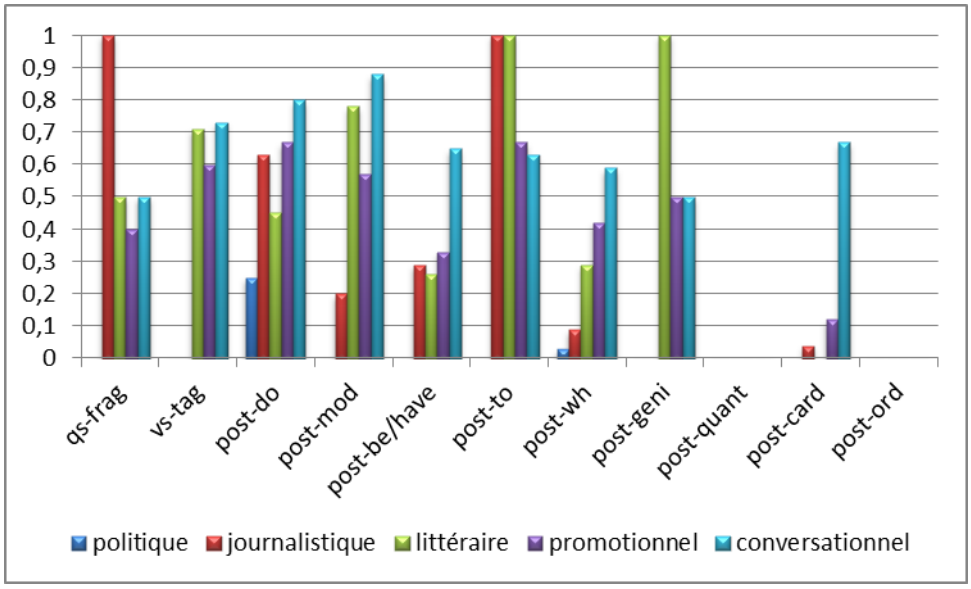



**Figure 14 : Précision par genre et type d'ellipse**

La précision de certains patrons (figure 14) ne figure pas dans le graphique en raison de l'absence des types d'ellipse dans certains genres (par exemple, {post-ord} et {post-quant} seulement présents dans le genre promotionnel) tandis que dans d'autres, elle est très basse notamment pour ce qui est de {post-wh} dans le genre conversationnel et de {post-mod} dans le corpus journalistique en raison, dans ces deux cas particuliers, des incises, à la source d'un étiquetage erroné.

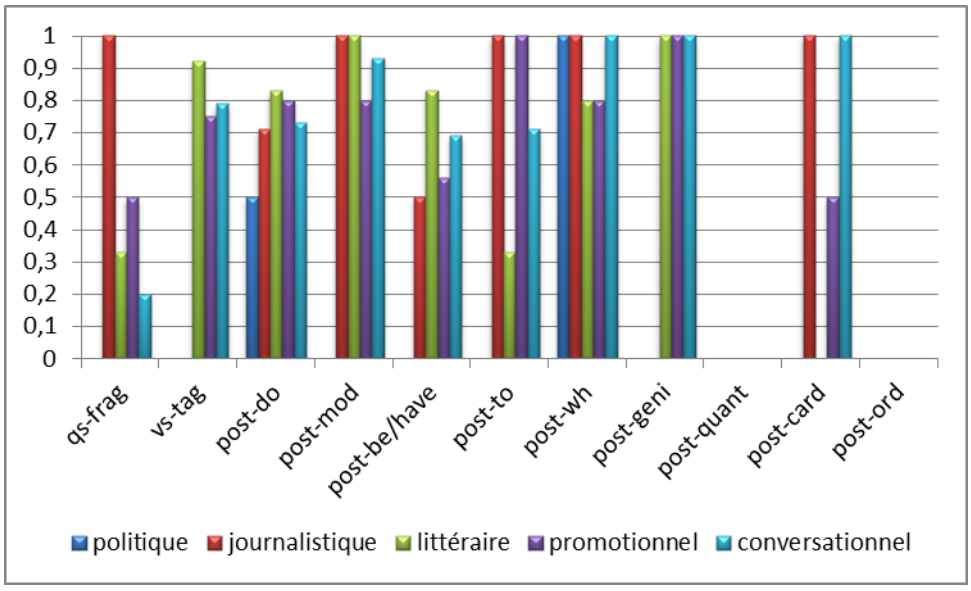



**Figure 15 : Rappel par genre et type d'ellipse**

En revanche, le rappel (figure 15) est relativement élevé pour tous les genres, à l'exception du patron {qs-frag} qui, comme nous l'avons vu, nécessite une amélioration, et de quelques autres patrons dans d'autres sous-corpus.

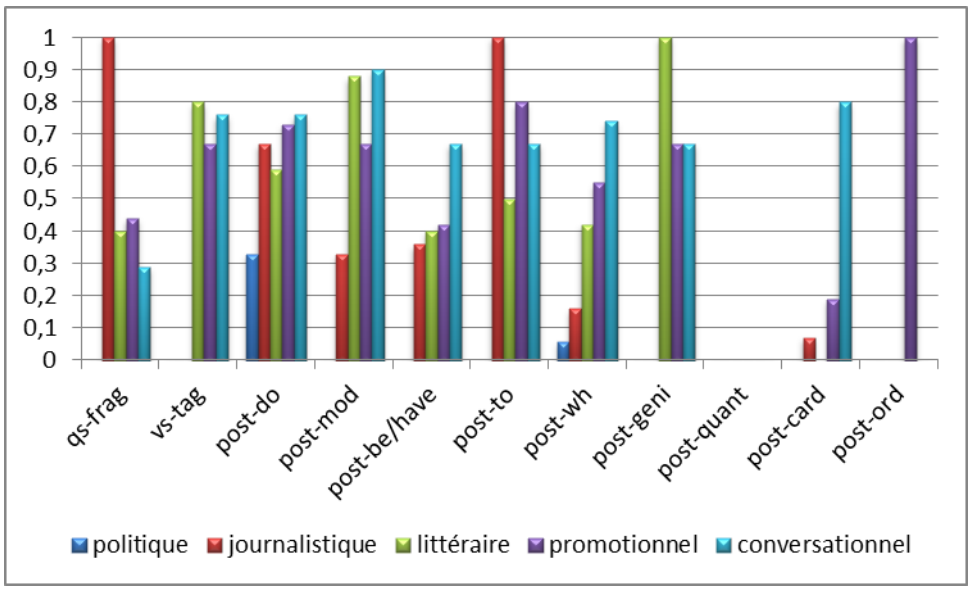



**Figure 16 : F-mesure par genre discursif et type d'ellipse**

La figure (16) synthétise les résultats de la détection et montre le taux de la F-mesure correspondant à la performance des patrons calculée à partir de l'établissement d'une moyenne du taux de précision et du rappel. Comme on le voit, certaines ellipses ne sont pas suffisamment représentées dans les corpus notamment pour ce qui concerne {post-quant} et {post-ord}. On remarque également une F-mesure élevée dans certains genres et dans certains types d'ellipses jusqu'à atteindre le maximum (1,00) comme c'est le cas de {post-geni} dans le genre littéraire et {post-to} ainsi que {qs-frag} dans le genre journalistique. Si ces taux ne sont pas représentatifs, ils permettent de mettre en avant l'efficacité des patrons à repérer les ellipses annotées jusqu'aux plus rares et évitent dans certains genres un nombre important de faux positifs.

À ce stade, le corpus de sous-titres, qui représente le genre conversationnel, se révèle être celui contenant le plus d'ellipses. Partant de notre observation initiale,

considérant l'ellipse comme une propriété du genre conversationnel, l'ellipse peut être également considérée comme stratégie utilisée par les scénaristes lors de la rédaction des scénarios pour imiter les conversations orales spontanées.

Il est vrai que les exemples d'ellipses présentés dans cette section ne manifestent pas vraiment de variation syntaxique lors du passage d'un genre de discours à un autre. En revanche, la variation peut se remarquer dans certaines catégories d'ellipses fréquentes dans un genre de discours spécifique. Prenons l'exemple de la question fragmentaire fréquente dans un corpus conversationnel, extraites de sous-titres de séries. Le sujet et son auxiliaire sont souvent ellipsés ensemble (*Ø Read them all?*) dans un dialogue, ce qui n'est pas le cas du genre journalistique. Ce type d'omission semble être contrôlé par les particularités inhérentes à chaque type de discours. Les participants à la conversation laissent des parties en attente, qu'ils pensent pouvoir reprendre immédiatement en cas d'ambiguïté. Or, dans le corpus journalistique, la plupart des articles sont écrits et se doivent d'être clairs et complets. Les nombreuses *question tags* qui caractérisent le genre conversationnel sont au contraire peu fréquentes dans le corpus journalistique, présentant moins de variation, et, de ce fait, moins de situations d'échanges effectifs.

Nous avons observé que la fréquence absolue des catégories varie ainsi d'un genre à l'autre et que certaines catégories ne sont pas suffisamment représentées. On pourrait alors, à partir de la figure (17), tirer trois observations :

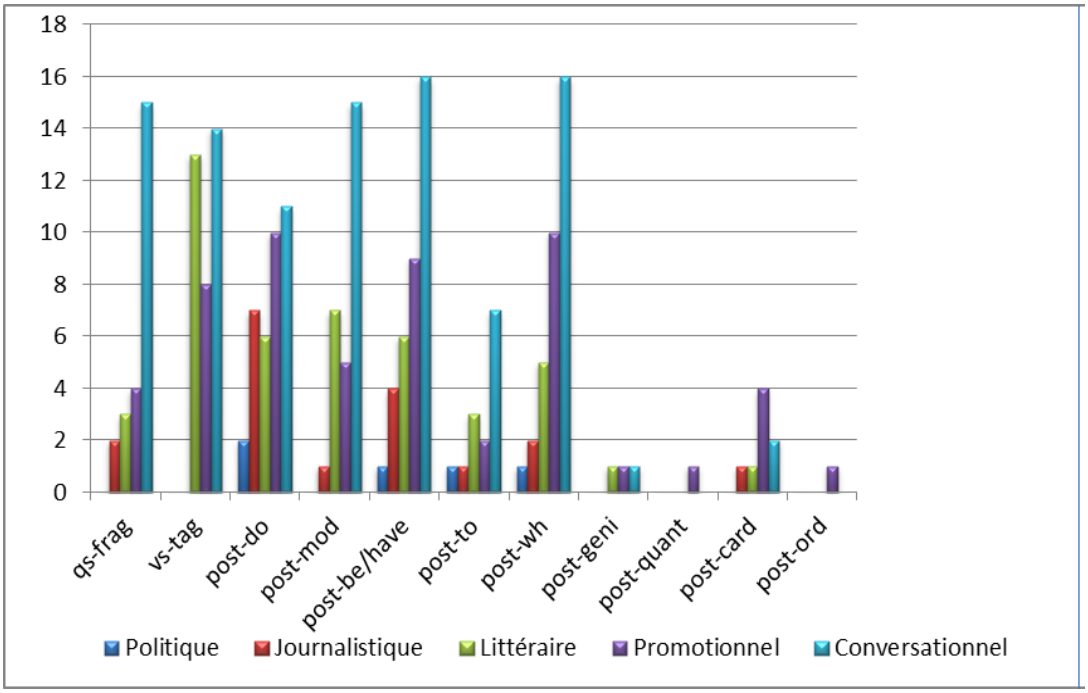



**Figure 17 : Distribution des ellipses par genre**

- parmi les ellipses annotées dans tous les genres analysés, les ellipses post-auxiliaires, les ellipses {post-wh} et {qs-frag} sont les plus représentées ;
- les ellipses nominales ({post-geni}, {post-quant}, {post-card} et {post-ord}, {vs-tag}) sont rares ;
- la distribution des catégories d'ellipses est différente dans les genres analysés, comme on peut le remarquer par exemple dans la figure (18) ci-dessous. Pour le sous-corpus journalistique, {qs-frag} correspond à 11,11% des ellipses repérées dans ce sous-corpus, {vs-tag} à 0%, {post-do} à 38.88% comparé à 0% pour ce qui est des {qs-frag} et {vs-tag}, et 17,85% {post-do} dans le sous-corpus politique.

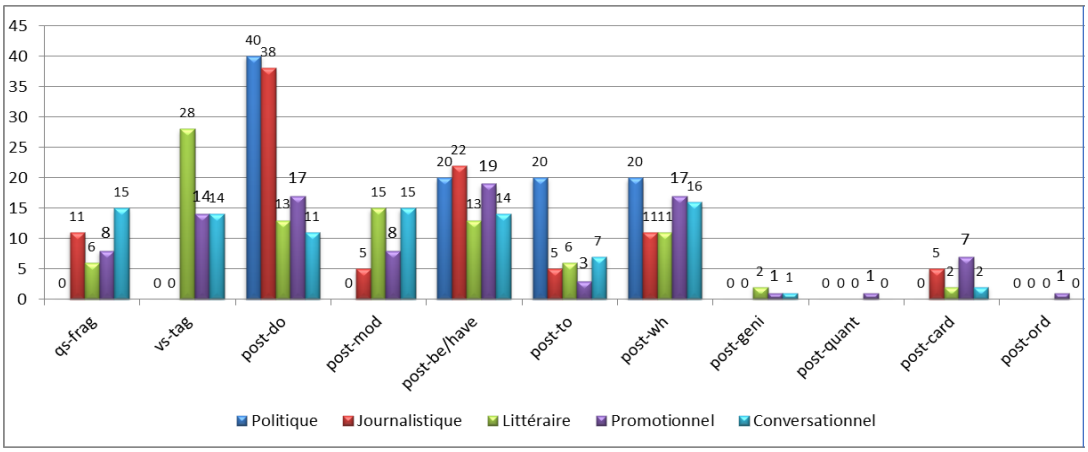



**Figure 18 : Distribution des ellipses par genre (en %)**

Ainsi observons-nous une différence manifeste dans la fréquence d'utilisation des ellipses en fonction du genre. Afin de nous assurer de la validité de cette observation, nous allons utiliser deux tests statistiques dans la partie qui suit pour nous assurer qu'elle n'est pas le fruit du hasard.

## 2. Analyse statistique de la distribution des ellipses dans les différents genres

Pour interpréter les résultats d'une recherche menée sur corpus, il est possible de s'appuyer sur la vérification de l'ensemble du corpus, et/ou de sélectionner des échantillons suffisamment représentatifs du corpus source, pour les soumettre à une évaluation. Lorsque le phénomène étudié est fréquent ou rare dans un corpus de grande taille, vérifier une hypothèse le concernant devient vite difficile à réaliser, sauf à disposer d'importants moyens humains. D'un point de vue statistique, il existe des tests qui permettent d'évaluer la significativité des résultats obtenus et de mesurer la possibilité de les généraliser sur un corpus plus large[128].

Notre analyse utilise le test du $\chi^2$, qui permet de comparer des distributions d'effectifs (Poudat & Landragin 2017, 192-193). L'objectif de ce test est de déterminer si la différence entre deux distributions est statistiquement significative.

[128] Nous remercions Rodrigo Wilkens pour son aide et pour ses conseils sur les questions statistiques ainsi que pour la relecture de ce chapitre.

Il consiste à comparer les effectifs réels observés (dans notre cas, le nombre d'occurrences des ellipses dans chaque genre) à des effectifs attendus sous l'hypothèse d'indépendance (dans notre cas, le nombre d'occurrences des ellipses dans chaque genre, s'il n'y avait aucune association entre le genre et la présence d'ellipses). Pour réaliser le test du $\chi^2$, nous avons utilisé la fonction `chisq.test` du logiciel d'analyse statistique R[129].

On considère l'hypothèse nulle H0 : la présence d'ellipses est indépendante du genre de texte. Si l'hypothèse nulle est rejetée, on peut considérer que la présence d'ellipses n'est pas indépendante du genre de texte. C'est ce que l'on appelle l'hypothèse alternative H1 : il y a une dépendance entre les ellipses et le genre.

Le résultat du test du $\chi^2$ nous donne une *p-value*, qui correspond à la probabilité que le hasard puisse expliquer à lui seul une différence au moins aussi importante que celle qui a été constatée entre nos observations. Cette *p-value* sert donc à déterminer la significativité. On utilise généralement un seuil assez bas de 0,05 ou 0,01, voire 0,005 ou 0,001. On peut rejeter l'hypothèse nulle H0 (et donc considérer que la différence est statistiquement significative) si la *p-value* se situe en-deçà de ce seuil. Dans le cas contraire, on doit considérer H0 comme possible.

### 2.1. Test 1

Notre premier test vise à vérifier s'il existe une association entre la présence ou l'absence d'ellipses et le genre analysé. Pour ce faire, nous comparons le nombre de phrases annotées contenant au moins une ellipse au nombre de phrases sans ellipse dans chaque sous-corpus :

| | Conversationnel | Littéraire | Journalistique | Promotionnel | Politique |
|---|---|---|---|---|---|
| **Présence d'ellipses** | 94 | 39 | 18 | 55 | 5 |
| **Absence d'ellipse** | 906 | 961 | 982 | 945 | 995 |

**Tableau 14 : Présence / absence des ellipses dans chaque corpus**

[129] https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/chisq.test (consulté le 23 juillet 2019 à 21:10).

Les effectifs attendus dans les différents genres sont les mêmes (44,2 pour ce qui concerne la présence d'ellipses et 955,8 pour ce qui concerne leur absence), puisque le même nombre de phrases a été annoté dans chaque cas (1000 phrases).

La *p-value* obtenue est inférieure à 2.2e-16, donc largement inférieure aux seuils de 0,01 ou 0,05 : on peut donc rejeter l'hypothèse nulle H0 et considérer que la présence ou non d'ellipses est bien liée au genre du corpus considéré.

Pour pousser plus avant cette analyse, nous analysons les résidus de Pearson (différence entre l'effectif observé et l'effectif attendu, divisée par la racine carrée de l'effectif attendu), afin de visualiser les informations qui contribuent le plus au $\chi^2$ global. Les résidus de Pearson sont présentés dans le tableau ci-dessous :

| | Conversationnel | Littéraire | Journalistique | Promotionnel | Politique |
|---|---|---|---|---|---|
| **Présence d'ellipses** | 7,974 | 0,493 | -3,725 | 1,970 | -5,726 |
| **Absence d'ellipse** | -1,674 | -0,103 | 0,782 | -0,414 | 1,202 |

**Tableau 15 : Résidus de Pearson**

Nous pouvons aussi visualiser ces valeurs sous forme graphique[130] dans le corrélogramme[131] ci-dessous :

[130] La présentation de ces valeurs en tableau et en graphique permet une meilleure visualisation.

[131] Visualisation disponible dans la librairie R corrplot qui « représente le graphique d'une matrice de corrélation […] La matrice de corrélation peut être […] réordonnée en fonction du degré de corrélation entre les variables ». Cette même visualisation peut être utilisée pour d'autres types de matrice, ce qui est notre cas ici.
http://www.sthda.com/french/wiki/visualiser-une-matrice-de-correlation-par-un-correlogramme (consulté le 15 juillet 2019 à 15:03).

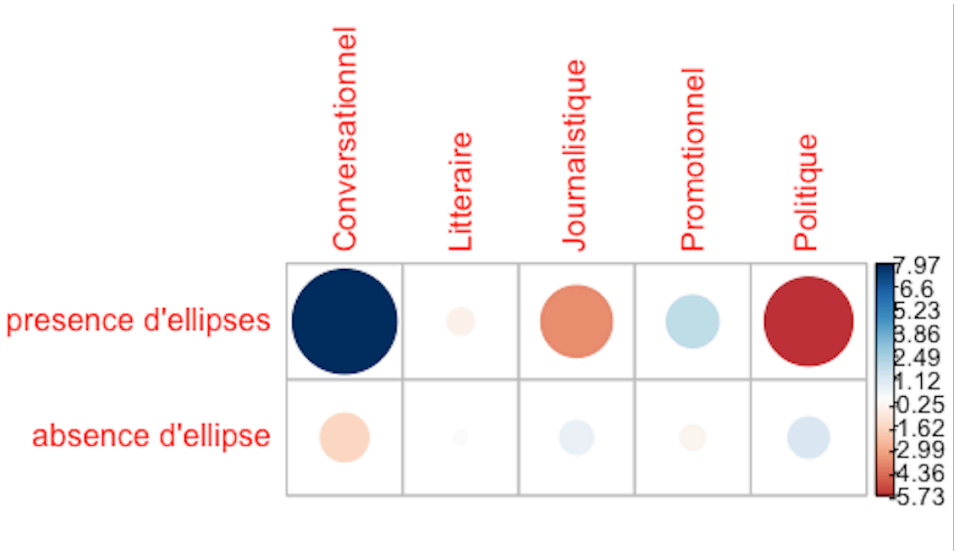



**Figure 19 : Corrélogramme des valeurs résidus de Pearson**

Les résidus positifs sont en bleu et indiquent une association *positive* entre la ligne et la colonne correspondante. On constate ici une forte association positive entre la présence d'ellipses et le genre conversationnel, ce qui est conforme à nos observations.

Les résidus négatifs sont en rouge et indiquent une association *négative* entre la ligne et la colonne correspondante. On constate ici une forte association négative entre la présence d'ellipses et le genre politique (-5,726) et, dans une moindre mesure, le genre journalistique. Ceci confirme nos observations sur ces deux derniers genres, dans lesquels l'ellipse reste un phénomène très rare.

Ce que nous pouvons interpréter à la suite de ces résultats est le fait que la présence d'ellipses est une caractéristique du genre conversationnel tandis que leur très faible nombre d'occurrences caractérise les genres journalistique et politique. Le genre littéraire, quant à lui, ne présente pas de différences marquantes par rapport à ces derniers. Par ailleurs, dans la mesure où les styles des romans varient[132], les données recueillies à partir de ce corpus sont hétérogènes.

[132] Il convient d'ajouter que le style des auteurs, qu'ils soient hommes politiques, écrivains, journalistes, conférenciers ou scénaristes, n'a pas été contrôlé en amont, ce qui remet peut-être en question l'équilibre des échantillons, et, par conséquent, la représentativité des résultats.

## 2.2. Test 2

Le deuxième test que nous avons effectué vise à vérifier s'il existe une différence entre les types d'ellipse en fonction du genre. En d'autres termes, la manifestation des catégories d'ellipses diffère-t-elle en fonction du genre analysé ?

À cette fin, comparons le nombre d'occurrences de chaque type d'ellipse en fonction du genre (à titre de rappel, ci-dessous le nombre d'ellipses annotées) :

| Type/Genre | Conversationnel | Littéraire | Journalistique | Promotionnel | Politique |
|---|---|---|---|---|---|
| qs-frag | 15 | 3 | 2 | 4 | 0 |
| vs-tag | 14 | 13 | 0 | 8 | 0 |
| post-do | 11 | 6 | 7 | 10 | 2 |
| post-mod | 15 | 7 | 1 | 5 | 0 |
| post-be/have | 16 | 6 | 4 | 9 | 1 |
| post-to | 7 | 3 | 1 | 2 | 1 |
| post-wh | 16 | 5 | 2 | 10 | 1 |
| post-geni | 1 | 1 | 0 | 1 | 0 |
| post-quant | 0 | 0 | 0 | 1 | 0 |
| post-card | 2 | 1 | 1 | 4 | 0 |
| post-ord | 0 | 0 | 0 | 1 | 0 |
| Total | 97 | 45 | 18 | 55 | 5 |

Après vérification, de nombreux effectifs attendus sont très faibles et inférieurs à 5, notamment pour les ellipses rares ({post-to}, {post-geni}, {post-quant}, {post-card}, {post-ord}), et pour certains genres (journalistique et politique), ce qui empêche l'utilisation du test du $\chi^2$. Avant de poursuivre, il convient de souligner que l'utilisation du test $\chi^2$ pourrait éventuellement être possible si l'on annotait davantage de données, ce qui pourrait augmenter les effectifs attendus et rendre les résultats exploitables.

Pour exploiter les données à disposition, nous avons remplacé $\chi^2$ par le test exact de Fisher (fonction `fisher.test` dans R) dont l'utilisation donne une *p.value* aux alentours de 0.34[133] : on ne peut donc pas rejeter l'hypothèse nulle H0 selon laquelle il n'y a pas de différence significative dans la distribution des types d'ellipses dans les différents genres. Pour visualiser ce résultat, nous utilisons le diagramme en

[133] Nous avons utilisé la simulation Monte Carlo pour calculer les *p.values*. Nous indiquons une valeur approximative, celle-ci variant légèrement à chaque essai.

mosaïque qui est « une représentation de la distribution marginale du tableau croisé » (Le Guen 2003, 10). La surface des mosaïques y est proportionnelle aux effectifs observés. Le diagramme en mosaïque ci-dessous ne montre que peu d'informations intéressantes. Seuls les effectifs de la catégorie {vs-tag} dans le genre littéraire, et celui de {post-do} dans le genre journalistique sont plus importants qu'attendus.

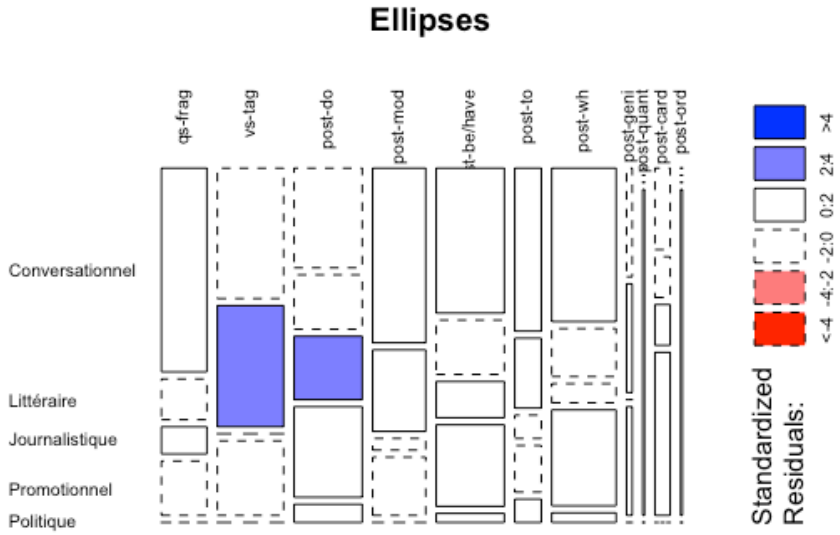



**Figure 20 : Mosaïque en diagramme pour comparer les types d'ellipse**

Nous avons poussé plus avant cette investigation en nous focalisant sur les 3 genres contenant le plus d'ellipses (conversationnel, littéraire et promotionnel) afin de vérifier s'il existe une différence entre la distribution des ellipses à l'intérieur de ces différents genres et ainsi caractériser plus finement chaque genre.

Par ailleurs, pour éviter les effectifs attendus inférieurs à 5, nous regroupons, dans la table ci-dessous, le nombre d'occurrences des ellipses {post-to}, {post-geni}, {post-quant}, {post-card} et {post-ord} sous l'appellation « rare ».

| | Conversationnel | Littéraire | Promotionnel |
|---|---|---|---|
| **qs-frag** | 15 | 3 | 4 |
| **vs-tag** | 14 | 13 | 8 |
| **post-do** | 11 | 6 | 10 |
| **post-mod** | 15 | 7 | 5 |
| **post-be/have** | 16 | 6 | 9 |
| **post-wh** | 16 | 5 | 10 |
| **rare** | 10 | 5 | 9 |

**Tableau 16 : Nombre d'ellipses après regroupement des ellipses rares**

Le test du $\chi^2$ ne permet pas de rejeter l'hypothèse nulle ($p = 0.4618802$). De la même manière, la comparaison 2 à 2 entre les divers genres ne fait pas apparaître de différences statistiquement significatives. On ne peut donc pas conclure définitivement, en l'état actuel de notre recherche, de l'utilisation de manière préférentielle de certains types d'ellipses dans certains genres.

Enfin, nous présentons les « boîtes à moustaches »[134] qui sont utiles pour représenter la distribution de variables aléatoires quantitatives[135] (dans notre cas, le nombre d'occurrences des ellipses). La figure (21) présente le nombre d'occurrences des ellipses par genre et la figure (22) le nombre d'occurrences par type d'ellipse. À l'intérieur de chaque boîte se trouvent 50% des données, avec une barre horizontale qui représente la médiane. 25% des données se trouvent donc au-dessus et aussi en-dessous de cette boîte. Enfin, les *outliers* sont représentés par un point en-dehors des « moustaches » : c'est le cas par exemple du point pour le genre journalistique qui correspond à l'ellipse {post-do} ou celui au-dessus de la boîte {qs-frag} correspondant au genre conversationnel (figures 21 et 22).

[134] « La boîte à moustaches, une traduction de *Box & Whiskers Plot*, est une invention de TUKEY (1977) pour représenter schématiquement une distribution. Cette représentation graphique peut être un moyen pour approcher les concepts abstraits de la statistique » (Le Guen 2001, 1).

[135] « Une variable aléatoire quantitative porte sur des grandeurs non numériques. L'étude de ce type de variable s'effectue par un tableau de dénombrement en donnée brute ou en pourcentage. » (Bourry & Saulnier 2009, cours en ligne) http://unt-ori2.crihan.fr/unspf/2009_Angers_Boury_stats/co/Patrick_Saulnier_web.html (consulté le 16 juillet 2019 à 17:59).

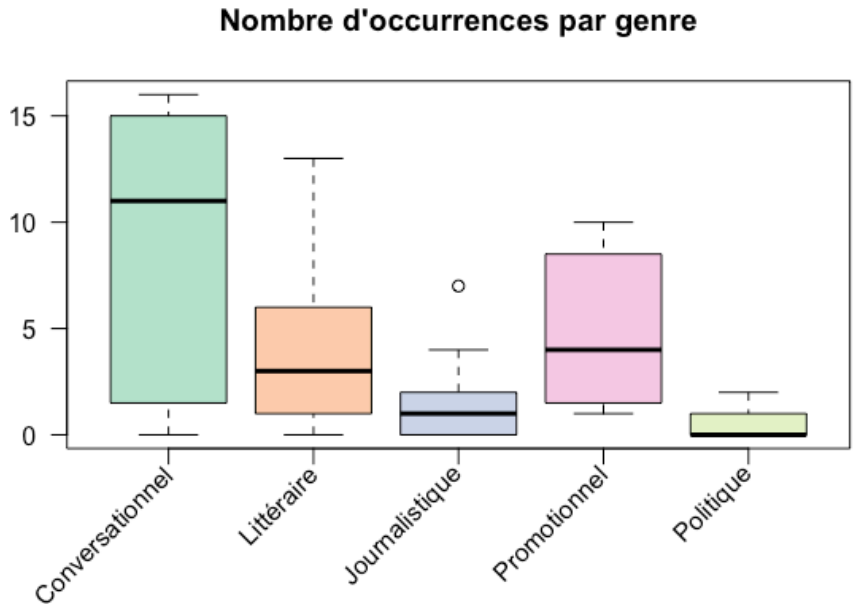



**Figure 21 : Distribution du nombre d'occurrences par genre**

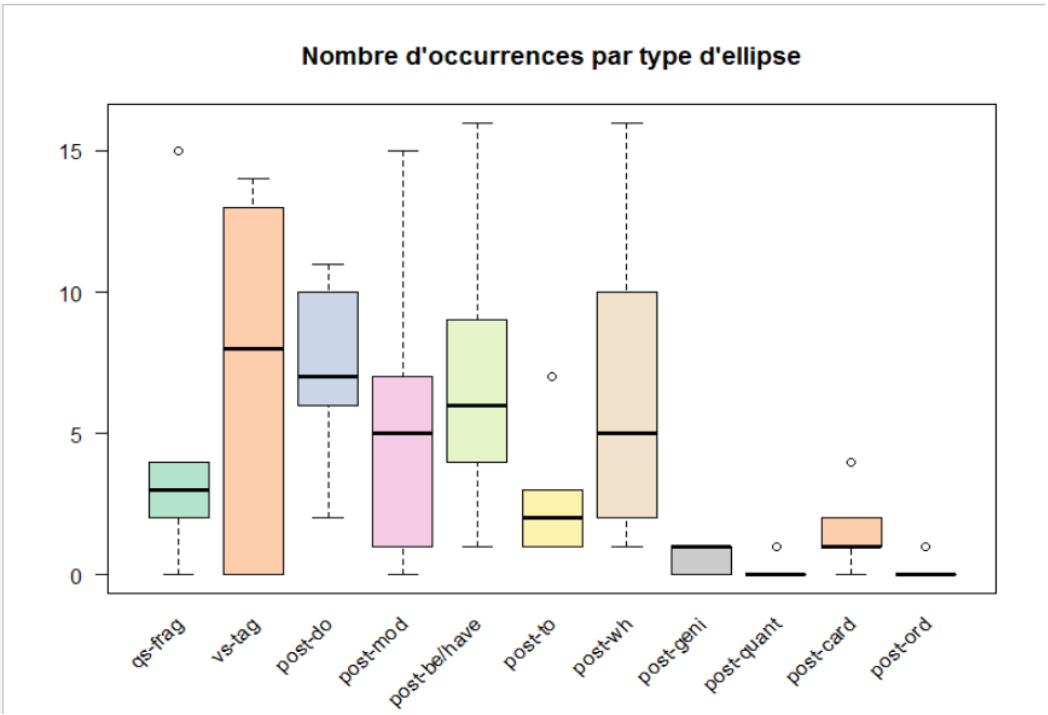



**Figure 22 : Distribution du nombre d'occurrences par type d'ellipse**

Afin de permettre une autre visualisation des résultats, les graphiques (23) et (24) qui suivent reprennent les mêmes informations et affichent en plus les niveaux de significativité pour la comparaison des moyennes[136]. Les moyennes de chaque groupe (types d'ellipse ou genre) sont comparées à la moyenne de l'ensemble des groupes à l'aide d'un test de Wilcoxon. Les niveaux de significativité sont indiqués par des caractères * ou ns :

- ns : p > 0.05 (donc non significatifs)
- *: p <= 0.05
- **: p <= 0.01

La ligne en pointillés indique la fréquence moyenne (4).

Dans le graphique ci-dessous, seules les ellipses très rares {post-ord} et {post-quant} se distinguent significativement des autres (*: p <= 0.05).

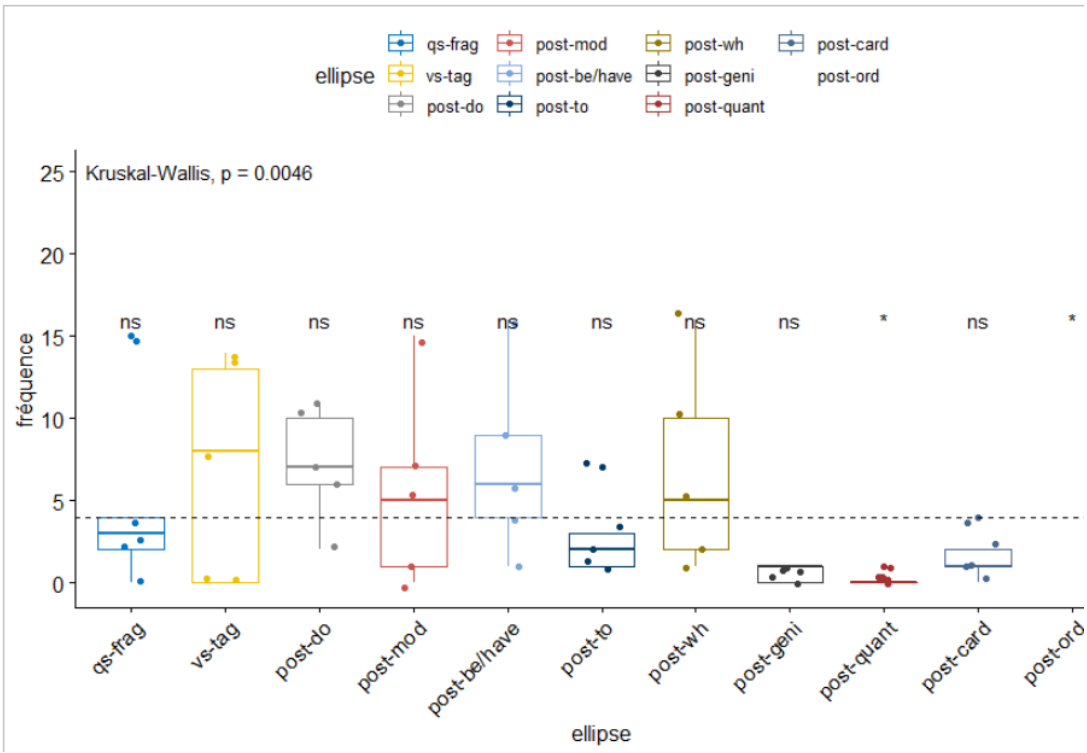



**Figure 23 : Distribution du nombre d'ellipses par type avec niveaux de significativité**

[136] Ces figures ont été obtenues à l'aide de la librairie ggpubr de R.

Pour ce qui est des genres, ce sont les genres conversationnel et politique qui se distinguent significativement des autres. Ce résultat rejoint les observations faites précédemment sur la présence d'ellipses dans le genre conversationnel et leur rareté dans le genre politique.

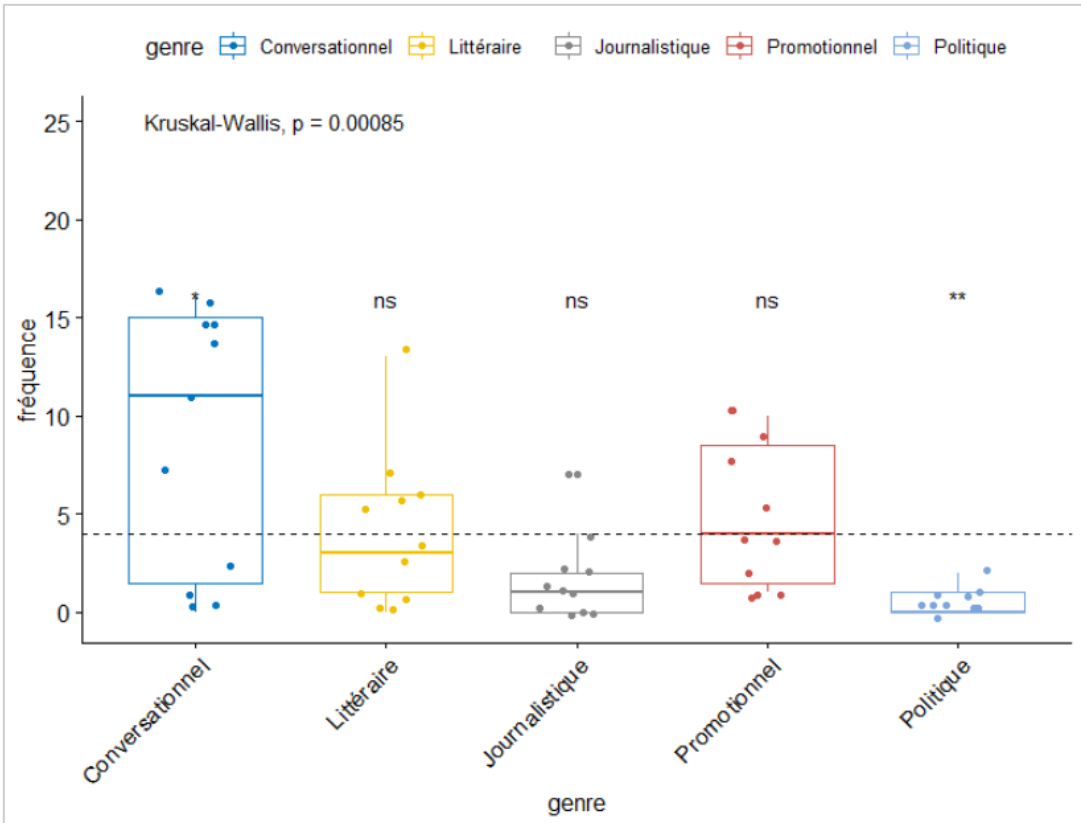



**Figure 24 : Distribution des ellipses par genre avec niveaux de significativité**

## 3. Bilan sur la distribution des ellipses dans les genres

Nous avons exposé dans ce chapitre les résultats de la détection automatique, effectuée à l'aide de patrons décrits dans le chapitre précédent, dans un corpus *genré* en illustrant les particularités caractérisant chaque échantillon exploité, caractéristiques que nous considérons d'une part à l'origine de la fréquence ou de la rareté des phénomènes elliptiques dans les genres et, d'autre part, à l'origine des erreurs produites par les patrons qui peuvent parfois être inadaptés à l'un ou l'autre genre. Nous avons ensuite introduit, à l'aide de deux tests $\chi^2$ et Fisher, quelques

vérifications statistiques pour caractériser la distribution des ellipses dans les échantillons analysés. Globalement, les ellipses {post-do}, {post-be/have}, {post-mod}, et {post-to}, {vs-tag} (ellipses post-auxiliaires) sont plus fréquentes tandis que les ellipses nominales apparaissent rares ({post-geni}, {post-quant}, {post-ord} et {post-card}). La taille du corpus, la *p-value* et les résultats obtenus statistiquement ne permettent pas de conclure en faveur d'une distribution différente, voire marquante, des catégories d'ellipses en fonction du genre. Les données peuvent néanmoins confirmer que l'absence ou la présence de l'ellipse peut être une caractéristique d'un genre spécifique. Ces analyses statistiques nous permettent également de mettre en avant les limites d'une analyse menée sur les genres dans le cas précis de l'ellipse. Comme nous le disions précédemment, nous avons pu remarquer, entre autres observations, que l'ellipse est plus fréquente dans le genre conversationnel que dans le genre politique et que, de la même manière, elle apparait plus souvent dans le genre littéraire que dans le genre journalistique, notre hypothèse étant que cette variation dépend alors directement des paramètres discursifs externes au phénomène même. En effet, le genre conversationnel, en particulier les sous-titres que nous traitons, correspond plus à un oral spontané où l'on s'attend presque naturellement à trouver des ellipses, tandis qu'à l'inverse, le genre littéraire dépendant de l'écrit, est supposé en présenter moins, sauf à relever de choix stylistiques, et, dans une moindre mesure, lors de transcriptions de dialogues.

L'exploitation d'un corpus de transcriptions orales pour évaluer la distribution des ellipses dans des registres de langues relâchés (souvent dans les conversations spontanées) aurait peut-être donné la possibilité de vérifier des paramètres que nous n'avons pas pu traiter ici. Cependant, comme nous l'avons expliqué dans le chapitre 3 consacré à la méthodologie, notre projet d'évaluer la traduction automatique en la comparant à la traduction humaine de l'ellipse a motivé le choix d'un corpus parallèle, à la place d'un corpus de transcriptions orales qui, par ailleurs et sauf erreur de notre part, n'existe pas sous forme de corpus parallèle anglais-français.

## 4. Bilan de la méthode appliquée à la détection et à la distribution des ellipses

À l'issue de cette présentation, il nous semble nécessaire de revenir sur certaines failles de notre approche expérimentale. La complexité du phénomène elliptique montre qu'un seul outil ne suffit pas à le traiter. Le seul axe linguistique (morphosyntaxique) n'est pas à même de le représenter entièrement et un corpus unique ne peut suffire à recenser toutes ses variations. La difficulté d'élaboration et d'écriture de patrons peut se solder par des erreurs dans les résultats obtenus automatiquement. Malgré les erreurs recensées et les difficultés rencontrées, notre démarche permet une méthode de détection simple à base de tokens et montre un taux de rappel élevé dans la plupart des catégories envisagées, ce qui met en avant l'adaptabilité des patrons aux genres sélectionnés. Par ailleurs, elle peut faciliter le repérage préliminaire des catégories elliptiques, étape que nous pourrons réaliser sur des corpus larges qui ne peuvent être exploités manuellement. Une vérification est ensuite nécessaire pour poursuivre les objectifs fixés.

Force est de constater que plus on croit cerner le phénomène elliptique, plus il résiste. Avec l'avancée des recherches dans le domaine de l'intelligence artificielle, il n'est pas impossible d'imaginer une meilleure reconnaissance et un traitement plus performant du phénomène elliptique dans son ensemble. En effet, à l'aide de plusieurs outils informatiques, annoter un corpus de développement de taille étendue, mettant en relief autant de variations véhiculées par une ellipse que possible, pourrait aider à entrainer des systèmes d'apprentissage automatique. Une approche à base d'apprentissage supervisé est concevable pour la classification des ellipses détectées par les patrons actuels en bons ou mauvais candidats. Ce travail pourrait se faire grâce aux annotations manuelles réalisées dans notre corpus et pourrait peut-être améliorer la précision et le rappel.

Pour faire suite à la présentation de la méthode de détection automatique que nous avons mise en place et appliquée au corpus *genré,* il nous faudrait dans un premier temps examiner les éventuelles retombées de cette tâche qui n'est en fait

qu'une étape préliminaire. Ainsi, à l'issue de notre démarche de détection, deux orientations nous ont semblé particulièrement pertinentes à suivre.

La première concerne la classification automatique des genres de discours, hypothèse que nous avons brièvement présentée dans les chapitres précédents. En effet, nous pensons que la détection automatique des ellipses, si l'on parvient à l'améliorer encore, peut également contribuer à la reconnaissance automatique du genre de discours dans lequel elles apparaissent, et donc à sa classification. À l'avenir, cette détection pourrait être effectuée à l'aide d'un apprentissage automatique suivant trois étapes importantes présentées ci-dessous :

- Identification des ellipses les plus courantes dans tel ou tel genre.
- Élargissement de la liste des genres à exploiter (médical, poésie, théâtre, autobiographie, conte, ...)
- Annotation des données d'entrainement en vue d'un apprentissage automatique pour la classification des genres selon la fréquence de l'une ou l'autre catégorie d'ellipse.

À partir de ce travail, et en fonction de la fréquence du phénomène, il s'agit d'attribuer son genre à chaque discours. Afin d'éviter de fausses classifications, l'apprentissage proposé ici n'est possible que si la détection automatique des ellipses montre un taux d'erreurs réduit grâce à une précision plus élevée.

La deuxième perspective est celle qui a motivé notre exploration d'une détection automatique de l'ellipse à savoir, la possibilité de traduire automatiquement des phrases elliptiques avec le moins d'erreurs possible. En effet, avec l'évolution de l'intelligence artificielle (IA) et le développement des outils de traitement automatique des langues, la place et le rôle du linguiste dans les études menées en sciences du langage, ou ceux du traducteur, se sont trouvés modifiés. Le recours à l'humain dans le passage d'une langue à une autre reste encore nécessaire pour tout ce qui relève des relations contextuelles, des figures de styles comme la métaphore,

ou des faits de langue comme l'ellipse. Les réflexions de François Yvon sur le phénomène illustrent ce constat[137] :

> La machine ne dispose pas de cette connaissance d'arrière-plan, ce qui rend la compréhension complète de la majorité des énoncés difficile, voire impossible, si l'on ne dispose pas de bases de connaissance additionnelles, donnant accès à la fois à un savoir sur le monde (ou le domaine) en général (connaissance statique) et sur le contexte de l'énonciation (connaissance dynamique). […] En l'absence de telles connaissances, bien d'autres problèmes de compréhension deviennent pratiquement insurmontables : pensez par exemple aux ellipses, aux métaphores, et, plus généralement, aux figures de style.

En effet, si le domaine de la TA ne cesse de progresser, notamment avec le développement de la traduction dite neuronale[138], ces progrès considérables apparaissent encore insuffisants pour traiter des phénomènes linguistiques tels que l'anaphore, la référence, où des éléments discursifs, non encore pris en charge en TAL, nécessitent une analyse *au-delà* de la phrase, ce qu'avaient montré les travaux de Benveniste. L'ellipse, bien sûr, ne fait pas exception d'autant plus qu'elle ajoute aux difficultés recensées dans la traduction du *dit* de la séquence non-elliptique, celles encore plus complexes du silence elliptique véhiculé par le dit *non représenté*.

Dans le domaine de la traduction automatique, l'absence d'un segment quelconque exprimée par un phénomène langagier attesté semble rendre sa traduction impossible. C'est ainsi que, suite à la l'observation des erreurs relevées dans la traduction de certaines catégories d'ellipse, notamment des ellipses post-auxiliaires, il nous est apparu qu'en tant qu'étape préliminaire à la traduction, la détection automatique du phénomène pourrait favoriser une réduction de ces erreurs et mettre en évidence des phrases potentiellement problématiques, là où l'ellipse est particulièrement source de confusions. Il nous reste à ajouter que, dans la mesure où chacun sait qu'on traduit des textes écrits, tandis qu'on interprète des conversations et des échanges dans des situations de « communication sociale », il

[137] Cours de François Yvon https://perso.limsi.fr/anne/coursM2R/intro.pdf (consulté le 12 octobre 2017 à 17:30).
[138] Nouveau paradigme de la traduction automatique. Voir annexe IV p.269.

peut paraitre paradoxal d'exposer les résultats d'une détection automatique des ellipses dans des corpus de genre conversationnel, relevant de l'oral spontané, pour ensuite traiter des erreurs générées dans leur traduction automatique, relevant logiquement de l'écrit. Or, notre corpus de genre conversationnel constitué de sous-titres est très particulier. Il s'agit de capter l'oral spontané et de le restituer par l'intermédiaire d'une version propre à être lue à un rythme aussi rapide que le perçoit l'oreille. Nous nous trouvons à mi-chemin : encore très proche de l'oral, mais pas encore approchant d'un écrit normé. Les trois étapes nécessaires à la production des sous-titres sont ainsi les suivantes :

- une interprétation/compréhension des échanges oraux entre les personnages,
- une transcription « économique » des échanges en tenant compte des contraintes spatiales inhérentes au format technique devant accueillir les sous-titres,
- une traduction de cette transcription pour transférer les sous-titres dans la langue cible choisie.

Ces étapes révèlent alors que la traduction de certaines occurrences orales n'est pas directement interprétée, mais s'inscrit dans la même démarche que celle adoptée pour la traduction d'un écrit quelconque. Par conséquent, dans le chapitre qui suit, l'objectif d'analyser les erreurs de la traduction automatique de l'ellipse appartenant à un corpus de genre conversationnel trouve ici pleinement sa justification puisqu'il s'agit d'une approche transférable et pouvant être appliquée à d'autres situations que celles explorées dans le présent travail.

# Chapitre 5

# Étude de la traduction automatique de l'ellipse post-auxiliaire

*Enjeu politique et culturel, linguistique et stylistique, la figure de style impose ainsi au traducteur de retrouver les processus qui la motivent, elle pose la question des équivalences formelles d'une langue à une autre, elle autonomise chaque effort expressif en sa singularité et libère enfin le potentiel d'invention de chacune des langues invitant in fine le lecteur à tracer dans le texte le chemin de sa propre interprétation.*

*Maryvonne Boisseau*[139]

Nous avons précédemment mis en avant les difficultés rencontrées lors de la détection automatique de l'ellipse et les résultats obtenus à partir d'une méthodologie de détection à base de tokens se sont avérés prometteurs en montrant que certaines catégories d'ellipses posaient moins de problèmes que d'autres. Nous abordons enfin, dans ce dernier chapitre, la question qui nous a préoccupée tout au long de ce travail conjointement à celle de la détection, puisque nous avons fait l'hypothèse qu'une détection préalable des ellipses était utile à la traduction automatique. Nous avons aussi suggéré qu'une détection efficace serait de nature à limiter les erreurs qui subsistent inévitablement après la traduction automatique d'un texte et, en l'occurrence, de phrases elliptiques. En effet, dépassant la question naïve qui consiste à se demander comment un système de traduction automatique peut traduire un segment de texte absent, nous pensons qu'une traduction automatique des ellipses dans des genres de textes particuliers est possible, et ce malgré les limites de la détection automatique à base de tokens telle que nous l'avons expérimentée et expliquée dans les chapitres précédents. Sans doute, faut-il, pour

[139] Boisseau (2011), voir la bibliographie pour la référence complète.

cela, parvenir à favoriser l'apprentissage automatique des systèmes pour les rendre capables de traduire l'ellipse détectée.

Considérant que notre travail préliminaire sur la détection automatique des ellipses pourra aboutir à leur reconnaissance systématique, et constatant par ailleurs que leur traduction pose encore problème, pouvons-nous envisager que la phase de détection puisse avoir un impact sur la traduction automatique de ces mêmes ellipses ? Plus précisément, pouvons-nous utiliser la détection automatique pour produire des données d'entraînement ? Nous espérons proposer des amorces de réponses à ces questions dans ce chapitre.

Nous poursuivons dans ce chapitre notre démarche « en mouvement » et partons de l'observation empirique d'exemples sélectionnés en comparant des traductions, de l'anglais vers le français, effectuées par des systèmes de traduction (Reverso, Systran, Google Traduction et DeepL[140]) avec celles produites par le traducteur humain, utilisées comme références (désormais notées TR dans nos exemples et tableaux), pour essayer de relever des récurrences d'erreurs, en particulier dans le cas d'ellipses. Ce relevé d'erreurs, pensons-nous, pourrait servir à construire un protocole, réunissant les procédures de détection automatique améliorée et les erreurs de traduction à éviter dans le cas de l'ellipse, afin de produire une traduction automatique *acceptable*. Pour ce faire, nous avons choisi de nous consacrer à l'analyse d'exemples présentant des ellipses post-auxiliaires qui introduisent de réelles ambiguïtés aussi bien dans leur détection automatique (voir chapitre 3) que dans leur traduction humaine. Au sein de l'ellipse post-auxiliaire, nous présentons les erreurs en fonction de l'élément déclencheur de l'ellipse (selon la classification suivie pour leur détection) *post-do*, *post-be/have*, *post-mod* et *post-to* avec une attention particulière portée à leur manifestation dans la *question tag* (catégorie que nous avons établie afin de faciliter la détection de l'ellipse).

Pour compléter notre analyse et pour initier nos investigations sur la traduction de l'ellipse, nous aborderons quelques exemples dans les questions et les réponses fragmentaires, ellipses dont la détection n'a pas atteint un taux de précision élevée.

[140] Une présentation brève de ces systèmes est donnée en annexe V p. 273.

Nous présentons les erreurs en fonction de l'élément déclencheur pour éviter une présentation *a priori* d'exemples (c'est-à-dire non classés synthétiquement) et pour mieux décrire notre démarche qui consiste à examiner d'abord, pour aboutir ensuite à des généralités potentiellement concluantes qui puissent être soumises à une validation ultérieure.

Un relevé des erreurs rencontrées par les systèmes de traduction et une synthèse des différentes stratégies utilisées par les traducteurs humains pour traduire l'ellipse seront présentées à l'issue des analyses. Nous proposerons, en guise de conclusion, quelques réflexions quant à la validité des observations faites dans ce chapitre en mettant en avant le lien étroit entre détection et traduction. Ceci nous permettra d'examiner à nouveau les notions d'*acceptabilité* et d'*ambiguïté*, notions-clés en traductologie.

## 1. La traduction de l'ellipse post-auxiliaire

Ni les recherches en linguistique portant sur l'ellipse, pour nombreuses qu'elles soient, ni même la traductologie n'ont véritablement abordé la traduction du phénomène. Fondamentalement, l'explication tient au fait que l'ellipse, comme en témoignent les exemples qui seront analysés dans ce chapitre, peut être si bien maîtrisée et exprimée qu'elle passe inaperçue lorsque l'on lit des traductions humaines. Mais si certaines ellipses sont très bien traduites, y compris en traduction automatique, d'autres, en revanche, posent un problème de résolution important. C'est le cas de l'exemple (106) qui présente une ellipse déclenchée par *to* où le segment *make myself*[141] *agreeable to young McClure* est omis.

(106)

| | |
|---|---|
| - Make yourself agreeable to young MacClure.<br>- I won't **fail to** Ø. | - Soyez gentil avec le jeune McClure<br>- Je n'**y** manquerai pas<br>(TR) |
| | - Faites-vous plaisir au jeune macclure.<br>**- Je ne manquerai pas.**<br>(Systran) |

[141] Ajustement syntaxique nécessaire pour copier l'antécédent dans le site elliptique.

Cet exemple n'a visiblement pas posé de problème au traducteur humain puisqu'il a repéré la présence d'un antécédent et l'a restitué par le pronom clitique *y*. Systran s'est contenté de reprendre mot-à-mot l'énoncé original rendant la traduction inacceptable dans ce contexte. On peut d'ores et déjà noter que les erreurs engendrées par la traduction automatique, parfois même par la traduction humaine, résultent d'un mauvais transfert syntaxique et sémantique qui nous semble étroitement lié aux fonctionnements différents des déclencheurs de l'ellipse dans les deux langues. Ces erreurs peuvent également être liées à la nature de l'antécédent qui peut parfois être très ambiguë et susciter plusieurs interprétations possibles, tâches que la traduction automatique ne parvient pas encore à gérer.

D'un point de vue contrastif, certaines remarques peuvent déjà être formulées :

- *Do* n'a pas d'équivalent en français,
- Les auxiliaires ont un fonctionnement différent (par exemple, le *pseudogapping* déclenché par un auxiliaire n'est pas autorisé en français),
- Les modaux sont nombreux en anglais et leur sens est fréquemment traduit par des verbes lexicaux ou des adverbes en français,
- L'antécédent de l'ellipse ne se trouvant pas toujours dans la même phrase, une analyse sémantique et pragmatique est requise pour interpréter le site elliptique.

À notre connaissance, aucune étude n'a été menée sur les erreurs provoquées par les éléments déclencheurs de l'ellipse dans la traduction automatique. Notons néanmoins, parmi les recherches menées sur la traduction du phénomène, celle de Bouillon *et al.* (2007, 54) lesquels abordent les ellipses dans **«** MedSLT », un système de traduction automatique de la parole dans le domaine médical qui « traduit des questions de diagnostic pour des patients étrangers […] en anglais, français, japonais, espagnol, catalan et arabe ». Les ellipses qui intéressent Bouillon *et al.* sont notamment celles relevant des questions et des réponses fragmentaires illustrées dans van Craenenbroeck & Merchant (2013). Deux versions de ce système sont

représentées : l'une à un niveau unidirectionnel (questions et réponses du type oui-non, où le patient répond non verbalement), l'autre à un niveau bidirectionnel (avec des réponses ouvertes et elliptiques). L'approche de Bouillon *et al.* permet une résolution de ces ellipses à l'aide d'une méthode d'apprentissage automatique entrainé sur des exemples. Cette recherche met en avant l'intégration du contexte du genre médical dans MedLST, permettant d'atteindre une qualité suffisante :

> [L]'approche contrôlée de MedSLT rend possible l'utilisation d'algorithmes de résolution très peu coûteux qui permettent une traduction en contexte intelligible et fidèle. Nous avons ainsi clairement contribué à intégrer le contexte dans une architecture comme la nôtre et, de manière plus générale, dans les systèmes de TAP[142]. (Bouillon *et al.,* 2007, 67)

Par ailleurs, si d'autres recherches ont été menées sur la traduction de la parole, elles ne traitent que rarement de la traduction du phénomène elliptique.

Pour ce qui nous concerne, nous présentons quelques exemples de traduction de phrases elliptiques ayant résisté à la traduction automatique. Ces exemples sont extraits de notre corpus de développement et de notre corpus d'évaluation, appartenant en particulier au genre conversationnel, témoignant de l'usage le plus fréquent de l'ellipse. En effet, bien qu'étant consciente des apports d'une analyse multi-genres dans la traduction du phénomène, nous avons choisi de laisser de côté les autres genres qui pourront être traités dans des recherches ultérieures.

### 1.1. Ellipses post-auxiliaires dans les *question tags*

Nous avons observé que les ellipses post-auxiliaires sont très fréquentes dans le corpus du genre conversationnel et nous avons d'abord remarqué leur présence dans les *question tags*. Pour rappeler brièvement son fonctionnement, la *question tag* est la reprise d'un énoncé initial sous forme de question dans laquelle le verbe lexical est omis et où l'auxiliaire ou le modal et le sujet sont inversés pour formuler une interrogation, comme dans l'exemple *He isn't eating, is he ?* Lorsque l'énoncé initial est affirmatif, la *question tag* adoptera une forme interro-négative et vice versa

[142] Traduction Automatique de la Parole

(polarité inverse). La *question tag* est une particularité de l'anglais, dans les conversations formelles ou informelles, dont l'intérêt est de vérifier l'attention du co-locuteur dans la conversation, au moyen d'éléments phatiques. Ces questions dont la réponse importe peu dans la plupart des cas, peuvent généralement être assimilées à des questions rhétoriques. Contrairement à l'anglais, le français s'en passe volontiers, en raison peut-être d'une plus grande sensibilité à l'égard de la répétition et du nombre limité de ses auxiliaires. Plus fondamentalement, cette particularité s'explique par le fonctionnement très différent de l'anglais et du français en matière d'interlocution[143].

Nous pourrions nous interroger ici de façon théorique sur l'aspect formel des *question tags*. Contiennent-elles ou non une ellipse dans la mesure où elles ne sont jamais formulées entièrement, selon le principe même d'une *question tag* ?

Dans cette étude, nous considérons la *question tag* comme porteuse d'ellipse verbale en raison du processus d'effacement que cette construction partage avec le phénomène elliptique et, plus spécifiquement, avec l'ellipse post-auxiliaire. En fait, les formes de ces *question tags* sont dépendantes de la proposition qui les précède, séparées d'elle par une virgule (avec présence ou non de la négation dans la proposition définissant le contexte). Il est à noter que nous considérons ici uniquement les *question tags* dont le contenu est dépendant de la proposition précédente et non les *question tags* dont le contenu est invariant quelle que soit la forme de la proposition précédente (comme par exemple les *question tags* formées avec *right ? no ? yeah ?*)[144].

En ne gardant que le modal ou l'auxiliaire, la *question tag* constitue un terrain favorable à l'ellipse post-auxiliaire. Compte tenu du fait que le français utilise rarement ce genre de question, le traducteur humain a deux possibilités : il peut, soit

[143] ce qu'a mis en lumière Guillemin-Flescher au fil de nombreux articles (voir la bibliographie).
Prenons par exemple la situation comique créée par l'interférence de l'anglais dans les dialogues en français de *Astérix chez les Bretons*, où l'auteur s'amuse à reproduire la forme littérale des *question tags* :
« Bonté gracieuse ! Ce spectacle est surprenant.
*Il est, n'est il pas ?...* »

[144] Sauf erreur de notre part, cette catégorie n'a pas fait l'objet d'études spécifiques en lien avec l'ellipse et pourrait donner lieu à une recherche ultérieure.

traduire en employant la formule *n'est-ce-pas ?* ou encore simplement en utilisant *non ? /oui ?*[145] soit ne pas traduire la *question tag*. L'exemple ci-dessous illustre ce constat :

(107)

| | |
|---|---|
| -[146] It's not that simple, though, **is it**?<br>- **It is** Ø. Cos I'm telling you it is Ø. | - C'est pas aussi simple.<br>- Si. Parce que je vous le dis.<br>(TR) |
| | - Ce n'est pas si simple, cependant **n'est-ce pas ?**<br>- Oui. Cos je vous dis ce sont.<br>(Reverso)[147] |
| | - C'est pas si simple, **non ?**<br>- C'est tout. Parce que je te le dis.<br>(Systran) |
| | - Ce n'est pas si simple, **n'est-ce pas?**<br>- **C'est**. Parce que je te le dis.<br>(Google Traduction) |
| | - Ce n'est pas si simple, **n'est-ce pas ?**<br>- Ça l'est. Parce que je te le dis.<br>(DeepL) |

Nous remarquons en effet dans cet échange une ellipse post-auxiliaire dans la *question tag* et deux autres dans la réponse où une partie du syntagme verbal *be+(adv+adj)* est omise. De prime abord, seule la traduction humaine n'a pas produit de traduction à la *question tag.* Le passage perd ainsi l'insistance phatique exprimée en anglais par ce *tag*. La réponse par *it is.* est traduite par l'adverbe *si,* tandis que la seconde occurrence dans la proposition complément de *tell* est évitée, le pronom *le* traduisant l'ensemble de la proposition *(that) it is*.

Curieusement, à l'exception de Systran, toutes les traductions automatiques ont opté pour le *n'est-ce pas ?* qui, en réalité, est inadapté ici. En fait, le rôle de la *question tag* dans ce cas n'est pas de demander l'avis de l'interlocuteur mais de lui

[145] Ces segments ont les mêmes valeurs illocutoires que les *question tags* et peuvent être considérés comme des traductions par équivalence.
[146] Chaque tiret (-), dans les exemples présentés, désigne un changement de locuteur.
[147] Sauf mention contraire, les exemples de ce chapitre ont tous été traduits pendant les mois de mai et juin 2018 et pendant les mois d'avril et juin 2019.

imposer (à titre d'information) ce qui précède étant donné la forme négative de la phrase précédente.

Dans certaines traductions automatiques, les réponses apportées à l'ensemble de la phrase semblent insolites, sans doute parce que la réponse *it is.* va à l'encontre de la réponse attendue : au lieu de consentir, d'aller dans le sens de cette question rhétorique, l'interlocuteur exprime son désaccord. Cet aspect de l'échange a été « mal interprété » par Reverso qui, dans ce cas précis, cherche à répondre positivement à une assertion négative suivie d'une *question tag* négative *n'est-ce-pas ?* Or, grammaticalement, la réponse affirmative à une question négative nécessite l'emploi de l'adverbe affirmatif *si* pour contredire la négation précédente, et non *oui*. Les traductions de Systran et Google sont elles aussi erronées, puisque dans ce contexte, les systèmes n'ont pas su anticiper le désaccord, et puisque l'emploi incongru de *c'est* (qui ne peut s'employer seul) ne peut être considéré comme une « bonne » traduction de la réponse *it is* Ø. Enfin, dans la traduction de DeepL, l'ellipse a été traduite à l'aide du pronom *l'* (anaphorique). Cette traduction pourrait être sémantiquement acceptable mais reste quelque peu éloignée du registre informel de cet échange. Il apparaît ainsi que la traduction humaine de l'ellipse de cet exemple, par le vide (effacement du verbe *être*) et l'assentiment, est, en dernière analyse, la plus pertinente du point de vue du sens, du registre et de l'interprétation. Elle peut donc être considérée comme une *traduction de référence*, parmi d'autres qui sont également possibles : *is it?* traduit par *hein ?* ou *tu crois pas ?* et *it is, cos I'm telling you.* par *Si, si ! Puisque je te le dis !*

Observons à présent à l'exemple (108) ci-dessous où l'on remarque deux ellipses du segment *managed so far* : une ellipse dans la *question tag*, et une autre, dans la réponse donnée, toutes deux déclenchées par *have*.

(108)

| | |
|---|---|
| | - Nous avons vécu bien pire.<br>- **Bien sûr.**<br>(TR) |

| | |
|---|---|
| - I won't let you down. We've managed so far, **haven't we** Ø?<br>- Yes, we **have** Ø. | - Je ne vous laisserai pas tomber. Nous nous sommes débrouillés jusqu'ici, n'est-ce pas ?<br>- **Oui.**<br>(Reverso) |
| | - Je ne te laisserai pas tomber. On a réussi jusqu'ici, n'est-ce pas ?<br>**- Oui, nous l'avons.**<br>(Systran) |
| | - Je ne te laisserai pas tomber. Nous avons réussi jusqu'ici, n'est-ce pas?<br>- **Oui.** (Google Traduction) |
| | - Je ne te laisserai pas tomber. Nous avons réussi jusqu'à présent, n'est-ce pas ?<br>- **Oui, nous l'avons fait.**<br>(DeepL) |

On repère dans l'exemple anglais la répétition, trois fois, du sujet et de l'auxiliaire *have* de la forme verbale perfective. La dernière apparition contient une insistance avec l'interjection *yes*. L'échange du locuteur se décompose ainsi en une déclarative à la forme négative *I won't,* suivie d'une assertion *we've managed so far*, elle-même suivie d'une *question tag* interro-négative qui est une demande de confirmation de l'assertion faite à l'interlocuteur. La traduction humaine proposée ici perd l'insistance exprimée dans la langue source puisque la *question tag* n'est pas traduite (tout comme *I won't let you down*). L'échange se limite ainsi à une assertion et une confirmation avec la locution adverbiale *bien sûr* qui marque l'insistance. Par conséquent, les marqueurs morphosyntaxiques renvoyant aux paramètres de la situation (sujet et temps) ont été effacés, ne laissant place à aucune répétition accentuant le propos.

Dans cet exemple, les systèmes ont tous traduit la *question tag* par *n'est-ce-pas?* y compris Systran qui dans l'exemple précédent a traduit avec *non ?* Sa traduction est un calque de la forme de la structure originale (négative puis affirmative), suivie de *n'est-ce pas ?* locution « vide » qui n'a pas d'impact sur la réponse mais qui insiste sur ce qui précède et oriente potentiellement la réponse. Quant aux systèmes Reverso et Google, dont les traductions sont plus pertinentes, ils ont omis dans la production de la réponse *Yes, we hav*e de traduire la proposition *I have,* et se sont contentés de *Oui* traduction qui a pour inconvénient de ne pas marquer l'insistance.

DeepL et Systran enfin ont produit une traduction incompatible avec le texte source. En effet, en grammaire française, *avoir* ne peut se passer de complément verbal et sa complémentation par le pronom *l'* et le verbe *faire* n'est pas compatible avec la reprise du verbe *réussir*. Si cet exemple n'illustre pas la difficulté des systèmes à traduire les *question tags*, il met en avant les erreurs de la traduction des ellipses déclenchées par *have* que nous détaillerons plus loin.

Nous abordons enfin un exemple d'ellipse post-auxiliaire (109) dans une *question tag* afin d'illustrer une erreur due à la mauvaise traduction des modaux et de *do*.

(109)

| | |
|---|---|
| - Don't worry, I'll see to this. I said I **would** Ø, **didn't I** Ø? | T'inquiète pas. Je m'en charge. **J'ai dit que je le ferais, non ?** (TR) |
| | Ne vous inquiétez pas, je m'en occuperai. **J'ai dit que je**, n'est-ce pas ? (Reverso) |
| | Ne t'inquiète pas, je vais y veiller. J'ai dit que je le ferais. (Systran) |
| | Ne t'inquiète pas, je vais voir ça. **J'ai dit que je le ferais, n'est-ce pas ?** (Google Traduction) |
| | Ne vous inquiétez pas, je vais m'en occuper. **J'ai dit que je le ferais, n'est-ce pas ?** (DeepL) |

Il s'agit dans cet exemple d'une première ellipse déclenchée par *would* où le segment *see to this* est effacé et d'une ellipse déclenchée par *do* à la forme négative, où *said I would* est omis[148]. La traduction de la *question tag*, *didn't I*, proposée par le traducteur humain se contente d'un *non* suivi du point d'interrogation, gardant ainsi la forme interrogative du texte source. Par sa traduction, on pourrait comprendre qu'il se rapproche davantage d'une question argumentative qui nécessite une réponse et qui peut soit exprimer l'insistance (agacée), soit l'incertitude du locuteur

[148] Élément restitué par l'interprétation : *didn't I say I would ?* En d'autres termes, l'antécédent n'est pas repris à l'identique dans le site elliptique, seule la partie porteuse du sémantisme indispensable à l'interprétation l'est en conformité, lors de la restitution, avec la syntaxe de l'élément déclencheur.

qui vérifie son assertion (en d'autres termes, sollicite son interlocuteur au cas où ce dernier douterait de ce qu'il a dit).

Les outils de traduction automatique, excepté Systran, s'accordent systématiquement sur la traduction de la *question tag* par *n'est-ce-pas ?* S'il existe suffisamment d'occurrences dans les mémoires de traductions qui montrent que dans ce type de contexte linguistique, *n'est-ce pas ?* est souvent possible, il reste difficile pour ces traducteurs de distinguer les *questions tags* qui requièrent des réponses de celles qui peuvent s'en passer. Par ailleurs, à l'instar de la traduction humaine, l'ellipse post-modale *would* qui précède la *question tag* a été traduite par le verbe anaphorique *faire*. Seul Reverso a complètement ignoré ce modal et s'est limité à la traduction du sujet *je,* rendant ainsi l'exemple incomplet, agrammatical et sémantiquement non acceptable.

Comme nous venons de le voir, traduire les ellipses dans les *question tags* par la locution *n'est-ce pas ?* n'apparait pas toujours adapté. Si *n'est-ce pas ?* est souvent proposé comme traduction possible d'une *question tag,* c'est vraisemblablement parce que les données d'entrainement contiennent suffisamment d'occurrences permettant d'identifier ce type de phrases. Car en effet, le schéma syntaxique de la *question tag* semble facile à automatiser du fait des repères concrets, voire visibles, qu'il présente dans l'inversion de la polarité et la ponctuation (virgule) qui sépare la phrase de base de la question même. Preuve en est, *Is he?* dans l'exemple (110) ci-dessous qui aurait pu être confondu avec une *question tag*, mais qui ne génère pas de traduction par *n'est-ce pas ?* du fait de l'absence de ces caractéristiques.

(110)

| | |
|---|---|
| - Is Stanley up yet?<br>- I don't know. **Is he** Ø? | - Stanley est levé ?<br>- Je ne sais pas. **Il est levé** ?<br>(TR) |
| | - Est Stanley en haut encore ?<br>- Je ne sais pas. **Est il?**<br>(Reverso) |

| | |
|---|---|
| | - Stanley est-il encore debout ?<br>- Je ne sais pas. **C'est lui ?**<br>(Systran) |
| | - Stanley est-il encore debout ?<br>- Je ne sais pas. **Est-il?**<br>(Google Traduction) |
| | - Stanley est levé ?<br>- Je ne sais pas. **Il l'est ?**<br>(DeepL) |

Il s'agit dans cet exemple d'une ellipse déclenchée par *be* dans une phrase interrogative. Le traducteur humain a interprété l'antécédent de l'ellipse *up* et l'a restitué en français par *levé*. Nous rappelons qu'en grammaire française, trois structures, correspondant à trois registres différents, sont récurrentes pour traduire les questions directes :

– Registre *familier*[149] : dans ce registre, la structure suit la forme régulière de la phrase affirmative, c'est-à-dire *sujet + verbe + complément,* mais s'en distingue par une intonation interrogative (ajout du *?*) comme dans *Tu as rangé ta chambre ?*

– Registre *standard* : la question est marquée par l'ajout du segment *est-ce que* à une phrase affirmative, comme dans *Est-ce que tu as pris tes médicaments ?*

– Registre *soutenu* : ce registre entraîne souvent une inversion des positions grammaticale du sujet et du verbe[150], comme dans *Veux-tu un thé ?*

La formulation *il est levé ?* relève ainsi du registre courant, quotidien, voire familier *Stanley est levé?* présentant une phrase affirmative suivie d'un point d'interrogation ce qui implique à l'oral l'intonation montante typique de l'interrogative. DeepL a ajouté le pronom *l'* qui se réfère à *levé* ellipsé, mais l'ensemble, malgré la complétude grammaticale, ne fonctionne pas en français : son authenticité[151] est douteuse. On

[149] La distinction est néanmoins sujette à caution dans la mesure où, par exemple, le registre *familier* pourrait bien être considéré comme aussi *standard* que la formule avec *est-ce que.*

[150] Le trait d'union (-) est obligatoire pour séparer le verbe et le sujet inversés. Aussi, pour des raisons phonétiques (hiatus), le **t** est ajouté pour séparer le verbe qui se termine par une voyelle et le sujet qui commence par une voyelle.

[151] Nous avons soumis cet énoncé à deux locutrices de langue maternelle française âgées respectivement de 30 et 32 ans), et nous avons obtenu les réponses suivantes : « Je comprends, mais ça sonne faux » et « Je dirais jamais ça, mais je vois ce que ça veut dire ». Ces réactions sont, d'un point de vue scientifique, sujettes à caution dans la mesure où la sensibilité au registre est très personnelle.

notera ici le fait que l'ajout de *l'* est le résultat d'un apprentissage automatique pour exécuter le remplacement de l'élément ellipsé par un pronom.

Les traductions de Google et Reverso, quant à elles, ne sont pas acceptables d'un point de vue grammatical. Elles sont toutes deux incomplètes mais ne peuvent être interprétées comme ellipses, ce qui pourrait signaler une fois de plus que le phénomène elliptique est soumis à des contraintes qui diffèrent dans les deux langues. Les systèmes de TA ne détectent aucune *question tag* en anglais puisque le contexte qui précède la question *Is he?* n'a pas la même configuration que la phrase de base dans la définition même d'une *question tag.* C'est la raison pour laquell*e n'est-ce pas ?* n'est pas proposé, prouvant par là même la réussite de l'apprentissage, dans une certaine mesure.

En conclusion, les exemples analysés précédemment montrent que les *question tags* (dont le contenu sémantique dépend d'une phrase qui les précède) sont toujours porteuses d'une ellipse post-auxiliaire enfouie en « structure profonde », preuve en est dans la réponse apportée à ces questions (voir les réponses dans les exemples 107 et 108). En effet, cette réponse contient une reprise systématique de l'auxiliaire, et se limite rarement à un simple *yes* ou *no.*

Par ailleurs, les ellipses contenues dans les *question tags* disparaissent dans la traduction, car le français ne les utilise que rarement et a recours aux adverbes, marqueurs de discours, *oui, non, n'est-ce pas ?* le cas échéant. Sur le plan discursif, lorsque le traducteur humain décide de maintenir l'intonation marquée par la *question tag*, la nécessité du repérage correct de l'antécédent devient primordiale. Pour cette tâche, l'outil de traduction rencontre une difficulté, notamment lorsque l'antécédent est éloigné du site elliptique, constituant ainsi deux phrases différentes comme illustré dans la question *Is he ?* dans l'exemple (110).

La traduction automatique des *question tags,* de l'anglais vers le français, révèle que les ellipses peuvent être sources de confusion parce que le recours à l'utilisation des modaux et des auxiliaires pour les former est une particularité de l'anglais, révélant de ce fait l'une des différences marquantes existant entre les deux langues.

Lors d'un travail sur un corpus de traduction, se pose alors la question de savoir si les *question tags* sont toujours traduites, et quelles sont les configurations linguistiques et les situations de locution pouvant déterminer leur traduction ou non.

### 1.2. Ellipses post-auxiliaires déclenchées par *do*

*Do* est également source d'ambiguïté en traduction. Pour pallier l'absence d'équivalent en français, les traducteurs humains mettent en place des stratégies adaptées au contexte linguistique, par exemple en ne traduisant pas systématiquement *do* par *faire*, et en restituant l'antécédent, ce que les outils ne parviennent pas toujours à réaliser.

Prenons l'exemple (111) ci-dessous :

(111)

| | |
|---|---|
| - There is a Gatling in that one. You **know what a Gatling gun is**, friend?<br>- I **don't** Ø. And I don't care neither. Tell you truth of it. | - Il y a une mitrailleuse, là-dedans. Une Gatling. Tu les connais ?<br>- Non je ne les **connais** pas – et je m'en fiche pas mal, pour tout dire.<br>(TR) |
| | - Il y a un Gatling dans cela un. Vous savez(connaissez) ce qu'une arme à feu Gatling est, ami ?<br>- **Je ne fais pas.** Et je ne me soucie aucun. Dites-vous la vérité de cela.<br>(Reverso) |
| | - Il y a un Gatling dans celui-là. Tu sais ce qu'est un revolver, ami ?<br>- **Non**. Et je m'en fiche. Dis-lui la vérité.<br>(Systran) |
| | - Il y a un Gatling dans celui-là. Vous savez ce qu'est un pistolet Gatling, mon ami ?<br>- **Je ne**. Et je m'en fous. Dites-en la vérité.<br>(Google Traduction) |
| | - Il y a un Gatling dans celui-là. Tu sais ce qu'est un pistolet Gatling, mon ami ?<br>- **Je n'en ai pas**. Et je m'en fiche aussi. Je vais vous dire la vérité.<br>(DeepL) |

Cet exemple est particulièrement intéressant dans la mesure où il peut révéler l'ambiguïté de l'élément ellipsé dans l'échange. On pourrait par exemple avancer que

la totalité de la *question tag, don't you ?* est ellipsée dans *You know what a Gatling gun is, Ø friend ?* On pourrait également parler d'une ellipse de *do* dans *Ø you know what a Gatling gun is, friend ?* puisque la particularité de *do*, indispensable à souligner dans un premier temps, est qu'il peut être effacé dans les questions (mais pas dans les réponses). La réponse *I don't* est elliptique et le segment *know what a Gatling gun is* est effacé.

De prime abord, la traduction de cet exemple peut paraitre problématique pour l'humain qui propose le verbe *connaitre* pour traduire *know* et établit une relation anaphorique entre *Gatling gun* (au singulier) et le pronom complément *les* au pluriel. Observons que cette traduction est proposée à la fois dans l'antécédent *know what a Gatling gun is* et le site elliptique suivant *don't*. Cependant, cette traduction peut être justifiée et perçue comme suit : le traducteur établit une forme de métonymie (plus précisément une éponymie) puisque *Gatling* est une marque de mitrailleuse du nom de son inventeur. Il semble demander à l'interlocuteur s'il connait ces armes. En d'autres termes, le traducteur humain a effectué le choix de clarifier ce que la désignation éponymique de l'arme en question rendait obscur en traduisant *gun* par le terme précis de *mitraillette* (procédé de traduction récurrent de l'anglais au français). Il est ainsi évident que ce choix a eu pour conséquence la reprise anaphorique de *mitraillette* par sa désignation éponyme suivie de la question à son interlocuteur, car tout le monde ne sait pas forcément ce qu'est une *Gatling (gun)*. Par conséquent, le pluriel s'impose pour marquer le « générique » et non cette *Gatling*-là en particulier. Cette interprétation nous éloigne du texte source par la configuration grammaticale et syntaxique mais s'en rapproche grâce aux facteurs discursifs du style relâché, courant et spontané adopté.

Par ailleurs, le traducteur humain résout et traduit *know* par le verbe *connaître* tandis que les systèmes ont opté pour *savoir*. Sans doute deux possibilités se sont-elles présentées : *tu sais ce que c'est* vs *tu les connais* ? La première jouerait sur l'ignorance possible de l'interlocuteur (ce qui est le cas, en réalité, au vu de la réponse), la seconde préfère sous-entendre qu'il sait ce que c'est mais qu'il n'en connaît peut-être pas le maniement. Comme on le sait, l'usage le plus fréquent de

*savoi*r est observé avec des compléments d'objet qui, d'un point de vue sémantique, intériorisent un apprentissage. Sachant que *connaitre* et *savoir* admettent tous les deux des compléments d'objet, il est difficile pour le système d'identifier si cet objet est le fruit d'un apprentissage ou de la reconnaissance d'une personne ou d'une chose, car les deux verbes ne sont pas interchangeables. De ce fait, seule la fréquence d'usage (le nombre d'occurrences) dans les données d'entrainement semblent guider la traduction automatique, principe de la TA basée sur des statistiques et de l'apprentissage automatique.

Quant à la réponse elliptique *I don't*, le traducteur aurait pu la traduire par *non* ou *aucune idée* en n'établissant aucune référence explicite aux paramètres linguistiques de la situation, ce qui se rapproche des traductions généralement proposées à l'égard des *question tags.* Il a cependant préféré répéter la construction de la proposition précédente avec le verbe *connaitre*, son sujet et son complément.

Toujours dans la même réponse, mais cette fois-ci traitée par la traduction automatique*, do* s'avère être source d'ambiguïté. En effet, à l'exception de Systran, qui a entièrement omis la phrase elliptique *I don't,* et l'a traduite par l'adverbe de négation *non,* rendant la traduction acceptable, toutes les autres traductions sont, soit des séquences grammaticalement incomplètes (Google Traduction), soit incompatibles avec le texte source, puisqu'elles ne répondent pas à la question posée. Reverso a traduit *do* comme s'il s'agissait d'un verbe lexical d'action (et non comme un auxiliaire). Il a alors simplement appliqué la règle concernant l'auxiliaire *do* qui, lorsqu'il apparait seul dans une phrase, implique qu'il soit interprété comme un verbe conjugué. DeepL a effectué une sorte de *modulation*[152] qui passe de la réponse indiquant l'ignorance (*I don't [know what a Gatling gun is]*) en anglais, à une réponse en français niant la possession (je n'en ai pas), ce qui apparait ni plus ni moins comme un contre-sens. L'analyse de cet exemple (111) s'applique également à

[152] « La modulation se définit de façon très générale comme un changement de point de vue. Celui-ci intervient au niveau du mot, de l'expression, ou de l'énoncé pris globalement ; il relève du lexique et/ou de la grammaire ». (Chuquet & Paillard 1987, 26)

l'exemple ci-dessous (112) qui présente deux ellipses déclenchées par *do,* dont l'antécédent est *miss me*.

(112)

| | |
|---|---|
| - Well, did you miss me?<br>- You bet **I did** Ø. I'll bet you **didn't** Ø. | - Je vous ai manqué ?<br>- **Bien sûr.** Je parie que **non.**<br>(TR) |
| | - Je vous ai manqué ?<br>- **Bien sûr**. Je parie que **non**.<br>(Reverso) |
| | - Tu m'as manqué ?<br>- Vous pariez. Je parie que **non**.<br>(Systran) |
| | - Tu m'as manqué ?<br>- Tu parles que je **l'ai fait**. Je parie **que vous ne l'avez pas fait**.<br>(Google Traduction) |
| | - Eh bien, je t'ai manqué ?<br>- Bien sûr que je **l'ai fait**. Je parie que **non**.<br>(DeepL) |

Comme on le remarque dans les deux séquences elliptiques, la traduction humaine a omis toute référence au sujet et au verbe et s'est contentée des locutions adverbiales *bien sûr* et *non* pour traduire *you didn't*. Le traducteur a ainsi estimé que la répétition était inutile et que ces locutions étaient suffisantes pour exprimer la même insistance que dans le texte source (il a donc traduit le sens).

Pour ce qui est des systèmes de traduction, à l'exception de Reverso qui a produit la même traduction que l'humain, et de Systran qui a omis de traduire la première séquence elliptique, DeepL et Google ont traduit *do* par le verbe *faire,* non acceptable dans ce contexte. C'est d'ailleurs le même type d'erreurs que nous relevons dans la traduction du *do* emphatique présenté dans les exemples (113) et (114) ci-après. *Do* est dit emphatique lorsqu'il est utilisé dans un énoncé pour marquer l'insistance et appuyer le propos. Dans l'exemple (113), le locuteur exprime l'emphase avec *do* pour reprendre *You earned it* et ajoute l'adverbe de certitude *sure.*

(113)

| | |
|---|---|
| -You're entitled to everything I can give you.<br>- You earned it, you **sure did**.<br>- Thanks. | - Tout ce que je peux vous donner vous revient.<br>-Vous le méritez, **c'est sûr.**<br>-Merci.<br>(TR) |
| | -Vous avez droit à tout ce que je peux vous donner.<br>- Vous l'avez mérité, **vous l'avez fait.**<br>- Merci.<br>(Reverso) |
| | -Tu as droit à tout ce que je peux te donner. - Tu l'as gagné.<br>- Merci.<br>(Systran) |
| | -Vous avez droit à tout ce que je peux vous donner. - Vous l'avez **bien mérité.**<br>- Merci.<br>(Google Traduction) |
| | - Vous avez droit à tout ce que je peux vous donner.<br>- Tu l'as **mérité**, tu l'as **bien mérité.**<br>- Merci. Merci.<br>(DeepL) |

Dans la traduction humaine de l'ellipse *you sure did*, tout repère à l' « animé humain », pour reprendre les termes de Guillemin-Flescher, *you* disparait et se voit remplacé par la tournure impersonnelle *c'est* ne laissant alors aucun vide elliptique. On remarque dans cette traduction que le traducteur semble avoir préféré mettre l'emphase dans la première occurrence *Vous le méritez* car une interprétation telle que *C'est sûr que vous le méritez* aurait également été possible.

Les traductions automatiques sont ici toutes intéressantes. Reverso traduit, par défaut, *do* par *faire* tandis que Systran ne le traduit pas du tout. En revanche, l'insistance et l'emphase exprimées par *do* dans cet énoncé sont aussi exprimées dans les traductions de Google et DeepL dans la mesure où toutes les deux ont ajouté *bien.* DeepL, dans cette séquence, a repéré l'antécédent du site elliptique et a restitué directement le verbe lexical *mériter* que *do* a accentué (et, au passage, a redoublé le *merci*, probablement une manière de traduire le pluriel). Dans certains exemples observés, lorsque le site elliptique se trouve dans la même phrase que

l'antécédent, DeepL ne semble pas rencontrer de difficultés à restituer l'élément omis. En effet, dans d'autres contextes où, par exemple, le site elliptique est éloigné de son antécédent, comme dans le cas suivant (114), DeepL réagit différemment.

(114)

| | |
|---|---|
| - Show me your papers<br>- I think we **know each other.**<br>- Yes, we **do** Ø. | - Montrez-moi vos papiers.<br>- Je pense que nous nous connaissons.<br>- **Oui, c'est vrai.**<br>(TR) |
| | - Montrez-moi vos papiers<br>- Je pense que nous nous connaissons.<br>- Oui, **nous le faisons**.<br>(Reverso) |
| | - Montrez-moi vos papiers.<br>- On se connaît.<br>- Oui, **nous le faisons.**<br>(Systran) |
| | - Montre-moi tes papiers<br>- Je pense qu'on se connait.<br>- **Oui.**<br>(Google traduction) |
| | - Montrez-moi vos papiers<br>- Je crois qu'on se connaît.<br>- **Oui, c'est vrai.**<br>(DeepL) |

Il s'agit d'une ellipse du syntagme *know each other* déclenchée par *do.* Le site elliptique Ø ne se trouve pas dans la même phrase et survient dans un échange entre deux personnes : l'une émet un constat et l'autre le confirme avec *yes* et le *do* que nous considérons comme emphatique, bien qu'il puisse simplement être vu comme la reprise anaphorique du verbe *do* (proform). Mais notre propos n'est pas là. Ce que nous pouvons observer, c'est que le traducteur humain emploie des stratégies différentes pour traduire les ellipses post-do et, en l'absence d'équivalent de *do* en français, il vise plutôt une fidélité sémantique au message original qu'une fidélité à la construction syntaxique. Pour ce faire, il ne maintient généralement pas l'ellipse du texte source mais utilise des marqueurs discursifs sans relation à un animé humain. Certains de ses choix sont le résultat de contraintes syntaxiques et lexicales dans la langue cible elle-même. En utilisant les mêmes marqueurs discursifs que l'humain,

les systèmes de traduction automatique (comme DeepL) parviennent à générer une traduction acceptable, et par là, à résoudre l'ellipse dont ils restituent l'élément omis. DeepL ne restitue pas ce qui manque mais l'interprète. Il emprunte alors une tournure impersonnelle *c'est vrai*, produisant la même traduction que l'humain. Ceci constitue une étape importante à la résolution de l'ellipse. D'autres (comme Systran), cependant, produisent des traductions agrammaticales et non recevables dans la langue cible notamment lorsque *do* est systématiquement traduit par *faire*. Afin d'améliorer les systèmes qui sont soumis dans ces cas aux contraintes syntaxiques et sémantiques, elles-mêmes liées aux propriétés primitives des verbes, il serait intéressant de répertorier des traductions *types* (même approximatives dans un premier temps) que le système pourrait utiliser lorsqu'il rencontre *do* déclencheur d'ellipse. C'est ici qu'une détection automatique de l'ellipse pourrait, selon notre hypothèse, contribuer à l'amélioration de la traduction automatique. Le contexte linguistique avec ou sans présence de verbes lexicaux, la ponctuation et la répétition sont des informations qui peuvent être détectées de manière automatique. Les occurrences-types mettraient en avant notamment les traductions de *do* comme auxiliaire qui, selon ces informations, pourrait être traduit soit par le verbe lexical qui le précède, soit par des locutions adverbiales (comme on l'a vu précédemment) ayant la capacité d'atteindre les mêmes valeurs sémantique et discursive que celles de l'énoncé initial.

### 1.3. Ellipses déclenchées par un modal

Comme l'a montré Guillemin-Flescher (2003), malgré la proximité de l'anglais et du français, les différences grammaticales entre ces deux langues sont subtiles et complexes. L'une des différences se situe dans la présence de modaux en anglais qui n'ont pas d'équivalents exacts en français car seuls *devoir* et *pouvoir* sont considérés comme de vrais modaux tandis que l'anglais en compte davantage, tels *shall, will, must, can, should, would, could, may, might.* De ce fait, le déficit du nombre de modaux en français a, parfois, pour conséquence, la disparition de l'ellipse lors de la traduction. En effet, face aux ellipses admises en anglais (exemple du *pseudogapping* entre autres), non autorisées en français car créant un vide agrammatical, voire

dépourvu de sens, le traducteur est contraint (par le système linguistique) de combler ce vide. Une première option consiste à répéter dans la langue cible, le syntagme verbal omis dans l'ellipse de la langue source, comme le montre l'exemple (115) ci-dessous :

(115)

| | |
|---|---|
| Get down! Get down off it, you old cuckold, I don't care who you are. I'll **put the first one through you**! I swear it, I **will Ø** now! One! Two! | 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 ! (TR) |
| | Descendez! Descendez-en, vous vieux cocu, je ne me soucie pas qui vous êtes. Je mettrai le premier par vous! Je jure cela, **je ferai(serai) maintenant** ! Un! Deux! (Reverso) |
| | Descends ! Descends, vieux vieux, je me fiche de qui tu es. Je vais mettre le premier à travers toi ! Je le jure, **je le ferai maintenant** ! Un ! Deux ! (Systran) |
| | Descendre! Descends, vieux cocu, je me fous de qui tu es. Je mettrai le premier à travers toi! Je le jure, je **le ferai maintenant**! Un! Deux! (Google Traduction) |
| | Baissez-vous ! Descends de là, vieux cocu, je me fiche de qui tu es. Je vais te faire passer le premier ! Je le jure, **je le ferai maintenant**! Un ! Deux ! (DeepL) |

Il s'agit ici, dans la version anglaise, d'une ellipse déclenchée par le modal *will* où le syntagme verbal *put the first one through you* est ellipsé. Le traducteur a choisi de traduire le syntagme contenant l'ellipse *I will Ø* par la périphrase *aller+inf* qui exprime un futur proche (*je vais tirer*). L'ellipse post-auxiliaire est ainsi rendue impossible **je vais Ø* puisque *aller* est, en quelque sorte un pseudo-auxiliaire qui ne peut apparaître seul dans cette configuration. La traduction avec *faire* aurait été possible, mais le traducteur semble avoir opté pour la périphrase *aller+inf* pour deux raisons. Sur le plan syntaxique, *will* est utilisé avec un localisateur temporel *now,* ce qui facilite le choix de la périphrase (Celle 1994, 105-106). Sur le plan sémantique, le traducteur a

interprété le modal *will* dans son sens radical de décision immédiate (il n'a donc pas la valeur temporelle de futur) et a répété le verbe *tirer*. Les outils de traduction automatique ont directement traduit *will* par *faire* (reprise anaphorique) au futur pour rendre, mais par erreur, la valeur de futur qui lui est souvent attachée, et sa valeur de décision entraînant une action immédiate, dans ce contexte.

L'exemple ci-dessous illustre un deuxième cas :

(116)

| | |
|---|---|
| - Can I be there, Billy? Will you be certain to call me? It's my colt.<br>- Sure I'**ll call you**. Of course I **will Ø.** | - Je pourrai être là, Billy ? Tu es certain que tu m'appelleras ? C'est mon poulain.<br>- Oui je t'appellerai, bien sûr que je **le ferai**.<br>(TR) |
| | - Puis-je être là, Billy ? Vous assurerez-vous de m'appeler ? C'est mon colt(poulain). «<br>- Sûr je vous appellerai. Bien sûr d'accord.<br>(Reverso) |
| | - Je peux être là, Billy ? Seras-tu certain de m'appeler ? C'est mon colt.<br>- Je t'appellerai. Bien sûr que je **le ferai**.<br>(Systran) |
| | - Puis-je être là, Billy ? Serez-vous certain de m'appeler ? C'est mon poulain.<br>- Bien sûr que je t'appellerai. Bien sûr.<br>(Google Traduction) |
| | - Je peux être là, Billy ? Tu seras sûr de m'appeler ? C'est mon poulain.<br>- Bien sûr que je t'appellerai. Bien sûr que je **le ferai**.<br>(DeepL) |

Comme on l'observe, cet exemple présente une ellipse post-auxiliaire déclenchée par le modal *will*, où le syntagme *call you* est ellipsé. La seconde option pour le traducteur humain est de se servir des reprises anaphoriques comme l'usage du verbe *faire* et des pronoms *le, la, l', les* le permet. Il a ainsi introduit le pronom *le* pour reprendre l'antécédent *call you* et ajouté *faire*. En réalité, *faire* est souvent utilisé pour traduire les reprises des actions précédentes, il prend ainsi en charge à lui tout seul ce qui a été énoncé. Les outils de traduction automatique Systran et DeepL ont traduit l'ellipse de la même manière que le traducteur humain, tandis que Google l'a

entièrement supprimée. Reverso, par contre, effectue une transposition[153], en passant de l'ellipse du syntagme verbal *call you* déclenchée par *will*, à l'utilisation des locutions adverbiales seules *bien-sûr* pour traduire *of course*, et *d'accord* pour accentuer le propos.

Cependant, l'utilisation de *faire* n'est pas toujours possible pour traduire les modaux qui déclenchent les ellipses. En effet, l'introduction de *faire* rend l'exemple ci-dessous (117) inacceptable.

(117)

| | |
|---|---|
| - Are they coming?<br>- Well, they said they **would Ø**. | - Ils vont venir?<br>- Enfin, ils ont **dit qu'ils viendraient**.<br>(TR) |
| | - Viennent-ils?<br>- Eh bien, ils ont **dit qu'ils**.<br>(Reverso) |
| | - Ils viennent ?<br>- Ils ont **dit qu'ils le feraient**.<br>(Systran) |
| | - Ils viennent?<br>- Eh bien, ils ont **dit qu'ils le feraient**.<br>(Google Traduction) |
| | - Est-ce qu'ils viennent ?<br>- Ils ont **dit qu'ils le feraient**.<br>(DeepL) |

L'ellipse dans cet exemple est déclenchée par le modal *would. Are they coming* renvoie à un futur implicitement daté signifiant *anytime soon.* Pour restituer cette ellipse, des ajustements syntaxiques sont nécessaires. En effet, l'antécédent *are coming,* étant à la forme progressive *be + ing*, ne pourra apparaître tel quel dans le site elliptique, puisque *would* nécessite une base verbale dans ce contexte (à savoir, le discours rapporté *We will come.*) Ainsi, seul *come* pourra être restitué. La résolution de cette ellipse est aussi possible avec *would be coming* pour marquer une insistance mais le sens en serait alors modifié.

[153] « La transposition est un procédé qui consiste à remplacer une catégorie grammaticale (traditionnellement appelée partie du discours) par une autre sans changer le sens de l'énoncé ». (Chuquet & Paillard 1987, 11).

La question qui se pose à l'issue de cet examen est de comprendre pourquoi le traducteur choisit de traduire l'ellipse *would Ø* par un conditionnel et non par un futur simple ? *Ils ont dit qu'il viendraient*, **ils ont dit qu'ils viendront* ou simplement *Ils ont dit que oui* sont-ils possibles ? Or, on peut constater ici que le système anglais et le système français fonctionnent de la même manière lorsqu'il s'agit d'un discours rapporté. En effet, *would* dans cet exemple révèle soit une visée exprimée par un autre locuteur soit un futur teinté d'une forme d'incertitude, car oui, ils ont prévu de venir, mais l'incertitude provient du temps qui s'écoule entre le moment où est énoncée la décision de *venir* (qui se situe dans le passé) et celui où est rapporté ce qui a été décidé (qui se déroule au présent). Dans ce laps de temps, tout changement est envisageable, contrôlé ou non par le sujet. Dans cette perspective, c'est donc cette incertitude qui rend la projection dans le futur plus hypothétique qu'un futur simple, et ceci est dû bien évidemment à la forme inhérente au discours rapporté, mais également à l'énonciation (l'action de dire est antérieure à la réalisation du projet). De ce fait, devant la marque conditionnelle tangible de *would*, les outils de traduction, excepté Reverso, le traduisent systématiquement par un conditionnel en français. Cependant, le choix du verbe (venir) au conditionnel, constitue un point de divergence entre la traduction humaine (qui répète ce verbe *venir*) et la traduction automatique (qui utilise *faire*).

En français, *venir* est un verbe de déplacement qui ne semble pas accepter *faire* comme équivalent. Il apparaît alors évident que le recours au verbe *faire* ne peut fonctionner dans tous les cas. Une explication tiendrait au type du verbe. Ainsi, dans le cas des verbes transitifs, l'emploi du verbe *faire* semble acceptable lorsque le complément est explicite :

> *Est-ce que l'enfant a mangé ? Oui, il l'a fait
> Est-ce que l'enfant a mangé sa pomme ? Oui, il l'a fait.
> (Pourrait éventuellement être accepté)

Cette piste de réflexion est certes très fragile, puisque la notion d'acceptabilité, encore une fois, est elle-même très relative d'un locuteur à un autre. Il nous paraît cependant important de proposer d'examiner les occurrences dans lesquelles des verbes de substitution comme *faire* prennent la place de l'ellipse, tout en sachant

que l'observation et la détection de ces phénomènes (accepter ou non le verbe *faire*), ne peuvent, à ce stade, mener à des conclusions générales.

Si l'utilisation de *faire* dans les traductions des ellipses post-auxiliaires en français s'avère parfois nécessaire pour pallier le vide non permis, celle des pronoms clitiques *en* et *y* permet également la résolution de l'ellipse. Dans l'exemple (118) ci-dessous, l'ellipse du syntagme verbal *go and get a bit of air* a été traduite par le traducteur humain qui a eu recours à l'usage du pronom clitique *y* placé entre le sujet *j'* et le verbe *vais* pour interpréter le syntagme manquant.

(118)

| | |
|---|---|
| - You go and get a bit of air.<br>- Yes, I **will Ø. I will Ø**. | - Va prendre un peu l'air.<br>- Oui, j'**y** vais. J'y vais.<br>(TR) |
| | - Vous allez chercher un peu d'air.<br>- Oui. **D'accord.**<br>(Reverso) |
| | - Tu vas chercher un peu d'air.<br>- Oui, **je le ferai. Je vais**<br>(Systran) |
| | - Vous allez prendre un peu d'air.<br>- Oui. **Je vais.**<br>(Google Traduction) |
| | - Va prendre un peu d'air.<br>- Oui, **je le ferai. Je vais le faire.**<br>(DeepL) |

C'est d'ailleurs l'absence du pronom clitique *y* dans les traductions de Systran et Google Traduction qui signale, dans le cas de l'ellipse en particulier, que la prise en compte du contexte n'est pas encore parfaite dans ces systèmes de traduction automatique. En effet, même le traducteur DeepL qui, dans certains exemples, a réussi à produire des traductions presque identiques à celles du traducteur humain, ne semble pas gérer cette contrainte puisque le pronom sélectionné est *le* et le vide sémantique est comblé, une fois encore par *faire*, traduction non pertinente ici car elle rend la phrase incomplète (pour les mêmes raisons que dans l'exemple 117 *supra*).

### *1.4.* Ellipses déclenchées par *have* et *be*

Les ellipses post-auxiliaires déclenchées par un auxiliaire *have* ou *be* sont, elles aussi, souvent sources d'erreurs dans les traductions. Par exemple, lorsque le traducteur humain rencontre une ellipse de l'attribut, il utilise généralement les pronoms *le, la, l', les* pour traduire le vide syntaxique du texte source. Prenons par exemple l'échange (119) présenté ci-dessous :

(119)

| | |
|---|---|
| - Only in the sense that we are all **cowards** about something.<br>- You **are not** Ø.<br>- Oh yes, I **am** Ø. | - Seulement dans la mesure où nous sommes tous lâches devant une chose ou une autre.<br>- **Vous, vous n'êtes pas lâche.**<br>**- Oh si, je le suis**<br>(TR) |
| | Seulement dans le sens que nous sommes tous lâches de quelque chose.<br>- **Vous n'êtes pas**.<br>- **Bien sûr, je suis**.<br>(Reverso) |
| | - Seulement dans le sens où nous sommes tous lâches sur quelque chose.<br>- **Tu ne le fais pas**.<br>- **Oh oui, je le suis.**<br>(Systran) |
| | - Seulement dans le sens où nous sommes tous lâches à propos de quelque chose.<br>- **Tu n'es pas.**<br>- **Oh oui je suis.**<br>(Google Traduction) |
| | - Seulement dans le sens où nous sommes tous lâches à propos de quelque chose.<br>- **Vous ne l'êtes pas.**<br>**- Oh oui, je le suis.**<br>(DeepL) |

Comme on peut le voir, le syntagme nominal (le nom *cowards* suivi d'un complément introduit par une préposition) à fonction adjectivale *cowards about something,* est ellipsé dans les deux réponses. En réalité, il y a ici d'abord un processus d'identification marqué par la copule *be* (*we = cowards*), et de ce fait, le nom, en position alors d'attribut du sujet, acquiert une fonction quasiment adjectivale.

Le traducteur humain a choisi de traduire le segment initialement ellipsé dans le texte original par l'adjectif *lâche* dans sa première occurrence et par le pronom complément *le* dans sa deuxième occurrence. On remarque ensuite que les systèmes de traduction proposent l'adjectif *lâche* pour traduire *cowards* dans l'occurrence non elliptique qui précède l'ellipse mais aucune ne propose le syntagme nominal *des lâches*. À l'exception de DeepL qui propose une traduction sémantiquement et syntaxiquement acceptable, tous les outils de traduction automatique ont échoué à traduire le vide engendré par l'ellipse, précisément parce qu'ils l'ont conservée faute de l'avoir « interprétée » (faute donc d'avoir analysé les relations de dépendance[154]), produisant ainsi une traduction incomplète et inappropriée (notamment Systran avec le verbe *faire* et Google Traduction avec l'absence d'un pronom complément).

Dans l'exemple (120) ci-dessous (traduit en 2018[155]), l'ellipse du verbe *hurt* déclenchée par *have* rejoint notre analyse de *do* emphatique.

(120)

| | |
|---|---|
| They've **hurt you**, I know they **have** ∅. | Ils vous ont blessé, **je le sais**. (TR) |
| | Ils t'ont blessé, je **le sais**. (Reverso) |
| | Ils vous ont fait du mal, **je sais**. (Systran) |
| | Ils vous font mal, **je sais qu'ils ont**. (Google Traduction) |
| | Ils t'ont blessé, **je le sais**. (DeepL) |

Comme précédemment avec *do*, bien que la reprise de l'auxiliaire paraisse tout à fait logique*, have* exprime ici, à notre avis, une insistance que l'intonation permet bien sûr de valider. Du point de vue théorique, il est généralement admis que l'emphase exprimée avec les modaux et *have* et *be* n'est pas source d'ambiguïté

[154] Nous faisons cette observation ici même si nous avons écarté la détection à partir d'une analyse de dépendance dans la présente recherche.
[155] Voir note de bas de page n° 147.

comme peut l'être, par contre, celle exprimée par *do*. Selon Larreya & Rivière (2005, 272) :

> Avec *BE, HAVE* et les modaux, la forme emphatique ne pose pas de problèmes de syntaxe, puisqu'elle se manifeste uniquement par une accentuation particulière de l'auxiliaire. En revanche, il y a un risque d'erreur en ce qui concerne l'utilisation du *DO* emphatique avec le présent et le prétérit simples.

Ceci ne semble pourtant pas être le cas dans la traduction automatique de Google qui opte pour un calque en maintenant la séquence elliptique bien qu'elle soit dépourvue de sens. Les autres systèmes, en revanche, traduisent comme l'humain qui a recours au pronom *le* pour traduire *they have Ø.* Cet exemple montre par ailleurs que le tutoiement et le vouvoiement est également source de confusion dans la TA.

Nous avons vérifié à nouveau, en avril 2019, la traduction de cet exemple par Google qui propose désormais une nouvelle traduction : *Ils vous ont fait mal, je sais qu'ils l'ont fait*. Malgré le recours au verbe anaphorique *faire* pour combler le vide initial et l'introduction du pronom *l'* pour reprendre l'antécédent, et malgré son acceptabilité syntaxique, cette traduction ne nous apparait pas davantage recevable au niveau de la compréhension.

Illustrant à présent une autre ambiguïté, l'exemple (121) ci-dessous présente une ellipse déclenchée par *be* où la forme aspectuelle [be] *trying to be nice* n'est pas prise en compte :

(121)

| | |
|---|---|
| - I was trying to be nice.<br>- **Were you** Ø?<br>- Yes. | - J'essayais d'être aimable.<br>- **Vraiment ?**<br>- Oui.<br>(TR) |
| | - J'essayais d'être gentil.<br>- Vous **l'**avez été ?<br>- Oui.<br>(Reverso) |
| | - J'essayais d'être sympa.<br>- Vous étiez ? (Systran)<br>- Oui.<br>(Systran) |
| | - J'essayais d'être gentil.<br>- Vous **l'**avez été ?<br>- Oui.<br>(Google Traduction) |
| | - J'essayais d'être gentil.<br>- C'est ce que **tu faisais** ?<br>- Oui, oui.<br>(DeepL) |

Le locuteur demande au co-locuteur de confirmer son assertion sur un ton moqueur. Le traducteur humain a choisi d'effacer toute référence à ce dernier et a opté pour une transposition en traduisant par l'adverbe *vraiment ?* introduisant implicitement le doute et, de ce fait, exprimant une connotation amusée, s'ajoutant à celle de l'intonation interrogative (renforçant le doute et par là, la moquerie). Ce choix semble donc répondre au mieux en termes d'équivalence à la question de l'original, laissant deviner une ironie sous-jacente sans pour autant surcharger le texte, ce qui donne une certaine marge d'interprétation.

Les systèmes de traduction automatique, quant à eux, ont tous conservé les mêmes paramètres linguistiques, temps verbaux et pronoms. Cependant, si les exemples traduits par Reverso, Systran et Google paraissent complets au niveau syntaxique, faute du repérage de l'antécédent, une confusion s'installe concernant le sens de la question elliptique. DeepL a traduit par *faire* l'action *d'essayer d'être gentil.* Hors contexte précis, cette interprétation semble également éloignée du texte source en mettant l'accent sur le verbe *essayer* (action), d'où l'emploi du verbe *faire,* alors que le message privilégie plutôt *être* (l'état) *gentil*. Les systèmes de traduction

ne savent pas reconnaître la valeur ponctuelle de la forme –*ing* et semblent ne retenir que la valeur dite progressive qui n'est compatible qu'avec les verbes d'action.

D'une manière générale, si le français possède bien des auxiliaires pour traduire ceux de l'anglais, la traduction automatique des ellipses révèle l'écart entre leur fonctionnement. Pour rappel, les vides syntaxiques autorisés après les auxiliaires en anglais ne le sont pas en français. Pour y remédier, dans les exemples que nous avons traités concernant les ellipses post-auxiliaires *have* et *be,* l'usage des pronoms anaphoriques constitue l'une des stratégies adaptées. Les outils de traduction automatique parviennent parfois à ajouter ces pronoms, mais dans la plupart des cas ils se contentent de reproduire le vide de la langue source, générant de ce fait des énoncés non acceptables. Comme il est difficile d'identifier les conditions sous lesquelles l'outil de traduction opère des ajouts ou non de pronoms, dont la distribution semble aléatoire, les conclusions à en tirer sont malheureusement prématurées et ne peuvent être généralisées à ce stade.

### 1.5. Ellipses déclenchées par *to*

Les ellipses déclenchées par *to* ressemblent à celles déclenchées par *have* et *be* dans la mesure où leurs traductions requièrent aussi des pronoms. Ces ellipses produisent un effacement de la base verbale qui suit le marqueur de l'infinitif *to*. Comme nous l'avons vu dans les chapitres précédents, certains verbes précédant *to* favorisent l'apparition de ces ellipses comme *want, allow, order.* Comme nous allons le constater, dans les exemples observés, le traducteur humain adopte trois stratégies pour traduire le vide : l'utilisation des pronoms *le, les, la, l', en* (122), la suppression entière de l'ellipse (123) ou le changement de la catégorie d'ellipse (124). Tandis que les deux premières relèvent des contraintes syntaxiques et grammaticales, la dernière dépend directement du choix du traducteur.

L'exemple (122) ci-dessous illustre le cas d'une ellipse du syntagme verbal *go back in the mountains again.*

(122)

| | |
|---|---|
| - Didn't you ever **go back in the mountains again**?<br>- No.<br>- Didn't you ever **want to** Ø? | - Vous n'êtes jamais retourné dans les montagnes ?<br>- Non.<br>- Et vous **n'en avez jamais eu envie** ?<br>(TR) |
| | - N'êtes-vous pas jamais retournés dans les montagnes de nouveau ?<br>- Non.<br>- **N'avez-vous pas jamais voulu ?**<br>(Reverso) |
| | - Tu n'es plus jamais rentré dans les montagnes ?<br>- Non.<br>- **Tu n'as jamais voulu ?**<br>(Systran) |
| | - Tu n'es plus jamais retourné dans les montagnes ?<br>- Non.<br>- **Tu ne l'as jamais voulu ?**<br>(Google Traduction) |
| | - Tu n'es jamais retourné dans les montagnes ?<br>- Non.<br>- Tu n'**en** as jamais eu envie ?<br>(DeepL) |

Le traducteur humain a eu recours au pronom *en* pour traduire le segment omis dans la version anglaise, ce que DeepL a parfaitement réussi à faire. La traduction de Google est acceptable avec le pronom *le* (parfaite syntaxiquement) tandis que celle de Systran est incomplète grammaticalement mais reste acceptable à l'oral. Celle de Reverso est agrammaticale en raison de la double négation *pas jamais*. L'absence de complément au verbe *vouloir* dans les traductions produites pas Reverso et Systran rend également leur acceptabilité douteuse.

Dans l'exemple (123), le traducteur humain a choisi de supprimer l'ellipse initiale où *like the fight game much* est ellipsé en anglais en traduisant l'aspect révolu et l'ellipse par l'adverbe *non*, ayant ainsi recours à la fois à une transposition et une modulation d'ordre temporel.

(123)

| | |
|---|---|
| You don't like the fight game much, **do you**?<br>I used **to Ø**. | - Tu n'aimes pas les jeux violents, n'est-ce pas ?<br>**- Non**<br>(TR) |
| | - Vous n'aimez pas beaucoup la partie de combat, n'est-ce pas ?<br>- Je **le faisais avant**.<br>(Reverso) |
| | - Vous n'aimez pas beaucoup le jeu de combat, n'est-ce pas ?<br>**- J'avais l'habitude**.<br>(Systran) |
| | - Vous n'aimez pas beaucoup le jeu de combat, n'est-ce pas ?<br>**- J'avais l'habitude de**.<br>(Google Traduction) |
| | - Tu n'aimes pas beaucoup le jeu du combat, n'est-ce pas ?<br>- **J'en avais l'habitude**.<br>(DeepL) |

Par cette traduction, il n'établit aucune référence linguistique au sujet de l'énoncé et insiste sur le présent de la situation, car en réalité l'interlocuteur aimait les jeux violents dans le passé mais plus maintenant. On pourrait ainsi considérer que la traduction humaine ne traduit pas « tout » ce que dit l'original dans ce cas.

Les systèmes de traduction automatique ont tous produit des séquences agrammaticales et sémantiquement incomplètes. Pourtant, la mention de ce sentiment qui appartient à un passé révolu est bel et bien présente : Reverso a eu recours au verbe *faire*, incompatible avec un verbe de sentiment, pour reprendre l'énoncé qui précède. Il le conjugue à l'imparfait et ajoute l'adverbe *avant* pour marquer l'antériorité, et donc le révolu. Systran a opté pour l'utilisation de l'imparfait et le nom (habitude) auxquels Google ajoute la préposition *de* suivi d'un point pour marquer la fin de la phrase et DeepL le pronom *en* pour reprendre le segment omis. En effet, les systèmes ont bien exprimé l'antériorité d'une situation ou d'un sentiment qui n'était pas ponctuel, d'où l'imparfait, les adverbes (avant) et le lexique même (habitude) mais ont traduit comme si un sentiment pouvait être une *habitude*, produisant par là une erreur sémantique. Syntaxiquement, Reverso a repéré le vide

du texte source et utilise donc systématiquement le pronom *le*. Systran calque la structure de la version originale et Google a interprété le rôle de *to* (marqueur d'infinitif) en utilisant *de* pour introduire le complément d'objet dans le groupe verbal.

Ces traducteurs suivent d'ailleurs exactement le même processus pour traduire l'exemple (124).

(124)

| | |
|---|---|
| - **Ø** Broken off that engagement?<br>- I was forced **to Ø**. | - Tu as rompu tes fiançailles ?<br>- **Bien obligé**.<br>(TR) |
| | - Vous avez rompu cet engagement ?<br>**- J'ai été** forcé **de le faire**.<br>(Reverso) |
| | - Rompre cet engagement ?<br>**- J'ai été** forcé.<br>(Systran) |
| | - Rompu cet engagement ?<br>- **J'ai été obligé de**.<br>(Google Traduction) |
| | - Tu as rompu ces fiançailles ?<br>- **J'ai été** forcé **de le faire**.<br>(DeepL) |

(124) présente deux ellipses en anglais : l'une dans la question fragmentaire où un auxiliaire et un sujet manquent, et l'autre déclenchée par *to,* où le segment *break off the engagement*[156] est omis. En ce qui concerne la première ellipse, tous les traducteurs à l'exception de Systran (infinitif inadapté) et de Google (calque non permis), ont traduit le sujet manquant pour former une question complète qui relève ici d'un registre standard (aucune inversion des positions sujet et verbe). L'ajout de *le faire* (Reverso et DeepL) ne perturbe pas l'acceptabilité de la traduction de l'ellipse déclenchée par *to*, et l'occurrence est alors complète.

Si les traductions automatiques des exemples (123) et de (124) sont similaires, le traducteur humain a, quant à lui, introduit une nouvelle catégorie d'ellipse en

[156] La restitution de cet élément dans « la structure de surface » requiert des ajustements.

donnant une réponse elliptique où le sujet *je* et l'auxiliaire *être* sont omis. Il opère ainsi le passage d'une ellipse post-to à une ellipse dans la réponse fragmentaire. Malgré un contenu grammatical incomplet, le contenu sémantique du texte source est transmis. Cette traduction ne relève pas de contraintes syntaxiques. Elle est acceptable et dépend du choix du traducteur qui aurait pu en proposer d'autres *j'étais obligé*, par exemple.

Les ellipses déclenchées par *to* sont fréquentes en anglais. Il s'agit d'une ellipse non permise en français puisque les prépositions, de par leur nature, (ici *de* qui introduit un infinitif) ne peuvent jamais se trouver en fin de phrase. Ces exemples montrent que le problème rencontré par les systèmes de traduction automatique dans ce cas relève de l'apprentissage de contraintes syntaxiques puisque le contenu de l'énoncé est resté fidèle à l'original. Si certains systèmes parviennent à respecter la règle et à utiliser des pronoms pour pallier le vide, d'autres se contentent de le reproduire. Chez Google, cette erreur persiste depuis 2014 et la même traduction continue à être proposée. Lorsque Google échoue à traduire cette occurrence, et que même Reverso, qui jusqu'à présent a produit des traductions erronées, parvient à contourner le problème évoqué, on ne peut manquer d'être surpris. Quelle est donc la source de cette erreur récurrente ? La réponse à cette question nous échappe encore.

En résumé, la fréquence de l'ellipse post-auxiliaire en anglais pourrait laisser croire qu'elle facilite la traduction automatique à travers des données d'entrainement contenant suffisamment d'occurrences et d'exemples de traduction. Cependant, le fonctionnement différent des déclencheurs dans chacune des deux langues joint à la difficulté de repérer l'antécédent du site elliptique font obstacle : pour l'instant, ils entravent la réalisation d'une traduction acceptable.

C'est ainsi qu'une détection automatique efficace des déclencheurs d'ellipses et la classification des ellipses mêmes (classification morphosyntaxique à partir de l'élément déclencheur, ou syntaxique à partir de la fonction des éléments ellipsés) pourraient contribuer à cerner les caractéristiques nécessaires à leur reconnaissance

par les systèmes de traduction pour les traduire en fonction de leur acceptabilité ou non dans la langue cible.

## 2. La traduction des questions et des réponses fragmentaires

D'autres cas que l'ellipse post-auxiliaire mériteraient d'être analysés, notamment les questions et les réponses fragmentaires (approche similaire à celle menée par Bouillon *et al.* 2005). Ainsi, le choix de traiter les questions fragmentaires est motivé par notre volonté d'élargir la définition de l'ellipse pour mieux cerner le problème qu'elles posent à sa traduction, dans la mesure où l'absence du sujet est grammaticalement inhabituelle, voire contraire aux règles de la grammaire de la langue écrite en anglais et en français. Cette ellipse se présente en effet assez fréquemment dans le discours spontané et les conversations quotidiennes, manifestant ainsi l'écart entre la règle et l'usage. Il arrive souvent de rencontrer dans un dialogue en anglais une question ou une réponse commençant par un participe passé ou présent (formes en *–en* ou *–ing*), élément d'un groupe verbal dans lequel le sujet et l'auxiliaire ont été ellipsés comme dans cet exemple :

(125) - What are you doing?
- Eating

En français, dans des échanges familiers, l'ellipse du sujet et de l'auxiliaire est parfois possible quand il y a un participe passé.

Les enfants, vous n'avez pas le droit de sortir du jardin. Compris ?

Ces constructions sont porteuses de questions théoriques que nous pensons intéressantes à soulever. Observons la traduction humaine de l'exemple (126) ci-dessous. En français, seul un complément d'objet, sans aucune forme verbale, est traduit. Pour ce cas précis, nous nous servons également des traductions françaises pour présenter quelques pistes de réflexions théoriques.

(126)

| | |
|---|---|
| - Where've you been?<br>- Just **doing stuff**, you know. | - Tu faisais quoi ?<br>- **Des trucs**, tu sais bien. |

D'un point de vue contrastif, la traduction de (126) pose problème car *tu faisais quoi ?* pourrait être considéré comme une traduction erronée de *Where've you been ?* dont la traduction attendue serait plutôt *T'étais où?* . Le traducteur a néanmoins su adapter sa traduction pour rendre le sens voulu (ou le « vouloir dire » du locuteur) : dans la phrase anglaise, *doing* ne répond pas à la question, mais en déplace l'objet qui était initialement un lieu. Il pourrait cependant y avoir un contenu implicite d'activités, non-dit, dans la question *where have you been?* ce que semble indiquer la réponse effective de l'interlocuteur en anglais. En effet, imaginons cet exemple comme étant issu d'un échange emprunté au discours quotidien : - « Tu étais où ? » - « Je faisais des trucs » ou bien « Je rangeais mes outils » : la réponse décrit une activité (non explicitement sollicitée) qui présuppose le lieu où cette même activité se déroule atelier/garage/cabane, immédiatement connu des deux locuteurs et répondant ainsi indirectement à la question posée explicitement. De ce fait, ne pourrions-nous pas nommer *ellipse de discours* ce type d'ellipses « croisées » à l'intérieur d'un dialogue (figure 25) ?

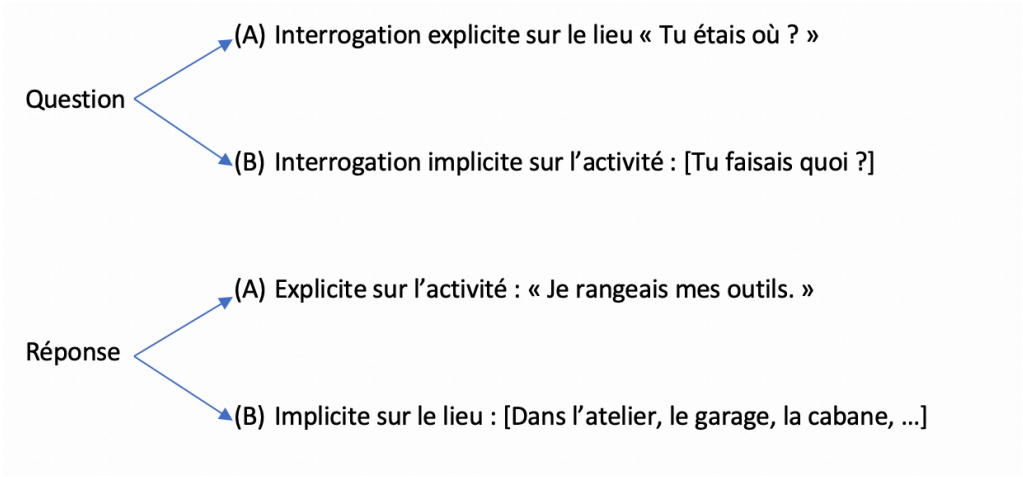



**Figure 25 : Proposition d'un schéma d'ellipse de discours (ici ellipses croisées)**

Ces constructions sont complexes, y compris pour le traducteur humain, dans la mesure où il ignore s'il s'agit d'une omission intentionnelle ou non-intentionnelle[157]. Dans quels cas doit-il les garder telles quelles ou les interpréter et les *restituer* ?

En tout cas, ce type d'ellipse (effacement du sujet et de l'auxiliaire) est interprété et restitué dans la langue cible dans les cas observés dans notre corpus. Ainsi, pour traduire ces occurrences dans les exemples que nous avons parcourus, le traducteur a parfois recours à l'usage d'autres structures qui s'éloignent grammaticalement du texte original. Prenons l'exemple de l'aspect *be+ing* que le français et l'anglais ne partagent pas entièrement : lorsque cette forme exprime la continuité d'une action, il n'est pas rare de la voir (sans doute maladroitement) traduite par la locution *en train de + infinitif* ou, plus naturellement, par l'ajout d'autres marqueurs comme l'adverbe *encore* pour marquer l'insistance, elle aussi véhiculée par la valeur dite modale de *be+ing*.

Prenons à ce stade l'exemple (127) ci-dessous :

[157] Nous pouvons défendre l'idée d'une omission intentionnelle puisque la réponse « des trucs » peut être une façon d'éluder la question *Where've you been ?*

(127)

| | |
|---|---|
| - Ø **Staying** here long?<br>- Not long. | - **Vous restez** longtemps ?<br>- Pas longtemps.<br>(TR) |
| | - **Séjour** ici longtemps ?<br>- Non longtemps.<br>(Reverso) |
| | - **Rester** ici longtemps ?<br>- Pas longtemps.<br>(Systran, Google, DeepL) |

La forme *–ing* de *staying* dans l'interrogation montre qu'il y a une ellipse de l'auxiliaire *are* et du sujet *you* (inféré dans la situation). Dans cette forme aspectuelle, sont transmis non seulement le temps présent (indiqué par l'auxiliaire effacé) et l'ancrage dans une situation particulière, mais aussi une question sur l'intention de l'interlocuteur. Une valeur d'intention (implicite) est couplée à la durée sur laquelle la question porte (*long*). La valeur d'intention peut alors rester implicite, ce qui se passe lors de la traduction. Pour ces raisons, et parce que le verbe *rester* est un verbe d'état, le traducteur est passé de la forme aspectuelle en anglais au présent simple *restez* (**êtes-vous en train de rester longtemps ici?* ne peut être acceptable). Le traducteur humain fait apparaître dans la traduction le sujet *vous,* effacé dans la version originale, choisissant le pronom à la seconde personne *vous* (pluriel ou politesse).

Dans ce cas précis, les systèmes de traduction automatique ont inclus la notion de durée dans les choix lexicaux : le verbe *rester* et le nom *séjour* impliquent, sémantiquement, une notion de durée. Dans l'exemple ci-dessus, la forme *–ing* est traduite automatiquement par l'infinitif *rester* ou par le substantif *séjour*, les systèmes réalisant de « mauvaises » transpositions, ce qui rend les traductions non acceptables. On peut donc se demander si le problème n'est pas tant la non reconnaissance de la valeur de durée que celui de la non reconnaissance de la forme grammaticale (verbale). Les erreurs relèveraient alors d'un problème de grammaire/syntaxe plutôt que de la sémantique.

Dans l'exemple (128) ci-dessous, un sujet et un auxiliaire ou un modal sont omis :

(128)

| | |
|---|---|
| - Ø **Pinch** anything ?<br>- No. | **- T'as chipé** quelque chose ?<br>- Non.<br>(TR) |
| | - **Pincement** quoi que ce soit ?<br>- Non<br>(Reverso) |
| | - **Pincez** tout ?<br>- Non.<br>(Systran) |
| | - **Pincez** quelque chose ?<br>- Non.<br>(Google Traduction) |
| | - **Vous avez pincé** quelque chose ?<br>- Non.<br>(DeepL) |

En l'absence d'un antécédent linguistique, la structure grammaticale de la phrase elliptique *Pinch anything?* pourrait sembler autoriser deux interprétations : *Modal + you pinch anything?* ou *Did/Do you pinch anything?* Ainsi, sont ellipsés soit le sujet et le modal*,* soit un sujet et l'opérateur *do*.

Le traducteur humain a choisi une traduction sans modal, compte tenu des valeurs sémantiques du verbe *pinch* incompatibles avec les modaux et opté pour un passé composé, *as chipé*, révélant ainsi que *Pinch anything* ne peut être perçu que comme la forme incomplète du SV *did you pinch anything ?* et, sémantiquement, référant à une action (répréhensible) que le protagoniste aurait pu faire (éventualité).

Pour ce qui est de la traduction automatique de cette ellipse, la base verbale *pinch* a posé problème en raison de sa polysémie et du registre familier de l'échange. On remarque d'abord que le verbe lexical *pincer*, utilisé par les systèmes pour traduire *pinch* correspond à une erreur de choix de la traduction automatique. En français ou en anglais, si l'on considère le sens du mot dans ce contexte précis, on « pince » le voleur mais pas l'objet dérobé. En anglais, dans cette acception et ce registre, son emploi est considéré comme démodé ou vieilli. Le mot *chiper*, choisi par le traducteur humain, révèle qu'il a connaissance du contexte puisque le verbe appartient au vocabulaire familier. *Pinch* est ensuite traduit en TA, par un nom *pincement* (Reverso) et par un verbe *pincez* (Systran et Google Traduction) et, dans tous les cas, par une

question marquée par le point d'interrogation (reproduisant sans problème la modalité énonciative de l'original). La comparaison des traductions donne l'avantage à celle de DeepL, intéressante dans la mesure où, hormis le faux-sens, sa proposition est proche de celle du traducteur humain dans sa forme : restitution de la modalité interrogative, du sujet, du temps et de l'aspect du verbe de la version source.

Enfin, un dernier cas, illustrant l'ambiguïté de ce type d'ellipse, dans la traduction automatique apparaît dans l'exemple (129) ci-dessous :

(129)

| | |
|---|---|
| - Ø Ever **been** anywhere near Maidenhead?<br>- No, never. | - **Vous ne vous êtes jamais trouvé** du côté de Maidenhead ?<br>- Non, jamais.<br>(TR) |
| | - **Jamais été** n'importe où près de Virginité ?<br>- Non, jamais.<br>(Reverso) |
| | - **J'ai jamais été** près de Maidenhead ?<br>- Non, jamais.<br>(Systran) |
| | - **Jamais été** près de Maidenhead ?<br>- Non jamais.<br>(Google Traduction) |
| | - **Vous avez déjà été** près de Maidenhead ?<br>- Non, jamais.<br>(DeepL) |

On remarque une ellipse du sujet *you* et de l'auxiliaire *have* (la flexion de *been* est la trace de cet effacement) dans *Ever been anywhere near Maidenhead?* Sans doute pour un effet de style, le traducteur humain a-t-il traduit cet exemple en changeant la structure de la phrase (procédé de l'explicitation), puisque, à notre avis, une traduction par *vous avez déjà* (ou *avez-vous déjà* pour un registre plus soutenu) *été à proximité/dans les environs de Maidenhead auparavant/avant ?* aurait pu également convenir.

Du côté de la traduction automatique, Systran a échoué dans la reconnaissance du sujet manquant qu'il a traduit par *je*, alors que les autres systèmes l'ont détecté même si Reverso et Google ne l'énoncent pas explicitement (il s'agit aussi d'une

ellipse dans la traduction). On soulignera enfin que DeepL explicite le sujet et propose une traduction aussi pertinente que celle d'un traducteur humain.

En résumé, l'ambiguïté résultant de l'ellipse de l'ensemble sujet-auxiliaire dans la question fragmentaire bloque, en traduction automatique, la réalisation adéquate des consignes de transfert d'une langue à l'autre. La non-reconnaissance de l'antécédent linguistique et la non-saisie des paramètres contextuels n'ont pas encore trouvé de résolution adaptée. Il reste encore trop de facteurs intervenant dans le processus de traduction que l'apprentissage automatique, pour performant qu'il soit, ne prend pas encore en compte. De plus, étant donné que le traducteur humain éprouve lui-même des difficultés à produire une *traduction-norme* pour ce type d'ellipse, on ne s'étonnera guère de constater dans les systèmes un dysfonctionnement à l'égard de l'énoncé source. En effet, si *n'est-ce-pas ?* à titre d'exemple, apparait comme la traduction préconisée pour traduire les *question tags*, il n'existe aucune autre traduction *normée* pour l'ensemble des occurrences.

Par ailleurs, on notera que les réponses fragmentaires sont elles aussi des occurrences qui paraissent incomplètes – au regard de la construction canonique d'une phrase – mais transmettent un sens que les locuteurs récupèrent à l'aide de la situation d'énonciation, linguistique et/ou extralinguistique. Ces fragments sont souvent associés à l'étude de l'ellipse. On rappelle que ces occurrences n'ont pas constitué un objet de détection dans la présente recherche compte tenu du nombre variable de mots pouvant constituer un fragment. En effet, un seul mot peut constituer un fragment : interjection, réponse courte ou question simple. Ces fragments sont caractéristiques du discours spontané, des conversations brèves et des dialogues informels. Dans les cas où un fragment est constitué d'un seul mot, on s'attend à ce que la traduction automatique ne rencontre aucune difficulté et se passe ainsi du contexte pour le traduire de façon adéquate. En effet, l'ellipse n'a nullement besoin d'être restituée. L'exemple (130) illustre en quelque sorte cette aisance, notamment dans la traduction des mots isolés.

(130)

| | |
|---|---|
| - What is it, then?<br>- Fried bread. | - Qu'est-ce que c'est alors ?<br>- Du pain frit.<br>(TR) |
| | - Qu'est-ce que c'est, alors ?<br>- Pain frit.<br>(Reverso, DeepL, Google Traduction) |

Par contre, pour être complètement réussies, les traductions auraient dû tenir compte des fonctionnements différents des langues en matière de détermination nominale, puisqu'en français, il serait nécessaire de trouver le partitif *du* devant *pain frit*. Toutefois, certains cas posent parfois problème : tels les fragments constitués d'un ensemble de mots liés syntaxiquement entre eux par une catégorie invisible, *Ø you yourself ?* dans (131) qui présentent un réel défi pour les systèmes de traduction automatique. Ces fragments sont identifiés par Reich (2011) comme *situation-based ellipsis*.

(131)

| | |
|---|---|
| He on the other hand **bought** his cheeses in the Rue de Tocqueville, only round the corner from the apartment.<br>- **Ø** you **Ø** yourself? | Lui, par contre, **achetait** ses fromages dans la rue de Tocqueville, juste au coin, près de son appartement.<br>- Vous les **achetez** vous-même ?<br>(TR) |
| | Il d'autre part a acheté ses fromages à Rue de Tocqueville, seulement au coin de la rue de l'appartement.<br>- Vous-même vous?<br>(Reverso) |
| | Il a d'autre part acheté ses fromages à la Rue de Tocqueville, seulement au coin de l'appartement.<br>- Vous vous êtes ?<br>(Systran) |
| | Il a acheté ses fromages dans la rue de Tocqueville, au coin de l'appartement.<br>- Toi toi-même ?<br>(Google Traduction) |
| | D'autre part, il achetait ses fromages dans la rue de Tocqueville, au coin de la rue de Tocqueville, à deux pas de l'appartement.<br>- Toi-même ?<br>(DeepL) |

En l'absence de contexte, source d'ambiguïté avec le passage de la narration au questionnement direct, le fragment *You yourself?* est lui-même équivoque. Le traducteur humain se fie à son intuition et met en œuvre sa créativité pour livrer un énoncé *presque* valide dans la langue cible[158]. Les outils de traduction automatique, malgré la traduction littérale qui semblait opérationnelle dans la plupart des cas illustrés par l'exemple (131), ne parviennent pas à produire une traduction acceptable.

Parvenue au terme de notre analyse empirique des exemples présentés, nous récapitulons ci-dessous nos observations concernant la traduction humaine et

[158] On pourrait aussi suggérer une autre traduction (dont on ne peut dire si elle est plus juste que l'autre, en l'absence de contexte) : Et vous-même ?

automatique des occurrences elliptiques. Nous relevons principalement les problèmes et éléments suivants :

- Le recours au mot-à-mot : souvent produit par les systèmes de traduction automatique, accepté dans certaines configurations, mais pas dans celles qui nécessitent une emphase particulière pour prendre en compte les dimensions discursives.
- La traduction automatique de la *question tag* est la plupart du temps *n'est-ce pas ?* Les réponses données à cette question sont souvent insolites puisque les systèmes ne semblent pas pouvoir prendre en compte le contexte précédant l'expression *n'est-ce pas ?* L'expression en elle-même vide remplace le point d'interrogation et devient de ce fait une expression neutre.
- L'absence de traduction d'occurrences elliptiques et l'élimination fréquente de la valeur d'insistance : cette remarque s'applique notamment aux exemples des auxiliaires et des modaux emphatiques.
- La traduction des séquences elliptiques par des locutions adverbiales en procédant à des transpositions et modulations.
- La restitution des verbes lexicaux auxquels le modal ou l'auxiliaire se substituent : seul le traducteur humain a recours à cette solution puisqu'il lui est plus aisé de repérer l'antécédent de l'ellipse, tâche impossible pour les systèmes de la traduction automatique. Seul DeepL arrive à résoudre ces ambiguïtés lorsque l'antécédent se trouve dans la même phrase que le site elliptique.
- Le recours à l'utilisation des anaphores en français : soit des pronoms compléments eux-mêmes anaphoriques, soit du verbe anaphorique *faire* (pour signifier un processus, il prend seul en charge le sens de ce qui précède), utilisé avec modération chez le traducteur humain, mais fréquemment par les outils de la traduction automatique. Ces reprises renvoient généralement aux antécédents ayant la forme d'un groupe verbal. Nous avons également remarqué le recours à l'utilisation de l'ensemble *le + faire,* ou même d'autres

pronoms comme *en* et y, selon les contraintes et les configurations syntaxiques.

- L'expression indirecte de la modalité à l'aide d'autres marqueurs (en raison de l'absence d'équivalents stricts).
- La traduction de l'ellipse par une autre catégorie d'ellipse.
- Le choix du traducteur : la traduction n'est conditionnée par aucune contrainte syntaxique. Plusieurs traductions sont possibles et acceptables.
- L'infidélité au texte source et le recours à d'autres stratégies pour répondre aux exigences du registre.

## 3. Discussion : relevé des erreurs

Dans la mesure où les nuances et les subtilités d'ordre extralinguistique dont le traducteur humain est conscient, par exemple, ne peuvent pas être intégrées à l'heure actuelle dans les outils de TA, nous sommes amenée à nous demander selon quelles *normes* nous pouvons juger la traduction comme étant *bonne* ou *mauvaise*[159]. Parmi les occurrences analysées, nous avons remis en question l'acceptabilité d'énoncés malgré leur complétude syntaxique (en tout cas dans le cas de l'ellipse). S'agit-il alors d'erreurs ?

Pour tenter d'apporter une amorce de réponse à cette question, nous avons étudié les typologies déjà dressées de la traduction automatique pour vérifier s'il était possible de les utiliser dans notre évaluation de la traduction de l'ellipse.

Dans une étude concernant la traduction automatique par le système Systran, Loffler-Laurian identifie douze types d'erreurs récurrentes de la traduction automatique qu'elle ramène ensuite à dix (1983, 65-78) : (forme de verbes, mots isolés, expression de modalité, etc.). Mais à cette époque (années 1980), l'approche adoptée par la plupart des systèmes automatiques reposait encore sur une traduction automatique à base de règles. Quelques années plus tard, et malgré les progrès que la traduction statistique apportait, Grass identifie treize autres erreurs

[159] Ces questions sont liées aux critères d'évaluation d'une traduction donnée et à son acceptabilité (peut ne pas comporter d'erreurs et pourtant être jugée « mauvaise »), problèmes qui rejoignent celui de l'attestation des exemples utilisés dans cette recherche.

qu'il nomme métaphoriquement « les treize péchés capitaux de la traduction automatique » (néologie, ambiguïté syntaxique, polysémie, *etc.*) (Grass, 2010). Ces erreurs relèvent d'un niveau purement linguistique[160] et ne sont pas le résultat de l'évaluation d'un phénomène précis.

Dans notre travail, l'évaluation est faite sur la traduction d'un phénomène précis qu'est l'ellipse dans le système de traduction neuronale. Les traductions erronées sont liées à la fois au type même de l'ellipse et à la différence existant entre les systèmes de langues. On pourrait ainsi difficilement parler d'erreurs liées aux mots isolés (vocabulaire, polysémie, néologie, *etc*.) dans le cas de l'ellipse, puisqu'elle est, comme nous l'avons définie, cette omission qui touche à la phrase entière. On peut, par contre, rencontrer des mots isolés dans les réponses fragmentaires où un seul mot prononcé remplit les conditions énonciatives d'un échange et, comme nous l'avons vu, ces occurrences semblent aujourd'hui poser moins de problèmes à la traduction.

Plutôt qu'une véritable typologie d'erreurs qu'il serait prétentieux de présenter car cela nécessiterait l'analyse d'un échantillon plus large, nous avons considéré l'origine dont relèvent les erreurs rencontrées dans la traduction automatique. Elles peuvent être réparties selon trois grands ensembles.

### 3.1. Erreurs relevant de l'ambiguïté morphosyntaxique

Dans le cas précis de l'ellipse, lorsque les deux systèmes de langues anglais/français diffèrent sensiblement au niveau morphosyntaxique, nous avons observé que les outils de traduction automatique produisent des erreurs, rendant la phrase non seulement agrammaticale mais aussi dénuée de sens dans la langue cible. L'impact est donc à la fois syntaxique et sémantique. Ces différences se manifestent notamment dans le fonctionnement des modaux et des auxiliaires. L'examen de ces erreurs nous a alors permis, dans un premier temps, d'observer les traductions humaines considérées comme traductions de référence et les stratégies mises en

[160] Nous ne détaillerons pas l'ensemble de ces erreurs dans la présente recherche, qu'elles soient issues d'une traduction à base de règles (Loffler-Laurian) ou d'une traduction statistique (Grass). Elles sont répertoriées dans l'annexe VI p. 277.

œuvre pour contourner l'ambiguïté du vide laissé par l'ellipse et aboutir à une traduction *acceptable*. Il s'est agi ensuite de comparer les deux types de traduction : le traducteur, avec sa connaissance extralinguistique et culturelle parvient le plus souvent à lever l'éventuelle l'ambiguïté des énoncés, à tel point qu'on pouvait, dans certains cas, douter de la présence d'une ellipse dans le texte original, puisqu'elle n'apparaissait plus dans le texte traduit. En revanche, ce n'est pas encore le cas des systèmes de traduction automatique qui, pour l'instant, ont traduit le plus souvent l'ellipse de façon littérale (reproduisant le vide de l'original non permis dans la langue cible). Cette traduction littérale peut convenir mais parait incompatible parfois avec le registre relâché caractérisant les dialogues. En effet, les données qui servent à entrainer un système de traduction appartiennent généralement aux traductions humaines de textes bien formés respectant la grammaire de l'écrit. Le système n'est alors pas entrainé à traduire les occurrences de l'oral.

Le cas que nous avons analysé et qui illustre l'ambiguïté morphosyntaxique est la traduction de questions fragmentaires entièrement erronée dans la mesure où elle ne respectait ni l'agencement syntaxique de la langue cible, ni le contenu sémantique de la langue source. De ce fait, une ambiguïté morphosyntaxique est apparue inévitable puisque les systèmes de traduction n'étaient pas entrainés sur la syntaxe adaptée de l'oral[161]. Ce sont ces directions que Google, par exemple, semble poursuivre aujourd'hui en travaillant de plus en plus sur la reconnaissance de la parole[162] et les traductions simultanées[163].

Ainsi, si les exemples présentés dans ce chapitre montrent qu'à cette échelle il est prématuré d'aboutir à des remarques générales portant sur les erreurs générées par l'ellipse lors de sa traduction automatique, ils permettent toutefois d'avancer que les outils de la TA produisent une traduction erronée du phénomène lorsqu'ils doivent

[161] Parmi les grammaires de l'oral qui existent dans les recherches théoriques, nous notons par exemple la publication en 1999 de la grammaire de l'oral, *Longman Grammar of Spoken and Written English,* une grammaire qui énonce des règles d'usage pour ce qui est de l'anglais et *Approches de la langue parlée en français,* ouvrage publié en 1997, pour le français (voir la bibliographie).

[162] https://ai.googleblog.com/2019/03/an-all-neural-on-device-speech.html (consulté le 24 juillet 2019 à 16:06)

[163] https://support.google.com/googlepixelbuds/answer/7573100?hl=en&ref_topic=7558549 (consulté le 24 juillet 2019 à 16:10)

traiter non des similitudes, mais des différences entre les systèmes de langues. Le problème apparaît dans les cas où il n'existe pas de correspondance entre un token en langue source et un token en langue cible. Par ailleurs, ce que nos analyses ont cherché à mettre en avant, ce sont les difficultés liées à des corpus d'entraînement au sein desquels la fréquence du phénomène est trop faible (ce que le chapitre 3 a permis de constater) pour que les traductions générées par les systèmes automatiques soient fiables.

### 3.2. Erreurs relevant de l'acceptabilité de la traduction

Il est question ici des erreurs qui relèvent non de contraintes syntaxiques mais des énoncés eux-mêmes soumis à l'appréciation du traducteur, du chercheur ou du locuteur. Comme nous avons pu le voir dans certaines occurrences, décider de la validité des traductions n'est pas chose aisée, puisque certaines suggestions de traduction peuvent apparaitre plus adaptées que d'autres alors même que le choix initial du traducteur ne peut être considéré comme une erreur, du fait d'une traduction restant *acceptable*. C'est d'ailleurs à ce stade qu'il nous faut revenir sur la notion d'acceptabilité – que nous avons déjà évoquée dans le chapitre 2 – nécessaire à l'évaluation de la traduction. Il s'agit bien de recourir à des normes établies pour pouvoir comparer et évaluer la traduction du phénomène elliptique. Si juger de l'acceptabilité en fonction d'une norme fixée, institutionnelle par exemple, ne pose pas de problème, il est très difficile de trancher lorsque ces normes sont liées à la langue en tant que moyen de communication en constante évolution soumise à la pratique de ses locuteurs. En effet, si la langue évolue en raison de contacts, d'emprunts, de réformes, et si la norme est définie comme un ensemble de règles que l'on doit appliquer à la lettre, comment peut-on juger de l'acceptabilité d'une traduction lorsqu'un écart se manifeste ? Dans ce cadre précis, si la norme se limite aux règles grammaticales d'une langue, il sera aisé de refuser une phrase dont le verbe est par exemple incorrectement accordé. Le cas se complique, en revanche, lorsque cette norme fixe un usage qui transmet une réalité sociale de l'énoncé. Il convient ainsi de déterminer à partir de quelle norme nous formerons notre

jugement lorsque nous aurons à évaluer une traduction et/ou à la comparer à une autre[164].

Dans notre recherche, lorsque la langue cible n'admet pas l'ellipse de la langue originale, l'humain met en place des stratégies de contournement de la difficulté, soit en ne tenant pas compte de l'ellipse initiale, soit par un ajout, en adaptant la structure de l'énoncé jusqu'à son acceptabilité. De plus, par son invisibilité partielle, l'ellipse a imposé une première évaluation au niveau syntaxique, avant de considérer les variations induites par le genre et le style de la traduction, pour aboutir enfin à une évaluation sémantique.

Par ailleurs, les systèmes de traduction automatique ont réussi à compléter les vides syntaxiques du texte source, à maintes reprises, par l'ajout de *le faire* par exemple dans le texte traduit. Mais cet ajout n'est pas toujours compatible et dépend du contexte, du co-texte et de l'énoncé même. Au cours de ce chapitre, nous avons eu recours à deux locuteurs natifs pour évaluer l'acceptabilité de la traduction à chaque fois qu'une difficulté à juger de l'acceptabilité des occurrences s'est manifestée. Sauf à lier cela aux données d'entrainements, il est difficile aujourd'hui de comprendre selon quels critères les systèmes fonctionnent avec l'ajout de *le faire* pour traduire l'ellipse. Il serait donc intéressant, pour juger de l'acceptabilité des traductions, d'effectuer des tests réunis dans des protocoles à réaliser auprès de locuteurs natifs (tous âge, sexe et catégorie socioprofessionnelle confondus), jusqu'à l'obtention de statistiques probantes pour développer des approches précédant (pré-traduction) ou suivant (post-édition) la traduction automatique, l'objectif étant la prédiction des erreurs et leur correction automatiques en proposant, entre autres, des substitutions par *faire* qui seraient acceptables. Parmi les recherches menées

[164] Bien que définir la norme se présente comme une tâche relativement complexe, Lavault-Olléon & Allignol (2014, 3), dans le cadre d'une traduction professionnelle, ont tenté d'éclairer le processus en établissant une distinction entre normes implicites et normes explicites qu'elles présentent comme très liées, complémentaires l'une de l'autre, permettant ainsi l'utilisation d'une sorte de grille dont le traducteur comme l'évaluateur peuvent se servir. Selon Lavault-Olléon & Allignol (2014, 8), en lien avec la norme européenne de traduction, « le traducteur doit prêter une attention particulière à sept éléments : la terminologie [...], la grammaire, le lexique [...], le style [...], les particularités locales ou régionales, la mise en forme [...], le groupe cible et l'objet de la traduction ». Chacun de ces éléments peut servir de point de départ aux évaluateurs qui, selon leurs formations et leurs objectifs, choisiront « où placer le curseur » ainsi que le formulent Lavault-Olléon & Allignol (2014, 10).

dans ce domaine, nous citons au passage l'étude de Wisniewski *et al.,* (2015) qui analysent la variabilité des post-éditions, tâche qui consiste à « corriger les sorties d'un système de traduction automatique (TA) afin de produire une traduction de qualité », ou encore celle menée par Ive *et al.*, (2018) sur le pré-traitement dans la traduction. Cette dernière étude envisage une procédure comme suit :

> (a) the detection of MT translation difficulties; (b) the resolution of those difficulties by a human translator, who provides their translations (pre-translation); and (c) the integration of the obtained information prior to the automatic translation.

Pour revenir à notre relevé d'erreurs, nous avons été amenée à juger de l'acceptabilité ou non d'un énoncé traduit sur la base d'une validation syntaxique et sémantique, reposant en partie sur le jugement des locuteurs natifs portant sur l'authenticité des phrases que nous leur avons soumises.

### 3.3. Erreurs relevant de la réception du site elliptique

Il s'agit dans ce dernier volet des erreurs qui surgissent lorsqu'une ambiguïté est non perçue par les systèmes de traduction automatique, ayant ainsi des conséquences sur la réception – compréhension et interprétation – des éléments omis (voir le propos de François Yvon à ce sujet p. 178).

D'une manière générale, la reconnaissance des erreurs dans la traduction automatique pourrait sembler une tâche beaucoup plus simple que dans celle de l'humain, dans la mesure où nombre de ces erreurs résultent souvent de la différence entre les systèmes de langues et qu'elles sont immédiatement repérables dans l'agencement syntaxique. Ce constat est vite remis en cause notamment dans la traduction des ellipses apparaissant dans les *question tags* où l'interprétation du site elliptique et le repérage de l'antécédent sont erronés, problème généralement résolu avec le *n'est-ce pas ?* « vide ». Nous avons pu constater que cette locution ne convient pas toujours et que le phatique exprimé par les *questions tags* en anglais n'est pas toujours traduit. Certaines configurations, comme lorsque l'antécédent se trouve dans la même phrase que le site elliptique, sont gérées par les systèmes de traductions en particulier DeepL qui a réussi dans plusieurs traductions à repérer

l'antécédent et à l'interpréter dans la traduction. En effet, malgré des ellipses qui résistent encore, le passage de la traduction statistique à des systèmes de traduction pouvant gérer diverses informations simultanément sur la base d'un fonctionnement neuronal a permis une amélioration remarquable dans les traductions automatiques.

De ce fait, puisque les problèmes des systèmes de traduction semblent se situer au niveau de l'antécédent, de la structure grammaticale dans le site elliptique, et de son interprétation, les erreurs que l'on peut rencontrer dans la traduction automatique des ellipses sont sans doute similaires à celles rencontrées dans la détection. En d'autres termes, quelle que soit la méthodologie adoptée pour établir un protocole et envisager une réduction d'erreurs dans la traduction, l'étape préliminaire est d'abord celle de la détection puis celle de la production ou de l'enrichissement des données d'entrainement. Comme énoncé, des étapes comme la pré-traduction pour prédire les erreurs ou la post-édition pour les corriger pourraient être envisagées. Dans l'analyse des exemples, afin de ne pas déborder du cadre de notre recherche, nous nous sommes limitée à leur traduction dans un seul genre de discours, précisément celui où l'ellipse est la plus fréquente. Il serait donc intéressant par la suite, d'étendre cette étude à d'autres genres et d'exploiter les pistes de réflexions et les remarques listées ci-dessus qui, de façon évidente, nécessiteraient une étude approfondie dont elles constitueraient l'objectif principal.

Enfin, il faut remarquer qu'en adéquation à nos objectifs de départ nous n'avons observé que la traduction des ellipses post-auxiliaires et, plus succinctement, quelques occurrences de questions fragmentaires de l'anglais vers le français. Il serait donc également intéressant d'inclure ultérieurement d'autres types d'ellipses pouvant être, de même, source d'erreurs, telle l'ellipse nominale ou propositionnelle. De plus, élargir ces études à d'autres langues, notamment celles moins dotées comme l'arabe, par exemple, apparait une perspective à envisager.

En définitive, ne butera-t-on pas toujours sur la non capacité des systèmes à *interpréter réellement un texte ?* Autrement dit, ne sommes-nous pas là, justement, dans une recherche tendant vers la modélisation d'une démarche heuristique (au sens de découverte et de compréhension) tandis qu'une modélisation d'une

démarche herméneutique (donc relative à l'interprétation) demeurera impossible ? C'est en fait toute la question traductologique de la compréhension d'un texte d'un côté et celle de son interprétation de l'autre qui est posée[165].

Pour conclure momentanément, il convient de dire qu'à ce jour, en raison des limites des outils et des algorithmes disponibles, la détection automatique s'est limitée à la morphosyntaxe. L'analyse des traductions révèle pourtant que la sémantique au sens large du terme, devrait en plus de la syntaxe prendre sa place tant au niveau du mot (transfert), qu'à celui du contexte immédiat (distribution) et du discours puisque l'ellipse est un phénomène linguistique, discursif et langagier. Ces deux derniers aspects sont souvent relégués à l'arrière-plan des procédures de reconnaissance et la traduction du phénomène.

Au terme de cet examen, s'il est clair que le traducteur humain continue d'avoir l'avantage sur les systèmes de la traduction automatique, il est certain que son rôle et son travail sont désormais profondément modifiés par le développement de l'IA : « la technologie ne va pas remplacer le traducteur, mais celui-ci sera remplacé par un traducteur qui utilise la technologie »[166], phénomène déjà à l'œuvre au sein des institutions européennes, par exemple.

[165] questions soulevées lors des échanges avec notre directrice de thèse, Maryvonne Boisseau.
[166] Entretien avec Pierrette Bouillon https://www.unige.ch/fti/ebulletin/entretien/entretien-1 (consulté le 24 avril 2019 à 10:05)

# Conclusion générale

Au terme de ce travail, la question se pose de savoir si nous avons atteint les objectifs que nous nous étions fixés, à savoir : la détection automatique de l'ellipse et la compréhension des erreurs engendrées par sa traduction automatique, les deux objectifs étant articulés l'un à l'autre.

Tout au long de notre cheminement, nous avons présenté cette recherche comme étant un travail expérimental, susceptible d'en initier d'autres qui prendraient en compte les avancées empiriques de notre démarche afin de réaliser une approche entièrement informatisée du phénomène elliptique, que ce soit lors de sa détection et/ou de sa traduction automatiques. Nos résultats et perspectives se déploient en trois volets – théorique, méthodologique et traductologique – selon l'enchaînement des chapitres de la thèse. Nous résumons ci-dessous les réserves que nous avons déjà pu formuler ainsi que les apports éventuels propres à chacun de ces volets.

Au plan théorique, l'aboutissement (provisoire) de notre recherche a révélé une fois encore la complexité du phénomène elliptique. Cette complexité se révèle à la fois dans les problèmes de classification des différentes catégories d'ellipses que nous avons eu l'occasion de soulever, dans les difficultés formaliser (sous forme de patrons) les règles nécessaires à sa reconnaissance automatique, enfin dans le repérage et la compréhension des erreurs générées dans sa traduction automatique. Par ailleurs, constatant que dans une situation sociale de communication (ou « situation d'énonciation », pour reprendre un terme culiolien), le co-locuteur comprend aisément le contenu du site elliptique et parvient parfois à produire des énoncés dont les réponses correspondent à des questions non formulées dans l'échange même, nous avons émis l'idée d'une autre catégorie possible que nous avons nommée *ellipse de discours*. Elle renverrait non à l'omission d'un segment de la phrase, mais à un énoncé implicite *sous-jacent* à un autre, pouvant relever de différentes modalités (interrogation, constat, commentaire), l'interlocuteur réagissant alors en rupture avec ce qui est explicite. Cette catégorie pourrait compléter la taxonomie de l'ellipse (peut-être même la remettre en question) et permettre éventuellement de lever l'ambiguïté existant entre elle et certains phénomènes linguistiques comme le sous-entendu. En effet, à l'inverse du sous-

entendu, l'ellipse de discours se perçoit lors d'un échange entre locuteurs en prenant en compte l'énoncé lui-même dans sa situation d'énonciation.

En préalable à nos travaux, nous avons donc tout d'abord proposé une définition de l'ellipse et dégagé les critères essentiels à son repérage dans un contexte donné, critères qui la différencient d'autres phénomènes. L'ellipse, en effet, a ceci de remarquable, qu'elle se caractérise par l'incomplétude syntaxique de la phrase sans pour autant perturber, dans la plupart des cas, son contenu sémantique. À cette incomplétude syntaxique, et afin que l'ellipse soit fonctionnelle, s'ajoute la récupérabilité de l'élément omis grâce à un antécédent.

Plusieurs études contemporaines ont abordé le phénomène à travers leurs différents champs d'approche, aboutissant à sa classification, l'une des stratégies possibles pour le définir. Ces classifications de l'ellipse ont ainsi permis d'envisager son traitement informatisé allant de sa reconnaissance jusqu'à sa résolution. Les chercheurs ayant travaillé sur ces questions ont mis en œuvre des procédures de TAL et ont suivi une démarche automatique ou semi-automatique, à base de patrons, de règles ou réalisée par apprentissage, appliqués majoritairement à la VPE et au *sluicing* telles que ces ellipses ont été identifiées dans les classifications syntaxiques. Ce traitement se déroule en trois étapes : (i) détection du site elliptique, (ii) délimitation et détection de l'antécédent et (iii) résolution de l'ellipse. À l'issue de l'examen des recherches qui nous ont paru les plus significatives par rapport à notre problématique, s'est donc dessiné notre choix d'envisager la détection du site elliptique. Afin de mener à bien nos travaux, nous avons établi une classification morphosyntaxique de l'ellipse, en fonction de l'élément déclencheur. Cette classification ne remet bien entendu nullement en cause les catégories existantes de l'ellipse dans les approches contemporaines, mais elle nous était simplement utile dans la phase de détection automatique à base de patrons. Toutefois, si une investigation théorique devait être conduite en vue d'une compréhension encore plus approfondie du phénomène, une classification syntaxique plus fine serait sans doute nécessaire.

Au plan méthodologique, notre étude a porté sur le couple de langues anglais/français, de l'anglais vers le français, s'appuyant sur une démarche contrastive, ou plus exactement semi-contrastive, et reposant sur l'exploitation d'un corpus parallèle. Nous avons en effet envisagé la détection des ellipses dans un corpus en anglais et avons utilisé les versions françaises pour analyser leurs traductions. Le recours à l'approche contrastive était ainsi appelé par la nécessité de comparer les résultats concrets obtenus par le traducteur humain à ceux réalisés par les systèmes de traduction automatique dans le cas des ellipses dont la complexité apparaît alors flagrante.

Les deux corpus constitués étaient de taille différente : un corpus de développement servant à l'élaboration de patrons de détection et un corpus servant à leur évaluation. Pour ce dernier, plusieurs échantillons ont été utilisés. En effet, partant de l'hypothèse que l'ellipse est plus fréquente dans un corpus conversationnel, nous avons sélectionné deux échantillons du corpus de sous-titres pour évaluer la performance des patrons. Pour parvenir ensuite à cerner d'éventuels paramètres de variations elliptiques, susceptibles d'être liés aux genres discursifs, nous avons également intégré d'autres échantillons issus de différents genres : politique, journalistique, littéraire et promotionnel. Les outils utilisés pour exploiter nos corpus et définir notre méthodologie de détection par patrons sont majoritairement inclus dans Stanford CoreNLP, à savoir : l'étiqueteur pour l'analyse morphosyntaxique et TokensRegex pour le développement des patrons.

Après avoir procédé à l'élaboration de patrons, nous avons exposé les apports, les limites et les contraintes techniques et fondamentales de la méthodologie que nous avons adoptée et affinée au fur et à mesure de sa mise en pratique. Ces dernières tiennent à la fois au corpus d'étude, aux outils utilisés et au phénomène lui-même qui, par sa forme absente, pose un défi au TAL. Parmi ces limites et contraintes, citons l'efficacité de la classification fortement dépendante des prétraitements (dans le corpus de développement), la nécessité de travailler sur des grands corpus annotés (annotation manuelle chronophage), et le déséquilibre dans la distribution des ellipses dans les corpus aboutissant de ce fait à l'impossibilité de généraliser les

observations à partir des résultats obtenus Après avoir procédé à l'élaboration de patrons, nous avons exposé les apports, les limites et les contraintes de la méthode que nous avons suivie et affinée au fur et à mesure de sa mise en pratique. Ces dernières tiennent à la fois au corpus d'étude, aux outils utilisés et au phénomène lui-même qui, par sa forme absente, pose un défi au TAL. Parmi ces limites et contraintes, citons la pertinence relative de la classification, fortement dépendante des prétraitements (dans le corpus de développement), la nécessité de travailler sur des grands corpus annotés (annotation manuelle chronophage), et le déséquilibre dans la distribution des ellipses dans les différents sous-corpus aboutissant de ce fait à l'impossibilité de généraliser les observations à partir des résultats obtenus.

Notre attention s'est alors concentrée sur la vérification de nos hypothèses concernant les défis à relever dans la phase de détection. À ce jour, il semblerait que seules les analyses morphosyntaxiques permettent d'envisager un traitement automatique des ellipses. Il apparait que certains paramètres de variations, sémantiques, pragmatiques et énonciatifs notamment, ne peuvent être pris en compte par les analyseurs automatiques. La détection automatique de l'ellipse ne réussit donc pas « à tous les coups » et, par conséquent, elle est limitée à la seule représentation syntaxique. L'analyse des résultats obtenus par les requêtes lancées à l'aide de patrons sur nos corpus s'est principalement effectuée à partir de la classification morphosyntaxique établie.

En présence des éléments déclencheurs, le taux du rappel a été particulièrement élevé pour la plupart des patrons, révélant les avantages que la détection à base de tokens peut apporter à la reconnaissance de l'ellipse. Nous avons néanmoins effectué une lecture des occurrences détectées par les patrons afin d'identifier certaines failles de la méthode et voir comment il serait possible d'y remédier. En effet, selon leur origine, nous avons pu identifier deux types d'erreurs : le premier relève de la précision insuffisante des patrons tandis que le deuxième tient à l'étiquetage.

Malgré le recours à la lemmatisation et aux entités nommées dans les expressions régulières pour pallier les lacunes rencontrées, le résultat – peu satisfaisant au demeurant – obtenu par certains patrons a montré qu'ils devraient être améliorés.

L'une des premières pistes d'amélioration que nous pouvons envisager consisterait à tester d'autres étiqueteurs ou bien à enrichir les étiquettes des outils de façon, par exemple, à établir une distinction entre auxiliaire et verbe plein (le cas de *do* ou encore *have* et *be*), entre *TO* préposition et *TO* marqueur d'infinitif. Certaines erreurs d'étiquetage peuvent également être prédites comme le —*s'* généralement étiqueté comme POS quelle que soit sa fonction. Cette étape d'enrichissement nous apparaît comme incontournable pour améliorer la détection automatique des ellipses à base de tokens à partir d'une analyse morphosyntaxique.

De plus, pour envisager une détection complète du phénomène, il serait également utile d'établir une classification à l'intérieur même des déclencheurs de l'ellipse, notamment pour les modaux. En effet, identifier les différentes valeurs sémantiques des modaux susceptibles de déclencher des ellipses serait un réel progrès dans la reconnaissance automatique de l'ellipse. Les études en français ont révélé par exemple que certaines valeurs sémantiques autorisent plus d'ellipses que d'autres. À l'aide d'une nouvelle classification de ces modaux et de leur annotation, un système par apprentissage automatique pourrait être entrainé pour être plus précis dans sa tâche de détection des ellipses, du moins à l'égard des catégories qui ne peuvent encore être prises en compte dans notre méthode de détection.

Par ailleurs, il n'est pas impossible d'imaginer, à partir de cette détection, le développement d'un système d'apprentissage automatique fondé sur une quantité de données importante, apte à prendre en charge un nombre conséquent de types d'ellipses. Ce système pourrait alors être exploité pour évaluer, voire améliorer, les applications informatiques ou les assistants virtuels proposés par les grandes entreprises : de transports (Ouibot de SNCF), sites de réservation (Booking), Siri (Apple), Alexa (Amazon), Ok Google, etc. – domaines nécessitant une gestion des dialogues – pour obtenir des échanges les plus naturels possibles (politesse, par exemple) et fournir une représentation *complète* de ce que souhaite l'utilisateur. Les recherches dans ce domaine sont actives depuis les années 1970 (ELIZA de Weizenbaum 1966), et continuent d'intéresser les chercheurs aujourd'hui (voir par exemple Dubuisson Duplessis *et al.,* [2015] et Campillos Llanos *et al.,* [2015]).

Le développement de ce type de système peut se révéler complexe notamment lorsque l'on change de domaine, lors du passage d'un genre à un autre, par exemple, ou encore d'une langue à une autre. À cet égard, il faudrait donc envisager l'enrichissement des corpus de développement et l'inclusion d'un nombre important de genres discursifs.

En tout cas, dans le corpus *genré* que nous avons analysé, l'ellipse apparait plus fréquemment dans le corpus de sous-titres que dans les autres sous-corpus. Les taux de précision et de rappel ont montré que les patrons étaient adaptés et parvenaient à repérer les types d'ellipses annotés manuellement dans des échantillons restreints. Le taux d'erreurs, tant au niveau de la précision que du rappel, a été en revanche particulièrement élevé dans le genre politique où la rareté de l'ellipse et la longueur des phrases ont induit les patrons en erreurs en détectant un nombre important de faux positifs. Les résultats obtenus à partir de notre méthode de détection ont montré qu'il est possible d'analyser, à l'aide du test statistique du $\chi^2$, la distribution des ellipses dans les genres analysés. Les résultats obtenus ont permis de confirmer les observations faites quant à la fréquence ou la rareté du phénomène au sein de chaque genre mais également de relever qu'il était difficile de conclure si l'une ou l'autre catégorie d'ellipse était une propriété spécifique d'un genre donné. À l'issue de l'évaluation des patrons dans ce corpus multi-genres, il nous a semblé pertinent de proposer, comme apport positif résultant de la détection automatique, la possibilité d'une classification des genres de discours en fonction de la fréquence ou de la rareté des ellipses.

Pour conclure ce volet, il convient d'ajouter que si les entités visibles du discours sont parfois difficiles à identifier et catégoriser en TAL, l'ellipse pose davantage de problèmes, encore insolubles à ce jour. De par son hétérogénéité et sa multi-dimensionnalité, elle ajoute aux difficultés du TAL les difficultés inhérentes à son traitement dans les corpus sélectionnés, difficultés qui proviennent de :

– son ambiguïté : chaque lemme répertorié dans notre classification n'est pas forcément déclencheur ;

- la variabilité du co-texte : la longueur des phrases pose problème à l'étiqueteur, ce qui se répercute ensuite sur la détection à l'aide de patrons, provoquant une augmentation dans la détection des occurrences non elliptiques ;
- la rareté ou l'apparition fréquente de l'ellipse : il est impossible d'inclure toutes les formes d'ellipses rencontrées dans un nombre limité de patrons, soit en raison de leur rareté (dans le corpus de développement), c'est le cas des ellipses {post-ord} et {post-quant} par exemple, soit en raison de leur fréquence (nombre important de formes {qs-frag}).

En résumé, la méthode que nous avons élaborée emprunte de toute évidence certains aspects de sa procédure à la linguistique théorique, à la linguistique de corpus, ainsi qu'à la linguistique outillée, de telle sorte que notre travail, dans ses grandes lignes, pourrait se résumer à deux étapes essentielles : (i) la recherche d'une détection efficace de l'ellipse (sans viser sa résolution) ; (ii) par nécessité pragmatique, une focalisation sur les erreurs de traduction observées sur un seul type d'ellipse. En effet, nous avons fait l'hypothèse qu'une éventuelle retombée de la détection automatique de l'ellipse réside dans la possibilité de prédire des erreurs susceptibles de se produire lors de la traduction automatique, raison pour laquelle nous avons consacré le dernier chapitre de notre travail à la traduction d'un phénomène en particulier, l'ellipse post-auxiliaire.

Du point de vue de la traduction**,** en effet, l'une des opérations menée dans la phase de détection relatée dans le dernier chapitre de notre étude et vers laquelle nous souhaitons orienter nos recherches à venir, consiste à comparer et à évaluer la traduction humaine et automatique de l'ellipse. Pour ce faire, des notions-clefs, telles que l'acceptabilité et l'ambiguïté, ont été prises en compte pour évaluer la qualité de la traduction produite, et ceci, en comparaison avec une traduction humaine considérée comme traduction de référence. La sélection des exemples a été faite à partir de l'élément déclencheur, modal ou auxiliaire. Les erreurs ont ensuite été analysées.

En vérité, au vu des exemples illustrant nos propos, s'il semble évident que les productions humaines, dans leurs traductions des faits elliptiques, restent la référence par rapport à celles proposées par les systèmes automatiques, mais de nombreux progrès ont été réalisés, particulièrement depuis mars 2016 où le passage à la traduction neuronale a modifié le paysage de la traduction automatique.

Nous reconnaissons au terme de ce travail que nombre de problèmes ne sont pas encore résolus. Malgré ces difficultés, en particulier celles liées à la traduction des modaux (déclencheurs d'ellipse comme on a pu le montrer) ou encore celles rencontrées par nos patrons lors des requêtes de détection, nous avons pu cerner un certain nombre d'obstacles entravant détection et traduction du phénomène, en comparaison des différentes stratégies utilisées par l'humain, ce qui constitue peut-être, bien que mené sur un échantillon restreint, l'un des apports essentiels de cette étude.

Par conséquent, les erreurs observées lors de la détection peuvent contribuer à la compréhension des sources d'erreurs (de l'ellipse) dans la traduction automatique du phénomène. Des procédures de prédiction ou de correction peuvent ensuite être établies pour produire une traduction acceptable de l'ellipse (qu'elles soient des procédures de pré-traduction ou de post-édition). Si l'évaluation de la traduction automatique reste problématique, c'est bien parce que la notion d'acceptabilité, nécessaire à son examen, est une notion très relative, tout comme l'est celle, au sein d'un contexte donné, de la grammaticalité d'un énoncé. Évaluer un énoncé comme valide dépend, en effet, techniquement, d'une évaluation syntaxique préalable à une validation sémantique, comme évoqué précédemment.

L'ellipse en tant que phénomène complexe liant syntaxe et sémantique pour être fonctionnelle, semble pourtant offrir un moyen pour l'évaluation des traductions en général, dans la mesure où traduire une ellipse ne revient pas seulement et systématiquement à faire apparaître un élément ellipsé de la langue source, mais nécessite une interprétation rigoureuse du site elliptique toutefois encore impossible à réaliser par les machines. À notre connaissance, seuls les traducteurs humains sont capables d'interprétations fines.

Parvenue à présent au terme d'une étude dont les contraintes de temps et de champ probablement trop vaste à embrasser, ont rendu l'exercice nécessairement limité, force est pour nous d'admettre avec lucidité et modestie qu'appréhender dans son entièreté l'ellipse en tant que fait de langue susceptible d'être abordé par les outils de détection et de traduction automatique, n'a pu être totalement réalisé. Des questions demeurent, tant dans l'appréhension de sa nature, et partant, de sa définition et de sa classification, qu'*a fortiori* dans la compréhension de ces domaines d'exploration nouveaux (traduction neuronale par exemple) nécessaires à la détection des phénomènes linguistiques et à leurs traductions automatiques. Cependant, les progrès de la science et de ses applications apportant chaque jour de nouvelles perspectives aux recherches en cours, leurs retombées patentes ont induit un réajustement continu des objectifs de notre recherche. Ainsi, notre préoccupation d'approcher au plus près un phénomène linguistique appartenant à une langue, en vue d'effectuer sa détection dans un corpus donné, pour aboutir à sa traduction dans une autre langue, a été en quelques mois confrontée directement au passage de la traduction statistique à la traduction neuronale aux performances démultipliées. De ce fait, nos travaux déjà en cours à ce moment-là, se sont naturellement inscrits dans une perspective évolutive, portés ainsi par une dynamique posant parfois problème lorsqu'il s'est agi de savoir où poser le regard pour observer tel ou tel résultat obtenu lors des différentes analyses du phénomène elliptique.

Avant de clore temporairement nos travaux, nous voudrions insister à ce stade sur le caractère pluridisciplinaire transversal de notre travail de recherche. La complexité même du phénomène, relevée par les différents courants de la linguistique au cours de son histoire, nécessite évidemment une étude comparative. Il s'agissait de permettre une approche ciblée dans une perspective appliquée à notre objectif initial, à savoir rendre possible la détection du phénomène elliptique en vue d'une avancée dans l'étape de sa traduction par les outils informatiques.

Par ailleurs, notre objectif, dérivé de celui concernant spécifiquement la traduction, visant la prédiction des erreurs engendrées par l'ellipse dans la traduction

automatique, ne pouvait être envisagé sans un travail d'investigation au plus près des exemples retenus, c'est-à-dire, en envisageant sa compréhension globale, allant des investigations théoriques à la difficulté de son repérage par l'humain ou par la machine, avant-même de réfléchir à comment la traduire. En effet, toutes les disciplines auxquelles nous avons eu recours dans ce travail se sont imposées du fait des objectifs que nous nous étions assignés et de la nécessité de comprendre le phénomène lui-même.

Si l'apport de disciplines liées au développement de l'informatique et de la linguistique de corpus, toutes nécessaires à la réalisation de ce travail, ne manque pas d'ouvrir de larges perspectives à la recherche, nous nous interrogeons sur la portée de notre thèse même dans la mesure où ce travail est essentiellement individuel, en dépit de son inscription dans le cadre général d'une équipe de recherche. Par ailleurs, si le caractère interdisciplinaire de cette étude lui fait courir le risque d'une dispersion, il apparait que le recours au corpus nous expose forcément à la croisée de plusieurs disciplines. Tout en restant modeste, nous pouvons invoquer comme « modèles », des « grands noms » tels Edgar Morin, ou Michel Serres, qui ont défendu les approches interdisciplinaires mettant en lumière la complexité des phénomènes, ou, plus proche de nous et directement lié à notre sujet, le chercheur Daniel Hardt, pionnier dans la recherche de la détection automatique de l'ellipse combinant le travail sur corpus et l'interdisciplinarité entre la linguistique théorique et le TAL.

En dépit de ces réserves, il nous revient de relever les apports de cette étude à la poursuite de futurs travaux, à savoir plus précisément, une classification simplifiée mais opérante des ellipses permettant leur détection, une méthodologie mise à l'épreuve de l'élaboration de patrons nécessaires à la détection du phénomène, un ensemble de corpus annotés suivant notre classification, le relevé d'une série d'erreurs et leur impact sur la lisibilité de la traduction, au regard des stratégies utilisées par le traducteur humain.

Parmi les perspectives que nous avons dessinées au fil de notre travail et, pour certaines évoquées à nouveau dans cette conclusion, nous en retiendrons trois, en

raison de leur importance et de la possibilité de leur réalisation à plus ou moins long terme :

– l'annotation d'une large quantité de données en vue d'améliorer la méthode de détection, ces données annotées pouvant également être utilisées pour entrainer un système par apprentissage, apte à classer les ellipses dans un large corpus ;

– l'extension de l'étude à d'autres langues moins explorées, comme l'arabe (voire même le berbère après la constitution d'un corpus électronique) ;

– l'établissement d'un protocole, élaboré en fonction des erreurs répertoriées en tenant compte des différences entre les systèmes linguistiques, et susceptible d'aboutir à une méthode automatique en post-édition, apte à corriger les structures erronées et les registres inadaptés dans les traductions du phénomène elliptique.

Le développement de la recherche dans le domaine de la traduction automatique et la diffusion de ses résultats devraient permettre l'avancée de ces différentes perspectives ; et malgré l'obsolescence rapide des résultats obtenus et l'aspect chronophage de l'annotation des corpus, toute recherche empirique, telle la nôtre, devrait pouvoir s'inscrire dans la poursuite de la dynamique en cours menant à des améliorations sensibles et concrètes.

Ainsi, tout en continuant à porter notre attention sur l'avancée permanente des outils de traduction automatique menant à des traductions de plus en plus fines, nous ne pouvons qu'affirmer, comme nous l'évoquions, que s'approcher du vide créé par le phénomène fascinant de l'ellipse nous donne encore le vertige et les tentatives pour en comprendre, sinon la vacuité – au sens premier du terme, « l'état de ce qui est vide » – du moins la forme du manque, sont gages, pour le chercheur, pour le linguiste, comme pour le traducteur humain, de lendemains professionnels encore riches de découvertes.

# Bibliographie


Abeillé, A. & Mouret, F. (2010). Quelques contraintes sur les coordinations elliptiques en français. *Revue de sémantique et de pragmatique*, 24, pp. 77–206.

Adam, J-M. (1997). Genres, textes, discours : pour une reconception linguistique du concept de genre. *Revue belge de philologie et d'histoire*. 75(3), pp. 665–681. En ligne : https://www.persee.fr/doc/rbph_0035-0818_1997_num_75_3_4188

Aelbrecht, L. (2008). Dutch Modal Complement Ellipsis. *Empirical Issues in Syntax and Semantics*, 7, pp. 7–33.

Aelbrecht, L. (2010). *The Syntactic Licensing of Ellipsis*. Amsterdam, Pays-Bas : John Benjamins Publishing.

Anand, P. & McCloskey, J. (2015). Annotating the Implicit Content of Sluices. *Proceedings of The 9th Linguistic Annotation Workshop,* Denver, Colorado, USA, pp. 178–187.

Antony, P. J., Warrier, N. J. & Soman, K. P. (2010). Penn Treebank-Based Syntactic Parsers for South Dravidian Languages Using a Machine Learning Approach. *International Journal of Computer Applications*, 7(8), pp. 14–21.

Arrivé, M., Gadet, F. & Galmiche, M. (1986). *La grammaire d'aujourd'hui*. Paris, France : Flammarion.

Bally, C. (1952). Le langage et la vie. Genève, Suisse : Droz.

Baird, A., Hamza A. & Hardt, D. (2018). Classifying Sluice Occurrences in Dialogue. *The LREC 2018 Proceedings, 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan, pp. 1580-1583.

Bakhtine, M. (1984). *Esthétique de la création verbale*. Paris, France : Gallimard.

Bernhard, D., Todirascu, A., Martin, F., Erhart, P., Steible, L., Huck, D. & Rey, C. (2017). Problèmes de tokénisation pour deux langues régionales de France, l'alsacien et le picard. *DiLiTAL 2017.* Orléans, France, pp. 14–23.

Beacco, J-C. (2004). Trois perspectives linguistiques sur la notion de genre discursif. *Langages*, 153 (1), pp. 109–119.

Beecher, H. (2008). Pragmatic Inference in the Interpretation of Sluiced Prepositional Phrases. UC San Diego : Department of Linguistics. En ligne https://escholarship.org/uc/item/2261c0tg

Benveniste, É. (1966). *Problèmes de linguistique générale*. Paris, France : Gallimard.

Blanche-Benveniste, C. (1997). *Approches de la langue parlée en français*. Paris : Ophrys.

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge, New York : Cambridge University Press.

Biber, D, Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, UK : Pearson Education.

Bích, N. (2011). Quelques réflexions sur le français parlé-oral spontané. ULIS.

Bîlbîie, G. (2013). *Grammaire des constructions elliptiques : une étude comparative des phrases sans verbe en roumain et en français* (Thèse de Doctorat, Lille, France : Atelier national de reproduction des thèses). En ligne http://www.theses.fr/2011PA070118

Boisseau, M. (2011). Présentation. *Palimpsestes*. En ligne http://journals.openedition.org/palimpsestes/776

Boisseau, M. (2016). Lire et relire Jacqueline Guillemin-Flescher. In Boisseau M. Chauvin C. & Delesse C. (éds.). *Linguistique et traductologie : les enjeux d'une relation complexe*. Arras, France : Artois Presses Université.

Bos, J. & Spenader, J. (2011). An Annotated Corpus for the Analysis of VP Ellipsis. *Language Resources and Evaluation*, 45(4), pp. 463–494.

Bouillon, P., Rayner, E., Chatzichrisafis, N., Hockey, B., Santaholma, M., Starlander, M., Nakao, Y., Kanzaki, K. & Isahara, H. (2005). A Generic Multi-lingual Open Source Platform for Limited-Domain Medical Speech Translation. *Proceedings of the 10th Conference of the European Association for Machine Translation (EAMT*), Budapest, Hongrie, pp. 50-58.

Bouillon, P., Rayner, E., Starlander, M. & Santaholma, M. (2007). Les ellipses dans un système de traduction automatique de la parole. *Actes de TALN/RECITAL,* Toulouse, France, pp. 53–62.

Busquets, J. & Denis, P. (2001). L'ellipse modale en français : le cas de *devoir* et *pouvoir*. *Cahiers de grammaire*, 26, pp. 55–74.

Cappeau, P., & Gadet, F. (2007). L'exploitation sociolinguistique des grands corpus. *Revue française de linguistique appliquée*, XII(1), pp. 99–110.

Campillos Llanos, L., Bouamor, D., Bilinski, É., Ligozat, A-L., Zweigenbaum, P. & Rosset, S. (2015). Un patient virtuel dialogant. *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles, Association pour le Traitement Automatique des Langues, 201*5.

Celle, A. (1994). La traduction de WILL. In Guillemin-Flescher, J. (éd.). *Linguistique contrastive et traduction* (3), pp. 87-139. Paris, France : Ophrys.

Chambers, N., Cer, D., Grenager, T., Hall, D., Kiddon, C., MacCartney, B., de Marneffe, M., Ramage, D., Yeh, E. & Manning, C. (2007). Learning Alignments and Leveraging Natural Logic. *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, Prague, République Tchèque, pp. 165–170.

Chanet, A. (1983). L'ellipse dans la tradition rhétorique grecque. H.E.L. (*Histoire Épistémologie Langage*) *5*(1), pp. 17–22.

Chomsky, N. (1980). *Rules and representations*. New York, USA : Colombia University Press.

Chomsky, N. (1993). *Lectures on Government and Binding: The Pisa Lectures*. Berlin, Allemagne : Mouton-Walter de Gruyter.

Chomsky N. (1995). *The Minimalist Program*. Massachussetts, USA : MIT Press.

Chung, S., Ladusaw, W. A. & McCloskey, J. (1995). Sluicing and Logical Form. *Natural Language Semantics*, 3(3), pp. 239–282.

Chuquet, H. & Paillard, M. (1987). *Approche linguistique des problèmes de traduction anglais-français.* Paris, France : Orphys.

Cloutier, R. A., Hamilton-Brehm, A. M. & Kretzschmar, W. A. (2010). *Studies in the History of the English Language V: Variation and Change in English Grammar and Lexicon : Contemporary Approaches*. Berlin, Allemagne : Walter de Gruyter.

Cori, M. (2008). Des méthodes de traitement automatique aux linguistiques fondées sur les corpus. *Langages*, 171(3), pp. 95-110.

Cornish, F. (2010). Anaphora : Text-based or Discourse-dependent ? Functionalist *vs.* Formalist accounts. *Functions of Language*, 17(2), pp.207–241.

Culicover, P. W. & Jackendoff, R. S. (2005). *Simpler Syntax*. Oxford, UK : Oxford University Press.

Culioli, A. (1974). À propos des énoncés exclamatifs. *Linguistique et enseignement du français.* 22(1), pp. 6–15.

Dagnac, A. (2008). L'Ellipse modale en français : arguments pour une ellipse du TP. *CMLF'08*, pp. 2453–2465.

Dalrymple, M., Shieber, S. M. & Pereira, F. C. N. (1991). Ellipsis and Higher-Order Unification. *Linguistics and Philosophy*, 14(4), pp. 399–452.

Depraetere, I. & Langford, C. (2002). *Advanced English Grammar : A Linguistic Approach*. Londres et New York : Bloomsbury Publishing.

Droganova, K. & Zeman, D. (2017). Elliptic Constructions : Spotting Patterns in UD Treebanks. *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies,* Gothenburg, Suède, pp. 48–57.

Droganova, K., Ginter, F., Kanerva, J. & Zeman, D. (2018a). Mind the Gap : Data Enrichment in Dependency Parsing of Elliptical Constructions. *Proceedings of the 2nd Workshop on Universal Dependencies,* Bruxelles, Belgique, pp. 47–54.

Droganova, K., Ginter, F., Kanerva, J. & Zeman, D. (2018b). Parse Me if You Can : Artificial Treebanks for Parsing Experiments on Elliptical Constructions. *LREC Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japon, pp. 1845–1852.

Dubuisson Duplessis, G., Béchade, L., Sehili, M., Delaborde, A., Letard, V. Ligozat, A-L., Deléglise, P. Estève, Y., Rosset, S. et Devillers, L. (2015). Nao is doing humour in the CHIST-ERA JOKER project. *16th Interspeech. Dresde*, Germany. pp. 1072–1073.

Eco, U. (2007). *Dire presque la même chose*. Paris, France : Grasset.

Fernández, R., Ginzburg, J. & Lappin, S. (2007). Classifying Non-sentential Utterances in Dialogue : A Machine Learning Approach. *Computational Linguistics*, 33(3), pp. 397–427.

Fiengo, R. & May, R. (1994). Indices and Identity. *Linguistic Inquiry,* (24), pp. xvii–315.

Fontanier, P. (1968). *Les figures du discours*. Paris, France : Flammarion.

Freyermuth, S. (2011). Un genre peut en cacher un autre : une histoire de détournement. *Linx*, (64-65), pp. 173–187.

Fuchs, C. (2008). L'incertitude interprétative dans l'activité de langage. *Actes de Savoirs,* (5), pp. 41–57.

Fuchs, C. (2009). L'ambiguïté : du fait de langue aux stratégies interlocutives. *L'ambiguïté*, 50, pp. 3–16.

Gandón-Chapela, E. (2016). Hunting for Post-Auxiliary Ellipsis in a Parsed Corpus of English. *Research in Corpus Linguistics*, (4), pp. 33–38.

Ginzburg, J. & Sag I. (2000). *Interrogative Investigations*. Stanford : CSLI publications.
Ginzburg, J. & Miller, P. (2019). Ellipsis in Head-Driven Phrase Structure Grammar. In van Craenenbroeck, J. & Temmerman, T. (ed.) *The Oxford Handbook of Ellipsis*, pp. 75–121. Royaume-Uni : Oxford University Press.
Godel, R**.** (1957). *Les Sources manuscrites du* Cours de linguistique générale *de F. de Saussure.* Genève, Suisse : Droz.
Goodall, G. (1987). *Parallel Structures in Syntax*. Cambridge : Cambridge University Press.
Grass, T. (2010). À quoi sert encore la traduction automatique ? *Les Cahiers du GEPE, Autres exploitations des outils électroniques*. Presses Universitaires de Strasbourg. En ligne : http://www.cahiersdugepe.fr/index.php?id=1367
Guillemin-Flescher, J. (1981). *Syntaxe comparée du français et de l'anglais : problèmes de traduction*. Paris, France : Ophrys.
Guillemin-Flescher, J. (1991). Représentation linguistique de l'activité, l'action et l'événement en français et en anglais. *Palimpsestes : Revue de traduction*, (5), pp. 51–69.
Guillemin-Flescher, J. (2003). Théoriser la traduction. *Revue française de linguistique appliquée*, VIII(2), pp. 7–18.
Halliday, M. & Hasan, R. (1976). *Cohesion in English*. Royaume-Uni : Longman.
Hamza, A. (2014). *A Contrastive Contribution to the Study of Ellipsis in English and French* (Master thesis, Unpublished, Université de Strasbourg).
Hamza, A. & Bernhard, D. (2019). Détection des ellipses dans des corpus de sous-titres en anglais. *Actes de la 21e édition TALN/RECITAL,* Toulouse, France.
Hankamer, J. & Sag, I. (1976). Deep and Surface Anaphora. *Linguistic Inquiry*, 7(3), pp. 391–428.
Hardt, D. (1992). An Algorithm for VP Ellipsis. *Proceedings of the 30th annual meeting on Association for Computational Linguistics,* Newark, Delaware, USA, pp. 9–14.
Hardt, D. (1993). *Verb Phrase Ellipsis: Form, Meaning, and Processing*. (PhD. USA : University of Pennsylvania). En ligne https://pdfs.semanticscholar.org/be80/bb604484910e4fc0a53bf903b653e3a7e754.pdf?_ga=2.142351922.1433708989.1561792350-1565405680.1561792350
Hardt, D. & Rambow, O. (2001). Generation of VP Ellipsis : A corpus-based Approach. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics,* Toulouse, France, pp. 290–297.
Harris, Z. (1976). *Notes du cours de syntaxe*. Paris, France : Éditions du Seuil.
Herment, S. (2011). Emphase prosodique et emphase syntaxique : le cas de « do » dans un corpus de parole naturelle. *Corela*. En ligne http://journals.openedition.org/corela/1059 ; DOI : 10.4000/corela.1059
Hobbs, J. R. & Kehler, A. (1998). A Theory of Parallelism and the Case of VP Ellipsis. *Proceedings of the 35th Conference of the Association for Computational Linguistics (ACL-97),* Madrid, Espagne, pp. 394–401.
Huddleston, R. (1986). *Introduction to the Grammar of English*. Cambridge : Cambridge University Press.

Ive, J., Max, A. & Yvon, F. (2018). Reassessing the proper place of man and machine in translation: a pre-translation scenario. *Machine Translation*, Springer Verlag, https://hal.archives-ouvertes.fr/hal-01908305

Johnson, K. (2009). Gapping Is Not (VP-) Ellipsis. *Linguistic Inquiry*, 40 (2), pp. 289-328.

Kehler, A. (2002). Another problem for syntactic (and semantic) theories of VP-ellipsis. *Snippets* - Issue 5. En ligne http://www.ledonline/snippets

Kenyon-Dean, K., Cheung, J. C. K. & Precup, D. (2016). Verb Phrase Ellipsis Resolution Using discriminative and margin-infused algorithms. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* Austin, Texas, USA, pp. 1734–1743.

Komur-Thilloy, G. & Trevisiol-Okamura, P. (2011). Les enjeux de l'ellipse dans l'écriture journalistique : quelques applications didactiques. *Synergies Pologne* (8), pp. 255–264.

Lallot, J. (1983). L'ellipse chez Apollonius Dyscole. H.E.L. (*Histoire Épistémologie Langage*), 5(1), pp. 9–16.

Lamy, B**.** (1712). *L'art de parler*. Paris, France : La veuve de Paul Marret.

Landheer, R. (1989). L'ambiguïté : un défi traductologique. *Meta : Journal des traducteurs / Meta : Translators' Journal*, 34(1), pp. 33-43.

Laporte, E. (2008). Exemples attestés et exemples construits dans la pratique du lexique-grammaire. In Jacques, F. (éd). *Observations et manipulations en linguistique : entre concurrence et complémentarité*. pp. 11-32. Louvain/Paris/Dudley : Peeters Editors.

Lappin, S. & Benmamoun, E. (1999). *Fragments : Studies in Ellipsis and Gapping*. New York, USA : Oxford University Press.

Lappin. S., (1999). *The Handbook of Contemporary Semantic Theory*. Oxford, Royaume-Uni : Blackwell.

Larreya, P. (2004). L'expression de la modalité en français et en anglais (domaine verbal). *Revue belge de philologie et d'histoire*, *82*(3), pp. 733–762.

Larreya, P. & Riviere, C. (2005). *Grammaire explicative de l'anglais*. Paris, France :
Pearson/Longman.

Lavault, E. (2007). *Traduction spécialisée : pratiques, théories, formations*. Berne, Suisse : Peter Lang.

Lavault-Olléon, É. & Allignol, C. (2014). La notion d'acceptabilité en traduction professionnelle : où placer le curseur ? *ILCEA. Revue de l'Institut des langues et cultures d'Europe, Amérique, Afrique, Asie et Australie*, (19). En ligne : http://journals.openedition.org/ilcea/2455 ; DOI : 10.4000/ilcea.2455

Leech, G. (1992). Corpora and Theories of Linguistic Performance. In Svartvik, J. (ed.). *Directions in Corpus Linguistics* : *Proceedings of Nobel Symposium 82 Stockholm (1991).* pp. 105-126. En ligne https://doi.org/10.1515/9783110867275

Lefeuvre, F. (2001). La phrase averbale en français. *L'Information grammaticale*, *88*(1), pp. 47–48.

Le Guen, M. (2003). Tableaux Croisés et Diagrammes en Mosaïque, pour visualiser les probabilités marginales et conditionnelles. *Bulletin de Méthodologie*

*Sociologique / Bulletin of Sociological Methodology, SAGE Publications*, pp. 62–79.

Le Guen, M. (2001). La boîte à moustaches de TUKEY, un outil pour initier à la Statistique. *Statistiquement Votre - SFDS*, pp. 1–3.

Levin, N. (1986). *Main-verb Ellipsis in Spoken English*. New York, USA : Garland Publishing.

Liu, Z., Pellicer, E. & Gillick, D. (2016). Exploring the Steps of Verb Phrase Ellipsis. *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, Michigan, USA, pp. 32–40.

Lobeck, A. C. (1995). *Ellipsis : Functional Heads, Licensing, and Identification*. New York, USA : Oxford University Press.

Loffler-Laurian, A. M. (1996). *La Traduction automatique*. Villeneuve-d'Ascq, France : Presses Universitaires du Septentrion.

Loock, R. (2016). *La traductologie de corpus*. Villeneuve-d'Ascq, France : Presses Universitaires du Septentrion.

Maingueneau, D. (2007). Genres de discours et modes de généricité. *Le français aujourd'hui*, (159), pp. 29-35.

Manning, C., Mihai, S., Bauer, J., Finkel, J., Bethard, S. & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit*. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* : *System Demonstrations,* Maryland, USA, pp. 55-60.

Martinet, A. (1955). *Économie des Changements Phonétiques. Traité de Phonologie Diachronique*. Berne, Suisse : Éditions A. Francke.

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based Language Studies : An Advanced Resource Book*. Abingdon-on-Thames, UK : Routledge.

McShane, M. & Babkin, P. (2016). Detection and Resolution of Verb Phrase Ellipsis. *Linguistic Issues in Language Technology*, 13, pp. 1-34.

McShane, M. J. (2005). *A Theory of Ellipsis*. New York, USA : Oxford University Press.

Mélanie-Becquet, F. & Landragin, F. (2014). Linguistique outillée pour l'étude des chaînes de référence : questions méthodologiques et solutions techniques. *Langages,* 195(3), pp. 117–137.

Merchant, J. (2001). *The Syntax of Silence : Sluicing, Islands, and the Theory of Ellipsis.* New York, USA : Oxford University Press.

Merchant, J. (2019). Ellipsis : A Survey of Analytical Approaches. In van Craenenbroeck J. & Temmerman, T. (eds.). *The Oxford Handbook of Ellipsis*, pp. 19-45. Oxford, UK : Oxford University Press.

Miller, P. (2011). The Choice Between Verbal Anaphors in Discourse. *Discourse Anaphora and Anaphor Resolution Colloquium,* Berlin/Heidelberg, Allemagne, pp. 82–95.

Miller, P. (2014). A Corpus Study of Pseudogapping and its Theoretical Consequences. *Empirical issues in syntax and semantics*, (10), pp. 73–90.

Miller, P. & Pullum, G. K. (2013). Exophoric VP Ellipsis. In Hofmeister, P. & Norcliffe, E. (eds.). *The Core and the Periphery : Data-driven perspectives on syntax inspired by Ivan A. Sag*, pp. 5-32. CSLI Publications.

Nielsen, L. A. (2004). Verb Phrase Ellipsis Detection Using Automatically Parsed Text. *Proceedings of the 20th international conference on Computational Linguistics*, Genève, Suisse, En ligne https://doi.org/10.3115/1220355.1220512

Nirenburg, S. & Raskin, V. (2004). *Ontological semantics*. Massachusetts, USA : MIT Press.

Nivre, J., Agić, Ž., Ahrenberg, L., Aranzabe, M., Asahara, M., Atutxa, A., Ballesteros, M., … (2017). Universal dependencies 2.0. *LIN- DAT/CLARIN digital library at the Institute of Formal and Applied Linguistics,* Prague, République Tchèque. En ligne https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1983

Nykiel, J. (2015). Constraints on Ellipsis Alternation : A View from the History of English. *Language Variation and Change*, 27(2), pp. 227-254.

Péry-Woodley, M. (2005). Discours, corpus, traitements automatiques. In Condamines, A. (éd.). *Sémantique et Corpus*, pp.177-210. Paris, France : Hermès.

Phillips, C. & Parker, D. (2014). The Psycholinguistics of Ellipsis. *Lingua*, (151), pp. 78–95.

Pilevar, M. T., Faili, H. & Pilevar, A. H. (2011). TEP: Tehran English-Persian Parallel Corpus. In Gelbukh, A. (ed.). *Computational Linguistics and Intelligent Text Processing*, pp. 68-79. Berlin, Allemagne : Springer.

Pitavy, J. & Bigot M. (2008). *Ellipse et effacement : du schème de phrase aux règles discursives : actes du colloque international de linguistique, 27 et 28 octobre 2005*. Saint-Étienne, France : Université de Saint-Étienne.

Poibeau, T. (2016). Traduire sans comprendre ? La place de la sémantique en traduction automatique. *Langages* (201), pp. 77-90. En ligne http://www.revues.armand-colin.com/lettres-langues/langages/langages-ndeg-201-12016/traduire-comprendre-place-semantique-traduction-automatique

Poibeau, T. (2017). *Machine Translation*. Cambridge, USA : MIT Press.

Poudat, C. & Landragin, F. (2017). Explorer un corpus textuel : Méthodes, pratiques, outils. Louvain-la-Neuve, Bélgique: De Boeck superieur.

Quirk, R., Greenbaum, S. & Leech, G. & Svartvik. J. (1985). *A Grammar of Contemporary English*. London, UK : Longman

Reiss, K. (2002). *La critique des traductions, ses possibilités et ses limites*. Arras, France : Artois Presses Université.

Reich, I. (2011). Ellipsis. In Maienborn, C., von Heusinger, K. & Portner, P. (eds.), *Semantics: An International Handbook of Natural Language Meaning*. Berlin, Allemagne : Mouton de Gruyter.

Rello. L. (2010). *Elliphant : A Machine Learning Method for Identifying Subject Ellipsis and Impersonal Constructions in Spanish*. (Master's Thesis, Université de Wolverhampton, Royaume-Uni). En ligne https://www.academia.edu/2677797/Elliphant_A_Machine_Learning_Method_for_Identifying_Subject_Ellipsis_and_Impersonal_Constructions_in_Spanish

Riegel, M., Pellat, J.-C. & Rioul, R. (2014). *Grammaire méthodique du français*. Paris, France : Presses Universitaires de France.

Robert, A-M. (2010). La post-édition : l'avenir incontournable du traducteur ?. *Traduire.* En ligne http://journals.openedition.org/traduire/460

Rønning, O., Hardt, D. & Søgaard, A. (2018a). Linguistic Representations in Multi-task Neural Networks for Ellipsis Resolution. *Proceedings of the 2018 EMNLP Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP*, Bruxelles, Belgique, pp. 66–73.

Rønning, O., & Hardt, D. & Søgaard, A. (2018b). Sluice Resolution without Hand-Crafted Features over Brittle Syntax Trees. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2)*, Louisiana, USA, pp. 236-241.

Ross, J. (1969). Guess who?. In Merchant J. & Simpson A. (eds.). *Sluicing: Cross-Linguistic Perspectives*, pp. 14-39. Oxford, USA : Oxford University Press.

Sag, I. A. (1976). *Deletion and logical form.* (PhD Thesis. Massachusetts Institute of Technology). En ligne https://dspace.mit.edu/handle/1721.1/16401

Schuster, S., Nivre, J. & Manning, C. (2018). Sentences with Gapping : Parsing and Reconstructing Elided Predicates. *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, New Orleans, USA, pp. 1156–1168.

Schwabe, K. & Winkler, S. (2003). *The Interfaces : Deriving and Interpreting Omitted Structures*. Amsterdam, Pays-Bas : John Benjamins Publishing.

Sinclair, J. (1996). Preliminary recommendations on Corpus Typology. T*echnical report, EAGLES, (Expert Advisory Group on Language Engineering Standards)*. En ligne www.ilc.cnr.it/EAGLES/pub/eagles/corpora/corpustyp.ps.gz

Svartvik, J. (1992). *Directions in Corpus Linguistics : Proceedings of Nobel Symposium 82 Stockholm*. Berlin/ New York : Mouton de Gruyter.

Toutanova, K & Manning, D.C. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000),* pp. 63–70.

van Craenenbroeck, J. & Temmerman, T. (éds.). (2019). *The Oxford Handbook of Ellipsis*. Oxford, UK : Oxford University Press.

van Craenenbroeck, J. & Merchant, J. (2013). Ellipsis Phenomena. In den Dikken M. (éd.). *The Cambridge Handbook of Generative Syntax*, pp. 701–745. Cambridge, UK : Cambridge University Press.

Vauquois, B. & Boitet, C. (1985). Automated Translation at Grenoble University. *Computational Linguistics*, 11(1), pp. 28–36.

Weizenbaum, J. (1966) : ELIZA – a computer program for the study of natural language communication between man and machine. *CACM*. 9, pp. 36-45.

Wisniewski, G., Pécheux, N., & Yvon, F. (2015). Why Predicting Post-Edition is so Hard? Failure Analysis of LIMSI Submission to the APE Shared Task. *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, Association for Computational Linguistics, pp. 222–227.

Winkler, S. (2005). *Ellipsis and focus in Generative Grammar*. Berlin, Allemagne : Mouton De Gruyter.

Zribi-Hertz, A. (1996). *L'anaphore et les pronoms : une introduction à la syntaxe générative*. Villeneuve-d'Ascq, France : Presses universitaires du Septentrion.

**Dictionnaire**

*Le Grand Robert de la langue française*. (1970). Clichy, France : Société du nouveau Littré.

**Autres**

Melville, H. [1853] (2003), *Bartleby the Scrivener*, *Bartleby le scribe*, trad. Leyris, P. Paris, France : Gallimard.

**Autres traductions consultées**

——————— (1976). *Bartleby*, trad. Causse, M. rep. (1989). *Bartleby,* suivi de Les îles enchantées. Paris, France : Flammarion.
——————— (1994). *Bartleby the Scrivener,* Bartleby, trad. Hoepffner, B. Paris, France : Mille et une nuits.
——————— (2003). *Bartleby the Scrivener, Bartleby, le scribe. Une histoire de Wall Street*, trad. Lacroix, J-Y. Paris, France : éditions Allia.
——————— (2007). *Bartleby the Scrivener*, *Bartleby, une histoire de Wall Street*, trad. Vidal, J. Paris, France : éditions Amsterdam.
——————— (2010). *Bartleby the Scrivener*, *Bartleby le Scribe, Billy Budd, Marin et autres romans*, trad. Jaworski, P. Paris, France : Gallimard.
——————— (2013). Bartleby, le scribe, Herman Melville – Stéphane Poulin, trad par Homassel, A-S. Paris, France : Éditions Sarbacane.

# Sitographie

**Presse en ligne en lien avec le sujet de la thèse**


Entretien avec Pierrette Bouillon
https://www.unige.ch/fti/ebulletin/entretien/entretien-1 (consulté le 24 avril 2019 à 10:05)

Yvon, F. (2017). http://www.lemonde.fr/pixels/article/2017/05/19/malgre-d-impressionnants-progres-la-traduction-automatique-a-encore-du-chemin-aparcourir_5130546_4408996.html#UcMKij3PFKr3Qs7P.99 (Consulté le 20 mai 2017 à 10 : 52)

Enquête de Barbara Vignaux : https://www.liglab.fr/files/qa_traduction_auto_bd.pdf (accès vérifié le 27 juillet 2018 à 10:30)

https://www.20minutes.fr/sciences/1934955-20161003-google-intelligence-artificielle-applique-ameliorer-traduction-automatique (accès vérifié le 2 mai 2018 à 11:00)

https://www.latribune.fr/entreprises-finance/tpe-pme/la-tribune-des-pme-avec-medias-france/systran-pionnier-de-la-traduction-innove-encore-grace-a-l-intelligence-artificielle-750094.htmm (accès vérifié le 2 mai à 2018 14:30)

http://www.systran.fr/systran/technologie/traduction-automatique/ (accès le 19 mai 2018 à 18:30)

http://www.lefigaro.fr/secteur/high-tech/2016/12/07/32001-20161207ARTFIG00005-l-intelligence-artificielle-au-service-de-la-traduction-automatique.php (accès le 13 avril 2018 à 17 :30)

https://www.actuia.com/actualite/la-traduction-neuronale-un-nouveau-standard-retour-sur-les-premieres-rencontres-de-la-communaute-internationale-opennmt/ (accès le 27 juillet 2018 à 20:00)

**Corpus exploités**

**Corpus d'exemples et de développement**

Ellipses repérées manuellement dans les sources ci-dessous :


Amis, M. Swift, G. & McEwan, I. (2006). *Contemporary English Stories, Nouvelles Anglaises Contemporaines*. France : Gallimard.

Arden, J. (1960). *Serjeant Musgrave's Dance*. London, UK : Methuen.

Arden, J (1963). *La Danse du Sergeant Musgrave*. Trad Pons, M. Paris, France : L'Arche.

Bradbury, R. (2011). *The Illustrated Man and Other Short Stories/ L'Homme Illustré et Autres Nouvelles*. France : Gallimard.

Coe, J. (1995). *What a Carve Up*. London, UK : Penguin Books.

Coe, J. (1995). Testament à l'anglaise. Trad. Pavans, J. Paris, France: Gallimard.

Forsyth, F. (2009). *The Shepherd/ Le Berger*. Paris, France : Gallimard.

Green, G. (2012). *Short Stories, Nouvelles*. Paris, France : Langues pour Tous.

Green, J. (2008). *Sud*. Paris, France : Flammarion.

Maugham, S. (2012). *Very Short Sstories/ Nouvelles Brèves*. Paris, France : Langues pour Tous.

Pinter, H. (1965). *The Birthday Party*. London, UK: Methuen.

Pinter, H. (1968). *The Birthday Party, L'Anniversaire.* Trad, Kahane, E. Paris, France: Gallimard.

Shields, C. (1980). *Happenstance*. Toronto, Canada: Vintage Books, Random House of Canada.

Shields, C. (2005). *Au Moment Même.* Trad. Saint-Martin, L. & Gagné, P. Canada : Éditions Québec Amérique Inc.

Steinbeck, J. (2003). *The Red Poney/ Le Poney Rouge.* Paris, France : Gallimard.


Sélection du corpus journalistique **PLECI** (partie non utilisée pour l'évaluation des patrons).

Pour compléter ce corpus, nous avons également utilisé des exemples issus de McShane & Babkin (2016) et Rønning *et al.* (2018b), *voir supra*.

**Corpus d'évaluation**

Extrait du **PLECI Presse + PLECI littéraire**

**PLECI** : corpus compilé par le Laboratoire FoReLL, Université de Poitiers, France et le *Centre for English Corpus Linguistics (CECL)*, Université Catholique de Louvain, Belgique.

**EUROPARL**

*Koehn* P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. *MT Summit* : http://www.statmt.org/europarl/ (accès vérifié le 3 janvier 2018)

**TED talks**

Cettolo M. Girardi C. & Federico M. (2012). WIT3: Web Inventory of Transcribed and Translated Talks. *Proc. of EAMT* (261-268) : https://wit3.fbk.eu/ (accès vérifié le 12 juin 2017 à 11:15).

**Sous-titres**

Percival, B., Bolt, B., Hall, E., Spiro, M. & Engler, M. (2018). *Downton Abbey - Saisons 1-6*. L'intégrale de la série. DVD. Production : Gareth Neame.

Strong, J. & Lyn, E. (2018). *Broadchurch Saisons 1 + 2*. DVD. Production: Kudos Film and Television.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. En ligne http://opus.nlpl.eu/OpenSubtitles-v2016.php (accès vérifié le 12 juin 2017 à 11:15)

**Outils d'exploitation de corpus**

Le CoreNLP *package* développé à l'université de Stanford.
https://stanfordnlp.github.io/CoreNLP/ (accès vérifié le 15 août 2018 à 13:40)

**POS Tagger : étiquetage morpho-syntaxique :**

Toutanova K. & Manning D. (2000). Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, pp. 63-70. https://nlp.stanford.edu/software/tagger.html (accès vérifié le 15 août 2018 à 13:40)

**Stanford Named Entity Recognizer (NER) : Reconnaissance des entités nommées**

Finkel J. Grenager T. & Manning C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370 https://nlp.stanford.edu/software/CRF-NER.html (accès vérifié le 15 août 2018 à 13:40)

**Stanford Parser : analyse syntaxique**

Chen D. & Manning C. (2014). A Fast and Accurate Dependency Parser using Neural Networks. *Proceedings of EMNLP 2014* https://nlp.stanford.edu/software/lex-parser.html (accès vérifié le 15 août 2018 à 15:03)

**TokensRegex**

Chang A. & Manning C. (2014). TokensRegex: defining cascaded regular expressions over tokens. *Stanford University Technical Report, 2014* https://nlp.stanford.edu/software/tokensregex.html (accès vérifié le 15 août 2018 à 15:07)

**Logiciels de traductions utilisés**

**Reverso** http://www.reverso.net/text_translation.aspx?lang=FR (accès vérifié le 5 mai à 17:30)

**DeepL** https://www.DeepL.com/translator (accès vérifié le 4 mai 2018 à 21:39)

**Google Traduction** https://translate.google.fr/?hl=fr (accès vérifié le 2 mai 2018 à 10:55)

**Systran** https://demo-pnmt.systran.net/production#/translation (accès vérifié le 2 mai 2018 à 13:00)

**Outils donnés à titre d'exemples**

Boersma P. & Weenink D. (2010) Praat: doing phonetics by computer. Amsterdam university : http://www.fon.hum.uva.nl/praat/ (accès vérifié le 17 avril 2018 à 4:49)

Landragin F. Poibeau T. & Victorri B. (2012). ANALEC : A New Tool for the Dynamic Annotation of Textual Data. *Proceedings of International Conference on Language Resources and Evaluation (LREC 2012)*: http://lattice.cnrs.fr/Telecharger-Analec (accès vérifié le 17 avril 2018 à 04:50)

**Autres**

https://reports.news.ucsc.edu/linguistics/ (consulté le 13 juillet 2018 à 15 :15)
http://www.cnrtl.fr/definition/suppl%C3%A9er (consulté le 19 juin 2018 à 15:39)
http://ohlone.ucsc.edu/SCEC/ (consulté le 14 avril 2017 à 17:20)
https://wanthalf.saga.cz/intertext (accès vérifié le 20 juillet 2018 à16 :30)
https://reports.news.ucsc.edu/linguistics/ (accès le 20 octobre 2017 à 20 :33)
http://forums.imore.com/off-topic-lounge/281999-programmer-jokes-just-another-off-topic-thread-imore.html (consulté le 30 Juillet 2014 à 19 :40)
www.boutique.afnor.org (consulté le 25 janvier 2018 à 13 :10)

# Annexe I : Corpus majoritairement exploités dans les recherches modernes sur l'ellipse

Le *Brown Corpus*[167] est le premier (1960) corpus compilé lisible par la machine. Il contient environ un million de mots de l'anglais américain réunis dans 500 échantillons de plusieurs genres discours : religieux, journalistiques, théâtraux, romanesques, et humoristiques. À ce jour, six versions de ce corpus sont disponibles avec ou sans annotations particulières.

Le *COCA*[168] est également une collection de textes de l'anglais américain contemporain, compilé pour la première fois en 1990 et mis à disposition de la communauté scientifique en 2008. Le *COCA* est considéré comme étant le corpus le plus équilibré de l'anglais américain et est aujourd'hui exploité par de nombreux chercheurs, traducteurs, linguistes et enseignants. Comme indiqué sur son site[169], il contient « plus de 560 millions de mots dans 220 225 textes auxquels se sont ajoutés 20 millions de mot, chaque année, de 1990 à 2017 ». Les textes collectés dans ce corpus appartiennent à plusieurs registres de langue, soutenu ou familier, et différents types de discours : académique, journalistique, fiction.

Le *BNC*[170] est un ensemble d'échantillons de l'anglais britannique écrit et oral des années 1990 contenant environ 100 millions de mots. Pour ce qui est de la partie écrite, le BNC contient des textes extraits d'articles de presse, d'ouvrages universitaires courants, des lettres et des essais publiés et non-publiés. La partie orale

[167] http://clu.uni.no/icame/brown/bcm.html (accès vérifié le 18 mai 2017 à 13:30)
[168] The Corpus of Contemporary American English : https://corpus.byu.edu/coca/ (accès vérifié le 18 mai 2017 à 13:05)
[169] "560 million words in 220,225 texts, including 20 million words each year from 1990-2017"
[170] http://www.natcorp.ox.ac.uk/ (accès vérifié le 18 mai 2017 à 15 :10)

comprend, quant à elle, les transcriptions de conversations informelles de personnes volontaires, d'âge, de sexe, et de religion variés, toutes classes sociales confondues.

Le *WSJ corpus*[171] est une collection d'articles de *Wall Street Journal,* d'environ 30 millions de mots, collectés sur une période de trois ans. La particularité de ce corpus est le fait qu'il est syntaxiquement étiqueté et analysé.

Le *New York Times corpus*[172] est une collection d'environ 1,8 million d'articles publiés dans le journal *New York Times* entre 1987 et 2007. La plupart des articles sont annotés et résumés manuellement.

[171] https://catalog.ldc.upenn.edu/ldc99t42 (accès vérifié le 18 mai 2017 à 15 :30)
[172] un sous-ensemble du corpus anglais *Gigaword* (*2nd edition*)
https://catalog.ldc.upenn.edu/LDC2005T12 (accès vérifié le 18 mai 2017 à 15 :45)

# Annexe II : Fiches des séries Broadchurch et Downton Abbey

| Broadchurch | Downton Abbey |
|---|---|
| **Genre** : série dramatique policière<br>**Création** : Chris Chibnal<br>**Pays** : Royaume-Uni<br>**Nombre de saisons** : 3 (seules 2 sont utilisées ici)<br>**Durée** : 25 min par épisode<br><br>Sous-titres récupérés des DVD de la série | **Genre** : série dramatique historique<br>**Création** : Julian Fellowes<br>**Pays** : Royaume-Uni<br>**Nombre de saisons** : 6 (seules 5 sont utilisées ici)<br>**Durée** :<br>- 50 min par épisode<br>- 66 min premier et dernier épisode par saison<br>- 92 min par épisode spécial noël<br>Sous-titres récupérés des DVD de la série |

## Annexe III : Expressions régulières utilisées

**Post-geni**

```
's[.,-][^\t\n]*$
```

**Post-be/have**

```
\b(are|am|is|was|were|have|has|had|'d|'m|haven't|hasn't|a
ren't|hadn't|weren't|isn't|'wasn't|'s|'re)( not)?[.!,-
][^\t\n]*$
```

**Post-do**

```
\b(do|does|did|don't|doesn't|didn't)( not)?[.!,-
][^\t\n]*$
```

**Post-mod**

```
\b(will|would|won't|wouldn't|can|could|couldn't|can't|can
not|'d|shall|should|shouldn't|shan't|must|mustn't|may|mig
ht)( not)?[.!,:;?-][^\t\n]*$
```

**Post-to**

```
\bto[.!,:;?-][^\t\n]*$
```

**Post-wh**

```
\b(why|how( many|
much)?|when|why|where|what|which|whose)(
(for|about))?[.!,?-][^\t\n]*$
```

**Qs-frag**

```
([.-] ?)[A-Z]\w+(ed|ing|e|er|t)( (it|him|us|you|her))?\?
^[A-Z]\w+(ed|ing|e|er|t)( (it|him|us|you|her))?\?
^[A-Z]\w+(ed|ing)\b.*\?[^\t\n]*$
(- )?(ever|already|never).*\?[^\t\n]*$
\b(ever|already|never)\b.*\?[^\t\n]*$
\b(remember|think)
^want
[.] want
[.] been
```

**vs-tag**

```
((are|am|is|was|were|have|has|had|'d|'m|haven't|hasn't|ar
en't|hadn't|weren't|isn't|'wasn't|'s|'re|do|does|did|don'
t|doesn't|didn't|will|would|won't|wouldn't|can|could|coul
dn't|can't|cannot|'d|shall|should|shouldn't|shan't|must|m
ustn't|may|might)( not)?) (I|you|he|she|it|we|they)
?[!?:]+[^\t\n]*$
```

**Post-ord**

```
the (first|second|third|fourth|fifth|sixth)[.;!?:,]+
```

**Post-card**

```
\b(one|two|three|four|five|six|seven|eight|nine|ten|eleve
n|twelve)[.;!?:,]+[^\t\n]*$
```

**Post-quant**

```
\b(some|any)[.;!?:,]+[^\t\n]*$
```

# Annexe IV : Bref aperçu historique de la traduction automatique

Les origines des recherches pour mettre en place ce paradigme datent des années 1950 (1948 au Royaume-Uni et aux États-Unis, et 1954 en Union soviétique et en France). Aujourd'hui, la traduction automatique est en pleine expansion, plaçant les investigations initiées par les recherches dans le domaine au centre d'une réelle dynamique. Pour parvenir à ce stade de réalisation, l'histoire de la traduction automatique s'est déroulée en six grandes périodes (Léon 2002)[173].

1948-1960 : période des idées et des expérimentations

1960-1966 : analyse syntaxique comme moyen de faire avancer la traduction

1966-1980 : arrêt des recherches sur la traduction automatique suite à une crise financière et économique.

1980-1990 : automatisation de la communication (traduction des modes d'emploi, et guides d'utilisation, …)

Dès les années 1990 : regain d'intérêt pour la traduction automatique et retour des traductions par règles et traductions statistiques.

2016 : Révolution et naissance de la traduction neuronale (le développement a sans doute été initié avant 2016 mais est devenu « visible » récemment).

La première réalisation d'une traduction automatique, à savoir la traduction de 49 phrases simples du russe vers l'anglais, a été faite grâce à un dictionnaire de 250 mots et six règles syntaxiques. Le choix de ces deux langues n'était pas anodin puisqu'on se trouvait alors en pleine guerre froide. Dès les années 1980, la traduction automatique connait une nouvelle étape avec l'automatisation de la communication et l'implication des Japonais dans la fabrication et la commercialisation des premiers micro-ordinateurs et des traitements de textes à l'échelle internationale. En effet, afin d'en assurer la vente, une traduction des modes d'emploi s'est avérée nécessaire. Dès lors, l'intérêt pour la traduction automatique trouve un nouvel essor.

[173] Rédigé à partir des notes de Léon (2002), de Gross (1972), Poibeau (2017) ainsi que des informations présentées sur le blog Systran http://blog.systransoft.com/how-does-neural-machine-translation-work/ (accès vérifié le 25 mai 2018 à 14:00)

Elle n'est désormais plus considérée comme une solution *secondaire* mais acquiert un statut à *part entière* dans le domaine du TAL, en pleine expansion. Trois approches majeures peuvent être identifiées comme issues de ces travaux préliminaires dans la traduction automatique : la traduction *directe*, la traduction à base de langue *intermédiaire* et la traduction s'appuyant sur le *transfert*. Ces trois approches sont illustrées dans le célèbre triangle de Vauquois (figure 26), en référence aux travaux menés par Boitet & Vauquois (1985).

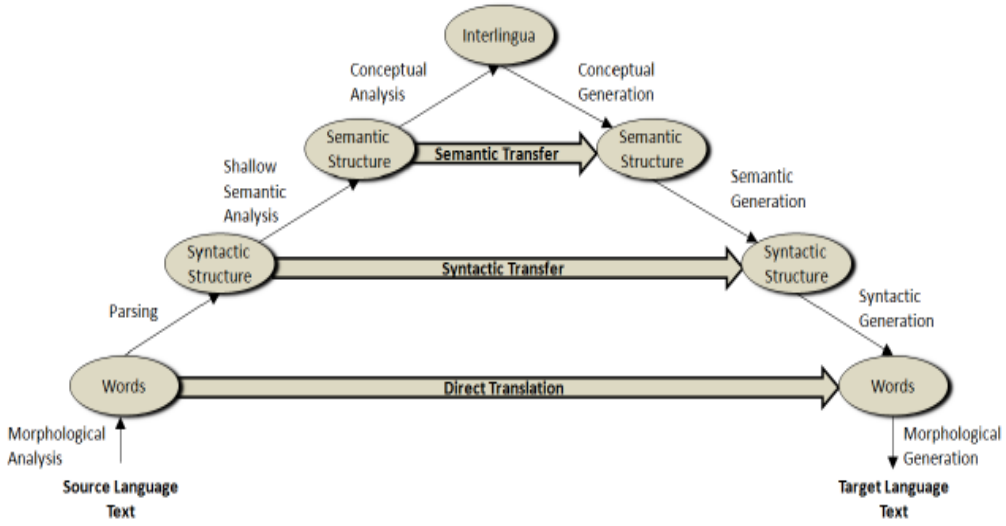



**Figure 26 : Le triangle de Vauquois 1985 illustrant les trois approches de traduction**[174]

Dans la traduction *directe*, des règles précises lors du passage d'une langue source à une langue cible sont appliquées. Seul le contenu sémantique est prioritaire, les règles utilisées ne relèvent pas de la syntaxe. La traduction à partir d'une langue *intermédiaire* consiste quant à elle à transcrire la langue cible dans une langue intermédiaire, appelée parfois *langue pivot*, qui n'appartient pas aux langues humaines, mais plutôt à un codage neutre ou *interlangue*, avant de générer la traduction dans la langue cible. La troisième approche s'appuie sur *le transfert*[175].

[174] https://www.researchgate.net/figure/The-Vauquois-triangle_fig1_233970127 (accès le 17 mai 2018 à 19 : 09)

[175] Opération qui consiste à produire des représentations syntaxiques et/ou sémantiques du texte en langue source et à les transférer dans la langue cible. Ces représentations résultent des analyses visant à désambiguïser la langue source.

Cette traduction se déroule en trois étapes : d'abord, l'analyse du contenu source, ensuite le transfert de l'analyse vers la langue cible et enfin la production de la traduction. Les analyses effectuées dans cette dernière approche sont beaucoup plus approfondies que les approches décrites précédemment. Elles comprennent des analyses morphologiques, morphosyntaxiques, en dépendances, en constituants, et des analyses sémantiques.

Les années 1990 sont marquées par l'avènement des corpus parallèles de grande taille et les méthodes d'exploitation mises en œuvre (alignement) permettent ainsi l'émergence de nouvelles approches en traduction automatique. L'insuffisance des données initiales est peu à peu comblée par l'importance des apports fournis par les corpus, générant de ce fait une orientation vers un nouveau paradigme fondé sur l'apprentissage automatique. Ainsi, pour répondre au besoin croissant de la traduction automatique, plusieurs approches computationnelles se développent. À ce jour, nous en retenons trois : la traduction automatique à base de règles, la traduction statistique et la traduction neuronale.

La traduction automatique à base de règles (*Rule-based machine translation*) est l'approche la plus ancienne et repose sur le triangle de Vauquois. Elle ne constitue pas une approche statistique et est indépendante des corpus. L'une de ses principales caractéristiques est l'avancée de la phase d'analyse dans la langue source au détriment de la phase de production qui reste très réduite. Cette traduction s'appuie donc sur un ensemble de règles et d'items lexicaux issus des grammaires et des dictionnaires et pouvant faire l'objet d'une modification par le linguiste. Ces règles sont réunies dans un programme exécutable afin de générer la traduction dans la langue cible.

La traduction automatique statistique à base de segments (*Phrase-based machine translation*) a fait son apparition dans les années 2000 et a été jusqu'en 2016 le paradigme majoritairement utilisé par les outils de traduction automatique. Cette traduction s'est inspirée de la méthode de déchiffrage de la pierre de Rosette et repose sur de larges quantités de corpus parallèles, textes traduits par l'humain, pour

envisager le calcul de la probabilité qu'un texte B soit la traduction adéquate d'un texte A. Trois ressources sont nécessaires à la traduction statistique :

- une table de segments[176] (*phrase-table*) : traductions possibles pour chaque segment et sa probabilité ;

- une table de réordonnancement (*reordering table*) : indique l'ordre des mots à suivre lors du passage vers la langue cible ;

- un modèle de langue (*language model*) : probabilités des séquences de mots dans la langue cible.

À partir de ces résultats, la traduction d'autres textes peut être générée[177].

La différence entre la traduction à base de règles et la méthode statistique à base de segments réside dans le fait que la première requiert une analyse linguistique préalable pour développer des règles qui constitueraient le système de traduction, tandis que la deuxième s'appuie entièrement sur le corpus parallèle. Il existe également des approches hybrides qui ont recours aux deux modes de traduction dans l'espoir de produire une traduction qui se rapprocherait le plus possible de celle du traducteur humain. Malgré ces efforts, les systèmes de traduction automatique rencontrent toujours des problèmes au niveau de la grammaire et de la syntaxe, des incompatibilités avec le contexte, des séquences redondantes, erreurs impactant la fluidité de la traduction.

[176] Un *segment* correspond à ce que les traductologues appellent « unité de traduction ».
[177] Toujours dans les mêmes langues.

# Annexe V : Logiciels de traduction automatique utilisés

Les principaux outils utilisés dans cette recherche afin de comparer les traductions des occurrences elliptiques, sont Google Traduction, Systran, Reverso et DeepL. Voici, ci-dessous, une présentation brève des éléments-clefs de ces outils :

**Google Traduction**[178] est un service visant la traduction de textes ou de pages web dans plusieurs langues. Il est proposé par Google (Alphabet Inc.), entreprise américaine de services technologiques fondée en 1998 et présente à l'échelle internationale. Le système de Google Traduction fonctionne à partir de la traduction statistique et est passé depuis peu à la traduction neuronale pour huit langues. Malgré cette évolution, les chercheurs de Google reconnaissent les lacunes de cette traduction. Ils expliquent :

> GNMT [Google Neural Machine Translation] fait encore de grosses erreurs qu'un traducteur humain ne ferait jamais, comme laisser des mots de côté et mal traduire des noms propres ou des termes rares, ou traduire des phrases de manière isolée au lieu de prendre en compte le contexte du paragraphe ou de la page. Il y a encore beaucoup de travail pour fournir un meilleur produit à nos utilisateurs[179].

Google Traduction est mis à la disposition de tous les usagers d'Internet en libre accès.

Le traducteur automatique **Systran**[180] est proposé par l'entreprise Systran et reste sans doute l'un des traducteurs les plus anciens (1968). Systran est passé depuis 2016 à la traduction neuronale et sa technologie, selon Jean Senellart[181], directeur technique et innovation, sert à :

[178] https://translate.google.fr/?hl=fr (accès vérifié le 2 mai 2018 à 10 : 55)
[179] https://www.20minutes.fr/sciences/1934955-20161003-google-intelligence-artificielle-applique-ameliorer-traduction-automatique (accès vérifié le 2 mai 2018 à 11 : 00)
[180] https://demo-pnmt.systran.net/production#/translation (accès vérifié le 2 mai 2018 à 13 : 00)
[181] https://www.latribune.fr/entreprises-finance/tpe-pme/la-tribune-des-pme-avec-medias-france/systran-pionnier-de-la-traduction-innove-encore-grace-a-l-intelligence-artificielle-750094.html (accès vérifié le 2 mai à 2018 14 : 30)

> Entrainer des agents conversationnels, ou chatbots. Il sera également possible de parler dans notre langue maternelle et d'être compris par un étranger, grâce à un objet connecté directement placé dans l'oreille. Grâce à la réalité augmentée, on pourra lire des panneaux ou des menus en langues étrangères.

Le logiciel de traduction Systran n'est pas le plus utilisé par les particuliers. Sa version complète est payante. De ce fait, il est plutôt destiné aux entreprises et aux grandes institutions comme le Parlement Européen, les gouvernements ou les organismes de défense. Pour la présente recherche, nous utilisons la version d'essai en ligne, forcément limitée dans la quantité à traduire. D'autres limites sont également énoncées sur le site de Systran[182] :

> - Some short sentences (or single words) are badly translated,
> - Very very long sentences may be truncated or reformulated,
> - We may experience some issues with Entities (both Numeric ones or Named ones). Named Entities may not be transliterated (e.g. in Arabic),
> - Some issues with quotes and numerical values with Japanese as target language.

**Reverso**[183] est un traducteur accessible en ligne, disponible en libre accès qui permet la traduction de mots et de textes. Il est proposé aux côtés d'autres outils comme le Reverso dictionnaire, correcteur d'orthographe, conjugaison et grammaire. Il est disponible en plusieurs langues.

**DeepL** est un tout nouveau traducteur automatique, libre d'accès dans sa version d'essai moins restrictive que celle de Systran, et qui utilise le système de traduction neuronale. Il a été développé en 2017 par l'équipe Linguee. Sept langues sont prises en compte par ce logiciel. DeepL se promeut sur Wikipédia de la façon suivante :

> DeepL dépasserait ses concurrents dans des tests à l'aveugle, entre autres Google Traduction, Microsoft Traduction et Facebook. Il serait aussi plus précis et plus nuancé pour une rapidité égale à ses concurrents[184].

[182] https://demo-pnmt.systran.net/information#/view (accès vérifié le 3 mai 2018 à 3 : 15)
[183] http://www.reverso.net/text_translation.aspx?lang=FR (accès vérifié le 5 mai 2016 à 17 :30)
[184] https://fr.wikipedia.org/wiki/DeepL (accès vérifié le 4 mai 2016 à 20 : 30)

Toute distance par rapport à Wikipédia mise à part, il n'en reste pas moins que l'expérience semble bien confirmer cette annonce. La traduction de certains de nos exemples d'ellipse a montré que DeepL était le logiciel qui produisait effectivement le moins d'erreurs, notamment dans la traduction de l'anglais vers le français. À ce sujet, le journal *Le Monde* écrit par exemple que :

> DeepL a également obtenu de meilleurs résultats que les autres services, grâce à des tournures de phrase plus « françaises »[185].

Au cours de cette recherche, nous avions souhaité travailler sur Microsoft car le logiciel permettait jusqu'en mars 2018 de faire une comparaison entre la traduction statistique et la traduction neuronale, ce qui correspondait à l'un de nos axes de recherche. Cette version n'est actuellement plus accessible en ligne depuis la dernière mise à jour[186]. Tenant compte de cet aléa lié directement à la dynamique du développement technologique, nous nous sommes alors fixée comme but essentiel, la comparaison des occurrences elliptiques au sein d'un même système, le système neuronal, tel qu'il est appliqué dans les logiciels définis ci-dessus.

Google Traduction, Reverso, Systran et DeepL constituent donc nos outils de base pour mener la comparaison entre différentes traductions d'occurrences elliptiques. Choisir ces outils et non d'autres parmi ceux disponibles sur le marché dans le domaine de la traduction automatique, correspond au fait qu'ils ont tous évolué vers la traduction neuronale et qu'ils sont tous aptes à travailler dans les deux langues analysées dans cette étude. La comparaison a fait l'objet du dernier chapitre de cette recherche et tente de montrer les enjeux et les défis que pose l'ellipse à l'épreuve de ces outils.

[185] https://www.DeepL.com/translator (accès vérifié le 4 mai 2018 à 21 : 39)
[186] 12 mars 2018

## Annexe VI : Erreurs de la TA répertoriées par Loffler-Laurian (1983) et Grass (2010)

| **Loffler-Laurian (1983)**<br>**Sur une étude du système *Systran***<br>**Anglais-Français** | **Grass (2010)**<br>**Sur une étude des systèmes *Systran* et *Google***<br>**Anglais-Français-Allemand** |
|---|---|
| **Erreurs liées aux mots isolés**<br><br>- Le vocabulaire et la terminologie (substantif, adjectif, verbe)<br><br>- Les sigles et les noms propres<br>- Les relateurs en particulier les prépositions :<br>- dans les déterminants nominaux<br>- dans les complémentations verbales<br>- Le déterminant de liste finie du nom : articles et démonstratifs, les modificateurs du verbe<br><br>Erreurs liées aux relations<br><br>- Les formes des verbes (temps)<br>- Les voix verbales (passif/actif) et la personnalisation de l'énoncé (passif/impersonnel par *il* ou par *on)*<br><br>- L'expression ou non de la modalité<br>- les négations<br>Erreurs liées aux structures et informations<br>- L'ordre des mots et des informations<br>- Les problèmes généraux d'incidence | - Polysémie et homonymie<br>- Ambiguïté syntaxique<br>- Ambiguïté référentielle<br>- Termes flous ou *Fuzzy hedges*<br>- Idiotismes et métaphores<br>- Néologie<br>- Noms propres<br>- Mots d'origine étrangère et emprunts<br>- Sigles et acronymes<br>- Séparateurs<br>- Synonymes<br>- Transposition<br>- Orthographe |

# Index de notions

## P



## R



## T

Université de Strasbourg

ÉCOLE DOCTORALE DES HUMANITÉS (ED 520)

**Anissa HAMZA**

# La détection et la traduction automatiques de l'ellipse

## Enjeux théoriques et pratiques

Cette thèse a pour objet le traitement automatique du phénomène elliptique. À la croisée de plusieurs disciplines – linguistique théorique, linguistique de corpus, linguistique outillée et traductologie –, elle s'inscrit dans une démarche expérimentale en poursuivant deux objectifs essentiels. Il s'agit tout d'abord de vérifier la possibilité de détecter automatiquement le phénomène elliptique en anglais pour explorer ensuite les procédures facilitant sa traduction automatique de l'anglais vers le français.
La détection automatique repose sur des analyses morphosyntaxiques qui paraissent suffisantes à la détection automatique de *certaines* catégories d'ellipse, puisqu'en décomposant le phénomène, elles permettent de l'identifier parmi d'autres. Un corpus parallèle et multi-genres, collecté et conçu pour répondre aux hypothèses de recherche, est utilisé. Afin d'élaborer des patrons de détection et exploiter le corpus, cette recherche utilise les outils CoreNLP développés à l'université de Stanford (USA) et met en lumière leurs limites lorsqu'ils sont confrontés à l'ellipse. Les résultats obtenus s'articulent autour du lien établi entre la détection et la traduction automatiques du phénomène elliptique, facteur déterminant dans la compréhension des erreurs de traduction générées lors de son traitement automatique.

**Mots-clefs** : Ellipse, détection automatique, patron, corpus, traduction automatique.

***The Automatic Detection and Translation of Ellipsis. A Theoretical and Practical study***

*This thesis examines the automatic processing of elliptical phenomena in English. Situated at the crossroads of various disciplines, such as theoretical linguistics, corpus linguistics, computational linguistics and translation studies, it implements an experimental approach in order to achieve two main goals. The first one consists in establishing that ellipsis can be detected automatically, and the second one in exploring procedures to facilitate its machine translation from English into French.*
*For these purposes, this thesis resorts to morpho-syntactic analyses, which seem sufficient for the automatic detection of some types of ellipsis, since through breaking down the phenomenon into its specific features it makes it possible to identify it among others. The investigation draws on a parallel, multi-genre corpus compiled to test the research hypotheses. In order to develop detection patterns and exploit the corpus, the CoreNLP package developed at the University of Stanford (US) is used, and its limitations in the processing of ellipsis are highlighted. Findings and results emphasize the close link between the detection of ellipsis and its machine translation as a key factor in understanding translation errors caused by its automatic processing.*

***Keywords**: Ellipsis, automatic detection, pattern, corpus, machine translation.*