Traitement automatique des langues

Varia

sous la direction de
Cécile Fabre
Emmanuel Morin
Sophie Rosset
Pascale Sébillot

Vol. 63 - n°1 / 2022

ATALA

Varia

ATALA

Revue de
l'Association
pour le Traitement
Automatique
des Langues

TAL
Vol.
63

n°1
2022

Varia

ATALA

Traitement automatique des langues

Revue publiée depuis 1960 par l'Association pour le Traitement Automatique des LAngues (ATALA), avec le concours du CNRS.

ISSN 1965-0906

https ://www.atala.org/revuetal

Traitement automatique des langues

Traitement automatique des langues

Volume 63 – n°1 / 2022

VARIA

Table des matières

Préface

Les préfaces des numéros non thématiques de la revue *TAL* permettent de faire chaque année le point sur la vie de la revue et de commenter les statistiques que nous présentons traditionnellement pour les numéros des trois années précédentes.

Nous avons, comme chaque année, procédé au renouvellement des membres du comité et, cette année encore, nous pouvons nous féliciter du maintien de la parité.

La régularité de publication de la revue est maintenue, avec moins de dix mois par numéro en moyenne sur les quatre dernières années. Rappelons que l'une des caractéristiques de notre revue est la définition d'un calendrier prévisionnel intégrant des réunions du comité de rédaction – en visioconférences – au cours desquelles, entre autres, nous décidons collégialement, à l'appui des relectures reçues, de l'acceptation ou du rejet des articles soumis.

La poursuite de la réflexion sur l'attractivité de la revue, entamée en 2021, a donné lieu à la diffusion d'un questionnaire sur la liste LN et lors de l'AG de l'ATALA. Ce dernier a été conçu par les rédacteurs en chef et le comité de rédaction pour recueillir des avis sur le fonctionnement actuel de la revue et sur les évolutions à envisager pour mieux répondre aux attentes de la communauté scientifique. De fait, la diminution des soumissions constatée dès 2020 s'est poursuivie, malgré la définition de thématiques plus larges que d'ordinaire en 2021 (voir tableau1, lignes 62-2 et 62-3). En 2022, les deux thématiques retenues ont été *TAL intermodal et multimodal* (63-2) et *États de l'art dans un domaine du TAL* (63-3). Le premier numéro thématique a eu peu d'échos malgré la pertinence du thème (soumissions en cours, trois reçues actuellement). Le second numéro – le premier de ce genre – a reçu dix-huit intentions de soumission et semble donc répondre à un besoin, mais seules huit soumissions ont finalement été déposées. Les réponses au questionnaire seront dépouillées et les résultats seront analysés et discutés au cours de prochaines réunions du comité éditorial de la revue, avec pour objectif de cerner les évolutions souhaitables de la revue.

Toujours dans un souci d'attractivité de la revue et de meilleure visibilité des articles qu'elle contient, il a été décidé, au cours de l'année 2022, de doter chacun de ceux-ci d'un DOI (*Digital Object Identifier*) individuel. Cette identification numérique individuelle devrait être effective pour le prochain volume.

Passons maintenant à nos statistiques. Elles considèrent toujours les dix derniers numéros sur les trois dernières années, en l'occurrence donc, du début de 2019 jusqu'à

Intitulé	Vol.	N°	Année	Soumis	Acceptés	% acceptés
Varia	60	1	2019	8	1	12,5 %
Corpus annotés	60	2	2019	6	3	50,0 %
TAL et humanités numériques	60	3	2019	13	5	38,4 %
Sous-total	**60**		**2019**	**27**	**9**	**33,3 %**
Varia	61	1	2020	8	1	12,5 %
TAL et santé	61	2	2020	4	3	75,0 %
Dialogue et systèmes de dialogue	61	3	2020	5	3	60,0 %
Sous-total	**61**		**2020**	**17**	**7**	**41,2%**
Varia	62	1	2021	7	2	28,6 %
Nouvelles applications du TAL	62	2	2021	2	1	50,0 %
Diversité linguistique en TAL	62	3	2021	3	2	66,7 %
Sous-total	**62**		**2021**	**12**	**5**	**41,7%**
Varia	63	1	2022	5	3	60,0%
Total	**Dix derniers n^{os}**			**61**	**24**	**39,3 %**

TABLEAU 1. *Taux de sélection aux appels de la revue TAL sur les dix derniers numéros de la période 2019-2022*

ce numéro *Varia* de 2022 inclus. Le tableau 1 donne les taux de sélection par numéro et par volume. La ligne du total synthétise ces chiffres sur l'ensemble des dix numéros considérés. Le taux de sélection sur l'ensemble de ces numéros s'élève à 39,3 % en moyenne, avec toutefois de grandes variations d'un numéro à l'autre (de 12,5 % pour les *Varia* de 2019 et 2020 à 75,0 % pour le numéro *TAL et santé*).

Rappelons qu'un numéro ne peut pas excéder cinq articles, pour des raisons liées au coût du processus d'édition. Ce seuil n'a pas été atteint depuis le numéro *TAL et humanités numériques* de 2019. Le comité de rédaction de la revue reste attaché à sélectionner les articles sur le seul critère de leur qualité, indépendamment du nombre d'articles soumis, et n'hésite pas à préserver cette exigence même en cas de nombre limité de soumissions.

Les statistiques que nous donnons sur l'origine des articles publiés considèrent le pays du laboratoire du premier auteur, hors de France ou pas, ainsi que la langue de la soumission, français en principe ou anglais si l'un des coauteurs n'est pas francophone. Les chiffres sont fournis dans le tableau 2 pour la même période de temps que le tableau 1. Le premier point que l'on constate est que les numéros *Varia* ont très majoritairement uniquement des auteurs francophones, et que pour tous ces derniers *Varia*, tous les (premiers) auteurs étaient en France. Ce constat ne s'applique en général pas aux numéros thématiques.

Le présent numéro contient les trois articles retenus lors de l'appel non thématique lancé en décembre 2021 et clos à la mi-juillet 2022. Cet appel portait comme d'habi-

Intitulé	**Vol.**	**N°**	**Année**	**% 1er auteur hors France**	**% en anglais**
Varia	60	1	2019	0,0 %	0,0 %
Corpus annoté	60	2	2019	0,0 %	0,0 %
TAL et humanités numériques	60	3	2019	40,0 %	40,0 %
Pourcentages par volume	**60**		**2019**	**22,2 %**	**22,2 %**
Varia	61	1	2020	0,0 %	0,0 %
TAL et santé	61	2	2020	33,3 %	33,3 %
Dialogue et systèmes de dialogue	61	3	2020	66,7 %	66,7 %
Pourcentages par volume	**61**		**2020**	**42,9 %**	**42,9 %**
Varia	62	1	2021	0,0 %	0,0 %
Nouvelles applications du TAL	62	2	2021	0,0 %	0,0 %
Diversité linguistique en TAL	62	3	2021	50,0 %	50,0 %
Pourcentages par volume	**62**		**2021**	**20,0 %**	**20,0 %**
Varia	63	1	2022	0,0%	33,3%
Pourcentages totaux	**Dix derniers nos**			**25,0 %**	**29,2 %**

TABLEAU 2. *Proportion des articles publiés d'un premier auteur d'un laboratoire hors de France et proportion des articles publiés rédigés en anglais sur les dix derniers numéros de la période 2019-2022. Attention, les pourcentages totaux ne sont pas de simples moyennes des chiffres donnés plus haut, car les dénominateurs changent.*

tude sur tous les aspects du traitement automatique des langues. Cinq articles ont été soumis, dont un en anglais, ce qui représente un très faible nombre de soumissions.

À l'issue du processus de sélection habituel à deux tours, trois articles ont été retenus pour publication :

– *Sous-titrage automatique : étude de stratégies d'adaptation aux genres télévisuels*, François Buet et François Yvon (LISN) : se situant dans le cadre de la traduction et du sous-titrage automatiques, cet article présente et compare des méthodes d'adaptation aux genres télévisuels pour la production de sous-titres ;

– *Biais de genre dans un système de traduction automatique neuronale : une étude des mécanismes de transfert cross-langue*, Guillaume Wisniewski (LLF), Lichao Zhu (LLF), Nicolas Ballier (CLILLAC-ARP), François Yvon (LISN) : cet article présente une analyse des biais de genre dans les systèmes de traduction automatique et recherche leurs causes en étudiant les différentes manières dont l'information de genre peut circuler entre le décodeur et l'encodeur ;

– *Survey on Narrative Structure : from Linguistic Theories to Automatic Extraction Approaches*, Aman Berhe (LISN), Camille Guinaudeau (LISN), Claude Barras (VOCAPIA) : cet article présente un état de l'art sur les modèles théoriques, les méthodes pour l'extraction automatique des structures narratives ainsi que les jeux de données utilisés.

On trouvera à la suite de ces articles des notes de lecture. Nous encourageons nos lecteurs à se faire mutuellement profiter de leurs lectures et à se mettre en contact avec Denis Maurel (denis.maurel@univ-tours.fr) pour les publier ici. Suit une liste de résumés de thèses ou d'habilitations à diriger les recherches en traitement automatique des langues préparée par Sylvain Pogodalla. Merci à Denis et Sylvain pour leur travail de veille et de collecte.

Merci aux membres du comité de rédaction de la revue qui ont participé aux différentes étapes d'élaboration de ce numéro, et en particulier à ceux qui ont pris en charge des relectures (voir la composition du comité sur le site de la revue : https://www.atala.org/content/comit%C3%A9-de-r%C3%A9daction-0), ainsi qu'aux relecteurs spécifiques de ce numéro :

– Nathalie Friburger, LIFAT, Université de Tours ;

– David Doukhan, INA ;

– Solen Quiniou, LS2N, Nantes Université ;

– Loïc Barrault, Meta ;

– Laurent Besacier, NeverLabs.

Enfin, rappelons que la revue TAL reçoit un soutien financier de l'Institut des sciences humaines et sociales (INSHS) du CNRS et de la Délégation générale à la langue française et aux langues de France (DGLFLF). Nous adressons nos remerciements à ces organismes.

Cécile Fabre
CLLE, Université Toulouse 2
cecile.fabre@univ-tlse2.fr

Emmanuel Morin
LS2N, Université de Nantes
emmanuel.morin@univ-nantes.fr

Sophie Rosset
Université Paris-Saclay, CNRS, LISN
sophie.rosset@lisn.fr

Pascale Sébillot
IRISA, INSA Rennes
pascale.sebillot@irisa.fr

Sous-titrage automatique : étude de stratégies d'adaptation aux genres télévisuels

François Buet — François Yvon

Université Paris-Saclay, CNRS, LISN,
Campus universitaire bât 508, Rue John von Neumann, F - 91405 Orsay cedex
{francois.buet,francois.yvon}@limsi.fr

RÉSUMÉ. *Les obligations légales concernant l'accessibilité des contenus audiovisuels conjuguées avec l'importance des volumes actuellement produits par diverses sources suscitent un intérêt croissant pour les systèmes de sous-titrage automatique. Traditionnellement, ces systèmes procèdent en enchaînant une étape de reconnaissance de la parole et une étape de « traduction » de la transcription vers les sous-titres. Pour le sous-titrage monolingue, la « traduction » correspond à une simplification et à une segmentation du texte, qui doivent notamment respecter des normes liées à l'affichage, et composer avec les erreurs issues de la reconnaissance vocale. Dans le cas des émissions télévisées, la forme et la teneur du flux audio initial comme des sous-titres à répliquer varient significativement selon les programmes. En prenant inspiration dans la littérature de la traduction automatique, cet article met en place et compare des méthodes d'adaptation aux genres télévisuels pour la production de sous-titres.*

MOTS-CLÉS : *sous-titrage automatique, simplification de textes, traduction automatique.*

TITLE. *Automatic closed captioning: a study of strategies for televisual genre adaptation*

ABSTRACT. *Interest in automatic closed captioning systems has risen on account of legal obligations concerning accessibility and the sheer amount of audiovisual content being produced by multiple sources. Such systems usually proceed by coupling Automatic Speech Recognition (ASR) and Machine Translation (MT) from transcript to captions. The "translation" task consist of a simplification and segmentation of the text, which must observe norms with respect to display, while handling ASR errors. In the case of TV shows, both the initial audio stream and the target captions vary significantly in form and content according to the program. Taking inspiration in MT literature, this paper implements and compare televisual genre adaptation methods for closed captioning.*

KEYWORDS: *automatic close captioning, text simplification, machine translation.*

1. Introduction

La production de sous-titres monolingues destinés à rendre accessibles au public sourd ou malentendant les émissions télédiffusées est depuis 2010 une obligation légale[1], qui a conduit à une augmentation considérable du nombre d'heures à sous-titrer. À côté de ces besoins réglementés, la demande de sous-titrage explose également sur Internet pour d'autres types de contenus : cours en lignes, vidéos de tutoriels, films promotionnels, etc. Dans ce contexte, le besoin de disposer d'outils performants pour assister à la production de sous-titres, voire de les réaliser de manière entièrement automatique, est de plus en plus pressant. Ces outils s'appuient typiquement sur des architectures en cascade, qui enchaînent une transcription vocale et des étapes de compression et simplification. Ces architectures sont aujourd'hui de plus en plus concurrencées par des architectures réalisant un apprentissage de bout en bout, un constat qui vaut également pour le sous-titrage en langue étrangère (Bérard *et al.*, 2016 ; Matusov *et al.*, 2019 ; Sperber et Paulik, 2020 ; Karakanta *et al.*, 2021).

Pour ce qui concerne le sous-titrage d'émissions de télévision, la génération de sous-titres intervient souvent en bout de chaîne du processus de production et est principalement réalisée selon deux modalités très différentes : le sous-titrage en direct, pour les journaux d'information, les émissions de plateau ou les évènements retransmis en direct ; le sous-titrage en différé pour les émissions de jeux, les documentaires et les fictions. Dans le premier cas, des contraintes de temps réel sont critiques, et le sous-titreur doit s'adapter à la spontanéité des prises de parole et plus généralement aux aléas du direct ; dans le second cas, il faut potentiellement faire face à une plus grande variété de la parole et des évènements sonores à prendre en compte : chansons, rires, bruits d'ambiance, interventions en langue étrangère, etc.

Dans cet article, nous étudions des stratégies de sous-titrage automatique en français inspirées des méthodes de traduction neuronales (Cho *et al.*, 2014 ; Bahdanau *et al.*, 2015 ; Vaswani *et al.*, 2017), la transcription vocale (automatique) jouant le rôle d'énoncé en langue source et le sous-titre (texte et segmentation) jouant le rôle d'énoncé en langue cible. En nous appuyant sur des données réelles, la principale question que nous essayons de traiter concerne la variabilité des genres télévisuels et son impact sur la qualité des sous-titres automatiques. Après avoir mesuré cette variabilité, nous comparons différentes méthodes, inspirées de travaux en traduction automatique, pour la prendre en charge à travers des modules de sous-titrage adaptés au genre. Trois approches sont implémentées et évaluées : une approche fondée sur la spécialisation par genre des taux de compression, une qui spécialise les représentations des énoncés sources (Kobus *et al.*, 2017), une troisième, enfin, qui repose sur l'affinage (Freitag et Al-Onaizan, 2016) des modèles de traduction. Nos évaluations permettent alors de conclure que ces deux dernières méthodes améliorent la qualité des sous-titres produits, y compris lorsqu'elles sont combinées avec d'autres stratégies d'autoapprentissage.

1. En application de la loi n° 2005-102 du 11 février 2005 pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées.

Cet article est organisé comme suit. La section 2 présente la problématique du sous-titrage, en étudiant en particulier sur un corpus la question de la variabilité par genre des sous-titres. Nous présentons à la section 3 les méthodes utilisées pour réaliser l'adaptation du sous-titrage au genre. Les éléments concernant le protocole expérimental, données et métriques, sont rassemblés dans la section 4. La section 5 présente enfin les principaux résultats expérimentaux et diverses analyses complémentaires. Nous concluons à la section 6.

2. Vers le sous-titrage automatique

2.1. *Contraintes sur la forme des sous-titres*

La production automatique de sous-titres implique de traiter un flux audiovisuel pour (a) y associer automatiquement un contenu textuel qui rend fidèlement compte des informations présentes dans le flux audio : principalement une retranscription des contenus parlés, mais également une description des changements de locuteurs et des segments non parlés (musique, évènements sonores) ; (b) afficher ce contenu textuel de manière synchrone avec l'image et le son.

À l'affichage, les sous-titres doivent satisfaire des contraintes spatiales (les tronçons de phrases doivent rentrer dans la largeur du moniteur, sans trop obstruer le champ de vision) et temporelles complexes (le texte doit être approximativement synchronisé avec les paroles ou l'image, et doit rester affiché suffisamment longtemps pour la lecture[2]). Diverses conventions et recommandations régissent également l'affichage et la position à l'écran des évènements sonores, le signalement des changements de locuteurs, etc. Ainsi, chaque sous-titre (ou bloc) doit contenir au plus deux lignes, si possible équilibrées en taille, et les césures entre lignes doivent préserver la cohérence des groupes syntaxiques.

2.2. *Comprimer et simplifier la parole*

La création de sous-titres monolingues nécessite en général d'effectuer une compression, voire une simplification du contenu, de manière à rendre le texte plus abordable pour certains des utilisateurs de sous-titres (monolingues ou bilingues). Ceux-ci sont en effet susceptibles de ne pas maîtriser parfaitement la langue écrite ; il peut s'agir par exemple de personnes ayant une autre langue maternelle, ou bien de personnes sourdes ou malentendantes locutrices de la langue des signes et pour qui l'écrit est assimilable à une langue étrangère (Daelemans *et al.*, 2004). Contrairement aux

2. La *Charte relative à la qualité du sous-titrage à destination des personnes sourdes ou malentendantes* du CSA préconise une fréquence moyenne d'affichage des caractères aux alentours de 12 à 15 car./s, et un écart maximum de 10 secondes entre le discours et le sous-titre correspondant (`https://www.csa.fr/content/download/20043/334122/version/3/file/Chartesoustitrage122011.pdf`, consultée le 31/10/21).

contraintes précédentes, le niveau de simplification ne fait pas l'objet de recommandations très précises, et les attentes des utilisateurs en la matière sont très dépendantes de leur niveau de langue ainsi que du type d'émission regardé.

2.3. *Variabilité des sous-titres d'émissions télévisées*

Un second type de variabilité est directement lié à la stratégie de production des sous-titres, qui peuvent être, selon les émissions, réalisés en direct, c'est-à-dire au fur et à mesure que l'émission se déroule, ou bien durant une étape de postproduction. Nous désignerons ces deux types de sous-titres par `direct` et `stock`. La production en direct est soumise à des contraintes temporelles et obéit à une très forte logique d'efficacité. Elle s'appuie sur des systèmes de transcription automatique et de correction d'orthographe, ce qui conduit en particulier à des sous-titres qui « collent » au plus près au contenu sonore, avec parfois des décrochages dus à l'impossibilité de suivre le rythme des échanges verbaux du direct. En comparaison, la génération de sous-titres en postproduction, qui subit d'autres contraintes (par exemple : économiques), peut prendre une plus grande distance avec le flux audio.

Ces différences sont objectivables et nous les illustrons dans le tableau 1, qui donne quelques résultats d'analyses lexicométriques du corpus d'apprentissage (détaillé en section 4.1, en particulier dans le tableau 3). Une première différence apparaît clairement entre les sous-titres `direct` et `stock` : les premiers sont plus verbeux avec environ 11,2 kmots/heure, alors que le rythme moyen de parole pour le `stock` est seulement environ 9,8 kmots/heure. Toutefois ces moyennes sont lissées sur la durée totale des émissions ; lorsque les périodes de silence sont exclues du calcul, nous observons que la vitesse d'élocution est comparable, voire légèrement supérieure pour `stock` (les émissions de `stock` reposent davantage sur l'image et contiennent davantage de silences). Les sous-titres en `stock` correspondent aussi à un plus grand nombre de phrases, qui sont donc plus courtes (7,7 mots par phrase en moyenne contre 12,7 mots pour les sous-titres `direct`), et probablement plus travaillées.

Au-delà de la stratégie de production, les sous-titres et les transcriptions sont affectés par des caractéristiques propres aux émissions et à leur format. La nature de certains programmes fait qu'une partie des phrases ont une structure assez spécifique (p. ex. énoncé d'une question de culture générale pour les jeux télévisés), et les thèmes abordés de façon récurrente ont une incidence sur le vocabulaire employé. D'autres différences très nettes se lisent dans le tableau 1 : ainsi, on constate que les jeux et séries ont un contenu textuel plus simple que les informations et les émissions politiques. Les journaux correspondent à une vitesse d'élocution intermédiaire, mais font l'objet d'une compression relativement faible, peut-être parce que la parole initiale est déjà formulée efficacement. En comparaison, les émissions politiques partent d'une vitesse d'élocution plus élevée, mais compriment davantage les sous-titres.

Le contexte de production de la parole conduit également à des divergences. Une prise de parole préparée à l'avance tend à se rapprocher du style écrit, tandis que la

Stratégie / Genre	Verbosité (mots/h)	Vitesse d'élocution (mots/h)	Taux de compression	Score BLEU	Flesch Reading Ease
Direct	11 222	12 649	76,0	46,8	77,0
Stock	9 767	13 260	74,6	34,9	87,8
Dessin animé [s]	6 326	9 190	0,95	38,3	88,5
Documentaire [s]	7 766	10 421	0,86	50,6	86,9
Fiction [s]	7 354	10 354	0,86	32,3	88,7
Jeu [s]	8 915	13 887	0,67	28,1	87,2
Journal [d]	10 511	11 593	0,87	58,9	73,5
Magazine [s/d]	11 545	13 825	0,71	36,1	85,0
Politique [s/d]	12 113	13 791	0,69	39,5	78,4
Vulgarisation [s]	10 164	12 012	0,83	51,7	87,1

Tableau 1. *Comparaison d'indices textuels pour plusieurs genres télévisuels. Les catégories « Magazine » et « Politique » contiennent un mélange d'émissions* `direct` *et* `stock`*, alors que les autres sont soit exclusivement* `direct` *[d] soit* `stock` *[s]. La verbosité est mesurée par le nombre de mots prononcés rapporté à la durée totale de l'émission. La vitesse d'élocution est mesurée en rapportant le nombre de mots prononcés à la durée de parole effective (sans compter les périodes de silence). Le taux de compression (CpR) est le ratio entre le nombre de mots dans les transcriptions automatiques et les sous-titres : un fort CpR implique un faible niveau de compression. BLEU compare superficiellement les transcriptions et les sous-titres (une valeur élevée dénote une forte similarité). Flesch Reading Ease (FRE) est une mesure de simplicité (une valeur élevée dénote un texte plus simple). Ces métriques sont décrites en section 4.2.*

parole spontanée suit des règles de syntaxe spécifiques à l'expression orale. Les émissions telles que les documentaires, les programmes de vulgarisation et les journaux, qui sont rédigées à l'avance, montrent une proximité entre la transcription et les sous-titres (score BLEU au-dessus de 50). Inversement, les jeux et les fictions qui sont constitués de parole spontanée (émulée, dans le cas des fictions) affichent une dissimilitude forte entre la transcription automatique et les sous-titres. Enfin, il faut prendre en compte la qualité des prédictions proposées par le système de reconnaissance vocale, inégale selon les émissions, qui se répercute sur la qualité des sous-titres engendrés en aval (section 4.1.4). La reconnaissance est notamment affectée par le débit de parole, la clarté de la prononciation, les dialogues avec recouvrement, et généralement la présence de bruits parasites.

Afin de prendre en considération cette variabilité, nous avons classé les programmes selon des genres identifiés par des experts des métiers du domaine, qui correspondent aux catégories utilisées par les fournisseurs de données : dessin animé, documentaire, fiction, jeu, journal, magazine, politique, et vulgarisation.

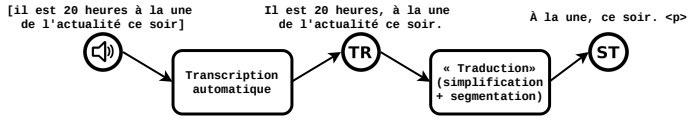


Figure 1. *Architecture globale pour le sous-titrage automatique. Nos expériences se concentrent sur la tâche de « traduction » de la transcription vers les sous-titres segmentés.*

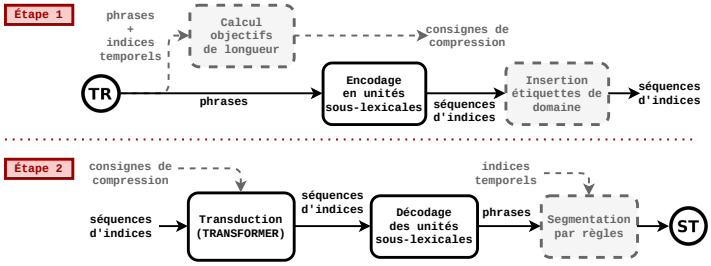


Figure 2. *Architecture détaillée pour la transduction de la transcription vers les sous-titres segmentés. En noir sont représentées les étapes partagées par tous les systèmes ; en gris et en pointillé celles réalisées seulement par certains systèmes : le calcul de consignes de compression (introduites dans le* TRANSFORMER*) est spécifique aux systèmes fondés sur le contrôle de longueur (sections 3.3.2 et 3.4.2) ; l'insertion d'étiquettes de domaine correspond à l'une des méthodes d'adaptation au genre (section 3.4.1) ; la segmentation du contenu textuel des sous-titres par un module à règles intervient uniquement dans les systèmes de base qui n'effectuent pas conjointement simplification et segmentation (section 3.3.1). Les indices temporels sont les périodes de prononciation des phrases (identifiées par l'outil de reconnaissance) ; les consignes de compressions sont des objectifs de longueur cible, modulés pour suivre un certain taux de compression (CpR), ou une fréquence d'affichage (CPS).*

3. Méthodes et modèles pour le sous-titrage

Nous décrivons dans cette section les méthodes utilisées pour construire nos systèmes de sous-titrage, dans leur version de base comme dans leur version adaptée. La figure 1 présente une vue générale de l'architecture en cascade commune à ces systèmes. La figure 2 apporte une vue détaillée sur la phase de transduction de la transcription vers les sous-titres segmentés.

3.1. *Architectures pour le sous-titrage automatique*

Un choix précoce de cette étude a été d'utiliser une architecture en cascade qui découple une phase de retranscription verbatim du contenu audio, de la phase d'élaboration et de génération des sous-titres. Cette décision est notamment motivée par la possibilité de disposer d'un outil de transcription vocale au meilleur état de l'art pour le français, dont l'utilisation nous a permis de nous focaliser sur la tâche de simplification et de compression. Récemment, la traduction automatique de paroles (Bérard *et al.*, 2016 ; Karakanta *et al.*, 2020a), ainsi que d'autres applications (Ghannay *et al.*, 2018) ont vu la progression d'architectures bout en bout (aussi appelées directes), qui s'abstiennent d'utiliser une représentation symbolique intermédiaire. Il convient néanmoins de noter que l'approche en cascade obtient encore en général les meilleurs résultats (Anastasopoulos *et al.*, 2021), surtout si des données indépendantes pour la transcription et la traduction sont utilisées (Etchegoyhen *et al.*, 2022).

Dans notre approche, la production de sous-titres repose sur la métaphore de la traduction, et s'appuie, à l'instar de nombreux travaux récents (Zhang *et al.*, 2017 ; Zhang et Lapata, 2017 ; Matusov *et al.*, 2019 ; Karakanta *et al.*, 2020a), sur des architectures neuronales encodeur décodeur. Selon cette métaphore, les échantillons de parole (pour les systèmes de sous-titrage de bout en bout) ou leur retranscription automatique (pour les systèmes en cascade) sont encodés sous la forme d'une suite de vecteurs numériques, qui sont ensuite décodés pour engendrer de proche en proche les séquences de mots correspondant aux sous-titres.

La tâche de sous-titrage demande non seulement de produire du texte, mais également d'engendrer des directives pour son affichage à l'écran et la resynchronisation avec la piste audio. Les indications de synchronisation portent sur des blocs entiers d'une ou deux lignes et correspondent à des indices temporels de début et de fin d'affichage du bloc : elles sont calculées dans notre processus à partir des périodes de parole identifiées par l'outil de transcription (en permettant à l'affichage de durer quelques secondes supplémentaires pendant les éventuels silences). Les directives d'affichage sont matérialisées par des balises qui sont insérées dans le flux textuel et signalent les fins de ligne (
) et les fins de blocs (<p>). Nous envisageons deux méthodes distinctes pour prédire la position des balises : la première l'envisage comme une étape séparée du calcul du sous-titre et repose sur un module à base de règles détaillé ci-dessous, la seconde méthode est une méthode intégrée qui permet d'engendrer simultanément le contenu linguistique et les marques de segmentation. Pour réaliser cet apprentissage de bout en bout, les balises de segmentation sont directement insérées dans le flux textuel lors de l'apprentissage et du décodage. Le système est alors libre de les produire comme il produirait n'importe quelle autre unité de son vocabulaire de sortie. Une illustration de ces sorties enrichies est donnée au tableau 2. Un dernier composant de notre architecture rassemble le contenu textuel accompagné de consignes de segmentation et de synchronisation temporelle en un fichier au format `ttml` qui peut être directement utilisé dans des systèmes de visualisation de vidéos.

3.2. *Un modèle encodeur décodeur à base de* TRANSFORMER

Nous nous appuyons sur l'architecture TRANSFORMER de Vaswani *et al.* (2017), qui constitue aujourd'hui l'état de l'art pour la traduction automatique comme pour de nombreuses autres tâches de traitement automatique du texte et de la parole. Nous avons réimplémenté cette architecture en Python. Les hyperparamètres ont été choisis en partie par imitation de la littérature (l'implémentation originale du TRANSFORMER notamment), et en partie par expérimentation. Les variations de dimensionnement ont été testées sur un corpus de développement correspondant à dix heures d'émissions (100 000 mots transcrits) échantillonnées aléatoirement dans nos données. La version qui est utilisée dans nos expérimentations utilise les paramètres suivants :

– dimension des représentations internes et des plongements lexicaux $d_{\mathrm{m}} = 512$;

– dimension du perceptron multicouche $d_{\mathrm{ff}} = 2\,048$;

– nombre de têtes d'attention $h = 8$;

– nombre de couches pour l'encodeur et le décodeur $N = 6$.

L'optimisation des paramètres du modèle est faite avec *Adam* (Kingma et Ba, 2015) en utilisant les paramètres suivants : $\beta_1 = 0{,}9$, $\beta_2 = 0{,}98$, $eps = 10^{-9}$. Nous avons également repris la méthode de variation du taux d'apprentissage proposée par Vaswani *et al.* (2017), en fixant le nombre d'étapes d'échauffement à 4 000.

Afin de pouvoir traiter d'un vocabulaire ouvert pour le sous-titrage, les phrases (aussi bien en source qu'en cible) sont segmentées en unités sous-lexicales avec *SentencePiece* (Kudo et Richardson, 2018), en prenant un modèle unigramme et un vocabulaire de 16 000 unités.

3.3. *Contrôle de la segmentation : règles et contraintes*

3.3.1. *Les règles de segmentation pour les systèmes de base*

Nous prenons comme architecture de référence un système qui utilise telle quelle la sortie de l'outil de reconnaissance vocale et qui calcule les balises de segmentation à l'aide d'un module à règles utilisant les heuristiques suivantes :

– une fin de phrase implique nécessairement un changement de bloc ;

– chaque bloc contient au plus deux lignes ;

– les lignes sont construites successivement en agrégeant progressivement les mots et les ponctuations ;

– si la ligne en cours dépasse en longueur une borne inférieure et se termine par une virgule, cela déclenche la fin de ligne (ou le cas échéant la fin de bloc) ;

– si la ligne en cours s'apprête à dépasser en longueur une borne supérieure[3], cela déclenche la fin de ligne (ou le cas échéant la fin de bloc).

Une seconde architecture insère une étape de compression des transcriptions réalisée par un modèle TRANSFORMER simple avant la segmentation par règles.

3.3.2. *Contrôler la verbosité du décodeur*

La version de base du TRANSFORMER utilise un *encodage positionnel* qui permet de différentier les positions d'entrée de l'encodeur. Cet encodage des positions est combiné avec le plongement de chaque mot de l'entrée (dans la partie encodeur) ou de l'amorce de phrase produite (dans la partie décodeur) selon :

$$\mathrm{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10\,000^{2i/d_m}}\right), \quad \mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10\,000^{2i/d_m}}\right), \qquad [1]$$

où pos est la position du mot dans la phrase, et $2i$ (resp. $2i+1$) correspond aux dimensions paires (resp. impaires) de l'encodage.

Pour contrôler de manière explicite la longueur des sous-titres produits par certains de nos modèles, nous avons également réimplémenté les variantes LRPE et LDPE, proposées par Takase et Okazaki (2019) et utilisées dans un cadre de sous-titrage également par Lakew *et al.* (2019). Ces encodages intègrent une consigne sur la longueur l du texte à produire. Cette contrainte peut être exprimée comme un ratio de compression entre entrée et sortie (LRPE) ou bien encore comme une différence relative entre la position courante pos et la fin attendue de la sortie (LDPE). Formellement, ces contraintes prennent la forme suivante :

$$\mathrm{LRPE}_{(pos,l,2i)} = \sin\left(\frac{pos}{l^{2i/d_m}}\right), \qquad \mathrm{LRPE}_{(pos,l,2i+1)} = \cos\left(\frac{pos}{l^{2i/d_m}}\right), \qquad [2]$$

$$\mathrm{LDPE}_{(pos,l,2i)} = \sin\left(\frac{l-pos}{10\,000^{2i/d_m}}\right), \quad \mathrm{LDPE}_{(pos,l,2i+1)} = \cos\left(\frac{l-pos}{10\,000^{2i/d_m}}\right). \qquad [3]$$

l est égal à la longueur de la séquence cible de référence pendant la période d'entraînement, mais est fixé par l'utilisateur pendant la période de test. LRPE caractérise à la fois la position courante pos et la longueur totale souhaitée l, tandis que LDPE exprime une distance à l'objectif de longueur.

Dans nos expériences, nous avons modulé les objectifs de longueur afin de contraindre les modèles LRPE et LDPE à engendrer des phrases respectant soit un taux de compression constant CpR (auquel cas l est égale à la longueur de la phrase d'entrée multipliée par CpR), soit une fréquence d'affichage des caractères constante CPS (auquel cas l est égale à la durée allouée à l'affichage des tronçons de la phrase multipliée par CPS).

3. L'intervalle d'acceptabilité de longueur est entre 13 et 32 caractères ; les bornes ont été choisies de manière à refléter la distribution de longueur dans les véritables sous-titres.

3.4. *Méthodes d'adaptation au genre*

Comme exposé en préambule, notre principal objectif est d'étudier l'apport de méthodes d'adaptation au domaine du système de sous-titrage. La question de l'adaptation au domaine est une question largement traitée en traduction automatique, que la visée soit d'adapter à un unique domaine cible (Chu et Wang, 2018 ; Saunders, 2021), ou bien de construire des systèmes multidomaines (Pham *et al.*, 2021). Nos travaux utilisent trois méthodes pour réaliser cette adaptation : l'utilisation d'étiquettes de domaine (Kobus *et al.*, 2017), l'utilisation d'objectifs de longueur par domaine, et l'affinage des paramètres (Luong et Manning, 2015 ; Freitag et Al-Onaizan, 2016). Les autres méthodes présentes dans la littérature reposent principalement soit sur des techniques d'apprentissage adverse qui neutralisent les différences entre genres au sein des représentations internes, soit sur l'affinage d'une sous-partie seulement du modèle, dont les paramètres sont adaptés à un domaine (Bapna et Firat, 2019).

3.4.1. *Utilisation d'étiquettes de domaine*

Cette méthode présente l'avantage d'être à la fois très simple, et relativement efficace dans de nombreuses situations expérimentales. Elle consiste à augmenter les représentations des segments sources par des informations relatives au domaine. Deux manières de procéder sont généralement considérées : soit insérer une balise de domaine en première position de chaque segment source, soit injecter cette information plus directement dans les représentations de chaque unité source. Nous avons implémenté cette méthode en utilisant la première de ces deux approches et en distinguant, en plus des deux grands types d'émissions (`stock` et `direct`), les huit domaines correspondant aux genres télévisuels (section 2.3). Chaque segment source est donc préfixé par deux symboles particuliers, qui permettent de spécialiser par domaine les représentations calculées par l'encodeur (tableau 2).

3.4.2. *Objectifs de longueur par domaine*

Les modèles LRPE et LDPE attendent une consigne sur la longueur de la phrase à engendrer. Disposant de la longueur de la phrase initiale, ainsi que de la période d'affichage disponible (donnée par l'outil de transcription automatique), nous avons mis en place une modulation de l'objectif de longueur de manière à ce qu'il corresponde soit à un taux de compression CpR, soit à une fréquence d'affichage CPS (section 3.3.2). En fixant des valeurs à suivre CpR et CPS spécifiques pour chaque domaine (choisies pour correspondre aux valeurs observées en pratique pour les émissions du corpus), une information sur le flux de sortie attendu est fournie au système.

3.4.3. *Affinage par genre*

L'affinage consiste à préentraîner un système générique de sous-titrage avec un ensemble divers de segments parallèles, représentant un mélange de tous les genres télévisuels. Dans un second temps, l'apprentissage se poursuit en réduisant les données au seul genre d'intérêt. Les paramètres résultants conduisent souvent à de meilleures performances que les systèmes utilisant des balises, mais présentent l'inconvénient

de conduire à l'apprentissage d'un modèle distinct pour chaque genre, au lieu d'un modèle unique capable de traiter tous les types d'émissions.

Dans nos expériences, nous avons utilisé les mêmes huit genres que pour les systèmes à base de balises, et implémenté l'affinage de la manière suivante : nous sauvegardons les paramètres obtenus à l'issue de l'apprentissage[4] d'un modèle TRANSFORMER classique et reprenons l'entraînement en restreignant les données au genre d'intérêt et en réduisant le taux d'apprentissage par un facteur 20. Nous appliquons les mêmes règles de convergence que pour le modèle de base.

4. Protocole et données expérimentales

4.1. *Corpus*

Nous avons à notre disposition un ensemble de vidéos, assorties de fichiers de sous-titres professionnels, correspondant à des programmes télévisés récemment diffusés en France. Le panel d'émissions auquel nous avons eu accès a été choisi de manière à représenter diverses catégories (dessin animé, documentaire, fiction, jeu, journal, magazine, politique, vulgarisation). Ces données ne peuvent cependant pas être partagées du fait des droits associés aux émissions[5].

4.1.1. *Transcription automatique*

Les instances de programmes collectées ont été transcrites automatiquement (mot pour mot) en utilisant un système de reconnaissance vocale préexistant. Ce système comporte un modèle acoustique hybride HMM-TDNN (*Hidden Markov Model*, *Time Delay Neural Network*) et un modèle de langue standard 4-gramme, entraînés sur de grandes quantités de données. Il produit des transcriptions automatiques qui sont segmentées en phrases selon les tours de parole, identifiés après un processus utilisant des modèles de mélange gaussien et un algorithme de regroupement des segments de parole. Les transcriptions sont aussi ponctuées automatiquement par un modèle 4-gramme, et elles respectent les principales règles typographiques (majuscule en début de phrase, pour les noms propres, etc.). Ce système délivre des performances à l'état de l'art pour la transcription du français, avec un taux d'erreur de mots (*Word Error Rate*) variant entre 10 et 40 % environ selon les émissions du corpus[6] (tableau 6) : les meilleurs scores correspondant à de la parole préparée (par exemple dans les journaux télévisés), et les moins bons à de la parole spontanée ou peu rédigée, potentiellement bruitée par l'environnement (par exemple dans les jeux télévisés).

4. C'est-à-dire après convergence de l'apprentissage.

5. Les données appartiennent au diffuseur pour la partie sous-titres, la propriété des enregistrements étant répartie sur les multiples acteurs de la chaîne de production.

6. Ces taux d'erreur ont été calculés par rapport à une transcription humaine de référence, considérée comme une version « idéale » de la transcription automatique.

TR	<direct> <journal> Tout au long de la journée, des orages violents, de fortes pluies et quelles conséquences pour la population, faisons le point ce soir sur cette soudaine montée des eaux et sur les vents violents qui ont soufflé cet après-midi, dans les Bouches-du-Rhône à Marignane et je vous le disais sur la Côte-d'Azur à à Valbonne Vence ou encore à Nice, Alexandre Christophe Larocca.
ST	Des orages violents, de fortes pluies et quelles conséquences pour <p> la population ? <p> Faisons le point sur cette soudaine montée des eaux et sur les vents <p> violents qui ont soufflé cet après-midi... <p>

Tableau 2. *Exemple d'apprentissage constitué d'un segment transcrit automatiquement TR (source) et d'un segment sous-titres ST (cible) produit par un sous-titreur professionnel. Les balises dans ST représentent la segmentation à l'affichage : `<br>` pour un saut de ligne au sein d'un bloc, `<p>` pour une fin de bloc (et changement d'écran). Les balises au début de TR indiquent la stratégie de sous-titrage et le genre de l'émission : ici `direct` et « journal » (de telles balises ne sont présentes dans les exemples que pour les systèmes d'adaptation avec étiquettes).*

4.1.2. *Alignement et « parallélisation » du corpus*

Le texte de la transcription ainsi obtenu a alors été aligné avec celui des sous-titres, afin de pouvoir reconstituer des paires de segments parallèles qui sont nécessaires à l'apprentissage et l'évaluation automatique du système. Cet alignement est principalement fondé sur la comparaison caractère par caractère des deux segments textuels sur la base des opérations d'édition usuelles (insertion, substitution, délétion, copie), complétées par diverses heuristiques pour prendre en compte, par exemple, les différences de capitalisation. Nous avons décidé d'utiliser la segmentation calculée par le système de transcription automatique comme base de l'alignement. Ces segments sont assez longs (environ quarante mots en moyenne) et correspondent généralement à plusieurs tronçons de sous-titres (tableau 2). Les autres informations délivrées par le système de transcription (locuteurs, pauses, etc.) n'ont pas été utilisées.

À l'issue de l'alignement, une partie des phrases transcrites n'étaient appareillée avec aucun sous-titre, soit parce que le texte de sous-titres correspondant était aligné avec le segment précédent ou suivant, soit parce que la phrase avait simplement été coupée lors du sous-titrage. Pour l'apprentissage des modèles de tels segments ont été filtrés du corpus. De même, les paires présentant une trop grande dissimilitude[7] ont été écartées avant l'apprentissage. Le corpus contient en tout environ 780 000 paires de phrases, soit environ 30 millions de mots transcrits, et près de 2 900 heures de vidéos (voir le tableau 3 pour la répartition par type d'émission). Toutefois, après filtrage selon la qualité de l'alignement, seulement 482 000 paires ont été utilisées pour l'apprentissage des modèles.

7. Si la distance d'édition (Levenshtein) entre le segment transcrit et le segment sous-titres excède 40 %.

Domaine	(h)	Blocs ST	Segments	Mots TR	Mots ST	%
Dessin animé [s]	8	7 k	2 k	0,05 M	0,04 M	0,2
Documentaire [s]	162	145 k	49 k	1,26 M	1,04 M	4,8
Fiction [s]	143	134 k	46 k	1,05 M	0,86 M	4,0
Jeu [s]	586	549 k	190 k	5,22 M	3,39 M	15,7
Journal [d]	587	681 k	157 k	6,16 M	5,36 M	24,9
Magazine [s/d]	1 285	1 293 k	317 k	14,83 M	9,98 M	46,3
Politique [s/d]	104	104 k	19 k	1,26 M	0,85 M	3,9
Vulgarisation [s]	4	4 k	1 k	0,04 M	0,03 M	0,1
Direct	1 217	1 284 k	272 k	13,65 M	10,06 M	46,7
Stock	1 662	1 633 k	509 k	16,23 M	11,49 M	53,3
Tout	2 878	2 917 k	780 k	29,88 M	21,55 M	100

Tableau 3. *Distribution par domaine des données du corpus d'apprentissage. Les pourcentages sont calculés par rapport au nombre de mots dans les sous-titres.*

4.1.3. *Sous-titres complémentaires et rétrotraductions*

La production de données d'apprentissage est un processus coûteux qui implique l'exploitation d'un système de reconnaissance vocale. En complément, nous avons également choisi d'utiliser des données pseudo-parallèles artificielles, qui sont directement dérivées des fichiers de sous-titres sans qu'il soit besoin de traiter la piste son. Nous nous inspirons des méthodes de rétrotraduction qui ont fait leurs preuves en traduction automatique neuronale (Sennrich *et al.*, 2016 ; Burlot et Yvon, 2018 ; Edunov *et al.*, 2018) et qui consistent à « inverser » le processus de traduction de manière à construire des données associant une transcription artificielle avec un sous-titre correct. Cette méthode permet en particulier d'améliorer l'apprentissage du décodeur du système de traduction. La rétrotraduction présente l'avantage (par rapport à d'autres méthodes de synthèse de données) de ne pas recourir à des données extérieures, et de conserver des phrases cibles syntaxiquement correctes, dont le genre télévisuel est connu.

La génération des pseudo-transcriptions est mise en œuvre de la manière suivante. En exploitant l'intégralité des données parallèles disponibles, nous avons entraîné un système encodeur décodeur qui inverse le processus de sous-titrage et produit des pseudo-transcriptions à partir des sous-titres. Ce système utilise la même architecture TRANSFORMER (nombre de couches, dimensions internes) que les systèmes de sous-titrage (section 3). Comme le système de reconnaissance de parole tend à produire de longs segments par rapport à ceux présents dans le texte des sous-titres (une phrase transcrite correspondant généralement à plusieurs phrases de sous-titres ; voir le tableau 2), nous concaténons aléatoirement les phrases de sous-titres en de plus longues séquences préalablement à la rétrotraduction. Le nombre de phrases à ras-

ST	Surtout pas Benjamin. <p> Mais je vous ai vus. Oui, tu nous as vus <p> en train de parler discrètement, à l'écart, peut-être, <p> mais pas parce qu'il y avait une histoire entre nous. <p>
TR̂	Surtout pas, Benjamin, mais je vous ai vu, oui, oui, oui, tu nous as vus en train de parler discrètement à l'écart, peut-être, mais pas parce qu'il y avait une histoire entre nous.

Tableau 4. *Exemple de pseudo-transcription obtenue par « rétrotraduction »*

Domaine	**(h)**	**Blocs ST**	**Mots ST**	**%**
Dessin animé [s]	9	8 k	43 k	0,5
Documentaire [s]	92	78 k	563 k	6,4
Fiction [s]	57	55 k	338 k	3,8
Jeu [s]	54	56 k	309 k	3,5
Journal [d]	185	176 k	1 425 k	16,1
Magazine [s/d]	637	602 k	4 623 k	52,2
Politique [s/d]	14	12 k	107 k	1,2
Vulgarisation [s]	10	10 k	77 k	0,9
Enseignement [s]	221	189 k	1 364 k	15,4
Prgm court [s]	2	1 k	12 k	0,1
Direct	782	737 k	5 758 k	65,0
Stock	496	452 k	3 102 k	35,0
Tout	1 278	1 189 k	8 860 k	100

Tableau 5. *Distribution par domaine des données rétrotraduites. Les pourcentages sont calculés par rapport au nombre de mots dans les sous-titres.*

sembler est échantillonné selon une loi normale centrée sur 3[8]. Ce système obtient un score BLEU[9] de 64,7 sur les données de test en comparant les énoncés artificiellement bruités aux sous-titres de référence. Cela suggère que les pseudo-transcriptions artificielles restent très proches des sous-titres de référence, et sont donc considérablement moins bruitées que les transcriptions réelles. Un exemple de pseudo-transcription est donné dans le tableau 4.

Pour nos expériences, nous n'avons pas rétrotraduit tous les sous-titres disponibles, mais avons effectué une sélection sur la base du genre des émissions. La répartition par genre d'émissions des données rétrotraduites est dans le tableau 5.

8. Valeur qui correspond au ratio observé en pratique entre les segments transcrits automatiquement et les phrases de sous-titres.

9. Les métriques sont décrites en détail à la section 4.2.

Domaine	(h)	Segments	Mots TR	Mots ST	%	WER
Fiction [s]	1,9	592	14 k	12 k	8,0	32,5
Jeu [s]	1,3	371	11 k	7 k	4,7	38,8
Journal [d]	3,5	901	37 k	32 k	21,9	11,5
Magazine [s/d]	8,5	1 998	90 k	57 k	39,2	22,3
Politique [s/d]	4,9	854	58 k	38 k	26,3	15,2
Direct	11,5	2 283	133 k	90 k	62,0	14,4
Stock	8,6	2 433	78 k	55 k	38,0	30,2
Tout	20,1	4 716	211 k	145 k	100	21,7

Tableau 6. *Distribution et taux d'erreur de mots (WER) par domaine des données du corpus de test. Les pourcentages sont calculés par rapport au nombre de mots dans les sous-titres.*

4.1.4. *Corpus de test*

Pour nos tests nous avons sélectionné au hasard vingt-six vidéos d'émissions, représentatives des programmes traités (les titres d'émissions du corpus de test font partie de ceux présents dans le corpus d'apprentissage, et il y a chevauchement des périodes de diffusion), la distribution des genres, donnée par le tableau 6, n'est toutefois pas identique à celle présente dans le corpus d'apprentissage (tableau 3). Le tableau 6 donne également le taux d'erreur de mots (WER) du système de transcription, qui agrège les erreurs correspondant à des ajouts, des omissions, et des substitutions, moyenné par catégorie d'émissions[10]. Cette mesure permet d'apprécier la qualité générale de l'entrée qui sera traitée par le système de sous-titrage. La durée cumulée de l'ensemble est d'environ vingt heures. Les segments correspondant aux sous-titres de référence ont été constitués par alignement automatique avec les phrases de la transcription automatique, de la même façon que pour le reste du corpus (section 4.1.2)[11].

4.2. *Métriques d'évaluation*

Nous avons suivi les précédents de la littérature (Matusov *et al.*, 2019 ; Karakanta *et al.*, 2020b ; Karakanta *et al.*, 2020a) concernant le choix des métriques pour l'évaluation de la segmentation des sous-titres, et de la conformité aux normes superficielles. Concernant la qualité des phrases produites, nous avons utilisé des métriques standard pour la tâche de simplification de texte. Enfin, nous avons mis en place une mesure de la précision de la longueur produite pour les systèmes reposant sur le contrôle de verbosité.

10. Le WER a été calculé pour chaque émission test en comparant avec une transcription humaine de référence.

11. Une partie des phrases transcrites (représentant environ 6 % des mots) n'ont pas pu être alignées avec les phrases des sous-titres ; nous avons décidé de les écarter pour l'évaluation.

4.2.1. *Qualité et simplicité des phrases*

BLEU (Papineni *et al.*, 2002) est une métrique standard pour la traduction automatique. Xu *et al.* (2016) ont montré que dans le cas de la simplification, BLEU corrèle les jugements humains pour le sens et la grammaticalité, mais pas pour la simplicité. Nous utilisons l'implémentation *SacreBLEU* de Post (2018).

SARI (Xu *et al.*, 2016) est une métrique pour la simplification de texte, qui compare les opérations d'édition (insertion, copie, suppression de n-grammes) observées entre l'entrée et la sortie, avec celles observées entre l'entrée et les références[12]. Nous utilisons l'implémentation de la bibliothèque EASSE (Alva-Manchego *et al.*, 2019).

Flesch Reading Ease (FRE) (Flesch, 1948) évalue la lisibilité, en se fondant sur le nombre moyen de mots par phrase et sur le nombre moyen de syllabes par mot. Nous reprenons la formule adaptée au français par Kandel et Moles (1958).

4.2.2. *Respect des normes superficielles de sous-titrage*

L'affichage de sous-titres nécessite des informations précisant certains aspects de la présentation à l'écran, tels que la segmentation du texte en blocs et en lignes, le temps d'apparition de chaque bloc, la couleur des caractères, ou encore le positionnement horizontal des lignes. Ce formatage doit se conformer à des codes et des normes qui assurent la lisibilité des sous-titres.

Le nombre de caractères par ligne (CPL) et le nombre de caractères par seconde (CPS, calculé à partir de la durée d'affichage des blocs) sont en particulier soumis à des recommandations. Pour rendre compte du respect de ces contraintes, nous calculons la proportion de lignes dont la longueur dépasse 36 car., CPL_{36+}, ainsi que la proportion de blocs qui dépassent une fréquence d'affichage de 15 car./s, CPS_{15+} (ces seuils correspondent aux limites préconisées en France).

4.2.3. *Qualité de la segmentation des sous-titres*

Nous reprenons deux métriques proposées respectivement par Matusov *et al.* (2019) et Karakanta *et al.* (2020a) pour évaluer la segmentation des sous-titres :

– BLEU, calculé en conservant les balises de fin de ligne et de fin de bloc dans les prédictions et les références ; cette mesure, notée $BLEU_{br}$, permet d'évaluer indirectement le positionnement des balises de sous-titrage dans les phrases ;

– TER (Snover *et al.*, 2006), calculé entre la sortie du système et la référence en masquant tous les mots à l'exception des balises de segmentation `<p>` et `<br>`.

12. N'ayant qu'une seule version de sous-titres pour les émissions, nous ne mesurons SARI qu'avec une référence.

4.2.4. *Précision du contrôle de longueur*

Pour estimer la précision du contrôle de longueur (opéré par les méthodes LRPE et LDPE, voir section 3.3.2), nous avons choisi de calculer l'erreur absolue moyenne (EAM) des taux de compression obtenus par rapport aux taux de compression visés :

$$\mathrm{EAM} = \frac{1}{n}\sum_{i=1}^{n} |\hat{r}_i - r_i|, \qquad [4]$$

où n est la taille de l'ensemble de test, et $\hat{r}_i$ et r_i sont respectivement le taux de compression obtenu et le taux de compression visé pour la i-ème phrase.

L'erreur absolue (EA) $|\hat{r} - r|$ peut aussi être vue comme la différence entre la longueur produite et la longueur visée $|l_{\hat{y}} - r \times l_x|$ rapportée à la longueur source l_x. Pour compléter nos métriques, nous avons évalué la proportion d'instances pour lesquelles l'erreur absolue est inférieure à 10 %.

5. Résultats

5.1. *Comparaison aux systèmes de base*

Comme indiqué à la section 3.3.1, la première approche mise en place pour insérer les balises
 et <p> dans les sous-titres consiste à utiliser un module séparé de segmentation par règles. Cette méthode a été testée d'une part, avec la sortie d'un modèle de simplification TRANSFORMER (appris sur des données pour lesquelles les balises
 et <p> avaient été filtrées) et, d'autre part, directement avec les segments transcrits automatiquement (résultant en un système qui se contente de segmenter les transcriptions automatiques). Le tableau 9 montre que l'ajout de l'étape de simplification entraîne un gain considérable pour toutes les métriques automatiques (le plus grand écart entre deux itérations de système pour les métriques BLEU et SARI).

L'intégration des balises de segmentation dans le côté cible des données d'apprentissage ne change que très peu les scores BLEU et SARI : la réalisation conjointe de la simplification et de la segmentation n'affecte pas la qualité de la simplification. Concernant l'apport pour la qualité de la segmentation, une amélioration peut être notée pour la métrique $\mathrm{BLEU_{br}}$; $\mathrm{TER_{br}}$ en revanche ne semble pas très sensible.

Contrairement aux modifications précédentes, l'ajout *via* LRPE ou LDPE de contraintes (non adaptées au genre) sur la longueur des sous-titres produits ne permet pas, dans l'ensemble, d'améliorer les métriques automatiques. La précision du contrôle de longueur en elle-même est relative, puisque la différence entre la consigne de longueur et sa réalisation représente en moyenne entre 16 et 20 % de la longueur source (EAM) (tableau 7) ; LRPE et LDPE sont ici comparables du point de vue de l'effectivité de ce contrôle. Pour ce qui est de la qualité des phrases engendrées (mesurée par SARI, BLEU, $\mathrm{BLEU_{br}}$), LRPE est supérieur à LDPE, et la poursuite d'une fréquence de caractères constante semble préférable à l'application d'un unique taux de

Systèmes	EAM	EA < 10 %
+ BS + $\text{LRPE}_{\text{CpR = 0,75}}$	17,2 %	13,4 %
+ BS + $\text{LRPE}_{\text{CPS = 14,5}}$	21,4 %	25,2 %
+ BS + $\text{LDPE}_{\text{CpR = 0,75}}$	18,1 %	12,0 %
+ BS + $\text{LDPE}_{\text{CPS = 14,5}}$	21,9 %	24,3 %

Tableau 7. *Résultats de l'évaluation du contrôle de longueur des modèles LRPE et LDPE (moyennés sur le groupe d'émissions de test)*

compression (ce qui paraît effectivement plus proche de ce que ferait un sous-titreur humain). Nous notons néanmoins un meilleur respect de la norme sur la fréquence d'affichage des caractères (CPS_{15+}), en particulier lorsque l'objectif de longueur est modulé pour suivre une fréquence constante, cas dans lequel le score TER_{br} est aussi plus bas (ce qui indiquerait un meilleur positionnement des coupures dans les phrases).

5.2. *Comparaison selon la stratégie de sous-titrage*

Un des axes d'analyse que nous considérons est l'évaluation sur des émissions appartenant aux catégories `direct` ou `stock`, qui correspondent respectivement aux stratégies de sous-titrage en simultané et en différé. Les principaux résultats sont détaillés dans le tableau 8. Les variations entre systèmes sont similaires au sein de chaque évaluation : les méthodes d'adaptation au genre, par balisage ou affinage, obtiennent des scores relativement meilleurs. De même l'utilisation de données rétrotraduites est toujours bénéfique, à l'exception du cas où elle est combinée avec l'utilisation de balises de genre, pour l'évaluation sur les émissions `stock`. Cette différence est potentiellement liée à la distribution des programmes dans les données complémentaires. Comme le montre le tableau 5, la proportion d'émissions `stock` est plus faible dans le corpus rétrotraduit que dans le corpus d'apprentissage (35 % au lieu de 53 %). En outre, nous avons introduit parmi les émissions `stock` un nouveau genre télévisuel (« enseignement », correspondant à des vidéos de cours de primaire, collège ou lycée), absent du corpus régulier, ce qui pourrait causer une inadéquation du système avec les exemples portant la balise <stock> à l'évaluation.

D'autres différences notables concernent la verbosité et la proximité par rapport à la référence : pour les sous-titres produits automatiquement pour les émissions `direct`, la norme sur le nombre de caractères par seconde est plus régulièrement dépassée (pour plus de 70 % des sous-titres), et la distance au texte de référence mesurée par BLEU et SARI est dans l'ensemble plus petite (ce qui se comprend dans la mesure où les conditions du direct contraignent les sous-titres de référence à être proches de la transcription dans l'absolu). Pour autant la qualité de segmentation représentée par TER_{br} paraît être meilleure pour l'évaluation sur `stock`.

Systèmes	$BLEU_{br}$	BLEU	SARI	TER_{br}	CPL_{36+}	CPS_{15+}
Évaluation sur les émissions de test de type `direct`						
+ BS	37,6	49,8	56,6	0,37	3,2	72,5
+ BS + RT	38,2	50,5	57,2	0,38	2,1	72,7
+ BS + BG	37,8	50,3	56,8	0,35	2,0	72,1
+ BS + RT + BG	39,0	51,5	57,9	0,34	1,7	69,7
+ BS + AF*	38,6	50,7	57,3	0,35	2,5	71,2
+ BS + RT + AF*	39,1	51,3	57,9	0,35	1,6	71,6
Évaluation sur les émissions de test de type `stock`						
+ BS	32,9	38,2	51,7	0,32	2,6	45,0
+ BS + RT	34,0	39,2	52,5	0,32	1,7	44,0
+ BS + BG	32,8	38,6	52,1	0,32	2,1	42,6
+ BS + RT + BG	33,3	38,5	52,5	0,31	1,7	41,0
+ BS + AF*	33,4	38,7	52,6	0,31	2,3	41,7
+ BS + RT + AF*	34,3	39,5	53,3	0,31	1,5	40,9

Tableau 8. *Résultats de l'évaluation de différents modèles sur les émissions* `direct` *ou* `stock`*. BS, RT et BG dénotent respectivement l'usage de balises de segmentation, de rétrotraduction, et de balises de genre. AF* indique que chaque émission a été traitée avec le modèle affiné sur le même genre (fiction, magazine, politique, etc.).*

5.3. *Effets de l'adaptation au genre télévisuel*

Nous évaluons trois stratégies pour l'adaptation au genre, détaillées à la section 3.4 : l'introduction de balises de genre, l'introduction de consignes de longueur spécifiques pour chaque type d'émission, enfin des stratégies d'affinage de systèmes. Les principaux résultats expérimentaux sont dans les tableaux 9 et 10. L'adaptation par balisage produit un effet positif modéré mais cohérent pour toutes les métriques (pour BLEU environ 0,5 point en moyenne). À l'inverse, le contrôle des longueurs produit des résultats très dégradés par rapport à l'utilisation d'une unique valeur cible pour le taux de compression, le contrôle du CPS s'avérant toujours bien meilleur que le contrôle du CpR. En combinant les deux types de contrôle, on aboutit à un point de fonctionnement relativement équilibré sur l'ensemble des indicateurs, avec une perte en score BLEU, mais un bien meilleur respect des contraintes de taille.

L'affinage des modèles produit des améliorations comparables, voire supérieures à celles obtenues avec des étiquettes de domaine, avec des variations en fonction des types de programmes. Cette observation est cohérente avec les résultats de Pham *et al.* (2021) qui montrent que l'affinage, qui spécialise un système différent pour chaque genre télévisuel, est une stratégie difficile à surpasser avec un seul système multi-genre. Les résultats détaillés par genre télévisuel (tableau 10) montrent des différences pouvant atteindre 1,3 point BLEU pour les émissions politiques.

Le tableau 10 permet également de voir l'incidence de la probabilité *a priori* des genres, pour le modèle générique (TRANSF + BS) comme pour les modèles adaptés.

Systèmes	BLEU_{br}	BLEU	SARI	TER_{br}	CPL_{36+}	CPS_{15+}
Systèmes de base et références						
Transcription + règles	20,4	33,6	17,9	0,54	0,0	80,0
TRANSFORMER + règles	32,0	44,7	54,4	0,37	0,0	61,5
Référence + règles	70,6	100	100	0,13	0,0	31,0
Référence	100	100	100	0,0	0,0	44,4
*Systèmes indifférents au genre (*TRANSFORMER*)*						
+ BS	35,4	44,4	54,3	0,35	2,9	59,8
+ BS + RT	36,2	45,3	55,0	0,35	1,9	59,5
+ BS + $\text{LRPE}_{\text{CpR = 0,75}}$	28,6	35,3	50,7	0,34	3,1	10,0
+ BS + $\text{LRPE}_{\text{CPS = 14,5}}$	31,6	39,2	52,2	0,30	3,3	0,5
+ BS + $\text{LDPE}_{\text{CpR = 0,75}}$	27,8	34,8	50,6	0,34	3,1	9,3
+ BS + $\text{LDPE}_{\text{CPS = 14,5}}$	30,8	38,4	51,9	0,31	3,1	0,4
*Systèmes adaptés au genre (*TRANSFORMER*)*						
+ BS + BG	35,5	44,9	54,6	0,34	2,0	58,5
+ BS + RT + BG	36,4	45,5	55,4	0,33	1,7	56,4
+ BS + $\text{LRPE}_{\text{CpR}^*}$	29,8	36,8	51,3	0,32	3,2	11,1
+ BS + $\text{LRPE}_{\text{CPS}^*}$	32,1	40,1	52,6	0,30	3,2	7,7
+ BS + $\text{LRPE}_{\text{CPS}^*}$ + BG	33,3	41,6	53,2	0,30	2,2	12,6
+ BS + AF^*	36,2	45,2	55,1	0,33	2,4	57,6
+ BS + RT + AF^*	36,9	45,8	55,8	0,33	1,6	57,4

Tableau 9. *Résultats de l'évaluation de différents modèles (moyenne sur le groupe d'émissions de test). BS, RT et BG dénotent respectivement l'usage de balises de segmentation, de rétrotraduction, et de balises de genre. CpR^* et CPS^* indiquent que les consignes de longueur (selon CpR ou CPS respectivement) sont adaptées pour chaque domaine. AF^* indique que chaque émission a été traitée avec le modèle affiné sur le même domaine.*

Les genres les plus représentés ont tendance à avoir de bons scores (« magazine » : 43 à 45 BLEU, 55 à 56 SARI) ; mais il n'y a clairement pas linéarité (ou même monotonie) pour la relation entre scores et représentation dans le corpus : « politique » ne compte que pour 4 % du corpus, mais est comparable à « magazine » (46 %) pour les résultats, tandis que « jeu » qui correspond à 16 % du corpus donne les pires performances (28 à 29 BLEU, 49 à 50 SARI). En revanche il y a une dépendance plus claire vis-à-vis de la qualité de la transcription automatique : nous avons calculé un coefficient de corrélation linéaire de – 0,77 entre le WER des émissions et les scores BLEU du système TRANSF + BS.

À titre indicatif, nous avons aussi évalué certains de nos systèmes sur des émissions de genres non vus à l'apprentissage : pour des leçons de MOOC (44 minutes) le système générique (TRANSF + BS) le score BLEU est de 45,8, et monte à 54,5 avec l'ajout de l'étiquette « vulgarisation » (représentant 0,9 % du corpus) ; pour une présentation TEDx (13 minutes) le système générique obtient BLEU = 63,8, mais des-

Systèmes	$BLEU_{br}$	BLEU	SARI	TER_{br}	CPL_{36+}	CPS_{15+}
Évaluation sur les émissions de test du genre « politique »						
+ BS	33,0	43,7	54,4	0,38	3,0	73,3
+ BS + RT	33,7	44,6	55,0	0,39	2,1	73,5
+ BS + BG	33,8	44,8	54,6	0,35	2,2	70,7
+ BS + RT + BG	35,5	46,1	55,8	0,35	2,4	67,3
+ BS + AF^{pol}	34,2	45,0	54,6	0,36	3,4	71,2
+ BS + RT + AF^{pol}	35,6	46,0	55,7	0,36	2,3	71,5
Évaluation sur les émissions de test du genre « magazine »						
+ BS	35,4	43,1	54,8	0,42	3,0	57,2
+ BS + RT	36,1	43,9	55,3	0,42	1,8	56,3
+ BS + BG	35,7	43,8	55,3	0,40	1,8	52,8
+ BS + RT + BG	36,6	44,3	55,8	0,39	1,5	51,4
+ BS + AF^{mag}	36,4	44,2	55,8	0,39	2,1	50,5
+ BS + RT + AF^{mag}	37,0	44,7	56,2	0,39	1,5	50,2
Évaluation sur les émissions de test du genre « jeu »						
+ BS	24,5	28,2	48,6	0,36	2,2	36,9
+ BS + RT	23,9	27,6	48,1	0,36	1,1	36,1
+ BS + BG	24,5	28,4	49,0	0,35	1,0	32,4
+ BS + RT + BG	24,9	28,5	49,4	0,34	1,3	31,4
+ BS + AF^{jeu}	25,0	28,9	49,7	0,33	1,8	32,4
+ BS + RT + AF^{jeu}	24,8	28,2	49,0	0,35	1,1	33,3

Tableau 10. *Résultats détaillés de l'évaluation de modèles d'adaptation au genre. BS, RT, BG et AF dénotent respectivement l'usage de balises de segmentation, de rétrotraduction, de balises de genre et de l'affinage.*

cend à 62,2 avec l'étiquette « vulgarisation ». Ces résultats témoignent d'une certaine robustesse des modèles appris.

Une dernière observation est que l'amélioration des performances obtenues par adaptation au genre reste dans tous les cas modeste. Ce résultat soulève la question de l'homogénéité réelle des émissions regroupées dans ces grandes catégories, qui, bien que relevant du même genre télévisuel, diffèrent sous de multiples autres aspects (contenus, thèmes abordés, intervenants, etc.).

5.4. *Utilité de la rétrotraduction*

Les motivations initiales pour utiliser des données rétrotraduites en complément des données alignées étaient (a) d'améliorer la qualité des plongements lexicaux ; (b) d'apprendre à mieux distinguer les styles de sous-titres avec davantage d'exemples ; (c) de renforcer la génération de la segmentation, dans la mesure où le système dispose d'un plus grand ensemble de sous-titres de référence correctement segmentés.

Dans le tableau 9, nous pouvons voir que l'ajout de données rétrotraduites par rapport à un système TRANSFORMER de base apporte un gain en BLEU (0,9 point) et en SARI (0,7 point). Il faut observer que l'utilisation de ces données complémentaires reste toujours bénéfique lorsqu'elle est combinée avec les méthodes d'adaptation au genre par balisage ou affinage : un gain de 0,6 point BLEU dans les deux cas, ce qui confirmerait une meilleure adaptation au genre. Une autre tendance régulière qui se dégage est le respect plus strict de la contrainte sur le nombre de caractères par ligne (CPL_{36+}) avec l'utilisation des données rétrotraduites, la quantité d'exemples de référence semblant bien importer pour cet aspect.

6. Conclusion

Dans cet article, nous avons présenté les travaux réalisés pour mettre en place un système entièrement automatisé de génération de sous-titres pour des émissions télévisuelles en langue française. S'appuyant sur les avancées récentes en traduction automatique neuronale, ce système repose sur la constitution d'un grand corpus associant transcriptions automatiques et sous-titres de référence pour des émissions variées, qui sert à l'apprentissage d'un modèle de traduction de type encodeur décodeur exploitant l'architecture TRANSFORMER. L'ajout de balises de segmentation explicites dans les textes générés permet de réaliser le sous-titrage en une seule étape, sans dégradation des performances par rapport à une architecture en pipeline.

Partant du constat que les genres télévisuels qui composent le corpus d'apprentissage présentent de fortes disparités quant à leurs sous-titres, nous nous sommes particulièrement focalisés sur l'étude de méthodes d'adaptation des modèles aux types de sous-titres et aux genres télévisuels. Nos expériences confirment l'apport des méthodes classiques d'adaptation, telles que l'utilisation d'étiquettes de genre et l'affinage, notamment quand elles sont combinées avec une technique d'augmentation de données ; les méthodes fondées sur le contrôle de longueur se sont en revanche montrées peu performantes dans l'ensemble. Dans ce contexte particulier les améliorations délivrées restent modestes, variables selon les genres et les types de sous-titres. Ceci suggère que la ventilation des données par genre télévisuel est loin de capturer toutes les sources de variation présentes dans les données et que des distinctions plus fines devraient être opérées pour tirer le meilleur parti des méthodes d'adaptation.

Dans nos travaux futurs, nous comptons poursuivre l'étude des méthodes d'adaptation en essayant d'exploiter au mieux la richesse de notre corpus d'apprentissage, pour lequel nous disposons de métadonnées très riches (par exemple : le nom de l'émission, la date de télédiffusion, l'identité des principaux intervenants). Ceci permet en particulier d'explorer la construction de modèles adaptés par émission, ou bien encore adaptés temporellement pour ce qui concerne en particulier les journaux. Nous avons aussi l'intention de poursuivre l'étude de l'adaptation multigenre à travers d'autres techniques, notamment les modules d'adaptation (*adapter layers*) (Bapna et Firat, 2019). Une autre question importante concerne les évaluations réalisées, qui s'appuient ici uniquement sur des métriques automatiques reflétant soit la similarité avec des ré-

férences, soit la conformité avec la charte du CSA : étudier également l'utilité des sous-titres automatiques du point de vue de leur utilisation par des sous-titreurs professionnels ou des spectateurs est également une perspective importante.

Remerciements

Nous remercions J.-L. Gauvain (LISN) et E. Florence (france.tv access) pour leur aide, ainsi que les relecteurs anonymes pour leurs remarques et suggestions constructives. Ce travail a bénéficié de calculs réalisés sur la plateforme LabIA. Le premier auteur est soutenu par un financement de la BPI dans le cadre du projet « Rosetta ».

7. Bibliographie

Alva-Manchego F., Martin L., Scarton C., Specia L., « EASSE : Easier Automatic Sentence Simplification Evaluation », *CoRR*, 2019.

Anastasopoulos A., Bojar O., Bremerman J., Cattoni R., Elbayad M., Federico M., Ma X., Nakamura S., Negri M., Niehues J., Pino J., Salesky E., Stüker S., Sudoh K., Turchi M., Waibel A., Wang C., Wiesner M., « FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN », *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, Association for Computational Linguistics, Bangkok, Thailand (online), p. 1-29, August, 2021.

Bahdanau D., Cho K., Bengio Y., « Neural Machine Translation by Jointly Learning to Align and Translate », *in* Y. Bengio, Y. LeCun (eds), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Bapna A., Firat O., « Simple, Scalable Adaptation for Neural Machine Translation », *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong, China, p. 1538-1548, 2019.

Bérard A., Pietquin O., Besacier L., Servan C., « Listen and Translate : A Proof of Concept for End-to-End Speech-to-Text Translation », *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain, December, 2016.

Burlot F., Yvon F., « Using Monolingual Data in Neural Machine Translation : a Systematic Study », *Proceedings of the Third Conference on Machine Translation*, Association for Computational Linguistics, Belgium, Brussels, p. 144-155, October, 2018.

Cho K., van Merrienboer B., Bahdanau D., Bengio Y., « On the Properties of Neural Machine Translation : Encoder–Decoder Approaches », *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Association for Computational Linguistics, p. 103-111, 2014.

Chu C., Wang R., « A Survey of Domain Adaptation for Neural Machine Translation », *Proceedings of the 27th International Conference on Computational Linguistics*, COLING 2018, Santa Fe, New Mexico, USA, p. 1304-1319, 2018.

Daelemans W., Höthker A., Tjong Kim Sang E., « Automatic Sentence Simplification for Subtitling in Dutch and English », *Proceedings of the Fourth International Conference on Lan-*

guage Resources and Evaluation (LREC'04), European Language Resources Association (ELRA), Lisbon, Portugal, May, 2004.

Edunov S., Ott M., Auli M., Grangier D., « Understanding Back-Translation at Scale », *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Brussels, Belgium, p. 489-500, October-November, 2018.

Etchegoyhen T., Arzelus H., Gete H., Alvarez A., Torre I. G., Martín-Doñas J. M., González-Docasal A., Fernandez E. B., « Cascade or Direct Speech Translation ? A Case Study », *Applied Sciences*, 2022.

Flesch R., « A new readability yardstick. », *Journal of applied psychology*, vol. 32, n° 3, p. 221, 1948.

Freitag M., Al-Onaizan Y., « Fast Domain Adaptation for Neural Machine Translation », *CoRR*, 2016.

Ghannay S., Caubrière A., Estève Y., Camelin N., Simonnet E., Laurent A., Morin E., « End-to-end named entity and semantic concept extraction from speech », *IEEE Spoken Language Technology Workshop*, Athens, Greece, December, 2018.

Kandel L., Moles A., « Application de l'indice de Flesch à la langue française », *Cahiers Etudes de Radio-Télévision*, vol. 19, n° 1958, p. 253-274, 1958.

Karakanta A., Gaido M., Negri M., Turchi M., « Between Flexibility and Consistency : Joint Generation of Captions and Subtitles », *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, Association for Computational Linguistics, Bangkok, Thailand (online), p. 215-225, August, 2021.

Karakanta A., Negri M., Turchi M., « Is 42 the Answer to Everything in Subtitling-oriented Speech Translation ? », *Proceedings of the 17th International Conference on Spoken Language Translation*, Association for Computational Linguistics, Online, p. 209-219, July, 2020a.

Karakanta A., Negri M., Turchi M., « MuST-Cinema : a Speech-to-Subtitles corpus », *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 3727-3734, May, 2020b.

Kingma D. P., Ba J., « Adam : A Method for Stochastic Optimization », *in* Y. Bengio, Y. LeCun (eds), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Kobus C., Crego J., Senellart J., « Domain Control for Neural Machine Translation », *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, Varna, Bulgaria, p. 372-378, September, 2017.

Kudo T., Richardson J., « SentencePiece : A simple and language independent subword tokenizer and detokenizer for Neural Text Processing », *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, Association for Computational Linguistics, Brussels, Belgium, p. 66-71, November, 2018.

Lakew S. M., Gangi M. D., Federico M., « Controlling the output length of neural machine translation », *Proceedings of IWSLT'2019*, 2019.

Luong M.-T., Manning C., « Stanford neural machine translation systems for spoken language domains », *Proceedings of the 12th International Workshop on Spoken Language Translation : Evaluation Campaign*, Da Nang, Vietnam, p. 76-79, December 3-4, 2015.

Matusov E., Wilken P., Georgakopoulou Y., « Customizing Neural Machine Translation for Subtitling », *Proceedings of the Fourth Conference on Machine Translation (Volume 1 : Research Papers)*, Florence, Italy, p. 82-93, August, 2019.

Papineni K., Roukos S., Ward T., Zhu W.-J., « BLEU : a method for automatic evaluation of machine translation », *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, USA, p. 311-318, 2002.

Pham M. Q., Crego J., Yvon F., « Revisiting Multi-Domain Machine Translation », *Transactions of the Association for Computational Linguistics*, vol. 9, n° 0, p. 17-35, 2021.

Post M., « A Call for Clarity in Reporting BLEU Scores », *Proceedings of the Third Conference on Machine Translation : Research Papers*, Association for Computational Linguistics, Belgium, Brussels, p. 186-191, October, 2018.

Saunders D., « Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation : A Survey », *CoRR*, 2021.

Sennrich R., Haddow B., Birch A., « Improving Neural Machine Translation Models with Monolingual Data », *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Berlin, Germany, p. 86-96, August, 2016.

Snover M., Dorr B., Schwartz R., Micciulla L., Makhoul J., « A Study of Translation Edit Rate with Targeted Human Annotation », *Proceedings of the seventh conference of the Association for Machine Translation in the Americas (AMTA)*, vol. 200, Cambridge, MA, Boston, Massachusetts, USA, p. 223-231, 2006.

Sperber M., Paulik M., « Speech Translation and the End-to-End Promise : Taking Stock of Where We Are », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 7409-7421, July, 2020.

Takase S., Okazaki N., « Positional Encoding to Control Output Sequence Length », *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, p. 3999-4004, June, 2019.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., « Attention is All you Need », *in* I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (eds), *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, p. 6000-6010, 2017.

Xu W., Napoles C., Pavlick E., Chen Q., Callison-Burch C., « Optimizing Statistical Machine Translation for Text Simplification », *Transactions of the Association for Computational Linguistics*, vol. 4, p. 401-415, 2016.

Zhang X., Lapata M., « Sentence Simplification with Deep Reinforcement Learning », *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark, p. 584-594, September, 2017.

Zhang Y., Ye Z., Feng Y., Zhao D., Yan R., « A Constrained Sequence-to-Sequence Neural Model for Sentence Simplification », *CoRR*, 2017.

Biais de genre dans un système de traduction automatique neuronale : une étude des mécanismes de transfert cross-langue

Guillaume Wisniewski* — **Lichao Zhu*** — **Nicolas Ballier**** — **François Yvon*****

* *LLF, Université Paris Cité & CNRS , 75 013, Paris France*
** *CLILLAC-ARP, Université Paris Cité, 75 013, Paris France*
*** *Université Paris-Saclay & CNRS, LISN, 91 403, Orsay France*

RÉSUMÉ. *Cet article a pour objectif de mettre en évidence les biais de genre dans les systèmes de traduction automatique et de rechercher leurs causes en étudiant les différentes manières dont l'information de genre peut circuler entre le décodeur et l'encodeur. Pour cela, nous décrivons un corpus minimal et contrôlé pour mesurer l'intensité de ces biais dans les traductions de l'anglais vers le français et du français vers l'anglais. Grâce à des méthodes de sondage et des interventions sur les représentations internes de l'encodeur, nos expériences montrent que l'information de genre est distribuée sur l'ensemble des représentations des tokens sources et cibles et que la sélection du genre en langue cible résulte d'une multiplicité d'interactions entre les diverses unités impliquées dans la traduction.*

ABSTRACT. *This paper describes a study on gender bias in French/English neural machine translation (MT) systems. We introduce a controlled corpus to measure the intensity of such biases in the two translation directions (from and into English). This corpus also allows us to investigate the information flow in a encoder-decoder architecture and to identify how gender information can be transfered between languages. Considering both probing as well as interventions on the internal representations of the MT system, we show that gender information is encoded in all token representations built by the encoder and the decoder and that there are multiple paths to transfer gender.*

MOTS-CLÉS : *biais de genre, traduction automatique neuronale, évaluation diagnostique en TAL.*

KEYWORDS: *Gender bias, Neural Machine Translation, Diagnostic Evaluation in NLP.*

TAL. Volume 63 – n° 1/2022, pages 37 à 61

1. Introduction

Il est largement admis (Callison-Burch *et al.*, 2006 ; Balvet, 2020) que les métriques automatiques telles que les scores BLEU ou METEOR sont inadaptées pour rendre compte des progrès de la qualité des traductions automatiques (TA) prédites par les systèmes neuronaux. Partant de ce constat, plusieurs protocoles ont été récemment proposés pour évaluer plus finement la traduction (Isabelle *et al.*, 2017 ; Burlot et Yvon, 2017 ; Burlot et Yvon, 2018). Ces protocoles reposent sur l'utilisation de jeux de tests élaborés (manuellement ou automatiquement) pour confronter les systèmes de TA à des problèmes de traduction spécifiques et bien caractérisés.

Les limitations de la traduction automatique ne se réduisent pas à leur incapacité à prendre en charge certains phénomènes linguistiques. Un autre problème important est l'existence de biais systématiques, en particulier de genre. Sous cette appellation, il faut distinguer plusieurs comportements problématiques : (a) le fait que des erreurs de traduction sont plus fréquentes pour des énoncés qui mettent en scène des actants de genre féminin ; (b) le fait que des traductions rendent linguistiquement explicite le genre des actants évoqués, alors que l'intention du locuteur peut être de le laisser ambigu ; (c) le fait que ces explicitations privilégient des assignations stéréotypiques, confortant, voire renforçant, des préjugés sexistes dans les textes traduits. Dans la typologie de Crawford (2017), affinée par Blodgett *et al.* (2020), ces problèmes sont susceptibles de fausser la manière dont certains groupes (ici, les femmes) sont représentés dans les textes *(representational harm)* ainsi que de conduire à un service (de TA) de moindre qualité pour les femmes *(allocational harm)*. Avec la massification de l'usage des technologies de TA, l'existence de tels biais est de plus en plus criante et dénoncée. Répondre à ces dénonciations exige à la fois des études précises (voir en particulier Savoldi *et al.* (2022) et les références citées), et des réponses appropriées de la part des fournisseurs de technologie[1].

Pour la paire de langues anglais français, ces problèmes peuvent être mis en évidence et quantifiés en observant la manière dont les marques de genre, qui peuvent être explicites ou non dans le texte source, se distribuent dans le texte cible. Une mesure de cet effet, mise en évidence dans plusieurs travaux, est proposée par Stanovsky *et al.* (2019), qui évalue les biais de genre à partir du décompte des erreurs portant sur la résolution d'anaphores pronominales. Ces travaux et d'autres études connexes visant à mesurer et à corriger les biais de genre sont passés en revue au § 2.

La première contribution nouvelle de cet article est d'étendre les analyses menées sur la traduction depuis l'anglais vers le français dans la direction inverse (§ 3). Nous proposons également de nouveaux contrastes pour mettre en évidence et quantifier les biais de genre grâce à la constitution d'un nouveau corpus contrôlé (§ 4). Nous nous intéressons, dans un second temps, à l'analyse des représentations internes d'un système de traduction neuronale à base de TRANSFORMER afin d'identifier plus finement la manière dont ces biais sont encodés dans les paramètres du réseau (§ 5). Nous pré-

1. Ainsi, les efforts récents de Google en la matière sont décrits dans ce billet.

sentons ensuite les premiers éléments d'une analyse causale permettant de mettre en évidence les différents mécanismes mis en œuvre dans le transfert de l'information de genre depuis le français vers l'anglais (§ 6).

2. Travaux connexes : mesurer et corriger les biais de genre

2.1. *Compter les erreurs et mesurer les biais*

La première étape pour étudier les biais de genre en TA consiste à les caractériser plus précisément, ainsi que les effets néfastes qu'ils peuvent produire auprès des utilisateurs de cette technologie (Blodgett *et al.*, 2020). Comme évoqué ci-dessus, ces auteurs distinguent en particulier les *biais de représentation*, qui conduiraient une TA à générer des textes véhiculant une représentation dénaturée des catégories sociales évoquées dans les textes traduits des *biais d'allocation*, qui se manifestent par un fonctionnement dégradé (des systèmes) pour certaines catégories d'usagers.

Lorsque l'on aborde ces questions sous l'angle quantitatif, à partir des observables que sont les sorties des systèmes de TA, deux situations sont à distinguer. Dans la première, le genre des personnes dont il est fait mention dans un texte source à traduire est indéterminé[2] et ne peut être déduit du contexte. Dans ce cas, on doit souhaiter que la traduction conserve cette ambiguïté, car tout autre choix impliquerait une interprétation non conforme aux intentions de l'auteur, tout en constatant que l'expression de cette ambiguïté est plus ou moins directe et transparente selon les langues, qui pour certaines disposent de formes neutres, ou bien ne marquent qu'exceptionnellement le genre, quand d'autres le marquent obligatoirement. À défaut, il semble souhaitable que les marques de genre qui seraient insérées le soient de manière équilibrée[3]. Lorsque ce n'est pas le cas, le système risque de créer, voire d'amplifier les biais de représentation, de fournir des informations faussées aux utilisateurs de la TA et de les propager dans les étapes de traitement ultérieures.

La seconde situation est celle dans laquelle l'information de genre[4] est explicite dans le texte source, auquel cas il est attendu qu'elle soit transférée correctement dans le texte cible, afin de toujours préserver les intérêts de l'auteur ainsi que celui des personnes qui seraient évoquées. Le système peut commettre deux types d'erreurs : (i) introduire dans le texte cible une ambiguïté qui est absente de la source ; (ii) se tromper dans l'expression du genre correct (complètement ou partiellement si le même genre est marqué sur plusieurs éléments du discours). En particulier entre dans cette

2. Cette formulation est simplificatrice, puisque, par exemple, il a longtemps été accepté en français dans certains usages que le genre masculin ait une valeur de générique – dans cette situation, il faudrait considérer que le genre des personnes représentées est indéterminé, alors même qu'une marque explicite de genre est présente.

3. Il est toutefois possible d'imaginer des situations ou des applications qui justifieraient de favoriser un genre (linguistique) plutôt qu'un autre dans les sorties.

4. Qu'elle soit encodée sous la forme d'une catégorisation binaire du genre ou bien qu'elle corresponde à des assignations plus fluides des identités de genre.

catégorie le fait de ne pas préserver l'ambiguïté ou la fluidité du genre alors que des pronoms sont disponibles pour éviter des assignations de genre binaire.

Même s'il est possible d'imaginer des situations dans lesquelles une traduction fidèle pourrait porter préjudice à certains usagers, il semble utile de mesurer les biais d'un système par des décomptes d'erreurs qu'il commet et la méthode que nous avons présentée *supra* s'inscrit dans cette démarche. Pour effectuer ces décomptes, la plupart des travaux analysant les biais de genre dans la traduction neuronale se sont concentrés sur le lexique de la profession (Kuczmarski et Johnson, 2018 ; Prates *et al.*, 2020), en étudiant aussi bien des corpus artificiels que des corpus réels (Gonen et Webster, 2020). Notons que la question du genre en TA peut être abordée sous d'autres angles : ainsi, Vanmassenhove *et al.* (2018) présentent des observations portant sur la distribution des verbes d'opinion en fonction du genre et du degré d'assertivité présumé chez les hommes et les femmes. Comme le montrent ces auteurs, qui étudient la traduction de dix langues vers le français, enrichir la phrase source (en anglais) par l'information explicite du genre du locuteur permet alors d'obtenir des traductions meilleures qu'un système qui ne dispose pas de cette information.

Une tentative de mesurer les biais dans la traduction vers l'anglais est détaillée par Cho *et al.* (2019), qui élaborent un indice du biais dans la traduction depuis le coréen *(translation gender bias index)*. Cet indice évalue la propension d'un système à traduire un pronom neutre en coréen en un masculin ou un féminin en anglais, ou bien encore en un groupe nominal non marqué pour le genre.

2.2. *Atténuer automatiquement les biais de genre*

Mesurer les biais permet aussi d'évaluer l'impact de travaux visant à les atténuer dans les TA. Ces travaux mobilisent principalement trois types de techniques (voir (Savoldi *et al.*, 2022) pour une étude récente). Une première consiste à manipuler les représentations lexicales. Elle est utilisée par Escudé Font et Costa-jussà (2019), qui injectent dans le système OpenNMT des plongements lexicaux entraînés avec l'algorithme *gender-neutral GloVe* (Zhao *et al.*, 2018). Ces auteurs testent ensuite la capacité à désambiguïser « *friend* » dans les traductions vers l'espagnol à partir des relations de coréférence ainsi que d'un nom de profession en attribut dans des phrases de la forme *I've known her for a long time, my friend works as a refrigeration mechanic.*

Les techniques de préannotation (Sennrich *et al.*, 2016) insèrent dans le texte source des marques explicites de genre, qui vont servir à orienter le système vers des traductions correctes. C'est, par exemple, l'approche suivie par Vanmassenhove *et al.* (2018), qui montrent que l'indication du genre des entités nommées dans l'anglais *(FEMALE Madam President, as a...)* permet d'améliorer les scores BLEU pour des traductions vers le français, l'italien, le danois et le finnois. Des résultats similaires sont obtenus par Basta *et al.* (2020) pour la traduction de l'anglais vers l'espagnol et des analyses complémentaires sont réalisées par Saunders *et al.* (2020). Cette tech-

nique est enfin utilisée par Kuczmarski et Johnson (2018) pour contrôler la traduction vers l'anglais de formes pronominales non marquées en turc dans des phrases telles que *« O bir doktor »* ou *« O bir hemşire »*.

Une troisième famille d'approche manipule les distributions des données d'apprentissage en s'appuyant sur des méthodes d'augmentation de données *(counterfactual data augmentation)*, CDA. Ainsi, Lu *et al.* (2020) engendrent automatiquement des corpus artificiels qui rétablissent l'équilibre en genre. Poursuivant cette direction, Saunders *et al.* (2020) montrent qu'il est plus simple et plus efficace de manipuler les distributions d'apprentissage en s'appuyant sur des méthodes d'adaptation au domaine. Ils utilisent un petit corpus artificiel équilibré en genre qui sert à adapter un système entraîné sur un corpus déséquilibré. Leur analyse de la traduction depuis l'anglais de trois langues montre que l'adaptation réduit les biais mesurés par les méthodes de Stanovsky *et al.* (2019).

3. Des jeux de tests contrôlés pour observer les biais de genre

Dans cette section, nous présentons la démarche suivie pour construire de nouveaux contrastes pour observer et quantifier les biais de genre en TA.

3.1. *Les corpus WinoGender et WinoBias*

Notre point de départ est l'étude de Stanovsky *et al.* (2019) qui formule des propositions concrètes pour évaluer les biais de genre, en s'appuyant principalement sur deux jeux de données : WinoGender (Rudinger *et al.*, 2018) et WinoBias (Zhao *et al.*, 2018), tous deux inspirés des schémas Winograd (Winograd, 1983)[5]. Un schéma Winograd repose sur une paire de phrases, chacune composée de deux propositions, qui ne diffèrent que d'un seul mot (ou une expression) prédicatif. Changer le verbe prédicatif induit un changement dans l'interprétation de la coréférence du pronom sujet dans la subordonnée, qui renvoie selon les cas soit au sujet soit à l'objet de la proposition principale, comme dans l'exemple suivant :

(1) *The city councilmen refused the demonstrators a permit because they [feared/advocated] violence*.

Dans cet exemple, la première partie de l'alternative (*feared*) conduit à interpréter *they* comme référant à *The city councilmen*, alors que la seconde (*advocated*) induit une coréférence avec *the demonstrators*. Ces schémas constituent des cas de test particulièrement difficiles pour les systèmes de TAL, car la résolution correcte de l'anaphore implique souvent une analyse profonde, voire des connaissances du monde.

5. On se reportera à Levesque *et al.* (2012) pour une discussion de ces schémas et à Amsili et Seminck (2017) pour leur adaptation au français.

Rudinger *et al.* (2018) décalquent ce schéma d'alternance pour cent vingt couples de phrases en mobilisant deux types de constructions pour constituer le corpus WinoGender :

(2) *[entity1] [interacts with] [entity2] [conjunction] [pronoun] [circumstances].*

(3) *[entity1] [interacts with] [entity2] and then [interacts with] [pronoun] for [circumstances].*

Les jeux de tests qui en dérivent reposent alors sur l'établissement de la relation de coréférence entre *he* ou *she* et son antécédent dans des phrases comme (l'antécédent escompté est entre crochets) :

(4) *[The developer] built a website for the tailor because [she] is an expert in building websites.*

(5) *The developer built a website for [the tailor] because [he] wants to sell cloths online.*

WinoBias est construit sur des principes similaires et comprend un ensemble équilibré de 3 160 phrases contenant des anaphores pronominales dont l'antécédent est un nom d'activité ou de profession. L'association entre un pronom et un nom est également répartie entre (a) des situations « stéréotypiques » (conformes aux distributions par genre de ces activités dans la population) et non stéréotypiques ; (b) des structures dans lesquelles l'anaphore peut être résolue à partir de la syntaxe, et des structures pour lesquelles il faut des connaissances supplémentaires.

3.2. *Une évaluation des biais de genre*

Les 3 880 phrases issues de ces travaux sont utilisées par Stanovsky *et al.* (2019) pour mesurer les biais de systèmes traduisant depuis l'anglais vers huit langues dans lesquelles le genre est grammaticalisé.

L'exemple (4) ci-dessus correspond à une situation non stéréotypique et sera jugé correct si *developer* est traduit par *développeuse*, incorrect sinon. Selon ces auteurs, le biais se manifeste par des erreurs de traduction qui privilégient des genres associés à des rôles stéréotypiques, plutôt que ceux qui sont attendus au vu de la relation de coréférence. Ils proposent donc de les mesurer en comparant les taux d'erreur des traductions des pronoms associés respectivement à des noms masculins et féminins.

La méthode de Stanovsky *et al.* (2019) pour mesurer le biais est problématique à plusieurs titres, et finalement peu appropriée pour notre étude. En effet, outre son utilisation d'un corpus artificiel, elle repose sur un repérage automatique du genre du nom choisi par le système. Or ce repérage n'est pas toujours possible (quand le groupe nominal ne porte aucune marque de genre), comme lorsque l'on traduit la phrase suivante dans laquelle le genre du groupe nominal sujet ne peut être déterminé :

(6) *L'analyste se dirigea vers la caissière car elle devait faire un dépôt.*

Ce repérage peut également être bruité du fait d'erreurs dans l'étiquetage morphosyntaxique ou dans l'alignement automatique avec la phrase anglaise, ou encore du fait d'une traduction très erronée[6].

Sémantiquement, l'absence d'alternance de prédicat peut conduire à des ambiguïtés de la portée du pronom. À la différence des énoncés des tests de Winograd, dont la robustesse de l'interprétation fait l'objet de tests d'évaluation par des humains, les énoncés de WinoGender n'ont pas été testés et sont parfois potentiellement ambigus entre ce qu'on appelle parfois une interprétation *de dicto* ou *de re*. Ainsi, dans l'exemple ci-dessous, la proposition introduite par *because* peut expliciter soit un point de vue sur les propriétés du référent du sujet grammatical et constitue un commentaire du prédicat de la principale (interprétation *de dicto*) soit expliciter et constituer une propriété du référent de l'objet indirect en tant qu'actant du prédicat de la principale (interprétation *de re*) :

(7) *The chief gave the housekeeper a tip because she was satisfied.*

(8) *The chief gave the housekeeper a tip because he was helpful.*

Dans cet exemple, l'alternance prévue de la coréférence dans ce modèle de test n'est ainsi pas certaine : dans la phrase (8), *he* pourrait renvoyer à l'objet indirect *(de re)* ou au sujet grammatical *(de dicto)*, de sorte que l'alternance en genre de *chief* n'est pas garantie dans ce couple de phrases.

Un second problème est que ce test est difficile à « inverser » pour évaluer ces phénomènes dans la traduction du français vers l'anglais. Nos premières tentatives pour construire un jeu de tests en post-éditant des traductions automatiques de WinoGender se sont rapidement heurtées à de nombreux cas d'ambiguïtés dans la détermination du genre correct français. Il apparaît enfin que ce corpus contient un trop grand nombre de sources de variabilité (des structures de phrase et du lexique) pour que l'on puisse facilement manipuler et visualiser les représentations internes calculées pendant la traduction. Nous avons préféré utiliser dans nos expériences un jeu de données plus simple, en nous inspirant des travaux de Saunders et Byrne (2020) présentés à la section 2.2.

3.3. *Une évaluation plus contrôlée du biais de genre*

À l'instar de Saunders et Byrne (2020), nous avons construit, pour étudier le transfert de genre entre le français et l'anglais, un corpus parallèle sur les patrons (9-10) :

6. Ainsi, les trois résultats du tableau 1 qui portent sur les 3 880 exemples de WinoGender, excluent chacun plusieurs centaines de phrases (près de 900 pour le système `fairseq`), pour lesquelles le script d'analyse échoue à prédire le genre.

(9) `[DET] [N]` a terminé son travail. (p. ex. : `L'acteur` a terminé son travail.)

(10) The `[N]` has finished `[PRO]` work. (p. ex. : The `actor` has finished `his` work.)

Dans ces patrons, `N` est un nom de métier qui est soit masculin soit féminin (p. ex. en anglais $actor_M$/$actress_F$; en français, $acteur_M$/ $actrice_F$), `DET` est le déterminant français qui s'accorde avec le nom (soit sous la forme du féminin la_F, soit du masculin le_M, soit de l'épicène *l'*) et `PRO` est le pronom possessif anglais her_F ou his_M. Dans ces phrases, la seule marque de genre est alors portée, en anglais, par le pronom[7] possessif et en français par le groupe sujet.

Nous utilisons la liste complète des noms de métier français collectée par Dister et Moreau (2014) afin de saturer la position `[N]` en français et de sélectionner le déterminant correspondant. Cette liste contient les formes féminine et masculine de 1 696 professions. Pour 274 de ces noms, ces deux formes sont identiques (noms épicènes). Au final, notre corpus contient donc 3 392 phrases suivant le patron 9, et est parfaitement équilibré en termes de genre. Ces phrases ont été traduites automatiquement et vérifiées manuellement pour produire la liste des phrases correspondantes en anglais. L'ensemble des corpus ainsi construits est librement téléchargeable sur le site du projet Neuroviz[8].

La plupart des noms de métier identifiés par Dister et Moreau (2014) sont des noms composés rares qui sont absents dans le corpus d'entraînement de notre système d'apprentissage (cf. § 4.1) : comme le montre la figure 1(a), environ 30 % des noms de métier utilisés pour créer le corpus n'apparaissent pas dans le corpus d'entraînement et moins de 6 % d'entre eux apparaissent plus de 10 000 fois. L'observation, à la figure 1(b), de la distribution du nombre de tokens résultant de la décomposition en unités sous-lexicales[9] des noms de métier illustre également la faible fréquence à laquelle ceux-ci sont observés : seuls 574 noms de métier ont une fréquence suffisante pour donner lieu à un mot du vocabulaire du système de TA, tous les autres sont décomposés, le plus souvent en deux ou trois unités sous-lexicales.

4. Évaluation du transfert de genre entre le français et l'anglais

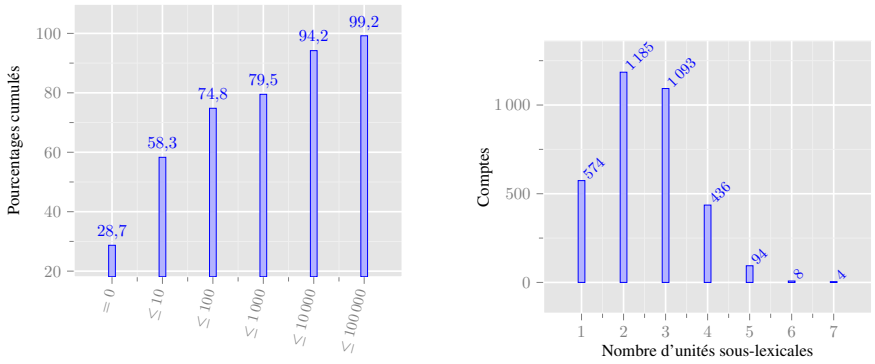
4.1. *Système de traduction*

Nous avons utilisé l'outil JOEYNMT, qui propose une implémentation « pédagogique » d'un système de traduction à base de TRANSFORMER (Vaswani *et al.*, 2017) permettant d'obtenir des résultats proches de l'état de l'art (Kreutzer *et al.*, 2019). Dans notre système, encodeur et décodeur sont composés de six couches, chacune

7. Nous suivons ici Huddleston *et al.* (2002) qui voient dans l'anglais une langue où le genre est peu grammaticalisé mais présent dans les relations de coréférence, comme avec les réfléchis *himself*, *herself* et *itself*.

8. `https://github.com/neuroviz/neuroviz/tree/main/gender_analysis_in_mt`.

9. Cette segmentation est détaillée à la section 4.1.



(a) Fréquences cumulées des noms de profession dans le corpus d'apprentissage.

(b) Distribution du nombre d'unités sous-lexicales dans les noms de métier.

Figure 1. *Rareté des noms de profession dans le corpus d'apprentissage*

avec huit têtes d'attention ; les couches de *feed-forward* comportent 2 048 paramètres et la dimension des plongements lexicaux est 512. Notre modèle comportait, au total, plus de 76 M de paramètres. Le système a été entraîné avec les données de la tâche « News » de la campagne WMT'15[10]. Les corpus `Europarl`, `NewsCommentary` et `CommonCrawl` sont utilisés pour l'apprentissage. Ils regroupent plus de 4 M de phrases et près de 141 M de tokens français. Tous les corpus ont été convertis en minuscules, tokenisés et segmentés en unités sous-lexicales en utilisant le modèle unigramme de l'outil SENTENCEPIECE (Kudo, 2018) et le vocabulaire résultant contient 32 000 unités. Le modèle est entraîné en optimisant l'entropie croisée à l'aide de la stratégie ADAM. Ce système obtient sur le corpus `newstest2014` un score BLEU de 34,0 pour la traduction du français vers l'anglais et de 32,7 pour la traduction de l'anglais vers le français.

Un autre point de comparaison est donné dans le tableau 1, qui reproduit pour ce système les mesures de biais de genre (Stanovsky *et al.*, 2019), en les comparant avec deux systèmes considérés dans cette étude, celui de `fairseq` (Ott *et al.*, 2018) et des traductions réalisées avec le système de Systran[11]. Il apparaît que notre implémentation de JOEYNMT atteint des performances conformes à celles des autres systèmes pour la prédiction du genre, avec une forte différence avec les prédictions pour le masculin et le féminin, et donc un fort biais de genre[12].

10. Il s'agit de la dernière campagne d'évaluation pour la traduction anglais français organisée dans le cadre de la conférence WMT (`http://statmt.org/wmt15`).

11. Dans ces deux derniers cas, nous utilisons les traductions de Stanovsky *et al.* (2019) et renvoyons à cette référence pour une description plus précise de ces deux systèmes.

12. *Précision* dénote la fréquence de la prédiction correcte du genre dans la phrase cible (complément du taux d'erreur à 1) ; ΔG_s dénote la différence de performances dans les traductions du masculin et du féminin.

	JOEYNMT	Fairseq	Systran
Précision	45,6	48,0	43,4
ΔG_s	30,1	4,4	41,8

Tableau 1. *Évaluation de la prédiction du genre de trois systèmes de traduction*

4.2. *Résultats expérimentaux*

Nous évaluons la capacité de notre système à prédire le genre des noms de métier en utilisant le jeu de tests décrit dans la section précédente et considérons comme point de comparaison les traductions engendrées par `e-translation`, le système de traduction développé par la Direction générale de la traduction de l'Union européenne[13].

4.2.1. *Mesure d'évaluation*

Nous nous intéressons, dans ce travail, à la capacité d'un système de traduction à transférer correctement l'information de genre du français vers l'anglais ou de l'anglais vers le français. C'est pourquoi notre évaluation ne porte que sur la capacité des systèmes à traduire correctement le groupe portant l'information de genre, indépendamment de la qualité de la traduction du reste de la phrase. Il est important de noter que, dans les deux directions, la position des mots portant l'information de genre est stable, ce qui simplifie l'évaluation de la correction des traductions par comparaison à celle de Stanovsky *et al.* (2019) qui implique une étape d'alignement automatique.

Vers le français, l'évaluation repose sur la prédiction du genre du groupe sujet traduisant `the [N]`, qui se trouve toujours en début de phrase. Quatre cas sont possibles, selon que le genre est porté par l'article et le nom (*la traductrice*), seulement par le nom (*l'actrice*), seulement par l'article (*la journaliste*), ou qu'il est complètement ambigu (*l'analyste*). Le tableau 2 décrit la distribution des différents cas dans notre corpus. Sauf mention contraire, nous considérerons que le genre du groupe sujet est correctement prédit lorsque le genre du déterminant et le nom sont tous deux corrects et évaluerons le taux de correction avec lequel le groupe sujet est prédit. Notons que cette mesure sous-évalue le nombre de phrases pour lesquelles le genre est correctement prédit : en effet, nous ne comptons comme « succès » que les phrases comportant la bonne traduction du nom de métier. Or, dans de nombreux cas, celui-ci est un mot rare (§ 3.3) et est mal traduit ; il est donc difficile d'évaluer si le genre est bien traduit ou non.

Dans l'autre direction, la vérification que le transfert du genre est correct en anglais est nettement plus simple et s'appuie sur le repérage du possessif (*her* ou *his*) dans la phrase cible. Il faut toutefois noter que pour les phrases sources dans lesquelles le déterminant et le nom de métier sont tous les deux épicènes (*l'analyste*) il est impossible de déterminer le genre et donc d'évaluer la qualité du transfert. Ce

13. `http://ec.europa.eu/cefdigital/eTranslation`

cas correspond à 272 exemples. Il est également envigeable qu'il ne soit pas possible de déterminer l'information de genre dans les traductions prédites par le système. Par exemple, dans certains cas, le système produit une traduction correcte n'utilisant pas les pronoms *her* ou *his* (*the programmer has finished working*) ; dans d'autres cas la traduction est complètement fausse (« L'inspectrice a fini son travail. » a été traduit en « *The young man bent on to work.* ») ou bien encore le possessif est traduit par *its* ou par *their*. Nous ne distinguons pas ces cas dans nos analyses et les regroupons tous sous la dénomination « autres » lorsque nous présentons les résultats.

4.2.2. *Résultats*

Pour la traduction du français vers l'anglais, il apparaît que notre système est capable de prédire correctement le pronom possessif dans seulement 52,4 % des phrases anglaises. *A contrario*, les résultats de `e-translation` atteignent presque la perfection : dans 90,9 % des hypothèses de traduction, le genre du pronom est correctement traduit. Pour la direction anglais-français, notre système est capable de prédire le déterminant correct pour 55,8 % des phrases et le nom pour 29,1 % des phrases. Ces observations suggèrent, qu'en plus de la difficulté de la tâche (traduction de noms de métier peu courants), le genre de la phrase est rarement correctement prédit. Quant à `e-translation`, il a été capable de prédire correctement le déterminant de 81,1 % des phrases et le nom de 57,8 % des phrases. La capacité de `e-translation` à correctement traduire les informations de genre a déjà été observée et celle-ci serait, *a priori*, due au choix de données d'apprentissage comportant majoritairement des énoncés sans biais de genre[14].

Pour la traduction vers l'anglais, le tableau 2 détaille ces scores : nous y indiquons, pour les différentes manières dont le genre peut être exprimé en français (§ 4.2.1), la proportion de phrases contenant *his* ou *her*. Ainsi, nous observons que, quand le déterminant et le nom ont tous les deux une forme spécifique au féminin, la traduction en *her* n'est choisie que dans 33,3 % des cas et *his* dans 18,5 % des cas. C'est d'ailleurs le seul cas, où le système génère plus souvent la forme féminine que la forme masculine. Ces résultats montrent que notre système, qui est entraîné à partir de corpus standard de traduction automatique, préfère nettement la traduction de *son* par un pronom masculin même dans des situations où il n'y a pas d'ambiguïté sur le genre du groupe nominal (p. ex. quand le déterminant et le nom ont tous les deux une forme spécifique au féminin). Dans l'ensemble, notre système n'obtient qu'une correction de 26,3 % pour les pronoms féminins, mais il prédit correctement le pronom dans 78,5 % des phrases au masculin. Ces conclusions corroborent celles obtenues par Saunders et Byrne (2020) sur les traductions de l'anglais vers l'allemand, l'espagnol ou l'hébreu. Des observations similaires ont été rapportées (Renduchintala et Williams, 2021) lorsque la traduction depuis l'anglais est testée sur un plus grand éventail de langues. À l'opposé, `e-translation` est capable de transférer correctement l'information de genre à partir du moment où celle-ci est marquée.

14. Le responsable du développement de `e-translation`, Markus Foti, a abordé cette question dans une interview accessible à ce lien.

Dét.	métier	nombre d'occurrences	pronom prédit	% phrases `JoeyNMT`	% phrases `e-translation`
l'	épicène	272	her	0,7 %	4,4 %
			his	80,1 %	94,1 %
			other	19,2 %	1,5 %
	fém.	251	**her**	**7,2 %**	**91,6 %**
			his	59,4 %	4,0 %
			other	33,4 %	4,4 %
	masc.	251	her	0,4 %	0,0 %
			his	**73,7 %**	**96,0 %**
			other	25,9 %	4,0 %
la	épicène	411	**her**	**31,6 %**	**93,0 %**
			his	43,9 %	0,3 %
			other	24,5 %	6,7 %
	fém.	898	**her**	**33,3 %**	**94,0 %**
			his	18,5 %	0,0 %
			other	48,2 %	6,0 %
le	épicène	411	her	0,7 %	0,0 %
			his	**84,4 %**	**95,4 %**
			other	14,9 %	4,6 %
	masc.	898	her	0,2 %	0,0 %
			his	**76,8 %**	**95,4 %**
			other	21,2 %	4,6 %

Tableau 2. *Pourcentage des hypothèses de traduction qui contiennent chaque type de pronom possessif selon la manière dont le genre est exprimé dans le sujet en français. Pour chaque cas de figure, le pronom anglais correct est en gras ; les scores de ces lignes correspondent alors à la correction du système.*

4.3. *Vers une analyse profonde du transfert du genre*

Les résultats de la section précédente mettent quantitativement en évidence les difficultés rencontrées par les systèmes de TA utilisés dans cette étude pour transférer l'information de genre entre langues, alors que celle-ci semble bien présente. Pour mieux comprendre les mécanismes à l'œuvre et mieux cerner les causes possibles de dysfonctionnements, nous poursuivons l'analyse en nous focalisant sur la direction français vers anglais, et donc sur les causes du choix d'un équivalent de traduction pour le mot français *son*. Ce choix est motivé par la structure systématique des phrases françaises et par la position relativement stable de ce mot dans la séquence d'entrée (il est toujours à la quatrième position à partir de la fin de la phrase), ce qui rend sa représentation interne assez facile à extraire et à manipuler. Notre principal objectif est alors de déterminer quels sont les éléments mis en jeu dans le choix du pronom anglais *his* ou *her* et d'étudier comment l'information de genre se propage dans le réseau pour construire les représentations numériques utilisées pour réaliser cette prédiction.

traduction	fréquence
_its	27,94 %
_his	**18,28 %**
_the	7,24 %
_her	**6,42 %**
_a	3,34 %
_their	2,92 %
_it	2,45 %
_sound	1,37 %
s	1,33 %
_he	0,76 %
autres	27,95 %

Tableau 3. *Les traductions les plus fréquentes du mot français* son *déduites des liens d'alignement de mots. Ainsi,* son *est aligné avec 3 658 types différents.*

Une première source de biais souvent mentionnée est liée aux données d'apprentissage. Comme reporté au tableau 3, il apparaît en effet que dans notre corpus d'apprentissage les traductions du pronom *son* par *his* sont trois fois plus fréquentes que les traductions par *her* (cette observation repose sur l'alignement des phrases sources et des phrases cibles avec `eflomal` (Östling et Tiedemann, 2016)). Nous pensons toutefois qu'il existe d'autres biais liés à la manière dont l'information de genre circule au sein de l'architecture TRANSFORMER.

Rappelons que, dans une architecture TRANSFORMER standard, le décodage s'effectue mot à mot de la gauche vers la droite. La sélection du mot anglais qui suit *has finished* s'appuie sur un vecteur contexte construit à partir des différentes couches d'attention de l'architecture. Trois mécanismes attentionnels sont simultanément à l'œuvre : l'auto-attention de l'encodeur, qui permet que les représentations des unités sources s'influencent mutuellement. L'auto-attention du décodeur, qui joue un rôle similaire côté cible, sous la contrainte que chaque unité n'ait accès qu'aux représentations des unités qui la précèdent. Enfin, l'attention croisée entre la source et la cible dans le décodeur, qui permet de contextualiser les représentations cibles en les combinant avec les représentations sources collectées sur la dernière couche de l'encodeur.

Pour traduire le genre du GN français, trois voies de propagation (non mutuellement exclusives) sont donc possibles : (a) une influence *directe* par le calcul de l'attention cross-langue ; (b) une influence *indirecte* passant par l'encodage (cross-langue) du genre dans la représentation du nom anglais, qui est propagée vers le pronom ; (c) une influence *indirecte* passant par l'encodage (monolingue) du genre dans la représentation d'autres mots français, en particulier du possessif français *son*, qui est ensuite propagée (cross-langue) vers le pronom anglais. Ces trois possibilités sont résumées dans la figure 2.

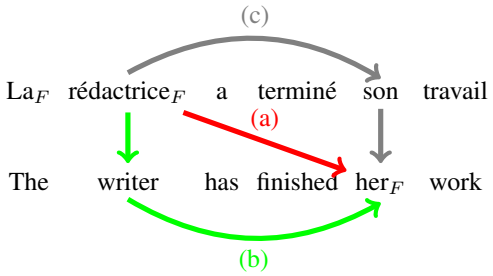


Figure 2. *Les différents mécanismes de transfert du genre du GN vers le pronom possessif en anglais*

Nous nous intéressons donc à démêler le rôle de ces différents mécanismes, par le truchement de plusieurs types d'analyses, qui vont permettre (a) de tester les représentations contextuelles en source et en cible par l'utilisation de sondes (section 5), une analyse que nous complétons en y intégrant des manipulations linguistiques (section 5.3) ; (b) de pondérer l'influence de ces trois mécanismes en utilisant des méthodes d'analyse causale (section 6).

5. Analyses par sondage

5.1. *Méthodes*

La première méthode considérée repose sur l'utilisation de sondes linguistiques *(probes)* (Belinkov et Glass, 2019) et consiste à tester la capacité à prédire le genre du GN français en observant seulement la représentation des mots sources construite par le système de TA. L'hypothèse sur laquelle repose cette méthode est que, s'il est possible de réaliser avec succès cette prédiction, c'est que les représentations correspondantes sont différentes pour les phrases comportant un GN masculin et pour les phrases qui comportent un GN féminin et peuvent donc influencer utilement le choix du pronom en anglais. Elles permettent de confirmer la possibilité d'une voie de transmission (b) et (c) de la figure 2.

En utilisant l'encodeur du système de traduction, nous calculons la représentation numérique associée à chaque mot du contexte source à droite du GN (soit : *a*, *terminé*, *son*, *travail*, . et `<eos>`), ainsi que le premier token (*the*) de la phrase en anglais dans toutes les couches du décodeur. Ces mots sont suffisamment fréquents pour donner lieu à une unique unité sous-lexicale et les représentations correspondantes sont à des positions stables. Nous considérons également une sonde qui est entraînée à partir de l'ensemble des tokens de la phrase cible (indépendamment de leur « valeur »), étant donné qu'il est impossible d'effectuer une analyse position par position en raison de la multitude des structures générées par la traduction automatique. Nous entraînons

ensuite un classifieur linéaire simple qui doit prédire le genre du GN à partir du vecteur représentant un token source ou un token cible.

Le classifieur utilisé est un modèle de régression logistique appris avec `scikit-learn` (Pedregosa *et al.*, 2011) en utilisant 75 % des données, et nous calculons le taux d'erreur sur les 25 % restant. Cette expérience est répétée cent fois pour pouvoir calculer l'intervalle de confiance de la mesure. Le classifieur est appris avec une régularisation ℓ_1 afin de contrôler la capacité des sondes (Hewitt et Liang, 2019).

5.2. *Résultats*

Les taux de correction *(accuracy)* obtenus par les diverses sondes sont dans le tableau 4. On observe en premier lieu que la représentation de *son* permet de prédire avec confiance le genre du nom, avec une correction supérieure à 80 %[15]. Comme la forme du mot n'est pas marquée en genre, on peut penser que cette information est apportée par le contexte qui va influencer la représentation interne. Cette influence du contexte se manifeste aussi par les corrections supérieures atteintes lorsque l'on exploite les couches les plus profondes de l'encodeur : la correction obtenue à partir des trois dernières couches de celui-ci dépasse les 90 %. Ce constat confirme qu'il existe bien un flux d'informations entre le pronom possessif et son antécédent, le nom de métier, ce qui correspond au chemin noté (c) dans la figure 2. On note que les scores de correction obtenus par la sonde lorsqu'elle prend en compte d'autres tokens sources sont également très élevés : ces scores sont comparables ou légèrement inférieurs à ceux obtenus avec *son*, ce qui montre que l'information de genre a un impact sur les représentations de tous les tokens sources, même lorsque ces tokens n'ont pas de relation syntaxique directe avec le groupe nominal sujet.

Les résultats de la sonde utilisant les représentations du décodeur comme caractéristiques (deux dernières colonnes du tableau 4) indiquent une tendance similaire : l'information de genre est encodée dans la représentation de *the*, même si le système génère toujours le même déterminant. Il apparaît également que la sonde est toujours capable de prédire le genre du nom de métier en français avec une assez grande correction à partir des représentations des tokens de la cible, illustrant, à nouveau, le caractère distribué de cette information.

Les résultats reportés dans la colonne *aléatoire* du tableau 4 correspondent à la prédiction d'une étiquette aléatoire à partir de ces mêmes représentations. Comme prévu, les résultats sont proches d'un tirage aléatoire et montrent que l'information présente dans les représentations est significative, puisque la sonde n'est pas capable d'apprendre des corrélations fallacieuses dans nos données (Hewitt et Liang, 2019).

15. Une analyse plus détaillée des prédictions montre que ce classifieur a un taux d'erreur similaire pour les phrases ayant un sujet féminin et pour les phrases ayant un sujet masculin.

	encodeur						aléatoire	décodeur	
couche	a	terminé	son	travail	.	eos	son	the	all
1	80,4	75,1	80,6	76,4	59,5	73,3	45,3	89,5	71,6
	±1,1	±0,3	±0,3	±0,6	±1,0	±1,0	±0,9	±0,2	±0,6
2	85,8	80,8	81,6	78,3	87,6	88,3	50,7	92,0	76,3
	±1,0	±0,2	±0,3	±0,7	±0,6	±0,7	±0,8	±0,1	±0,7
3	89,5	88,2	89,2	82,0	86,5	87,6	48,8	91,8	78,1
	±0,6	±0,2	±0,2	±1,1	±1,0	±0,6	±0,9	±0,1	±0,6
4	90,8	89,3	90,6	85,9	85,7	85,6	48,6	90,9	79,1
	±0,4	±0,2	±0,2	±0,9	±1,0	±0,7	±0,8	±0,2	±0,6
5	90,4	89,3	90,4	85,5	86,4	85,2	49,6	89,3	82,4
	±1,0	±0,2	±0,2	±0,8	±0,8	±1,2	±0,8	±0,2	±0,5
6	91,0	89,3	90,0	86,0	86,4	85,1	49,2	87,7	84,7
	±0,6	±0,2	±0,2	±1,0	±1,1	±0,8	±0,8	±0,2	±0,3

Tableau 4. *Correction (en %) de la prédiction du genre du syntagme sujet ou correction lorsque l'on cherche à prédire des étiquettes choisies aléatoirement à partir de la représentation de* son *(colonne aléatoire).*

5.3. *Sondage de représentations manipulées*

Nous étendons les tests par sondage de la section précédente en manipulant linguistiquement l'information de genre entre le déterminant, le nom, le pronom possessif et d'autres composantes de la phrase, afin de mesurer à quel point l'information de genre présente dans la sortie de l'encodeur est robuste à des variations du contexte d'occurrence. Le tableau 5 dresse la liste des manipulations considérées.

Les manipulations portent sur les caractéristiques suivantes. La première consiste à affaiblir le marquage du genre en remplaçant le DET (qui peut varier en genre) par *chaque*, qui est épicène. Nous varions ensuite le genre du nom déterminé par *son* : dans le patron initial, il est masculin (*travail*), nous proposons également une version avec un nom féminin (*activité*). Dans un troisième temps, nous augmentons la distance entre le groupe sujet et le possessif en insérant une proposition relative (*qui a chanté formidablement hier*) qui modifie le sujet. Cette manipulation est réitérée en insérant dans la relative un nom distracteur susceptible d'introduire du bruit dans la propagation du genre du groupe sujet (« homme » ou « femme »). Cet effet est encore renforcé lorsque l'on affaiblit le déterminant initial en le remplaçant par « chaque ». À l'inverse, nous considérons également la possibilité de renforcer le marquage du genre en introduisant un troisième composant adjectival dans le groupe sujet de manière à ce que le sujet en français soit toujours marqué explicitement.

La correction d'une sonde appliquée aux représentations des phrases manipulées est reportée dans le tableau 5. Une première observation est l'effet net des manipulations d'affaiblissement et de renforcement qui induisent respectivement des baisses et des hausses très sensibles de la qualité de la prédiction. Ces observations mettent en évidence le fait que l'encodage du genre dans les représentations en sortie de l'encodeur résulte d'un processus cumulatif dans lequel toutes les positions sources impliquées au sein du groupe sujet jouent un rôle effectif. À l'inverse, les autres manipula-

	couche	encodeur a	terminé	son	travail	.	eos
Affaiblissement							
Chaque surveillant a terminé son travail.	1	73,1	73,6	65.7	63,5	53,9	56,7
	6	71,0	71,4	70,4	68,2	71,2	69,7
Renforcement							
Le surveillant *français* a terminé son travail.	1	99,9	98,5	95,0	80,6	62,0	80,4
	6	100,0	99,7	99,7	98,9	98,8	96,9
Genre du complément							
Le surveillant a terminé son *travail*.	1	79,4	74,6	79,0	75,0	58,8	72,0
	6	90,3	88,8	89,2	85,3	86,2	83,3
Le surveillant a terminé son *activité*.	1	80,5	75,5	78.6	62,6	57,6	67,2
	6	89,7	88,3	89,6	84,3	86,1	84,1
Éloignement							
Le surveillant *qui a chanté formidablement hier* a terminé son travail.	1	71,1	66,3	68,8	81,1	56,8	65,4
	6	91,5	91,0	90,5	86,8	81,2	82,1
Distracteur							
.sans affaiblissement							
Le surveillant *que cette femme critiquait* a terminé son travail.	1	65,7	66,6	69,3	79,50	62,8	68,5
	6	90,6	89,6	89,1	85,91	81,9	80,2
Le surveillant *que cet homme critiquait* a terminé son travail.	1	65,4	67,0	68,7	80,0	63,4	68,2
	6	90,3	89,3	89,7	86,6	81,0	79,9
.avec affaiblissement							
Chaque surveillant *que cet homme critiquait* a terminé son travail.	1	63,1	63,5	64,3	62,4	56,2	55,8
	6	72,1	71,4	69,7	69,9	71,8	69,2
Chaque surveillant *que cette femme critiquait* a terminé son travail.	1	63,3	64,6	65,9	63,4	55,4	55,2
	6	71,8	71,8	70,0	69,2	70,2	69,5

Tableau 5. *Correction des sondes pour les phrases transformées*

tions visant à dégrader l'encodage du genre, soit en insérant du matériel linguistique, soit en ajoutant des distracteurs, n'ont qu'un effet limité : si la qualité des sondes est globalement moins bonne lorsque l'on utilise la première couche, la correction des prédictions sur la dernière couche reste très stable, proche de 90 % pour les mots *a*, *terminé* et *son*. On voit ici à l'œuvre le mécanisme progressif par lequel les représentations contextuelles se construisent dans les différentes couches de l'encodeur : l'empilement de couches permet de filtrer les cooccurrences accidentelles au profit d'informations plus structurelles, moins dépendantes de l'éloignement entre mots.

En conclusion, cette première salve d'expériences a mis en évidence la présence d'une information distribuée portant sur le genre du groupe sujet en français, qui est disponible dans les sorties de l'encodeur et permet en théorie de prédire le genre du possessif en anglais avec une bonne correction (supérieure à 90 %). Or, les performances du système de traduction sont bien moindres, ce qui nous amène à explorer plus précisément, dans la section suivante, l'utilisation qui est faite de ces représentations dans le décodeur.

6. Éléments d'analyse causale

Dans cette section, nous nous attachons à mesurer si le genre est directement transféré depuis le groupe sujet ou bien si ce transfert passe aussi par les autres mots du contexte, dont nous avons vu (§ 5) qu'ils encodent cette information. Pour cette expérience, nous utilisons des techniques de manipulation des représentations internes du TRANSFORMER, en nous inspirant des méthodes d'analyse causale (Pearl, 2001 ; Vig *et al.*, 2020). L'analyse causale s'intéresse à quantifier les effets directs et indirects d'une variable X sur une variable Y, potentiellement médiatisés par une variable Z. Dans notre cas, on veut savoir si le genre grammatical du groupe sujet en français a un effet direct sur le choix de la forme du pronom possessif en anglais ou bien s'il y a un effet indirect du contexte de la phrase source ou du contexte cible.

6.1. *Une nouvelle mesure des biais de genre*

Nous commençons par décrire deux nouvelles mesures pour quantifier la préférence d'un système de TA pour une forme féminine ou masculine. Ces mesures reposent sur une comparaison de la probabilité d'engendrer la forme masculine ou féminine lors d'un décodage forcé[16] de la traduction de référence, plutôt que sur l'analyse de l'hypothèse de traduction prédite par le système de TA. Cette approche présente deux avantages : d'une part, elle n'est pas affectée par les éventuelles erreurs de la traduction automatique, d'autre part, comme elle s'appuie sur des comparaisons de probabilités d'engendrer certains tokens, elle permet de détecter et de quantifier des variations plus fines dans le modèle, même lorsqu'elles ne se traduisent pas par des modifications de l'hypothèse de traduction. Pour rendre plus concrètes nos explications et nos analyses, nous ne considérons ici que la traduction du français vers l'anglais, mais notre méthode se généralise à d'autres problèmes de traduction.

6.1.1. *Définitions*

Chaque exemple de notre corpus de test est représenté par un contexte de traduction source, $u(\texttt{DET,N})$ où `N` et `DET` correspondent au nom du métier et à son déterminant. Le genre d'un exemple est noté $G(u)$ et peut prendre les valeurs masculin (M), féminin (F) ou indéterminé (I). Nos mesures reposent sur le calcul de $P(\text{his}|G(u))$ et de $P(\text{her}|G(u))$, soit la probabilité de produire *his* ou *her* sachant le contexte de traduction. Ces quantités sont impossibles à calculer exactement, puisqu'il faudrait sommer sur tous les débuts de phrases en anglais susceptibles de figurer devant le pronom. Nous réalisons donc une première approximation en considérant plutôt $P(\text{his}|G(u), e\triangleright)$, où $e\triangleright$ est le préfixe cible de la traduction de référence, et

16. Dans un décodage forcé, le i^{e} token prédit par le modèle n'est pas celui dont la probabilité est le plus élevée selon le modèle, mais celui correspondant au i^{e} token de la référence. Cette méthode de décodage permet de garantir que lors de la prédiction du pronom possessif anglais le contexte cible ne contienne pas d'erreur et de limiter ainsi le bruit dans nos mesures.

$P(\text{her}|G(u), e\triangleright)$ est défini de manière analogue. Ces deux quantités peuvent être calculées pendant le décodage forcé de la référence.

Nous considérons deux mesures du biais, dérivées respectivement de l'étude des contextes français pour lesquels le contexte permet de déduire le genre ($G(u) = M$ ou F), ou bien au contraire le laisse indéterminé ($G(u) = I$). Pour chacune des mesures, on s'intéresse aux valeurs moyennées sur tous les contextes u. La première mesure quantifie les biais de genre dans les contextes ambigus et est définie par :

$$b(u) = 1 - \frac{2 \times P(\text{his}|G(u) = I, e\triangleright)}{P(\text{his}|G(u) = I, e\triangleright) + P(\text{her}|G(u) = I, e\triangleright)} \quad [1]$$

Plus $b(u)$ est proche de 0, plus les probabilités de *his* et *her* sont proches, ce qui est attendu pour un contexte dans lequel le genre est indéterminé. Plus la probabilité d'engendrer *his* devient grande devant celle de *her*, plus $b(u)$ sera négatif. À l'inverse, une valeur proche de 1 indique une préférence pour la génération du pronom *her*.

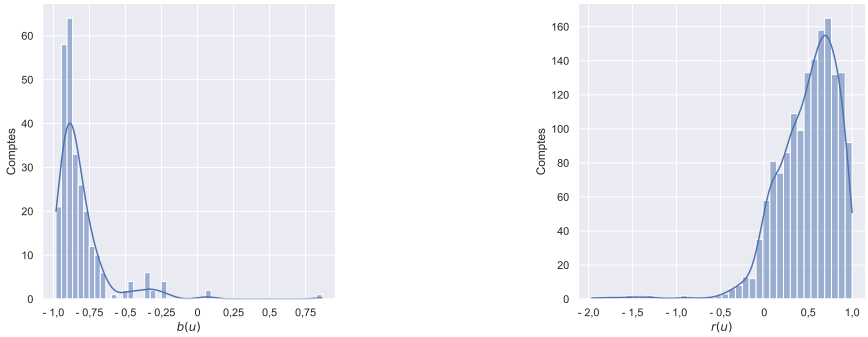
La seconde mesure est définie pour les contextes non ambigus par :

$$r(u) = 1 - \frac{P(\text{his}|G(u) = F, e\triangleright)}{P(\text{his}|G(u) = M, e\triangleright)} \quad [2]$$

Cette mesure compare, pour chaque contexte u, la probabilité d'engendrer le pronom *his* lorsque u correspond à la forme féminine du nom de métier, avec la probabilité d'engendrer *his* lorsque c'est la forme masculine qui est utilisée. La quantité $r(u)$ est proche de 0 quand les deux probabilités sont proches, c'est-à-dire lorsque la variation en genre du groupe sujet (le seul élément qui change dans le contexte) n'a pas d'influence sur les probabilités de sortie, ce qui manifeste une anomalie. Au contraire, si $r(u)$ est proche de 1, la probabilité de produire *his* lorsque le groupe sujet est masculin est très grande devant celle de le produire lorsque le groupe sujet est féminin, ce qui est le comportement souhaité ; une valeur fortement négative correspond à la situation (paradoxale) inverse : il est plus probable d'engendrer *his* avec un groupe sujet féminin plutôt qu'avec un groupe sujet masculin.

6.1.2. *Analyse des valeurs de base*

Nous avons représenté à la figure 3 la distribution des valeurs de b et de r calculées pour les phrases de notre corpus. Ces observations confirment les conclusions de la section 4.2 et montrent que notre système présente un biais très fort vers le masculin : pour les contextes indéfinis, la quasi-totalité des valeurs de $b(u)$ sont plus petites que $-0{,}75$ et cette mesure n'est positive que pour quelques rares contextes (p. ex. *autodidacte*, *aide* ou *interprète*), alors que la valeur moyenne de $b(u)$ devrait être nulle pour un système non biaisé. Pour les contextes non ambigus, $r(u)$ atteint une valeur moyenne de $0{,}46$, plus de 40 % des valeurs sont inférieures à $0{,}5$, et seules $0{,}7$ % des valeurs sont négatives. Ces observations montrent que la probabilité d'engendrer *his* est plus grande quand le groupe sujet est masculin que lorsqu'il est féminin, ce qui est attendu. En moyenne, passer d'un sujet masculin à un sujet féminin rend moins probable *his* d'un facteur 2 au bénéfice des alternatives : *her*, *it*, etc.



(a) Contextes ambigus (b) Contextes non ambigus

Figure 3. *Distributions de r et b sur les phrases de notre corpus*

6.2. *Importance du contexte dans la prédiction du genre*

6.2.1. *Trois manipulations des représentations*

Nous présentons maintenant une série d'expériences pour quantifier l'influence directe et indirecte des éléments du contexte source dans le choix de la forme du pronom possessif. Cette influence peut prendre diverses formes : en premier lieu par l'effet de l'attention cross-langue du décodeur qui peut intégrer directement des informations relatives au sujet français ; ce même mécanisme peut également s'appuyer indirectement sur les représentations contextuelles des mots d'autres sources, qui dépendent (*via* l'encodeur) du sujet ; la même dépendance indirecte peut être médiatisée par les représentations du début de phrase cible. Il peut enfin exister une influence directe entre le GN *cible* et le pronom qui s'exerce indépendamment du contenu de la source.

Pour essayer de démêler ces effets, nous construisons trois contextes alternatifs visant à neutraliser l'une ou l'autre des voies de transfert de l'information de genre identifiées dans la figure 2. Dans le premier contexte, noté u_1, la représentation lexicale de *son* est remplacée dès la première couche de l'encodeur par un plongement lexical « indifférencié » (dans nos expériences, nous avons choisi celui associé aux mots inconnus). Ce contexte réduit l'influence lexicale de *son*, sans empêcher toutefois que le genre soit transféré par la voie (c), puisque les couches supérieures de l'encodeur construisent une représentation contextualisée intégrant des informations sur le genre du groupe sujet. Les deux mesures $b(u_1)$ et $r(u_1)$ évaluent alors l'effet du transfert par les autres voies. Remarquons que même quand le token *son* est masqué dans la représentation de l'encodeur, il reste possible de calculer b et r, puisque ces deux mesures reposent sur un décodage forcé (qui produit toujours soit *his*, soit *her*).

Dans le second contexte (u_2), nous cherchons inversement à neutraliser l'effet du contexte source sur la représentation de *son* construite par l'encodeur. Pour cela, pour

chaque paire de phrases (forme féminine, forme masculine), nous substituons, dans chacune des phrases, la représentation de *son* sur la couche de sortie de l'encodeur par la moyenne des représentations construites pour ce token dans des contextes féminin et masculin. Nous supprimons ainsi l'influence possible de la contextualisation de *son* et empêchons que le genre soit transféré par la voie (c). À la différence de u_1, le biais intrinsèque de *son* est pris en compte. La mesure $r(u_2)$[17] quantifie alors l'effet cumulé du transfert par la voie (a), la voie (b) et l'influence lexicale directe de *son*.

La troisième modification (contexte u_3) a pour objectif de rendre les représentations du groupe sujet indifférentes au genre du nom . Le biais ne peut alors plus être dû qu'à des effets indirects (voies (b) ou (c)). Pour cela, nous considérons, comme précédemment, les paires (forme féminine, forme masculine) formées à partir d'un nom de métier donné et traduisons celles-ci après avoir remplacé, en sortie de l'encodeur, la représentation du nom de métier par la moyenne de la représentation construite avec un contexte féminin et de celle construite avec un contexte masculin[18]. Notons que, l'intervention étant réalisée en sortie de l'encodeur, les informations de genre peuvent toujours être présentes dans les représentations contextualisées des autres mots de la source. En faisant ce changement dès la première couche de l'encodeur, il serait possible d'empêcher la diffusion du genre en source et, d'une certaine manière, de rendre tous les noms de métier épicènes. Notons que comme pour le contexte u_2, cette intervention n'a pas d'effet pour les phrases dans lesquelles le genre n'est pas marqué.

6.2.2. *Résultats et analyses*

L'effet des interventions décrites ci-dessus est représenté à la figure 4. Ces résultats montrent que la neutralisation de *son* (intervention u_1) a un impact très fort sur les deux mesures que nous étudions : la figure (b) montre que la valeur moyenne[19] de $r(u)$ passe de $0{,}49$ à $0{,}18$ et sa médiane de $0{,}56$ à $0{,}34$ après modification du contexte. Modifier le contexte en gommant l'influence purement lexicale de *son* amplifie fortement le déséquilibre entre les deux pronoms puisque le $r(u)$ moyen s'éloigne de 1 et que le $b(u)$ moyen s'éloigne de 0 et, plus généralement, toute la distribution de ces deux quantités est déplacée vers les valeurs négatives. Dans ce contexte, il est permis de penser que les deux alternatives deviennent toutefois moins probables que d'autres déterminants (notamment *the*). La prédiction du déterminant anglais dépend donc fortement de la présence ou non du possessif *son* dans la source française. En revanche, une fois supprimée l'information lexicale liée à ce mot, l'influence indirecte du genre sujet dans la prédiction du pronom anglais se trouve réduite (voie (c)).

17. $b(u_2)$ et $b(u)$ sont identiques, la modification sur la couche de sortie de l'encodeur n'ayant aucun impact pour les phrases dans lesquelles le genre n'est pas marqué.

18. Cette intervention ne peut être effectuée que lorsque les segmentations de la forme féminine et de la forme masculine d'un nom de métier contiennent le même nombre d'unités sous-lexicales. Nous remplaçons alors, dans les deux phrases, la représentation de chaque unité par la moyenne des représentations féminines et masculines, position par position.

19. Les valeurs des moyennes et des médianes ont été calculées après avoir supprimé sept points aberrants qui accentuaient plus encore la baisse observée.

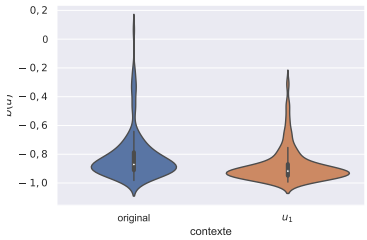
Les deux autres interventions que nous proposons ont un impact nettement plus faible sur $r(u)$: sa valeur moyenne (resp. médiane) passe de $0{,}46$ à $0{,}48$ (resp. de $0{,}558$ à $0{,}559$) après la modification du contexte u_2 et à $0{,}50$ (resp. $0{,}549$) après la modification u_3. Ces résultats agrégés masquent le fait que les interventions que nous proposons ont un fort impact sur quelques phrases pour lesquelles les valeurs de $r(u)$ sont anormalement faibles : les valeurs moyenne et médiane de $r(u)$ sont sensiblement différentes et surtout ne varient pas toujours dans le même sens (l'intervention u_3 entraîne une augmentation de la moyenne $r(u)$, mais une baisse de sa médiane).

Si l'on considère seulement l'évolution des valeurs médianes (pour limiter l'effet des valeurs extrêmes), il apparaît que la neutralisation du nom de métier (contexte u_3) entraîne une diminution faible de l'influence du genre du groupe sujet : quand l'information de genre n'est pas marquée sur le nom, la probabilité de générer *his* dans un contexte féminin et celle de le générer dans un contexte masculin deviennent un peu plus proches, avec toujours une très forte préférence pour le masculin. Ceci met en évidence l'existence de la voie (b), qui contribue toutefois peu à la sélection du genre du pronom. Gommer l'information de genre encodée dans la représentation de *son* (intervention u_2) n'a qu'un effet marginal, ce qui suggère de nouveau que la voie (c) ne joue qu'un petit rôle dans le transfert de l'information.

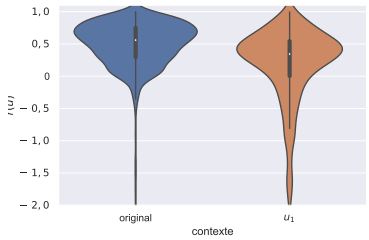
L'interprétation la plus réaliste de ces résultats est alors que l'essentiel du déséquilibre entre *his* et *her* s'explique par les corrélations entre mots anglais, capturées par l'auto-attention du décodeur : peu importe le genre du GN français, l'association entre nom et genre du pronom semble principalement due à la distribution inégalitaire observée dans la partie cible du corpus d'apprentissage. Cette interprétation, à confirmer par d'autres manipulations, dessine deux pistes non mutuellement exclusives pour réduire les biais observés : d'une part, rééquilibrer les statistiques du corpus d'apprentissage (sans nécessairement chercher à les associer à la réalité du genre correct en français, qui compte pour peu), d'autre part, renforcer explicitement les mécanismes de transfert cross-langue (voies (a) et (b)), par exemple en intégrant les dépendances syntaxiques dans le mécanisme d'attention intra ou inter-langue.

7. Conclusions

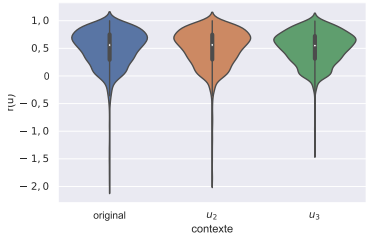
Nous avons présenté dans ce travail un nouveau jeu de tests permettant de mettre en évidence les biais de genre dans les systèmes de traduction automatique neuronale. Ce jeu de tests offre de nombreuses possibilités pour analyser finement les échanges d'informations entre les différentes composantes du réseau de neurones. En particulier, nous avons pu mettre en évidence que l'information de genre était distribuée dans les représentations de l'ensemble des tokens de la phrase source et de la phrase cible, et ce, même lorsque le patron générant les phrases fait intervenir des dépendances complexes. Nous avons également montré, grâce à des expériences consistant à intervenir sur les représentations internes du réseau de neurones, que le transfert de l'information de genre entre le français et l'anglais était complexe et reposait sur de multiples facteurs, même si le contexte de la phrase source semble jouer un rôle prépondérant.



(a) Phrases avec genre indéterminé



(b) Phrases avec genre non ambigu



(c) Phrases avec genre non ambigu

Figure 4. *Impact des interventions décrites dans la section 6.2 sur la distribution des deux mesures de biais de genre. La valeur moyenne des différentes mesures est représentée par un point blanc. Pour rendre le graphe plus lisible, les sept plus petites valeurs de $r(u)$ pour le contexte u_1 ont été supprimées de la figure (b) et les trois plus petites valeurs pour le contexte original de la figure (c).*

Pour prolonger ce travail, nous souhaitons confirmer nos observations expérimentales, notamment en étudiant de nouvelles interventions visant à découpler le décodeur de l'encodeur et à intégrer d'autres informations lexicales, comme la fréquence d'apparition des tokens étudiés. Nous espérons également parvenir à mieux caractériser les biais du corpus d'apprentissage et à étendre les méthodes décrites ici à d'autres langues et à d'autres phénomènes.

Remerciements

Ce travail a été partiellement financé par le projet NeuroViz soutenu par la Région Île-de-France dans le cadre d'un financement DIM RFSI 2020. Nous remercions le CNRS/TGIR HUMA-NUM et le Centre de calcul IN2P3 (Lyon - France) pour la fourniture des ressources informatiques et de traitement des données nécessaires à ce travail ainsi que les trois relecteurs pour tous leurs commentaires.

8. Bibliographie

Amsili P., Seminck O., « Schémas Winograd en français : une étude statistique et comportementale », *TALN 2017*, Orléans, France, p. 28-35, June, 2017.

Balvet A., « Métriques d'évaluation en Traduction Automatique : le sens et le style se laissent-ils mettre en équation ? », *in* T. Milliaressi (ed.), *La Traduction épistémique : entre poésie et prose*, Presses Universitaires du Septentrion, p. 315-356, 2020.

Basta C., Costa-jussà M. R., Fonollosa J. A. R., « Towards Mitigating Gender Bias in a decoder-based Neural Machine Translation model by Adding Contextual Information », *Proc. of the The Fourth Widening Natural Language Processing Workshop*, ACL, Seattle, USA, p. 99-102, July, 2020.

Belinkov Y., Glass J., « Analysis Methods in Neural Language Processing : A Survey », *TACL*, vol. 7, p. 49-72, 04, 2019.

Blodgett S. L., Barocas S., Daumé III H., Wallach H., « Language (Technology) is Power : A Critical Survey of "Bias" in NLP », *ACL*, ACL, Online, p. 5454-5476, July, 2020.

Burlot F., Yvon F., « Evaluating the morphological competence of Machine Translation Systems », *WMT*, ACL, Copenhagen, Denmark, p. 43-55, September, 2017.

Burlot F., Yvon F., « Évaluation morphologique pour la traduction automatique : adaptation au français », *Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN*, p. 61-74, 2018.

Callison-Burch C., Osborne M., Koehn P., « Re-evaluating the role of BLEU in machine translation research », *EACL*, 2006.

Cho W. I., Kim J. W., Kim S. M., Kim N. S., « On Measuring Gender Bias in Translation of Gender-neutral Pronouns », *Proc. of the First Workshop on Gender Bias in Natural Language Processing*, ACL, Florence, Italy, p. 173-181, August, 2019.

Crawford K., « The Trouble with Bias », *Keynote at NeurIPS*, 2017.

Dister A., Moreau M.-L., *Mettre au féminin : guide de féminisation des noms de métier, fonction, grade ou titre*, 3e édition edn, Fédération Wallonie-Bruxelles, 2014.

Escudé Font J., Costa-jussà M. R., « Equalizing Gender Bias in Neural Machine Translation with Word Embeddings Techniques », *Proc. of the First Workshop on Gender Bias in Natural Language Processing*, ACL, Florence, Italy, p. 147-154, August, 2019.

Gonen H., Webster K., « Automatically Identifying Gender Issues in Machine Translation using Perturbations », *EMNLP*, ACL, Online, p. 1991-1995, November, 2020.

Hewitt J., Liang P., « Designing and Interpreting Probes with Control Tasks », *EMNLP*, ACL, Hong Kong, China, p. 2733-2743, November, 2019.

Huddleston R., Pullum G. K. *et al.*, *The Cambridge Grammar of English*, Cambridge University Press, 2002.

Isabelle P., Cherry C., Foster G., « A Challenge Set Approach to Evaluating Machine Translation », *EMNLP*, ACL, Copenhagen, Denmark, p. 2486-2496, September, 2017.

Kreutzer J., Bastings J., Riezler S., « Joey NMT : A Minimalist NMT Toolkit for Novices », *EMNLP, Demonstrations*, ACL, Hong Kong, China, p. 109-114, November, 2019.

Kuczmarski J., Johnson M., Gender-aware natural language translation, Technical report, 2018.

Kudo T., « Subword Regularization : Improving Neural Network Translation Models with Multiple Subword Candidates », *ACL*, ACL, Melbourne, Australia, p. 66-75, July, 2018.

Levesque H., Davis E., Morgenstern L., « The Winograd schema challenge », *KR*, 2012.

Lu K., Mardziel P., Wu F., Amancharla P., Datta A., « Gender bias in neural natural language processing », *Logic, Language, and Security*, Springer, p. 189-202, 2020.

Östling R., Tiedemann J., « Efficient word alignment with Markov Chain Monte Carlo », *Prague Bulletin of Mathematical Linguistics*, vol. 106, p. 125-146, October, 2016.

Ott M., Edunov S., Grangier D., Auli M., « Scaling Neural Machine Translation », *WMT*, ACL, Brussels, Belgium, p. 1-9, October, 2018.

Pearl J., « Direct and Indirect Effects », *UAI*, UAI '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, p. 411–420, 2001.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., « Scikit-learn : Machine Learning in Python », *JMLR*, vol. 12, p. 2825-2830, 2011.

Prates M. O. R., Avelar P. H., Lamb L. C., « Assessing gender bias in machine translation : a case study with Google Translate », *Neural Computing and Applications*, vol. 32, nº 10, p. 6363-6381, 2020.

Renduchintala A., Williams A., « Investigating Failures of Automatic Translation in the Case of Unambiguous Gender », *CoRR*, 2021.

Rudinger R., Naradowsky J., Leonard B., Durme B. V., « Gender Bias in Coreference Resolution », *in* M. A. Walker, H. Ji, A. Stent (eds), *NAACL-HLT*, ACL, p. 8-14, 2018.

Saunders D., Byrne B., « Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem », *ACL*, ACL, Online, p. 7724-7736, July, 2020.

Saunders D., Sallis R., Byrne B., « Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It », *Proc. of the Second Workshop on Gender Bias in Natural Language Processing*, ACL, Barcelona, Spain (Online), p. 35-43, December, 2020.

Savoldi B., Gaido M., Bentivogli L., Negri M., Turchi M., « Gender Bias in Machine Translation », *Transactions of the Association for Computational Linguistics*, vol. 9, nº 0, p. 845-874, 2022.

Sennrich R., Haddow B., Birch A., « Controlling Politeness in Neural Machine Translation via Side Constraints », *NAACL*, ACL, San Diego, California, p. 35-40, June, 2016.

Stanovsky G., Smith N. A., Zettlemoyer L., « Evaluating Gender Bias in Machine Translation », *ACL*, ACL, Florence, Italy, p. 1679-1684, July, 2019.

Vanmassenhove E., Hardmeier C., Way A., « Getting Gender Right in Neural Machine Translation », *EMNLP*, ACL, Brussels, Belgium, p. 3003-3008, October-November, 2018.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. u., Polosukhin I., « Attention is All you Need », *NeurIPS*, p. 5998-6008, 2017.

Vig J., Gehrmann S., Belinkov Y., Qian S., Nevo D., Singer Y., Shieber S., « Investigating Gender Bias in Language Models Using Causal Mediation Analysis », *NeurIPS*, vol. 33, Curran Associates, Inc., p. 12388-12401, 2020.

Winograd T., *Language as a cognitive process : Volume 1 : Syntax*, Addison-Wesley Pub. Co., Reading, MA, 1983.

Zhao J., Wang T., Yatskar M., Ordonez V., Chang K.-W., « Gender Bias in Coreference Resolution : Evaluation and Debiasing Methods », *NAACL*, ACL, New Orleans, Louisiana, p. 15-20, June, 2018.

Survey on Narrative Structure: from Linguistic Theories to Automatic Extraction Approaches

Aman Berhe* — **Camille Guinaudeau*** — **Claude Barras****

* *Université Paris-Saclay, CNRS-LISN*
** *Vocapia Research*

ABSTRACT. *Narration is an essential element in the transmission of written and oral stories and corresponds to both who is telling the story and how the story is told. Philosophers and structuralists have analyzed and defined different narrative structures. Recently, researchers in Machine Learning (ML) or Natural Language Processing (NLP) have been particularly interested on the extraction and understanding of narrative structure in different collections. On this work, we present a survey on theories, research and techniques around narrative structure: from linguistic theories to automatic approaches used for the extraction and analysis of narrative structure in different multimedia collections and annotation tools.*

KEYWORDS: *Narratives, Narratives structure, Multimedia collection.*

RÉSUMÉ. *La narration est un élément essentiel à la transmission d'histoires écrites et orales et correspond à la fois à la personne qui raconte l'histoire et à comment l'histoire est racontée. Philosophes et structuralistes ont analysé et défini différentes structures narratives existantes. Récemment, des chercheurs en apprentissage automatique ou en traitement automatique des langues se sont particulièrement intéressés à l'extraction et à la compréhension de la structure narrative dans divers contenus. Dans ce travail, nous présentons une revue des théories, recherches et techniques autour de la structure narrative. Cette description recouvre à la fois les théories linguistiques et les approches automatiques utilisées pour l'extraction et l'analyse de la structure narrative dans divers contenus multimédias ainsi que les outils d'annotation mis en place pour développer des corpus.*

MOTS-CLÉS : *Narration, Structures narratives, Ressources multimédias.*

1. Introduction

Narrative is a way to tell an information or a story from a particular point of view. It is used to tell stories, facts or scientific results in the form of texts, audios and videos for the purpose of entertainment, education or history preservation. The way narratives progress gradually is referred to as narrative structure. Narratives are audience interactive as they include high level themes related to deep human emotions and create a strong connection and motivation among the audience. This connection is also established through a dramatic element that helps to capture the audience, the narrative-hook, that is a core point of the structure and progress of the narrative. Onega and Landa (2014) defined narrative as the semiotic representation of a series of events meaningfully connected in a temporal and casual way. They further classified films, plays, comic strips, novels, newsreels, diaries, chronicles and treatises of geological history as narratives. Movies, TV series, fictional books, audio recordings that focus on telling a story follow a complex sequence of steps to mesmerize audiences from the start to the end of the intended story. These sequences of steps are referred to "narrative structure".

Since the 19th century, some renowned philosophers (Lucas, 1968), formalists (Propp, 1958) or structuralists (Leach, 1970) have studied narratives and their structural development in an intensive manner, from the point of view of literature. Their work has been the basis for the development of the study of storytelling in general and narrative structure specifically. The importance of narrative is undoubtedly clear on conveying information, entertaining, history preservation or documenting, starting from the pictorial stone scripture of the Stone Age period until the most sophisticated storytelling tools of this current era. That is why the literature-wise study of narratives and their structure was vastly explored.

Narratology (Genette, 1983; Onega and Landa, 2014) is, etymologically, the science of narratives as defined by the structuralists. The concept of narratology has been evolving through the years and now studies the narrative aspects of literary and non literary genres, such as poems, films, drama, history or advertisement. Narratives come in different mediums, linguistic narrative (history, novel, short stories) or audiovisual narratives (films, TV series, TV shows). Each medium allows for a specific presentation of the stories, different points of view, various degrees of narratorial intrusiveness and different time-handling techniques. Consequently, each narrative medium requires an analytical approach to narrative structures.

Narrative structure can be used for the reorganization of a huge collection of multimedia contents (Berhe *et al.*, 2020). Hence, automatic methods should take advantage of the narrative structure found in multimedia contents for better accessibility, management and understanding of these collections. On the other hand, automatic analysis, understanding and extraction of narrative structure may benefit writers of short skits, comedies or folktales to structure their writings and capture their audience. Additionally, it is an important asset for producing long and continuous episodes without contradictions or continuity problems for sequel film makers and TV show producers.

Recommendation systems, more specifically entertainment recommendation systems, may also benefit from the extraction of narrative structures with regard to recommending contents according to narrative structures rather than just meta-data and content similarity. Hence, the giant video content platforms, such as YouTube, Netflix, etc., may use narrative structure for better retrieval and searching. Furthermore, the gaming industry can apply narrative structures on programming different steps and situations that are connected narrative-wise, for example in video games (Vargas, 2017). The education system may also utilize narrative structures to present contents of different courses so that it can be entertaining and engaging for children and teachers at different levels of studies. Finally, social media posts and tweets are also used to identify narratives and understand events that happened and are going to happen in the future (Brogan, 2015; Sadler, 2018).

Computational narrative is the research domain that involves Artificial Intelligence (AI), Machine Learning (ML) or Natural Language Processing (NLP) approaches for automatic extraction or analysis of narratives. The representation, analysis, extraction and manipulation of narrative structure according to the existing narrative theories and narrative structures have been studied. The advancement of NLP research and results in many areas has inspired researchers to continue working on narrative structure understanding using automatic methods. In order to elaborate and evaluate the proposed algorithms, annotations and visualization tools have a vital role. Narratives have been annotated manually and using annotation tools (Finlayson, 2013). Visualization tools made annotation and understanding of narratives less difficult considering automatic methods.

In this survey, we investigated published articles and books that focus on the automatic understanding of narrative and their structures. Besides, available databases and annotations that focus on automatic approaches are described. The papers presented in this work focus on narrative content extraction, analysis and understanding. To our knowledge, this survey is one of its kind. There are surveys (Finlayson, 2013) on short narrative texts, such as folktales, and their focus was on morphologies and linguistic characteristics. However, our survey covers all modalities (textual, audio and visual) of narrative content, and the investigation starts from the narrative theories and gradually develops into the state-of-the-art ML and NLP algorithms used for the understanding of narratives and their structures. Therefore, the survey discusses work from the basic narrative theories until the development of automatic methods for the understanding of narratives and their structure. Hence, we believe this work can pave the way for future studies in automatic processing of narrative contents and facilitate the road with already available methods and resources, and included missing points that can motivate the advancement of research towards narrative structure understanding, extraction, analysis from a machine learning and natural language processing point of view.

The paper is organized in the following manner. Section 2 briefly discusses the history of narrative theories and the evolution of narrative structures' definition. Section 3 presents features utilized to automatically understand narratives and their structures,

based on different modalities. Section 4 describes the available annotated datasets and visualization tools concerning narratives and their structure. Section 5 details algorithms and computational models applied on narrative structure. Finally, Section 6 concludes the paper and recommends possible future research on narrative structure.

2. Short history of narratives

Narratology has been dominated by structuralist approaches since the 1990s, and has been developed into a variety of theories, concepts, and analytical procedures. The term "narratology" was introduced in the structuralist study of narratives by Tzvetan Todorov in 1969. Schmid (2010) believed that narrativity can be identified by two distinct concepts. The first one is the classical narrative theory, long before the term "narratology" was first used, and the second one is the structuralist concept of narrative. Genette (1988) developed a theory of narratological poetics that may be used to address the entire creation of narrative processes in use. Structuralism was further shaped by Lévi-Strauss (1958) who claimed that myths found in various cultures can be interpreted in terms of their repetitive structure, which leads to the study and formulation of narrative structures.

In Aristotle's approach [1], a narrative is classified into three main parts which are the beginning, the middle, and the end (Lucas, 1968; Whalley *et al.*, 1997). The beginning is where the characters and main settings are introduced. In the middle, the conflict starts and the protagonists get acquainted with the problem. At the end, the problem is solved and the life of the protagonists goes back to normal. Many narrative content writers follow Aristotle's three-stage structure in different mediums of narrative, including Hollywood movies (Field, 2009).

Propp (1958) focused on repeated plot elements in his studies of the morphology of folktales, which he called "functions", and their associated character roles. He defined function as "an act of a character, defined from the point of view of its significance for the course of the action". Each function involves a set of characters who filled certain roles, the *dramatis personae* of the morphology. He identified seven *dramatis personae* classes: Hero, Villain, Princess, Dispatcher, Donor, Helper and False Hero. Propp (1958) identified 31 elements of stories that can be categorized into four spheres, namely the introduction, the body of the story, the donor sequence (the sequence of actions by a provider to the hero) and the hero's return, in his studies. This categorization was first developed by Todorov and Weinstein (1969) who state that there are five steps that most narrative stories or plots follow. These are equilibrium (starting the story where the life of characters are normal), disruption (the life of a character or characters is disrupted), realization (characters are informed about the situation and chaos occurs), restored order (characters resolve the disruption) and equilibrium again (equilibrium is restored, new equilibrium).

1. Aristotle's *Poetics*, 347-342 B.C., is a little collection of lecture notes, yet for many centuries it served as the foundation of narrative theory.

Novelist Freytag (1872) developed a narrative pyramid as a description of the narrative structure in fictions. In dramatic narratives, he proposed a dramatic structure containing five parts (Exposition, Rising Action, Climax, Falling Action and Denouement), which were also shared by Todorov. Many films and dramatic fictions use Freytag's pyramid of dramatic sentiments to present narratives of any kind.

Comparatively, many narrative theories and structures share at least one common point which originates from Aristotle's theory. But Todorov and Propp shared most of the steps in their narrative theory and structure, even if they differed on the story and the content of the narrative. They also agreed with Aristotle's structure in a more general way. The beginning step in Aristotle is equivalent to introduction and equilibrium in Propp's and Todorov's narrative theory, respectively. The middle step in Aristotle is equivalent to the body of the story, the donor sequence step, in Propp's theory, and disruption and realization steps, in Todorov's. Finally, the end stage of Aristotle's theory is equivalent to the hero's return and new equilibrium steps, in Propp and Todorov respectively.

In his narrative approach, Lévi-Strauss found out, through his studies of hundreds of myths, that we, as human beings, make sense of the world or the people, or events, as binary opposites (Lévi-Strauss, 1958). He indicated that binary opposites are the center of narratives, so that narratives are organized around the conflict between such opposites (e.g. good vs evil, man vs woman, peace vs war, wisdom vs ignorance, etc.).

Some researchers took advantage of the contents of the narratives and tried to find structure within the contents. For example, Cohn (2013) applied narrative grammar to verbal discourse and movies, claiming that "the narrative structure orders information into a particular pacing, from which a reader can extract a sequence's meaning—both the objects that appear across panels and the events they engage in". Bordwell (2013), Berger (1997), and Chatman (1980), on their side, have proposed to split the narratives depending on the structure and the contents. Bordwell (2013) and Chatman (1980) divided the narrative into *histoire* and *discours*, which literally mean "story" and "discourse", or plot, respectively. "Story" is the content of the narrative. It can also be described as the raw material of dramatic actions which is made up of events, characters, entities, etc. Plot (sometimes called "discourse") is the way a story is presented. Berger (1997) divided the narrative into "fabula" and "syuzhet" which are equivalent to story and plot respectively. Van Dijk (1981) considered episodes to be semantic units of discourse and characterized an episode of a discourse as a specific "sequence of propositions". Furthermore, he specified such a sequence must be coherent according to conditions of textual coherence.

Most of the above narrative theories were established from text books as stories or fairy tales. Research has been done based on these structures and morphologies in the literature domain. Many modern writers used Campbell's theory (Campbell, 2008) of mythological structure of the journey of a hero. Screen and story writers—for movies, theatres or TV series for example—have also different ways of writing the narrative that goes on through the media pieces by pieces. The most common techniques followed by film makers (mainly in Hollywood) are three-act (III-act) and

five-act (V-act) structures (Field, 2009) formed by decomposing the concept of Lucas (1958), Todorov and Weinstein (1969), and Freytag (1872). An act, as defined by McKee (1997), is a "series of sequences that peaks in a climatic scene which causes a major reversal of values, more powerful in its impact than any previous sequence or scene". All the theories and approaches on narrative structure, for both textual and audiovisual contents, can be summarized in Figure 1. Hauge (2011) proposed a structure with six stages known as "Michael Hauge's Six Stage Plot Structure" that is based on the three-act (III-act) narrative structure.

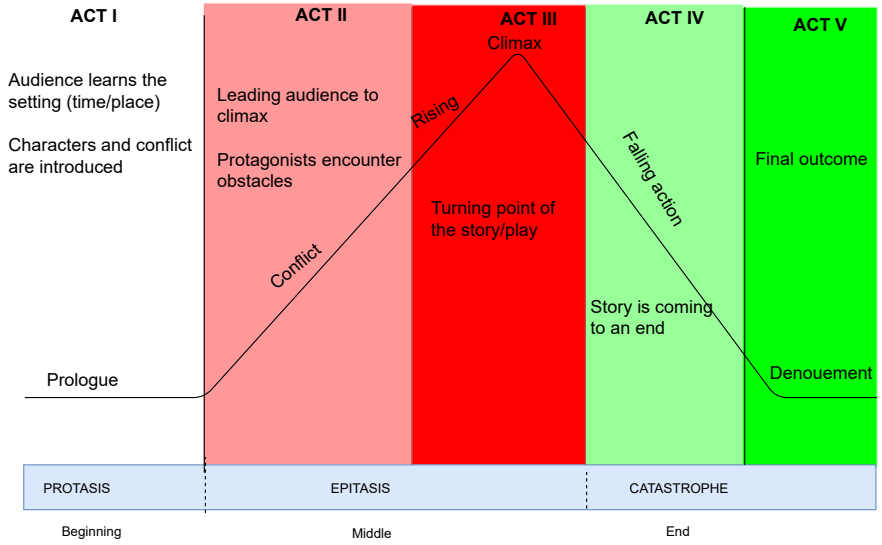


Figure 1. *Narrative structure summary adapted from Freytag's (1872) pyramid.*

Figure 1 divides a narrative into five parts, the V-act structure, but also embeds the three-part structure (III-act). Act I and II in the V-act structure are equivalent to act I of the III-act structure. Act III in the V-act structure is act II in the III-act structure. Act IV and V, in the V-act structure, are equivalent to the act III of III-act structure. The III-act structure (Whalley *et al.*, 1997) is mostly a replica of Aristotle's narrative theory and the V-act structure is coined from Freytag's pyramid (Freytag, 1872).

3. Modalities for analysis and extraction of narrative (structures)

Since the 1970s, narratives and storytelling have been investigated using scientific methods in the field of artificial intelligence for understanding and evaluating human cognition theories (Vargas, 2017; Finlayson, 2012; Andersen and Slator, 1990). The field of computational narrative links the daily human activities (narratives) and the computing world (machines computations) by analyzing and modeling narratives,

narrative understanding and machine-readable representations of narratives with the purpose of enabling computers to tell a story.

The understanding and extraction of narrative structure is a difficult problem for computers due to the complex nature of the different inputs that constitute a narrative and different ways of constructing a narrative structure. Most of the research that focus on the extraction of narratives and their structures use simple stories or folktales in the form of text input. Recent work, however, also focuses on feature films, video shots and long and progressive videos (i.e. TV series). In the context of multimedia collections, narratives can also be carried by several modalities. Researchers have been dealing with different modalities, used jointly or independently, to extract narratives and their structure.

In the following subsections, textual and audiovisual features are presented. Subsection 3.1 presents the works that dealt with the textual modalities of narratives and the features used and Subsection 3.2 introduces audiovisual modalities used for automatic computation of narrative contents.

3.1. *Textual modality*

Most of the research on the extraction of narratives and their structures used simple stories or folktales in the form of text input, based on Vladimir Propp's influential theory of the structure of folktale plots. Finlayson (2012), Valls-Vargas *et al.* (2014), Bod *et al.* (2018) and Malec (2010) focused their study on narrative discourse of Propp's folktales, and Elson (2012) concentrated his work on a set of fables as well as longer texts including literary fiction and epic poetry.

Finlayson (2012), in his work on Russian folktales, uses semantic level descriptions, known as "event timeline", and abstracts them to the next higher level, structures, such as Villainy, Struggle, Victory, and Reward. Event timeline corresponds to the order of the events as expressed in the text, which is different from their temporal order within the story world. It includes straightforward things that happen ("He fought the dragon."), times and dates ("He fought the dragon on Tuesday."), aspectual constructions ("He began to fight the dragon.") and subordinate expressions ("He promised he would fight the dragon, but never did."). The author believed that each of these types of occurrences has a particular meaning in the timeline of the story world. He used specifically Propp's morphology and proposed an Analogical Story Merging (ASM) algorithm that extract plot patterns. The algorithm is discussed in Section 5 in detail.

Malec (2010) uses a semantic markup language, PftML, to automatically analyze and parse folk narratives. The language proposed by Malec (2001) is similar to an XML grammar and enables to decompose tales into Propp's folktales functions.

Lô *et al.* (2020) apply machine learning and natural language processing approaches to identify, analyze and generate West African and West European folktales

by using semantic and syntactic coherence of the narratives. The authors use explicit West African features such as culture-specific protagonists, other characters or objects to appear in the generated texts. They utilize word embeddings and term frequency-inverse document frequency (TF-IDF) to build a character-level and a word-level recurrent neural model. Additionally, they search for recurring patterns that could indicate the existence of a distinctive narrative structure between West African and Western European folktales. To do so, they employ linguistic markers to divide the folktales following the III-act structure and classify the obtained segments.

Concerning literary fictions, in order to model the narrative structure of fables and literary fictions, Elson (2012) extracts textual features vectors based on word distance, type of punctuation and character mention and constructs conversational networks that are then analyzed with regard to literary theories.

Valls-Vargas *et al.* (2014) extract entities from a narrative text and compute a feature vector using both linguistic information and external knowledge (WordNet and ConceptNet) for each extracted entity. The features are computed from the parse tree of a sentence where an entity was found, the sub-tree representing the entity, the leaves of the sub-tree (i.e., word-level tokens with POS tags) and the dependency lists that contain a reference to any node in the entity's sub-tree. The features are automatically extracted thanks to *Voz* [2], a system that explores techniques for automatic extraction of narrative information from text.

Film scripts are also among the studied sources of textual modality for narratives. A lot of them, including several genres, are available and openly accessible in different websites and fan pages, e.g. IMSDb. [3] Murtagh *et al.* (2009) study the narrative structure of the *Casablanca* film script and of six episodes of *CSI* (*Crime Scene Investigation*) TV series. They show that film scripts are important to analyze and understand narrative structures based on the theories of McKee (1997) on principles of screenwriting. To do so, they take all words into account in semi-structured texts that were extracted from film scripts.

Finally, Kim and Monroy-Hernandez (2016) and Barbieri (2007) use the theory of narratives to do task-specific application which are based on textual information. Kim and Monroy-Hernandez (2016) extract social media content based on narrative theory. For that, they annotate the beginning, middle and end of an event according to a narrative structure and then automatically cluster the text into parts based on the sentences of the event. Similarly, in social-media narrative content, Tekiroğlu *et al.* worked on generating counter-narrative (CN) against online hate speech (HS). CN is a non-aggressive response that offers feedback through fact-bound arguments to withstand HS. To this end, Tekiroğlu *et al.* (2020) build a dataset based on quality decomposed into two terms, conformity and diversity (lexical and semantic).

2. https://sites.google.com/site/josepvalls/home/voz.

3. The Internet Movie Script Database, www.imsdb.com.

3.2. *Audiovisual modalities*

The video and audio of a film or a TV series (or any type of multimedia document) are essential and core information narrators. Little research is conducted on audiovisual data for the analysis, extraction, understanding and description of narrative structure from TV series or other multimedia collection.

Nonetheless, some studies focus on the computation of narrative structure using videos of Hollywood movies. Zhao and Ge (2010), for example, work on the calculation of a structure-model for Hollywood movies, applying film-making rules and film grammars to their data. Other works investigate narrative structure in educational videos in order to help and motivate students (Dorai *et al.*, 2003; Phung *et al.*, 2002). Dorai *et al.* (2003) built a hierarchical structure that decomposes the video into sections based on the contents presented to the students. Their method exploits variations in color distributions in frames to separate sections containing some form of text from narrator content. In their study, Phung *et al.* (2002) use three narrative structure parts (narration, conversation/discussion and linkage sections) in the domain of educational videos.

TV series or sequential TV shows can be treated as a large collection of meaningful episodes, scenes or events. They can also be seen as a collection of short pieces of logical narrative units ordered chronologically or casually. The narratives in this kind of large collection can vary in type and length, but this collection has at least one main narrative that goes on from the beginning to the end of the series that can be dynamically captured by creating links between the smallest logical story unit, mostly known as a scene (Bost, 2016; Tapaswi *et al.*, 2014; McKee, 1997; Zhao and Ge, 2010). Narrative units are the smallest portions of a narrative content or story (Berhe *et al.*, 2019; Ercolessi *et al.*, 2011).

As mentioned in Section 2, narrative structure is made up of two components, the plot and the story. Story refers to the raw material of dramatic action and answers the story questions like who, what and where. It also corresponds to the description of settings, characters and events. Plot refers to how the story is told. It is a sequence of events that drives the story forward from beginning to end. It is more concerned with the characters and their interaction and answers the questions like how and when actions/events have occurred. Some researchers (Bost *et al.*, 2016; Tapaswi *et al.*, 2014; Ercolessi *et al.*, 2012; Berhe *et al.*, 2020) rely on such narrative elements (characters interactions, events, entities) to perform narrative structure extraction on TV programs.

Entities are important element of narratives. They refer to the mentions of places, names and organizations. Piskorski *et al.* (2020) use entities to extract events where the target entities have participated or been mentioned for a structured news data.

Events can be used to connect narratives and form a structure of the narrative. They can come in different granularity, for example in a scene, in an episode (mostly in books, films, TV series) or a phrase (mostly in folktales). Chambers and Jurafsky

(2009) propose to learn narrative schema, coherent sequences or sets of events using unsupervised techniques. To do so, they extract chains of events to obtain narratives from a document (Chambers and Jurafsky, 2009; Chambers and Jurafsky, 2008). Similarly, Regneri *et al.* (2010) and Finlayson (2012) propose to learn event scripts from lists of actions using multiple sequence alignment techniques.

Narrative structure extraction can rely on different kind of features coming from the various modalities of the document under consideration: linguistic or syntactic features, audio information (pitch, rythm) or visual clues (shots, movement). In the next section, annotations, available dataset and visualization tools for the different narrative contents are discussed.

4. Annotations and visualization tools for narrative structure

In modeling techniques for narratives understanding, annotation is a necessary step to represent the story from a text to machine-readable format and evaluate the automatic extraction methods developed.

Table 1 presents datasets specifically used for narrative content analysis, extraction and understanding and discussed in this survey. It describes the name of the datasets (some of them do not have a name), the annotation type, size (number of narrative units in the dataset), number of words (number of scenes for video datasets), modality, language and the availability of the dataset (where "available" refers to the online availability of the dataset and "NA" refers to "not available"). Furthermore, some research works have also presented a summary of datasets for narrative contents (see, for example, Doukhan *et al.* (2015)).

Different annotation schemas and environments have been proposed. Some of the most famous and reliable annotation environments are the Story Workbench by (Finlayson, 2008) (2008) and the Scheherazade system by Elson (Elson, 2012). Both dealt with the annotation of folktales and short narrative texts. Finlayson worked on 15 Russian folktales, a subset of Propp's original tales, annotated for 18 aspects of meaning by 12 annotators using the Story Workbench, a general text-annotation tool developed for this work (Finlayson, 2012). Each aspect was annotated by two annotators with an inter-annotator agreement between 0.7 and 0.8. Finally, Sloetjes and Wittenburg (2008) provide a multimedia annotation tool, ELAN, which makes it possible to annotate multiple category of annotations on the same multimedia documents.

The annotation of narrative documents, particularly multimedia narratives, is a very time-consuming task. Li *et al.* (2018) and Eisenberg and Finlayson (2019) have worked on the annotation of narrative elements of short stories in two different ways. Li *et al.* (2018) produce a guideline to annotate directly the narrative structure based on Freytag's (1872) pyramid, and Eisenberg and Finlayson (2019) provide a guideline for narrative characteristics annotation to collect human judgments on narrative characteristics.

Dataset	Annotation	Size	# of words or scenes (*)	Modality	Lang.	Access
GV-LEx (Doukhan *et al.*, 2015)	Lexical and structural	101	66,935	text/speech	Fr	online
Propp's folktales (Bod *et al.*, 2018)	Lexical and structural	450	-	text	En	NA
(Lô *et al.*, 2020)	West African & Western European tales	742	406,403	text	En	online
French tales (Garcia-Fernandez *et al.*, 2014)	Narrative classification	107	85,600	text	Fr	online
Russian folktales	Linguistic	15	18,862	text	En	online
Movies (Guha *et al.*, 2015)	ACT boundary	9	27*	video	En	NA
ScriptBase (Gorinski and Lapata, 2015)	Summaries, loglines & taglines	1,276	-	text	En	NA
NarrativeQA (Kočiskỳ *et al.*, 2018)	Question answering	1,572	15,406	text	En	online
FSD (Liu *et al.*, n.d.)	Stories for each scene	60	1,569*	video	En	online
TRI-POD (Papalampidi *et al.*, 2019)	Turning points of screen-play	99	54,600	text	En	online
CSI (Frermann *et al.*, 2018)	Entities	39	>500,000	text/video	En	online
Casablanca (Murtagh *et al.*, 2009)	Structural	1	77*	text	En	NA

Table 1. *Summaries of available datasets regarding narrative content. "*" shows the number of scenes available in a dataset when the number of words are not present and the dataset is described only by scenes.*

Garcia-Fernandez *et al.* (2014) propose the digitization and annotation of a tales corpus from a narrative point of view (only the French tales corpus is available) and classify it according to the Aarne & Thompson (1961) narrative classification of folktales. Doukhan *et al.* (2015) provided GV-LEx, a corpus of French folktales annotated using textual and audio modalities, during their study of the relationships between the textual structures of tales and speech prosody. They annotated 89 text and 12 speech corpora with the targeted application of an expressive text-to-speech synthesis system embedded in a humanoid robot. They performed lexical level (extended definitions of enumerations, time, place and person named entities, and part of speech (PoS) tags) and supra-lexical level (the segmentation of tales into a sequence of episodes, the localization and attribution of direct quotations, together with tale protagonists coreferences) annotations. Lô *et al.* (2020) provide two corpora of West African and Western European folktales that are used in three experiments on cross-cultural folktales analysis. They collected a total of 742 English narratives, 252 of which were West African, and the other 490 Western European, to predict the next words that continues the narrative based on an input seed from the narratives in the corpora. The West African folktales were written by authors from Anglophone West African countries such as Gambia, Ghana, and Nigeria, while the Western European folktales were from countries such as Netherlands, Germany, France, and the UK (Lô *et al.*, 2020). Finlayson (2013) presents a survey of corpora for the advancement of scientific understanding of narrative. He identifies 167 unique text collections (155 with some sort of annotation) that could be considered a "corpus" and contained 17 different broad types of narratives, 5 different modalities, and approximately 42 different types of annotations. From the annotations presented, the most common and complex are events, named entities, and roles annotations.

Concerning the annotation of multimedia documents, Ercolessi *et al.* (2011), Bost (2016) and Liu *et al.* (2020) have annotated several seasons of TV series with scene boundaries. Bost annotated 5 seasons of *Game of Thrones*, 2 seasons of *Breaking Bad* and 1 season of *House of Cards*. Ercolessi *et al.* annotated *Buffy The Vampire Slayer* and *Mac and Alice*. Liu *et al.* collected 60 episodes of *The Flintstones* TV series (which are composed of 1,569 scenes) and annotated the dataset into story. To this end, 105 undergraduate engineering students in data science were invited to annotate the scene labels and each student annotated 4 episodes. They have provided that dataset as Flintstones Scene Dataset (*FSD*).[4] Liu *et al.* constructed the dataset on the assumption of the "three-act" structure (see Figure 1). Furthermore, Tapaswi *et al.* (2014) annotated face tracks, shots and scene boundaries, script-to-video alignment in *The Big Bang Theory* TV series, and some story-line in *Game of Thrones*. Papalampidi *et al.* (2019) developed the TuRnIng POint Dataset (TRI-POD)[5] composed of 99 annotated screenplays. Their work focused on identifying turning points of screen plays based on textual information. Similarly, Frermann *et al.* (2018) built a dataset of

4. Flintstones Scene Dataset (FSD) is available at https://github.com/llafcode/The_FSD_dataset.git.
5. https://github.com/ppapalampidi/TRIPOD.

episodes of *Crime Scene Investigation* TV series [6] for natural language understanding. The dataset is composed of 39 episodes (seasons 1-5) with screenplays and entities annotations (perpetrator/s in a crime scene). For these annotations, they hired three post-graduate annotators that were not regular fans of the TV series.

Movies have also been used as main sources of audiovisual and linguistic analysis. Guha *et al.* (2015) annotated 9 movies according to the III-act narrative structure. They used film experts to annotate 2 act boundaries as they believe that accurate detection of act boundaries requires knowledge of screenwriting and narrative structure. The dataset ScriptBase (Gorinski and Lapata, 2015) compiles a collection of 1,276 movie scripts (from 1909 to 2013) divided into 23 genres; each movie is on average accompanied by 3 user summaries, 3 loglines (one-sentence summary of a movie), and 3 taglines (short snippets used to promote a movie). Kočiskỳ *et al.* (2018) developed a dataset, NarrativeQA, of stories on books (collected from project Gutenberg [7]) and movie scripts based on question answering using summaries. NarrativeQA is composed of 1,572 stories, evenly split between books and scripts, and 46,765 question-answer pairs. Lewis *et al.* (2017) collected a large scale dataset of 10,945 subtitles files associated with movies metadata that were pre-processed so subtitles contain only linguistic information. Finally, the last movie dataset was proposed by Murtagh *et al.* (2009). It is composed of the *Casablanca* film script divided into 77 successive scenes. The source text for the 77 scenes, containing in total 6,710 words, including metadata, varies between 5 and 1,017 words.

In order to visualize important information or evaluate automatic systems, visualization and annotation tools were proposed: StoryFlow, StoryGraph, Yarn, MovieGraph and StoryCake. StoryFlow (Liu *et al.*, 2013), StoryGraph (Tapaswi *et al.*, 2014) and Yarn (Padia *et al.*, n.d.) were developed to visualize a succession of events in a narrative using merging and diverging timelines, with the temporal continuity of these events in mind and less concern about the exactitude of their temporal placement. MovieGraph (Vicol *et al.*, 2018) proposed a graph-based visualization of a video clip for the annotation and visualization of social situations in a movie clip. Kim *et al.* (2017) developed a visualization technique to explore non-linear narratives in movies. They introduced Story Explorer, an interactive tool that visualizes narrative patterns of a movie via portraying events of a story in chronological order. Story Explorer displays a story curve together with information such as characters and settings. Finally, StoryCake (Qiang *et al.*, 2017) proposed a hierarchical plot visualization method according to the story elements and the hierarchical relationships of entities. Table 2 describes annotation and visualization tools used in narrative contents.

The available datasets and their annotations discussed above are used in different tasks of automatic narrative content extraction, analysis and understanding. In the next section, the algorithms designed to obtain the narrative structure and better processing of narratives from various types of narrative contents are described.

6. CSI dataset is available at https://github.com/EdinburghNLP/csi-corpus.
7. http://www.gutenberg.org/.

Tool	Purpose	Source	Description
StoryFlow	Visualization	(Liu *et al.*, 2013)	Visualization of sequences of events
StoryGraph	Visualization	(Tapaswi *et al.*, 2014)	Visualization of sequences of events
Yarn	Visualization	(Padia *et al.*, n.d.)	Visualization of sequences of events
MovieGraph	Visualization & Annotation	(Vicol *et al.*, 2018)	Visualization and annotation of social situations in a movie clip
Story Explorer	Visualization	(Kim *et al.*, 2017)	Interactive display of a story curve with different narrative elements information
StoryCake	Visualization	(Qiang *et al.*, 2017)	Plot visualization of a story using entity relationships
Story-Workbench	Annotation	(Finlayson, 2008)	Annotation of folktales and short narrative texts for many linguistic aspects of narratives
Scheherazade	Annotation	(Elson, 2012)	Annotation of folktales and short narrative texts for many linguistic aspects of narratives
ELAN	Annotation	(Sloetjes and Wittenburg, 2008)	Annotation of multiple categories on the same multimedia document

Table 2. *Summary of available annotation and visualization tools used in narrative documents.*

Finally, few research papers have used recent natural language processing (NLP) methods, such as attention mechanisms, for the understanding of narrative content. For example, Lô *et al.* (2020) train a bag of words (BoW), an LSTM (long-short term memory neural network), and other classifiers in order to identify, analyze, and generate West African folktales for better extraction of the organization of folktales according to the III-Act narrative structure.

5. Methods of automatic narrative structure extraction

Natural language Processing (NLP) has been vastly investigated for the understanding and extraction of important information from textual documents. Understanding and analysis of narratives and their structure has been dealt according to multiple tasks using different machine learning and natural language processing approaches.

Table 3 summarizes the main tasks and describes the algorithms, type of narrative content and modalities used in the papers. The following subsections discuss the main

Task	Algorithm	Narrative	Modality
Representation (2)	Co-occurrence, bootstrapping	Fables	Text
Semantic analysis (3)	Correspondence analysis		Text
Content Extraction (7)	PMI, CRF, decision-tree, NER, graph analysis	Movies, speech, folktales	Text, Speech, Video
Event chains (4)	PMI, NER	Folktales	Text
Morphology (Propp's) (2)	ASM	Folktales	Text
Decomposition & linking (10)	Community clustering, recursive algorithms	Movies & TV shows	Video
Hyperlinking (10)	NER, classification	Social media, folktales, movies	Text, Video

Table 3. *Summary of tasks on narratives and algorithms used. Numbers in the parenthesis, in the task column, refers to the number of articles included in this work that use the task on narrative contents.*

methods (some of the methods overlap with each other) used to understand and extract narratives and their structure.

5.1. *Key narrative elements extraction*

Narrative elements of narrative contents constitute an important information. One way of using narrative elements for a better understanding of narratives is the creation of a character network, i.e. a graph that illustrates the interaction of the characters, as suggested in the survey of fictional character networks (Labatut and Bost, 2019). The extraction of social networks (Agarwal and Rambow, 2010; Elson, 2012; Bost *et al.*, 2016) explaining the connection of characters to understand the story between them showed promising results. Valls-Vargas *et al.* (2017) build a graph which captures the narrative entities, such as characters, organization and places, and in turn depict the story between the entities, referred to as story graph. Similarly, Bost *et al.* (2016) take advantage of the plot properties of narratives in TV series to construct a character network and represent the dynamics of the characters.

Content extraction techniques have been used to extract important features of events, stories or entities (Chen *et al.*, 2015; Tapaswi *et al.*, 2015; Yu *et al.*, 2016; Arnulphy *et al.*, 2015; Ghannay *et al.*, 2018). As an example, Tapaswi *et al.* (2015) work on the alignment of plot synopsis to video to provide story-based retrieval from videos. To do so, they consider shots and sentences as atomic units and extract named entities from the text and person identification from the video to create alignments between

synopsis and sets of shots (scenes). Arnulphy *et al.* (2015) work on event extraction from textual documents in the TimeML [8] challenges for the French and English languages. They used event descriptors to assign every word to a label that indicates whether it is an event or not by using conditional random field (CRF) and decision-tree based algorithms. Ghannay *et al.* (2018) use an end-to-end entity recognition (NER) approach for a slot filling task which is a semantic concept extraction in speech, in the framework of a human/machine spoken dialog dedicated to hotel booking.

Chambers and Jurafsky (2009), Vargas (2017) and Finlayson (2012) describe in their research works the importance of event chains, that is to say, connections of events found in a narrative. Valls-Vargas *et al.* (2014) extract entities from events and compute a feature-vector for each of them using linguistic information and external knowledge. They propose the *Continuous Jaccard* measure to estimate the similarity between the extracted entities to decide if they represent characters or not. Chambers and Jurafsky (2008) extract narrative event chains from raw texts through a three-step process. First, they learn basic information about the narrative chain (the protagonists and constituent sub-events). Then, they use the Pointwise Mutual Information (PMI) measure to compute how often events share grammatical arguments. Finally, they build a global narrative score such that all events in the chain provide feedback on the event. In other words, they find the next most likely event to occur, given all narrative events in a document, by maximizing the PMI score. Chambers and Jurafsky (2009) learn narrative schemas, coherent sequences or sets of events (e.g. arrested (Police, Suspect), convicted (Judge, Suspect)) using unsupervised techniques based on co-referencing arguments in chains of verbs. Finally, Reagan *et al.* (2016) use plot sequences or event sequences to construct the story arcs of English fiction books.

Finlayson (2012) introduce the Analogical Story Merging (ASM) algorithm, based on Bayesian model merging, a machine learning technique for learning regular grammars, in order to learn Propp's morphology. To do so, he takes descriptions at the semantic level and abstracts them to a higher level, i.e. structures such as Villainy, Struggle, Victory and Reward. Model merging used to derive a regular grammar is the foundation for ASM with two key differences: filtering and analogical mapping. The filtering process constructs another model from the final merged model from which all states that do no meet certain criteria are removed. The states that survive this cutting become the alphabet, or for Propp's morphology, the functions. In the work of Finlayson (2012), it is shown that the ASM algorithm learns a big part of Propp's theory of folktales structure.

5.2. *Decomposition*

Another approach to extract narrative (structure) consists in the decomposition of the stories contained in the narrative into story-lines and the re-connection of the decomposed stories. For example, Park *et al.* (2012) work on detecting some story lines,

8. ISO-TimeML is an International Standard for time and event markup, and annotation.

organized around characters, from narratives of a movie. They propose a Character-net that can represent the relationships between characters using dialogues, and a method that can extract the sequences via clustering communities of characters based on heuristic algorithm. Many research works (Li *et al.*, 2001; Zhao and Ge, 2010; Adams *et al.*, 2002; Guha *et al.*, 2015) also decompose movies into acts using computational methods for better understanding of the narrative act boundaries and the semantics of a narrative in movies. Adams *et al.* (2002) study film grammar and decomposition of a movie with the goal of automatically locating dramatic events and section boundaries. In their work, they are able to reconstruct the dramatic development of films, focusing on the filmmakers' point of view. Film grammars or Hollywood film-making strategies can work on full movies and standalone episodes of TV series. They use the attributes of motion and shot length to define and compute a measure of the tempo of a movie. They applied Deriche's recursive filtering algorithm to detect the edges of a movie to locate significant tempo pace changes. Similarly, Guha *et al.* (2015) deconstruct a movie into a III-act structure (act I (exposition), act II (conflict) and act III (resolution)) thanks to a popular movie grammar, followed by most screen writers. They detect the act boundaries based on the knowledge of film grammar and features from three modalities (text, video, music). They address the problem of automatically detecting the III-act narrative structure in movies in an unsupervised manner. They cast the problem as one dimensional edge detection in the story intensity curve P and used probability distribution to find the act boundaries. Finally, Lee *et al.* (2021) decompose narrative multimedia plots into story-lines based on the estimation of the personality of the characters. They estimate a character's personality from the average length of dialogues and the ratio of out-degree for in-degree in a graph of characters relationships.

5.3. *Hyperlinking*

When the narrative document (movie, fiction, TV series or TV shows) is quite large, it becomes very complicated to extract its narrative. Character interactions and stories that flow through, from the beginning till the end, are intertwined. Therefore, in order to better understand the narratives and their structure from a large collection, documents need to be reorganized in a more sensible way and in smaller narrative units such as scenes. This can be done by creating links between the narrative units according to a narrative point of view. Many researchers try to link multimedia documents, coined by the term multimedia hyperlinking (Bois *et al.*, 2017a; Bois *et al.*, 2017c; Budnik *et al.*, 2018; Chaturvedi *et al.*, 2018).

Multimedia hyperlinking is a way to navigate videos inside a collection by jumping from one video to another, using different techniques. Bois *et al.* (2015); Ordelman *et al.* (2015); Kim and Monroy-Hernandez (2016), etc. design some linking categories or typologies for multimedia hyperlinking and build graphs to easily explore news by following links that lead to similar news. Kim and Monroy-Hernandez (2016) use narrative theory as a framework to identify the links between social media contents. They

first identify and fill narrative gaps in a social media record, then they link content to narrative categories with respect to storytelling roles. Ordelman *et al.* (2015) present a video-hyperlinking method based on named entity identification. They investigate an unconstrained, multimodal perspective on the identification of anchor points, and a perspective based on the detection of entities from existing metadata and/or automatic audiovisual analysis.

Graph based hyperlinking methods have also shown promising results (Vicol *et al.*, 2018; Li *et al.*, 2017; Valls-Vargas *et al.*, 2017; Bois *et al.*, 2017b; Ercolessi *et al.*, 2012). Valls-Vargas *et al.* (2017) and Vicol *et al.* (2018) produce graphs of stories using narrative elements such as entities. To do so, Valls-Vargas *et al.* (2017) use co-referenced entities, entities and the role of the entity to build their story graph from short textual documents (Russian folktales in English). They extract entities and classify their roles and finally build the graph with the help of a common sense knowledge database. On their side, Vicol *et al.* (2018) propose a method for querying videos and text with graphs, and show that: first, their graphs contain rich and sufficient information to summarize and localize each scene and second, subgraphs allow them to describe situations at an abstract level and retrieve multiple semantically relevant situations. Their graphs capture people's interactions, emotions and motivations which must be inferred from a combination of visual cues and dialog. Vicol *et al.* (2018) use 8 different node types (such as characters, relationship, topics, etc.) while constructing the graph and prepare the Moviegraph dataset. Bois *et al.* (2017b) generate links between news documents surrounding a specific event. They propose a set of intuitive properties that a graph should exhibit to be explorable and used nearest neighbor approaches to create the graph.

Goyal *et al.* (2010) explore NLP techniques to automatically generate plot unit representation. To this end, they develop a tool that produces plot unit representations, known as AESOP, and use it to affect projection rules in order to connect situations with the respective characters. They identify affect states and map the affect states onto characters in a story using "projection rules". They use co-occurrence with Evil/Kind Agent patterns, and bootstrapping over conjunctions of verbs.

5.4. *Clustering*

Long and progressive multimedia contents, such as TV series, TV shows or documentaries have interesting intertwined narratives and they follow different narrative structure. In TV series, Ercolessi *et al.* (2012), Bost *et al.* (2016) and Chaturvedi *et al.* (2018) linked scenes using the concept of multimedia hyperlinking and used these links to tie different videos together and recreate one whole narrative. Chaturvedi *et al.* (2018) identify instances of similar narratives from a collection of narrative texts of movies. They found correspondences between narratives in terms of plot events and resemblances between characters and their social relationships. They coin the term story-kernel to quantify the correspondence similarity. Ercolessi *et al.* (2012) apply clustering methods for plot de-interlacing. They aim at grouping semantically related

scenes into stories or sub-stories of a TV series episode (*Ally McBeal* and *Malcolm in The Middle*). They cluster scenes using traditional agglomerative clustering and graph based community detection algorithm, known as Louvain (Blondel *et al.*, 2008), to group scenes of the TV series.

Hierarchical clustering is one of the machine learning algorithms used to capture the semantics of narrative documents. For example, Murtagh *et al.* (2009) use hierarchical clustering through a sequence of agglomerations of successive scenes or temporal segments or intervals of successive scenes. They take into account the sequential nature of the scenes, ensured through the requirement that agglomerations must be adjacent. The scenes are compared through Correspondence Analysis using euclidean embeddings of film scripts. Murtagh *et al.* (2009) take all words into account in the semi-structured texts that composed film scripts. Each scene is cross-tabulated by the set of all words so that, in this cross-tabulation table, at the intersection of scene i and word j, the value corresponds to a presence (1) or absence (0) value.

6. Conclusion

Narrative structure is an important pattern found in narratives contents (text, audio, video) and AI, ML and NLP approaches have been used to extract, analyze and represent it. In this work, we have presented different kinds of narrative theories and structures, starting from Aristotle. Most of the structures can be summarized using the III-Act (Three stage) more general structure and the V-Act (Five stage) more specific one, see Figure 1. Folktales have been the main source of studies on narratives and their structures using computational methods. However, due to the fast and ever growing multimedia contents research have also been dealing with films, TV shows or TV series, recently. To this end, research relies on textual and audiovisual modalities as well as narrative elements (characters, entities or events) extracted from these modalities.

Machine learning and NLP algorithms, such as clustering, correspondence analysis and deep learning are used for the analysis and understanding of narrative and their structures in different narrative content modalities (textual and audiovisual). Recently, pre-trained deep learning models (including language models) are applied to represent and extract narratives for different purposes.

Annotation and visualization tools have been developed to alleviate the problem of the lack of annotated corpora for narrative structure analysis. Visualization tools can also provide easier overviews of narratives for a better understanding and representation of very long and complicated narratives. However, there is no standard dataset for narrative structure besides the Propp's for folktales (Finlayson, 2013). In order to accelerate the research on narrative structure extraction or analysis, the release of a (multimodal) corpora is, therefore, necessary. Annotation guidelines presented should be standardized so that every one could annotate any narrative content.

Concerning applications, the domains of audiovisual content analysis, summarization, event extraction and recommendation could benefit greatly from the extraction of narrative structure as a narrative structure could quickly and effectively represent a content. Therefore narrative elements could be used for the task of multimedia summarization, video recommendation and production of narrative contents (either multi-modal or mono-modal).

In our studies, only few research papers use deep learning approaches and the current advancement of NLP algorithms and pre-trained models, such as BERT (Devlin *et al.*, 2019) and GPT (Floridi and Chiriatti, 2020). A recent work on the detection of the most salient scenes within a set of semantically related scenes, Berhe *et al.* (2020), shows that time distributed LTSM produce promising results. We believe that current advancements of NLP may greatly help the future works on narrative content extraction, analysis and understanding. Furthermore, state-of-the-art neural network architectures, such as transformers, that have remarkable advancement on NLP, could take narrative content understanding to the next level.

7. References

Adams B., Dorai C., Venkatesh S., "Toward Automatic Extraction of Expressive Elements from Motion Pictures: Tempo", *IEEE Transactions on Multimedia*, vol. 4, n° 4, p. 472-481, 2002.

Agarwal A., Rambow O., "Automatic detection and classification of social events", *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, p. 1024-1034, 2010.

Andersen S., Slator B. M., "Requiem for a theory: the 'story grammar'story", *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 2, n° 3, p. 253-275, 1990.

Arnulphy B., Claveau V., Tannier X., Vilnat A., "Supervised Machine Learning Techniques to Detect TimeML Events in French and English", *Proceedings of the 2015 International Conference on Applications of Natural Language to Information Systems*, p. 19-32, 2015.

Barbieri M., Automatic summarization of narrative video, PhD thesis, Eindhoven University, 2007.

Berger A. A., *Narratives in Popular Culture, Media, and Everyday Life*, Sage, 1997.

Berhe A., Barras C., Guinaudeau C., "Video Scene Segmentation of TV Series Using Multimodal Neural Features", *Series-International Journal of TV Serial Narratives*, vol. 5, n° 1, p. 59-68, 2019.

Berhe A., Guinaudeau C., Barras C., "Scene Linking Annotation and Automatic Scene Characterization in TV Series.", *Proceedings of the 2020 Text2Story Workshop at European Conference on Information Retrieval*, p. 47-53, 2020.

Blondel V. D., Guillaume J.-L., Lambiotte R., Lefebvre E., "Fast unfolding of communities in large networks", *Journal of statistical mechanics: theory and experiment*, vol. 2008, n° 10, p. P10008, 2008.

Bod R., Fisseni B., Kurji A., Löwe B., "Objectivity and Reproducibility of Proppian Narrative Annotations", *Proceedings of the 3rd Workshop on Computational Models of Narrative*, p. 15-19, 2018.

Bois R., Gravier G., Jamet E., Morin E., Robert M., Sébillot P., "Linking multimedia content for efficient news browsing", *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, p. 301-307, 2017a.

Bois R., Gravier G., Jamet E., Robert M., Morin E., Sébillot P., "Language-based construction of explorable news graphs for journalists", *Proceedings of the Workshop on Natural Language Processing meets Journalism in Empirical Methods in Natural Language Processing*, p. 31-36, 2017b.

Bois R., Gravier G., Sébillot P., Morin E., "Vers une typologie de liens entre contenus journalistiques", *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, p. 515-521, 2015.

Bois R., Vukotić V., Simon A.-R., Sicre R., Raymond C., Sébillot P., Gravier G., "Exploiting multimodality in video hyperlinking to improve target diversity", *Proceedings of the International Conference on Multimedia Modeling*, p. 185-197, 2017c.

Bordwell D., *Narration in the fiction film*, Routledge, 2013.

Bost X., A Storytelling Machine?: Automatic Video Summarization: the Case of TV Series, PhD thesis, Université d'Avignon, 2016.

Bost X., Labatut V., Gueye S., Linares G., "Narrative smoothing: dynamic conversational network for the analysis of TV series plots", *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, p. 1111-1118, 2016.

Brogan M. K., "How Twitter Is Changing Narrative Storytelling: a Case Study of the Boston Marathon Bombings", *Elon Journal of Undergraduate Research in Communications*, vol. 6, nº 1, p. 28-47, 2015.

Budnik M., Demirdelen M., Gravier G., "A study on multimodal video hyperlinking with visual aggregation", *Proceedings of the 2018 IEEE International Conference on Multimedia and Expo*, p. 1-6, 2018.

Campbell J., *The Hero with a Thousand Faces*, New World Library, 2008.

Chambers N., Jurafsky D., "Unsupervised learning of narrative event chains", *Proceedings of the 2008 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 789-797, 2008.

Chambers N., Jurafsky D., "Unsupervised learning of narrative schemas and their participants", *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, p. 602-610, 2009.

Chatman S. B., *Story and Discourse: Narrative Structure in Fiction and Film*, Cornell University, 1980.

Chaturvedi S., Srivastava S., Roth D., "Where have i heard this story before? identifying narrative similarity in movie remakes", *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 673-678, 2018.

Chen Y., Xu L., Liu K., Zeng D., Zhao J., "Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks", *Proceedings of the Annual Meeting of the Association for*

Computational Linguistics and the International Joint Conference on Natural Language Processing, p. 167-176, 2015.

Cohn N., "Visual narrative structure", *Cognitive science*, vol. 37, n° 3, p. 413-452, 2013.

Devlin J., Chang M.-W., Lee K., Toutanova K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the 2019 Meeting of the North American Chapter of the Association for Computational Linguistics*, p. 4171-4186, 2019.

Dorai C., Oria V., Neelavalli V., "Structuralizing Educational Videos Based on Presentation Content", *Proceedings of the 2003 IEEE International Conference on Image Processing*, p. II-1029, 2003.

Doukhan D., Rosset S., Rilliard A., d'Alessandro C., Adda-Decker M., "The GV-LEx Corpus of Tales in French", *Language Resources and Evaluation*, vol. 49, n° 3, p. 521-547, 2015.

Eisenberg J., Finlayson M. A., "Annotation Guideline No. 1: Cover Sheet for Narrative Boundaries Annotation Guide", *Journal of Cultural Analytics*, vol. 4, n° 3, p. 11199, 2019.

Elson D. K., *Modeling narrative discourse*, Columbia University, 2012.

Ercolessi P., Bredin H., Sénac C., Joly P., "Segmenting TV series into scenes using speaker diarization", *Proceedings of the Workshop on Image Analysis for Multimedia Interactive Services*, p. 13-15, 2011.

Ercolessi P., Sénac C., Bredin H., "Toward plot de-interlacing in tv series using scenes clustering", *Proceedings of the 10th international workshop on content-based multimedia indexing*, p. 1-6, 2012.

Field S., *Selling a Screenplay: The Screenwriter's Guide to Hollywood*, Delta, 2009.

Finlayson M. A., "Collecting Semantics in the Wild: The Story Workbench.", *Proceedings of the AAAI Fall Symposium: Naturally-Inspired Artificial Intelligence*, p. 46-53, 2008.

Finlayson M. A., Learning Narrative Structure from Annotated Folktales, PhD thesis, Massachusetts Institute of Technology, 2012.

Finlayson M. A., "A Survey of Corpora in Computational and Cognitive Narrative Science", *Sprache Und Datenverarbeitung*, vol. 37, n° 1–2, p. 113–141, 2013.

Floridi L., Chiriatti M., "GPT-3: Its Nature, Scope, Limits, and Consequences", *Minds and Machines*, vol. 30, n° 4, p. 681-694, 2020.

Frermann L., Cohen S. B., Lapata M., "Whodunnit? Crime Drama as a Case for Natural Language Understanding", *Transactions of the Association for Computational Linguistics*, vol. 6, n° 1, p. 1-15, 2018.

Freytag G., *Die Technik des Dramas*, Autorenhaus Verlag, 1872.

Garcia-Fernandez A., Ligozat A.-L., Vilnat A., "Construction and Annotation of a French Folkstale Corpus", *Proceedings of the 2014 International Conference on Language Resources and Evaluation*, p. 2430-2435, 2014.

Genette G., *Narrative discourse: An essay in method*, vol. 3, Cornell University Press, 1983.

Genette G., *Narrative Discourse Revisited*, Cornell University Press, 1988.

Ghannay S., Caubrière A., Estève Y., Camelin N., Simonnet E., Laurent A., Morin E., "End-To-End Named Entity and Semantic Concept Extraction from Speech", *Proceedings of the 2018 IEEE Spoken Language Technology Workshop*, p. 692-699, 2018.

Gorinski P., Lapata M., “Movie Script Summarization as Graph-Based Scene Extraction”, *Proceedings of the Annual Meeting of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 1066-1076, 2015.

Goyal A., Riloff E., Daumé III H., “Automatically Producing Plot Unit Representations for Narrative Text”, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, p. 77-86, 2010.

Guha T., Kumar N., Narayanan S. S., Smith S. L., “Computationally Deconstructing Movie Narratives: an Informatics Approach”, *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 2264-2268, 2015.

Hauge M., *Writing Screenplays That Sell*, Bloomsbury Publishing, 2011.

Kim J., Monroy-Hernandez A., “Storia: Summarizing social media content based on narrative theory using crowdsourcing”, *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, p. 1018-1027, 2016.

Kim N. W., Bach B., Im H., Schriber S., Gross M., Pfister H., “Visualizing Nonlinear Narratives with Story Curves”, *Proceedings of the 2017 IEEE Transactions on Visualization and Computer Graphics*, p. 595-604, 2017.

Kočiskỳ T., Schwarz J., Blunsom P., Dyer C., Hermann K. M., Melis G., Grefenstette E., “The Narrativeqa Reading Comprehension Challenge”, *Transactions of The Association for Computational Linguistics*, vol. 6, nº 1, p. 317-328, 2018.

Labatut V., Bost X., “Extraction and Analysis of Fictional Character Networks: A Survey”, *ACM Computing Surveys*, vol. 52, nº 5, p. 1-40, 2019.

Leach E. R., *Claude Lévi-Strauss*, Viking Press, 1970.

Lee O.-J., You E.-S., Kim J.-T., “Plot Structure Decomposition in Narrative Multimedia by Analyzing Personalities of Fictional Characters”, *Applied Sciences*, vol. 11, nº 4, p. 1645, 2021.

Lévi-Strauss C., *Anthropologie Structurale*, Plon Paris, 1958.

Lewis R. J., Grizzard M., Lea S., Ilijev D., Choi J.-A., Müsse L., O’Connor G., “Large-Scale Patterns of Entertainment Gratifications in Linguistic Content of US Films”, *Communication Studies*, vol. 68, nº 4, p. 422-438, 2017.

Li B., Cardier B., Wang T., Metze F., “Annotating High-Level Structures of Short Stories and Personal Anecdotes”, *Proceedings of the 11th International Conference on Language Resources and Evaluation*, p. 1-7, 2018.

Li R., Tapaswi M., Liao R., Jia J., Urtasun R., Fidler S., “Situation Recognition with Graph Neural Networks”, *Proceedings of the 2017 IEEE International Conference on Computer Vision*, p. 4173-4182, 2017.

Li Y., Ming W., Kuo C. J., “Semantic Video Content Abstraction Based on Multiple Cues”, *Proceedings of the 2001 IEEE International Conference on Multimedia and Expo*, p. 159-159, 2001.

Liu C., Shmilovici A., Last M., “Towards Story-Based Classification of Movie Scenes”, *PloS one*, vol. 15, nº 2, p. e0228579, n.d.

Liu S., Wu Y., Wei E., Liu M., Liu Y., “StoryFlow: Tracking The Evolution of Stories”, *Proceedings of the IEEE Transactions on Visualization and Computer Graphics*, p. 2436-2445, 2013.

Lô G., de Boer V., van Aart C. J., "Exploring West African folk narrative texts using machine learning", *Information*, vol. 11, n° 5, p. 236, 2020.

Lucas D. W., "Aristotle Poetics", *The Classical Review*, 1968.

Malec S., "Autopropp: Toward the Automatic Markup, Classification, and Annotation of Russian Magic Tales", *Proceedings of the 2010 International Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts*, p. 68-74, 2010.

Malec S. A., "Proppian Structural Analysis and XML Modeling", *Computers, Literature and Philology*, vol. 4, n° 1, p. 68-74, 2001.

McKee R., *Story: Substance, Structure, Style and the Principles of Screenwriting*, Harper-Collins Publishers, 1997.

Murtagh F., Ganz A., McKie S., "The structure of narrative: the case of film scripts", *Pattern Recognition*, vol. 42, n° 2, p. 302-312, 2009.

Onega S., Landa J. A. G., *Narratology: an Introduction*, Routledge, 2014.

Ordelman R., Aly R., Eskevich M., Huet B., Jones G. J., "Convenient discovery of archived video using audiovisual hyperlinking", *Proceedings of the 3rd Edition Workshop on Speech, Language & Audio in Multimedia*, p. 23-26, 2015.

Padia K., Bandara K. H., Healey C. G., "A System for Generating Storyline Visualizations Using Hierarchical Task Network Planning", *Computer Graphics*, n.d.

Papalampidi P., Keller F., Lapata M., "Movie Plot Analysis via Turning Point Identification", *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, p. 1707-1717, 2019.

Park S.-B., Oh K.-J., Jo G.-S., "Social network analysis in a movie using character-net", *Multimedia Tools and Applications*, vol. 2, n° 59, p. 601-627, 2012.

Phung Q. D., Dorai C., Venkatesh S., "Narrative structure analysis with education and training videos for e-learning", *Proceedings of the 16th International Conference on Pattern Recognition*, p. 835-838, 2002.

Piskorski J., Zavarella V., Atkinson M., Verile M., "Timelines: Entity-Centric Event Extraction from Online News", *Proceedings of the 2020 Text2Story Workshop at European Conference on Information Retrieval*, p. 105-114, 2020.

Propp V., *Morphology of the Folktale*, University of Texas Press, 1958.

Qiang L., Bingjie C., Haibo Z., "Storytelling by the Storycake Visualization", *Visual Computer*, vol. 33, n° 10, p. 1241-1252, 2017.

Reagan A. J., Mitchell L., Kiley D., Danforth C. M., Dodds P. S., "The Emotional Arcs of Stories are Dominated by Six Basic Shapes", *European Physical Journal Data Science*, vol. 5, n° 1, p. 1-12, 2016.

Regneri M., Koller A., Pinkal M., "Learning script knowledge with web experiments", *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 979-988, 2010.

Sadler N., "Narrative and Interpretation on Twitter: Reading Tweets by Telling Stories", *New Media and Society*, vol. 20, n° 9, p. 3266-3282, 2018.

Schmid W., *Narratology: an introduction*, Walter de Gruyter, 2010.

Sloetjes H., Wittenburg P., "Annotation by Category-ELAN and ISO DCR", *Proceedings of the 2008 International Conference on Language Resources and Evaluation*, p. 816-820, 2008.

Tapaswi M., Bäuml M., Stiefelhagen R., "StoryGraphs: Visualizing Character Interactions as a Timeline", *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, p. 827-834, 2014.

Tapaswi M., Bäuml M., Stiefelhagen R., "Aligning Plot Synopses to Videos for Story-Based Retrieval", *International Journal of Multimedia Information Retrieval*, vol. 4, n° 1, p. 3-16, 2015.

Tekiroğlu S. S., Chung Y.-L., Guerini M., "Generating Counter Narratives Against Online Hate Speech: Data and Strategies", *Proceedings of the 2020 Annual Meeting of the Association for Computational Linguistics*, p. 1177-1190, 2020.

Todorov T., Weinstein A., "Structural analysis of narrative", *NOVEL: A forum on fiction*, vol. 3, p. 70-76, 1969.

Valls-Vargas J., Ontanón S., Zhu J., "Toward automatic character identification in unannotated narrative text", *Proceedings of the 7th intelligent narrative technologies workshop*, p. 38-44, 2014.

Valls-Vargas J., Zhu J., Ontañón S., "Towards Automatically Extracting Story Graphs from Natural Language Stories", *Workshops of the 31st AAAI Conference on Artificial Intelligence*, 2017.

Van Dijk T., "Episodes as Units of Discourse Analysis. Analyzing Discourse: Text and Talk", 1981.

Vargas J. V., *Narrative Information Extraction with Non-Linear Natural Language Processing Pipelines*, 2017.

Vicol P., Tapaswi M., Castrejon L., Fidler S., "Moviegraphs: Towards Understanding Human-Centric Situations from Videos", *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, p. 8581-8590, 2018.

Whalley G., Baxter J., Atherton P. *et al.*, "Aristotle's Poetics: Translated and with a Commentary By George Whalley", *Dramatic Theory and Criticism (DTC): Greeks to Grotowski*, vol. 26, n° 1, p. 36-37, 1997.

Yu H., Zhang S., Morency L.-P., "Unsupervised Text Recap Extraction for TV Series", *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 1797-1806, 2016.

Zhao Z., Ge X., "A computable structure model for hollywood film", *Proceedings of the 2010 IEEE International Conference on Image Processing*, p. 877-880, 2010.

Notes de lecture

Rubrique préparée par Denis Maurel

Université de Tours, LIFAT (Laboratoire d'informatique fondamentale et appliquée)

Jimmy LIN, Rodrigo NOGUEIRA, Andrew YATES. Pretrained Transformers for Text Ranking: BERT and Beyond. *Morgan & Claypool publishers*. 2021. 308 pages. ISBN : 978-1-636-39230-1.

Lu par **Marilyne LATOUR**

Entreprise ReportLinker

Le but du « text ranking » *ou classement de textes est de générer une liste ordonnée de textes extraits d'un corpus en réponse à une requête. Bien que l'utilisation la plus courante du classement de textes soit pour la recherche d'informations, d'autres applications du traitement automatique des langues (TAL) peuvent également en bénéficier. Ce livre donne un aperçu du classement de textes avec les architectures de réseaux de neurones appelées transformateurs, dont BERT (*Bidirectional Encoder Representations from Tranformers*) est l'exemple le plus connu. Ce livre fournit une synthèse des travaux existants en tant que point d'entrée pour les praticiens qui souhaitent mieux comprendre comment appliquer les transformateurs aux problèmes de classement de textes et les chercheurs qui souhaitent poursuivre leurs travaux dans ce domaine. Il couvre un large éventail de techniques modernes, regroupées en deux grandes catégories : les modèles de transformateurs qui effectuent le reclassement dans des architectures multi-étages (*multi-stage architectures*) et les modèles qui effectuent directement le classement (*dense retrieval – DR – techniques*). Deux thèmes principaux transparaissent dans ce livre : les techniques permettant d'aborder le compromis entre l'efficacité (le résultat, la qualité) et l'efficacité (la latence des requêtes, le modèle et la taille de l'index).*

Dans le chapitre 1, les auteurs introduisent la notion de *deep learning* (*apprentissage profond* ou *apprentissage en profondeur*) à travers plusieurs modèles : les modèles basés sur la représentation (cette classe de modèles apprend des représentations vectorielles des requêtes et des documents qui peuvent être comparées au moment du classement pour calculer les scores de pertinence requête document), les modèles basés sur les interactions (cette classe de modèles capture explicitement les « interactions » entre les termes de la requête et les termes du document), les modèles hybrides (qui se composent d'un des éléments des deux premiers modèles).

Le chapitre 2 commence par caractériser plus formellement le problème de classement de textes, en énumérant explicitement les hypothèses sur les caractéristiques des données en entrée et en sortie. Les auteurs adoptent la

perspective de l'accès à l'information en se concentrant spécifiquement sur le problème du classement de textes en fonction de leur pertinence par rapport à une requête particulière. Du point de vue de l'évaluation, ce livre se concentre sur ce que l'on appelle communément le paradigme de Cranfield, une approche de l'évaluation orientée système des systèmes de recherche d'informations (SRI).

Le chapitre 3 met en évidence les aspects de BERT qui sont importants pour expliquer ses applications au classement de textes mais n'est pas un tutoriel du modèle. La formulation la plus simple et la plus directe du classement de textes consiste à convertir la tâche en un problème de classification de textes, puis à ordonner les textes à classer en fonction de la probabilité que chaque élément appartienne à la classe souhaitée. Pour les problèmes d'accès à l'information, la classe souhaitée comprend des textes qui sont pertinents pour le besoin d'informations de l'utilisateur, et cette approche peut être utilisée en tant que classification de pertinence. Plus précisément, l'approche consiste à entraîner un classifieur pour estimer la probabilité que chaque texte appartienne à la classe « pertinente », puis au moment du classement (c'est-à-dire de l'inférence) à trier les textes en fonction de ces estimations. Cette approche représente une réalisation directe du *Probability Ranking Principle*, qui stipule que les documents doivent être classés par ordre décroissant de la probabilité estimée de pertinence par rapport au besoin d'informations. C'est notamment la façon dont a été utilisé BERT dans sa première application en 2019.

Dans le chapitre 4, les auteurs présentent un certain nombre de techniques basées sur des transformateurs préentraînés qui opèrent sur les représentations *textuelles* des requêtes et des documents. Celles-ci peuvent être caractérisées comme des techniques d'interrogation et d'expansion de documents, qui ont une riche histoire dans la recherche d'informations, remontant à plusieurs décennies. La section 4.1 commence par un bref aperçu, mais les auteurs ne se veulent pas exhaustifs. Au lieu de cela, ils se concentrent uniquement sur les préalables nécessaires pour comprendre l'expansion des requêtes et des documents dans le contexte des modèles de transformateurs préentraînés. La discussion sur l'expansion des requêtes et des documents dans ce chapitre se déroule comme suit : remarques générales puis dans la section 4.2, techniques d'expansion des requêtes en utilisant des commentaires de pseudo-pertinence qui tirent parti des modèles basés sur les transformateurs. Les auteurs présentent ensuite quatre techniques d'expansion de documents : doc2query, DeepCT, HDCT et DeepImpact. Toutes ces techniques se concentrent sur la manipulation de représentations terminologiques (*i.e.* textuelles) des requêtes et des textes du corpus. Enfin, en dernière section de ce chapitre, les auteurs portent leur attention sur les techniques qui manipulent les requêtes et les représentations textuelles qui ne sont pas basées directement sur le document textuel.

Le chapitre 5 traite des transformateurs pour générer des représentations de textes qui conviennent au classement dans un cadre supervisé (*representation learning*). Les auteurs commencent ce chapitre en identifiant les liens entre les problèmes de pertinence et de similarité textuelle. En particulier, dans une perspective de classement, le principal défi reste le problème de l'estimation de la relation entre deux morceaux de texte. De la même manière que la recherche par

mot-clé nécessite des index et une infrastructure associée pour prendre en charge le classement *top-k* en utilisant des correspondances exactes sur un grand corpus, le classement *top-k* en termes d'opérations de comparaison vectorielle simples telles que les produits internes sur des représentations denses nécessite également une infrastructure dédiée. Les auteurs traitent de ce problème de la recherche du plus proche voisin (*nearest neighbor search*) dans la section 5.2. Comme pour les techniques de reclassement neuronal, il est utile de discuter du développement historique en termes de modèles pré-BERT et post-BERT dans la section 5.3. La section 5.4 introduit la notion de *bi-codeurs :* cette section se concentre sur les *bi-encodeurs* « simples », où chaque texte du corpus est représenté par un seul vecteur, et le classement est basé sur des opérations de comparaison simples telles que les produits scalaires. La section 5.5 présente des techniques qui améliorent la conception de base du *bi-codeur* de deux manières : chacun des textes du corpus peut être représenté par plusieurs vecteurs et le classement peut être effectué à l'aide de comparaisons plus complexes entre les représentations. Ces techniques visent des compromis entre efficacité et rendement différents par rapport aux *bi-codeurs* « simples ». La section 5.6 traite des techniques de récupération dense qui tirent parti de la distillation des connaissances. Enfin, les auteurs discutent des défis ouverts et des spéculations sur ce qui nous attend.

Le chapitre 6 conclut sur des questions de recherche ouvertes. Bien que les architectures de transformateurs et les techniques de pré-entraînement soient des innovations récentes, de nombreux aspects de leur application au classement de textes sont relativement bien compris et représentent des techniques établies. Cependant, il reste encore de nombreuses questions de recherche, et donc en plus de jeter les bases des transformateurs préentraînés pour le classement des textes, ce livre tente également de prévoir où peut se diriger le domaine.

L'ouvrage de Rodrigo Nogueira, Jimmy Lin et Andrew Yates est une étude fouillée, précise et extrêmement bien documentée du classement de textes (*text ranking*). La couverture de l'ensemble du domaine, ainsi que la mise en perspective des techniques utilisées rendent la lecture très attrayante. Nous recommandons donc chaudement la lecture de ce livre à tout chercheur en traitement automatique des langues, quel que soit son domaine de spécialité.

Anders SØGAARD. Explainable Natural Language Processing. *Morgan & Claypool publishers*. 2021. 108 pages. ISBN : 978-1-636-39215-8.

Lu par **Marie CANDITO**

Université Paris Cité / LLF, UMR CNRS Paris Cité

Le TAL neuronal a drastiquement accru l'opacité des modèles. Les tentatives pour comprendre le comportement d'un système de TAL, ou plus généralement un modèle neuronal, sont devenues un domaine de recherche, dont le nom même est fluctuant : l'auteur utilise les termes explainable (neural) NLP *(TAL (neuronal) explicable), avec* explainable

explicitement synonyme de interpretable, *plus que* model explainability *(explicabilité des modèles), ainsi que de* model explanation *(explication de modèles), terme que nous retiendrons. L'émergence de ce domaine s'accompagne d'une redondance, entre chercheurs issus de traditions scientifiques légèrement différentes, qui peuvent présenter les mêmes idées sous des termes différents. D'où la proposition d'une taxonomie des méthodes d'explication de modèles neuronaux utilisés en TAL, avec l'objectif de faciliter les rapprochements entre différents travaux et de pointer les problèmes qui demeurent ouverts dans ce domaine, ce qui, pour l'auteur, devrait permettre* in fine *d'accélérer les recherches dans ce domaine. Cet objectif est important à souligner, car il explique pourquoi l'ouvrage n'est pas très pédagogique : il s'agit d'un ouvrage pour des chercheurs du domaine, destiné à mieux éclairer leurs recherches.*

Dans le chapitre d'introduction, l'auteur commence par présenter deux distinctions usuelles concernant les méthodes d'explication. La première distingue les méthodes « locales » *versus* « globales » : une méthode est dite globale si elle nécessite d'utiliser forcément tout un échantillon d'exemples. Au contraire, une méthode sera dite locale si elle peut fournir des explications pour des exemples individuels. Il s'agira typiquement d'expliquer les sorties pour un exemple particulier (par exemple pourquoi tel verbatim a été classé comme négatif). Une méthode globale cherchera plutôt à mettre au jour des caractéristiques globales d'un modèle (par exemple, s'il comporte des biais, quelle est la densité des représentations qu'il construit, etc.).

La deuxième distinction concerne les méthodes d'explication intrinsèque (prédictions conjointes des sorties du modèle et des explications) *versus* « *post hoc* » (qui utilisent des techniques orthogonales aux modèles à expliquer). L'auteur va ensuite motiver une nouvelle taxonomie. La section 1.2 décrit rapidement huit taxonomies existant dans la littérature très récente, et en pointe les incohérences et incomplétudes. Là encore, cette partie s'adresse à qui connaît déjà très bien les travaux en TAL explicable.

L'auteur ne retient pas comme caractéristique première la forme des explications (par exemple, explication sous forme de visualisation graphique, de coefficients, de règles), qui est en réalité orthogonale aux méthodes, une même méthode pouvant fournir des explications sous différentes formes. La nouvelle taxonomie est ensuite présentée rapidement : l'auteur reprend la distinction locale et globale, mais plutôt que la distinction « intrinsèque » *versus* « *post hoc* », il choisit la distinction entre méthodes « avant » *versus* méthodes « arrière », faisant référence à la propagation avant des réseaux, *versus* la rétropropagation : les méthodes arrière utilisent essentiellement les gradients de la fonction de perte. Les méthodes « avant » sont subdivisées en trois sous-cas, selon qu'elles expliquent des représentations intermédiaires, des représentations continues ou des représentations discrètes. Cette subdivision n'est pas étanche et est destinée à structurer la présentation des méthodes. Ceci donne au final 2 (local/global) * 4 = 8 catégories de méthodes.

Le chapitre 2 introduit très rapidement les architectures principales du TAL neuronal (classification non linéaire, encodage récurrent et encodage à l'aide de *transformers*) et détaille les caractéristiques des huit catégories de méthodes.

Ensuite, chacune des huit catégories donne lieu à un chapitre (chapitres 3 à 10). Pour chaque chapitre, l'auteur liste diverses méthodes, en présentant rapidement la technique formelle, puis un certain nombre de travaux de TAL ayant utilisé ladite méthode. La description technique est très succincte, la valeur ajoutée est clairement dans la liste des travaux cités pour chaque type de méthode.

Les méthodes « arrière » (utilisant une propagation arrière, pour calculer les gradients ou calculer des valeurs de « pertinence » de chaque neurone) :

– locales (applicable à un exemple précis) : la technique de base consiste simplement à étudier le gradient de la perte par rapport à l'exemple d'entrée (*vanilla gradients*), pour en déduire les traits de l'entrée expliquant le mieux la sortie. (Bien que l'auteur n'utilise pas ce terme, on retrouve le concept de *feature attribution*, qui prend un exemple en entrée et assigne un score aux traits selon la contribution que ce trait apporte dans la sortie du modèle.) Avec la *layer-wise relevance propagation*, des valeurs de pertinence de *chaque* neurone sont rétropropagées de couche en couche ;

– globales (nécessitant un échantillon d'exemples) : les méthodes globales arrière sont majoritairement fondées sur l'idée que des modèles plus parcimonieux seront mieux interprétables. Aussi, l'auteur inclut-il ici des méthodes d'élagage de modèles (où une partie des poids sont supprimés). On suppose que toute méthode d'explication sera plus parlante si elle est appliquée sur un modèle élagué. L'idée la plus ancienne est de supprimer les poids correspondant à de faibles dérivées partielles. L'élagage est fait soit après apprentissage (*post-hoc unstructured pruning*), soit par réapprentissage de candidats modèles élagués (*lottery tickets*), soit conjointement à l'apprentissage (*dynamic sparse training*).

Les méthodes « avant » (utilisant une propagation avant au sein du réseau complète ou partielle) :

– expliquant des représentations intermédiaires :

- locales : étude des *gates* ou des poids d'attention pour un exemple donné,

- globales : étude de quel *gate* ou quelle tête d'attention code telle ou telle information. Élagage de *gate* et de têtes d'attention ;

– expliquant des représentations continues :

- locales : l'auteur classe ici d'une part, les études des vecteurs statiques de mot (corrélation avec jugements humains de similarités, analogies entre mots...) et, d'autre part, l'étude de la dynamique des valeurs d'activation, au sein d'une séquence. La distinction d'avec les représentations intermédiaires n'est pas très claire ici,

- globales : les méthodes qui analysent tout un nuage de points, chaque point étant une représentation continue d'un exemple de l'échantillon (comme un nuage de mots). L'analyse peut être un *clustering*, ou bien la corrélation avec un autre nuage de points, issus d'un autre système ou de mesures humaines (par exemple des points issus d'enregistrements fMRI). Sont également classées ici les sondes

linguistiques (des classifieurs appris sur des représentations issues de réseaux, pour prédire des propriétés linguistiques). Les *concepts activation vectors* dépassent l'attribution de traits en permettant une attribution de concepts. Il s'agit de généraliser les exemples d'entrée, en entraînant des classifieurs prédisant des concepts prédéfinis, puis de quantifier pour chaque concept dans quelle mesure il explique la prédiction de telle ou telle classe ;

– expliquant des sorties discrètes :

- locales : l'auteur classe ici les méthodes utilisant des jeux de données linguistiques difficiles ou d'intérêt, comme des constructions linguistiques particulières. Se retrouve classée ici, par exemple, la technique pour tester la capacité à encoder l'accord sujet verbe, utilisant directement les probabilités d'un modèle de langue du verbe s'accordant *versus* la forme violant l'accord. La taxonomie montre ses limites, ainsi le lien avec les autres sondes linguistiques citées *supra* n'est pas fait. Une autre technique célèbre classée ici est LIME (*Local Interpretable Model-agnostic Explanations*), qui permet d'expliquer un modèle boîte noire, au moyen de classifieurs locaux, intrinsèquement interprétables (comme des arbres de décision), entraînés sur des paires entrée et prédiction du modèle à expliquer. Enfin, sont citées ici les techniques pour identifier les « exemples influents », c'est-à-dire ceux dont la suppression à l'apprentissage a le plus d'impact sur le modèle appris,

- globales : la technique d'*uptraining* peut être utilisée pour entraîner un modèle *m'*, intrinsèquement interprétable, sur un grand volume de données annotées *via* le modèle *m* à expliquer. L'interprétation du modèle *m'* servira d'interprétation de *m*.

L'auteur cite également des techniques d'analyse des performances du modèle sur différents jeux de données, ce qui peut aider à caractériser le comportement d'un modèle.

Le chapitre 11 traite des techniques pour évaluer les méthodes d'explication. Le chapitre 12, intitulé « *Perspectives* », commence par des observations générales, mais très techniques, sur les différentes catégories de méthodes, qui ne sont pas vraiment des perspectives.

Dans la dernière (courte) section du chapitre, l'auteur prend du recul, en abordant la question des motivations du TAL explicable, avec un parallèle entre l'explication de décisions « neuronales » et l'explication de décisions humaines. Selon l'auteur, parmi les motivations pour une explicabilité des modèles neuronaux (analyse d'erreurs, maintenance des modèles, amélioration de l'efficacité, détection de la vulnérabilité à des attaques), la motivation la plus citée est le « droit à l'explication ». L'auteur s'interroge sur les motivations de ce droit à l'explication : d'un point de vue « moral », qu'est-ce qui le justifie ? L'auteur semble sceptique quant au bien-fondé de ce droit, en défendant rapidement l'idée que, malgré leurs limites, les méthodes d'explication des modèles neuronaux permettent finalement un niveau d'explicabilité que n'ont pas les décisions humaines elles-mêmes.

Pour conclure sur l'ouvrage, la présentation de la taxonomie et des méthodes relevant de chaque catégorie est rapide et technique, et s'adresse à un public connaissant bien le domaine. Plutôt qu'une taxonomie cherchant à assigner une seule catégorie à chaque méthode, une classification multifacette serait peut-être mieux adaptée. En particulier, typer les méthodes d'explication en fonction des objectifs généraux (Quels traits expliquent une sortie ? Quels concepts expliquent une sortie ? Quels concepts sont encodés dans des paramètres ?) aiderait la lectrice ou le lecteur à aborder plus facilement ce domaine foisonnant.

Beata BEIGMAN KLEBANOV, Nitin MADNANI. Automated Essay Scoring. *Morgan & Claypool publishers*. 2021. 294 pages. ISBN : 978-1-636-39224-0.

Lu par **Laurie ACENSIO**

Lexiane Formation (Paris)

Considéré comme étant une alternative à l'évaluation manuelle des enseignants, la notation automatisée des essais permet de fournir aux apprenants une évaluation instantanée. En complément des techniques d'apprentissage automatique, cet ouvrage aborde les techniques de traitement du langage (TAL) davantage adaptées pour la notation automatisée des productions écrites. Des travaux scientifiques à travers des cas pratiques (tests de langue, tâches ouvertes, questions à réponse élaborée) sont exposés dans un cadre universitaire et industriel en mettant en évidence les aspects multidimensionnels (conceptuel, méthodologique et technique) pour automatiser l'évaluation des compétences de l'expression écrite au sein d'un système de notation.

L'ouvrage est structuré autour de cinq parties : l'introduction, les bonnes pratiques pour construire un système de notation automatisé simple associé à des cas d'expérimentation, un état de l'art scientifique sur les différents modèles de notation, l'implémentation et l'évaluation des systèmes, les méthodes d'argumentation (rétroaction, analyse des contenus et discours), puis une dernière partie est consacrée aux discussions et aux perspectives de recherche.

La première partie aborde l'approche historique de ce domaine de recherche à travers les travaux de Page qui a développé le premier système de notation automatisée nommé PEG (*Project Essay Grade*). La technique est basée selon une approche simple : une phase d'entraînement et une phase de notation en utilisant un ensemble de coefficients de corrélation pour attribuer un score comparé par la suite à une notation humaine. Acquis par l'entreprise Measurment Inc, le PEG a progressivement intégré de nombreuses caractéristiques intrinsèques liées à la qualité de l'écriture (fluidité, diction, grammaire) en utilisant des techniques d'analyse sémantique et syntaxique. Peu utilisé à ses débuts, cet outil l'est massivement aujourd'hui au sein des écoles et universités américaines, notamment lorsque les productions écrites se sont informatisées à partir des années 1990. Ainsi, la composante arbitraire réelle ou perçue de l'évaluation manuelle s'est atténuée progressivement afin de garantir davantage d'objectivité à travers des critères d'évaluation instrumentés par des outils d'aide à la notation. Néanmoins, il apparaît

que la notation automatisée reste un outil d'assistance lors du jugement d'évaluation humain, mais ne peut se substituer pleinement à celui-ci.

La deuxième partie décrit les différentes étapes pour construire un système de notation informatisée : la collecte de données à partir d'un corpus écrit (par exemple des résultats de tests de langue comme le TOEFL[1] ou ESOL[2]), la phase de modélisation dont le jugement humain s'avère être déterminant, les critères d'évaluation ainsi que les expérimentations. L'automatisation du processus de notation implique de définir des attributs observables et des critères d'évaluation. Or, il apparaît que les approches statistiques basées sur une corrélation élevée entre les attributs ne sont pas forcément pertinentes au niveau de la qualité de l'écriture. De plus, l'analyse textuelle (fréquence et longueur des mots) et celle au niveau de la surface de texte (grammaire et orthographe) s'avèrent rapidement insuffisantes pour évaluer les compétences associées à l'expression écrite. Progressivement, de nouveaux critères linguistiques ont été mis en place pour s'adapter à la complexité des réponses, l'un des auteurs les plus cités dans l'ouvrage propose une analyse multicritère pour déterminer un score global (ou holistique) lors de l'analyse de la production écrite (style, clarté, argumentation, cohérence et pertinence).

La troisième partie constitue la majeure partie de l'ouvrage avec une description des modèles et des techniques utilisées : la régression linéaire ainsi que l'analyse sémantique latente (*Latent Semantic Analysis*) sont les techniques les plus implémentées dans les logiciels. Plus récemment, les techniques d'apprentissage profond (*Deep Learning*) démontrent une nette amélioration des résultats, dont les travaux de Taghipour qui exploite un réseau neuronal pour améliorer la précision des scores holistiques en comparant les résultats avec d'autres techniques (classification hiérarchique et classification de texte bayésienne). Néanmoins, la complexité de traitement des réseaux de neurones est un obstacle pour expliciter les scores prédits, notamment à travers les commentaires afin d'indiquer les points d'amélioration pour l'apprenant.

Les chapitres suivants abordent les fonctionnalités génériques, les différents types d'essais (l'écriture argumentative, l'écriture narrative et l'écriture de réflexion), la phase de production d'un système de notation à travers un exemple d'architecture illustré par un schéma global et des extraits de code source.

La quatrième partie aborde les systèmes de rétroaction (*feed-back*) notamment en relation avec le chapitre 5 de la partie précédente concernant la qualité de l'écriture argumentative et la qualité de réflexion. Puis, la notation automatisée du contenu et du discours est abordée à travers quatre chapitres dans la perspective d'une personnalisation des rétroactions. Cette phase de rétroaction pédagogique est une étape déterminante constatée à partir d'une méta-analyse dont l'ensemble des travaux démontre que cette forme de révision a des effets particulièrement positifs sur la qualité d'écriture des apprenants lors de l'évaluation des reformulations (ou corrections). Néanmoins, la rétroaction doit être explicitée afin de s'assurer une

1 *Test of English as a Foreign Language*.
2 *English for Speakers of Other Languages*.

bonne compréhension de l'apprenant. Il apparaît que les corpus annotés nécessaires et qui impliquent d'être construits dans un domaine de connaissance spécifique sont peu disponibles pour optimiser cette étape d'évaluation.

La cinquième partie finalise l'ouvrage en mettant en évidence les défis de la notation automatique des essais, dont l'évaluation de l'écriture dans plusieurs langues, la standardisation des tests, la validité des textes, l'interprétabilité du modèle, la recherche de l'équité lors de l'attribution des scores, l'omniprésence et l'évolution constante de la technologie. Il apparaît que la principale utilité de la notation automatisée des essais est le gain de temps pour les évaluateurs, l'élimination de préjugés humains et la différence de perception afin d'assurer l'équité dans la notation.

Cet ouvrage démontre que la notation automatisée des essais est un sujet de recherche actif s'appuyant sur de nombreuses références bibliographiques essentiellement issues de revues internationales, mais sans la présence d'équipes de recherche en France. Il peut s'adresser à des chercheurs en TAL intéressés au domaine d'application liée à l'éducation impliquant des questions de recherche tout autant en didactique qu'en linguistique. En effet, la rétroaction pédagogique est un défi considérable afin d'identifier les erreurs et, par conséquent, de mieux cibler les points d'amélioration de l'apprenant qui ne sont pas détaillés lors de l'attribution d'un score de type holistique. Actuellement, les travaux restent focalisés sur les techniques utilisées pour améliorer la qualité de prédiction des scores holistiques avec notamment le potentiel de l'apprentissage profond (ou réseaux neuronaux). Néanmoins, ils abordent peu les enjeux de l'IA explicable afin d'aboutir à des systèmes de notation automatisés transparents et impartiaux. Les techniques de TAL se sont imposées du fait de la complexité croissante des tâches d'évaluation d'apprentissage notamment à travers les questions ouvertes ou les questions à réponse élaborée. Les problématiques sont d'ordre syntaxique, lexical et sémantique afin de détecter principalement des erreurs de l'apprenant, mais demeurent largement étudiées préalablement dans d'autres contextes applicatifs (par exemple l'aide au diagnostic médical). Néanmoins, les enjeux de ce domaine de recherche impliquent de prendre en considération le style et le type d'écriture (argumentation, réflexion, narratif et créatif) lors du traitement du contenu textuel soulevant ainsi des questions originales liées au processus cognitif humain associées à une représentation du monde et des connaissances. En effet, l'évaluation de la qualité argumentative d'une production écrite soulève des défis relativement nouveaux consistant à prendre en compte la créativité et la subjectivité humaine. Au-delà du cadre éducatif, cet axe de recherche aborde des questions communes avec l'extraction et l'analyse automatique des arguments (*argument mining*) ou des opinions (*opinion mining*) au sein de corpus textuels.

Philippe BARBAUD. L'instinct du sens – Essai sur la préhistoire de la parole. Éditions *Des auteurs des livres*. 2021. 342 pages. ISBN : 978-2-9570999-9-3.

Lu par **Georgeta CISLARU**

Paris Nanterre / MoDyCo

Cet ouvrage se propose de répondre à la question, complexe, des origines du langage, que l'auteur articule à la non moins complexe question des origines du sens. Malgré cette double complexité, le texte se laisse lire avec aisance, en raison à la fois de l'avant-propos, qui pose explicitement les jalons du raisonnement, et de la clarté du style. Les deux grands axes que l'auteur met en avant sont la référenciation et l'énonciation, en tant que terreau de la construction du sens et moteurs de l'émergence du langage articulé, et l'abstraction, en tant qu'artefact culturel sous-tendu par les capacités mémorielles de l'humain, rendant possibles l'encodage morphologique, la grammaticalisation et les relations grammaticales, entre autres.

Le langage sert à parler du monde, mais, avant tout, il sert à vivre, son essence même est de signifier, selon E. Benveniste[3]. C'est pourquoi la question des origines du langage n'a pour lui aucun sens : « *le langage est aussi ancien, ou aussi primordial, que la signification elle-même, et l'on ne saurait imaginer un homme qui ne posséderait pas la faculté fondamentale de donner un sens aux choses, c'est-à-dire de parler* »[4]. Dans son ouvrage, P. Barbaud s'emploie à donner du sens au chemin parcouru par l'humain depuis ses origines, sens qui se trouve sémiotisé par le langage.

L'ouvrage est constitué d'un avant-propos et de trois chapitres détaillés qui s'appuient sur des publications issues de nombreux champs des sciences du langage[5]. Les enjeux et les positionnements théoriques sont explicitement formulés dès l'avant-propos, ce qui facilite la lecture.

Le premier chapitre esquisse l'évolution de 2,5 millions d'années et la rengaine simienne pour les placer dans une perspective darwinienne en termes d'adaptation, de sélection naturelle et d'évolution. L'instinct animal de la communication régit des comportements réflexifs en réponse à des signaux, tandis que le délitement de la

3 Benveniste É., « La forme et le sens dans le langage », *Problèmes de linguistique générale II*, Paris, Gallimard, p. 215-238, 1974.

4 Mosès S., « Émile Benveniste et la linguistique du dialogue », *Revue de métaphysique et de morale ;* n° 32 (4), p. 509-525, 2001.

5 Comme cela arrive souvent, le fait d'impliquer, au fil de l'argumentation, des problématiques diverses conduit parfois à des « points aveugles », par exemple à ne pas convoquer les travaux de Willy Van Langendonck pour défendre une conception catégorielle des noms propres qui ont un rôle à jouer dans l'interpellation comme étape de l'émergence du langage.

rengaine, relevant de l'hérédité régressive, et l'exaptation des organes phonatoires répondent à la pression adaptative exercée par la nécessité de faire sens, donnant lieu à ce que l'auteur appelle la « naissens » de la parole. L'auteur s'intéresse plus particulièrement à la période transitoire constituée d'imitations de sons de la nature, de babils et autres émergences phoniques précédant l'invention des premiers mots (et, donc, des premiers symboles). C'est la curiosité humaine et le discernement de référents pertinents, détachés perceptivement et émotionnellement du contexte environnant, qui permettent cette invention, qui font aussi que, selon l'auteur (suivant en cela P. Ricœur), le sens est inséparable de la référence.

La thèse principale du deuxième chapitre est l'évolution, grâce à l'hérédité régressive, de la rengaine animale en signes dotés de sens. Les protolangues seraient ainsi des produits de l'esprit-cerveau. L'auteur distingue quatre phases du langage : interjective, vocative, objective et énonciative. Dans le processus évolutif, la production des outils comme la taille des galets s'accompagnant d'interjections exprimant des émotions (douleur, joie, surprise) et des interpellations de l'autre.

L'avant-signe interjectif aurait ainsi contribué à installer un habitus phonatoire, pour ensuite implémenter le sens dans la cognition humaine, en passant par l'interjection vocative suivie de ce que l'auteur appelle la conquête du référent objectal détaché du moi et qui deviendra l'autre. C'est donc dans l'altérité que le sens émerge et, avec lui, le langage tel qu'on le connaît. La prégnance et la saillance[6], en tant que principes de discrimination perceptuelle (système DISPER), constituent des éléments déterminants de la stabilisation du référent objectal. Le passage du référent au signifié implique le passage d'une mémoire autobiographique à une mémoire sémantique.

S'ensuit l'émergence de la première et de la deuxième personne, communes à toutes les langues, comme pendant de la fonction conative du langage : moi s'énonce et interpelle l'autre, en faisant ainsi évoluer les interjections vers des formes pronominales, puis des noms propres. Dans cette optique l'autre est le non-allocutaire, la non-personne qui peut, de ce fait, être intégrée à la catégorie des personnes. L'émergence des toponymes viendrait par ailleurs configurer les dyades *moi-ici* (et *eux-là-bas*). La diversité des référents saillants conduit à la généralisation des noms communs. En dénommant, l'humain présapiens confère « *son signifié descriptif à un signifiant référentiel appartenant à l'univers du "ça" dérivé de celui de la "non-personne"* ». Pour reprendre les termes de l'auteur, la référence argumentale vient compléter la référence temporelle induite par le dessein. En parallèle, aurait eu lieu une implémentation psychocérébrale de la symbiose entre la perception et la phonation, ouvrant ainsi la voie aux développements articulatoires, du monosyllabique au plurisyllabique. L'auteur fait coïncider l'émergence de l'articulation monosyllabique avec la maîtrise du feu, intervenue il y a ± 500 000 ans, qui a permis à l'humain présapiens de dépasser sa condition animale.

---

6 Que l'on retrouve, par exemple, dans certaines théories de la référenciation et de l'anaphore.

Dans le troisième chapitre, l'auteur tient à dissocier les dynamiques de l'émergence de la parole de la description de l'état actuel du langage. On perçoit néanmoins, dans les analyses et les spéculations qui sont proposées en marge de l'encodage syntaxique, des éléments permettant de comprendre des préférences structurelles, comme l'ordre SOV dans la phrase.

La récursivité apparaît comme un moteur de l'évolution et de la structuration des langues. La récursivité gouverne la mémoire et, grâce à des outils lexicaux tels les hyperonymes, domine la grammaire. Dans le même esprit, P. Barbaud met en avant le rôle de la mémoire plutôt que celui de la grammaire universelle dans l'enseignement et l'apprentissage des langues. C'est le même mécanisme qui entre en jeu lorsque des lexèmes commencent à fonctionner comme des morphèmes, à commencer par les classificateurs propres à plusieurs langues asiatiques, amérindiennes, etc. La morphologie est ainsi, à la différence du lexique[7], le produit artéfactuel du système et non du cerveau-esprit ; il s'agit de la manifestation d'une évolution culturelle qui favorise l'abstraction et qui entraîne dans le sillage la grammaticalisation.

En filigrane, on reconstitue une représentation du langage et des langues ancrés dans une sémantique se détachant des formalismes. Pour l'auteur, la combinaison de signifiants est régie par des mécanismes indépendants de ceux qui régissent la combinaison des signifiés qui leur sont associés, tandis que la structure n'est pas garante de sens.

Au terme des 327 pages riches en notes et commentaires, complétées par un glossaire et deux index, on apprécie la diversité des références bibliographiques, couvrant plusieurs champs de la linguistique et des disciplines connexes comme la psychologie, les sciences cognitives, etc. Des encadrés, des notes de fin de chapitre et des schémas accompagnent le propos, de sorte que certaines pages donnent à l'ouvrage une valeur ressource indéniable. Que l'on soit ou non en accord constant avec le propos développé par l'auteur, on apprécie également la réflexion de P. Barbaud, qui n'hésite pas à scruter sous des angles inédits des aspects ponctuels ou des affirmations plus générales issus des différents travaux cités.

7 Dont elle reste partie prenante, comme objet de mémoire collective.

Résumés de thèses et HDR

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Ygor GALLINA : ygor.gallina@univ-nantes.fr

Titre : Indexation de bout-en-bout dans les bibliothèques numériques scientifiques

Mots-clés : indexation automatique, mots-clés, évaluation extrinsèque, recherche d'information, génération de mots-clés, méthodes de bout en bout.

Title: *End-to-End Indexation in Digital Scientific Libraries*

Keywords: *automatic indexing, keywords, extrinsic evaluation, information retrieval, keyword generation, end-to-end method.*

Thèse de doctorat en informatique, Laboratoire des Sciences du Numérique de Nantes, UMR 6004, UFR Sciences et Techniques, Université de Nantes, sous la direction de Béatrice Daille (Pr, Université de Nantes) et Florian Boudin (MC, Université de Nantes). Thèse soutenue le 28/03/2022.

Jury : Mme Béatrice Daille (Pr, Université de Nantes, codirectrice), M. Florian Boudin (MC, Université de Nantes, codirecteur), M. Richard Dufour (Pr, Université de Nantes, président), Mme Josiane Mothe (Pr, Université de Toulouse, rapporteuse), M. Patrick Paroubek (IR, CNRS, rapporteur), Mme Lorraine Goeuriot (MC, Université Grenoble Alpes, examinatrice).

Résumé : *Le nombre de documents scientifiques dans les bibliothèques numériques ne cesse d'augmenter. Les mots-clés permettant d'enrichir l'indexation de ces documents ne peuvent être annotés manuellement étant donné le volume de documents à traiter. La production automatique de mots-clés est donc un enjeu important. Le cadre évaluatif le plus utilisé pour cette tâche souffre de nombreuses faiblesses qui rendent l'évaluation des nouvelles méthodes neuronales peu fiables. Notre objectif est d'identifier précisément ces faiblesses et d'y apporter des solutions selon trois axes. Dans*

un premier temps, nous introduisons KPTimes*, un jeu de données du domaine journalistique. Il nous permet d'analyser la capacité de généralisation des méthodes neuronales. De manière surprenante, nos expériences montrent que le modèle le moins performant est celui qui généralise le mieux. Dans un deuxième temps, nous effectuons une comparaison systématique des méthodes états de l'art grâce à un cadre expérimental strict. Cette comparaison indique que les méthodes de référence comme TF#IDF sont toujours compétitives et que la qualité des mots-clés de référence a un impact fort sur la fiabilité de l'évaluation. Enfin, nous présentons un nouveau protocole d'évaluation extrinsèque basé sur la recherche d'information. Il nous permet d'évaluer l'utilité des mots-clés, une question peu abordée jusqu'à présent. Cette évaluation nous permet de mieux identifier les mots-clés importants pour la tâche de production automatique de mots-clés et d'orienter les futurs travaux.*

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03667015

Timothee MICKUS : timothee.mickus@univ-lorraine.fr

Titre : Du statut des plongements lexicaux en tant qu'implémentations de l'hypothèse distributionnelle

Mots-clés : TAL, génération automatique de texte, sémantique distributionnelle, plongements lexicaux, génération de définition, lexicographie.

Title: *On the Status of Word Embeddings as Implementations of the Distributional Hypothesis*

Keywords: *NLP, NLG, distributional semantics, word embeddings, definition modeling, lexicography.*

Thèse de doctorat en informatique, ATILF, UMR 7118, Université de Lorraine, sous la direction de Mathieu Constant (Pr, Université de Lorraine, ATILF, UMR 7118) et Denis Paperno (universitair docent, Universiteit Utrecht, Pays-Bas). Thèse soutenue le 31/03/2022.

Jury : M. Mathieu Constant (Pr, Université de Lorraine, ATILF, UMR 7118, codirecteur), M. Denis Paperno (universitair docent, Universiteit Utrecht, Pays-Bas, codirecteur), M. Benoît Crabbé (Pr, Université de Paris, rapporteur, président), M. Nabil Hathout (DR, CNRS, rapporteur), Mme Gemma Boleda (Research Professor, Universitat Pompeu Fabra, Barcelone, Espagne, examinatrice), Mme Vera Demberg (Pr, Universität des Saarlandes, Sarrebruck, Allemagne, examinatrice), Mme Claire Gardent (DR, CNRS, examinatrice), M. Alessandro Lenci (Pr, Università di Pisa, Italie, examinateur), M. Kees van Deemter (Pr, Universiteit Utrecht, Pays-Bas, examinateur).

Résumé : *Cette thèse s'intéresse au statut des plongements lexicaux (ou «* word embeddings *»), c'est-à-dire des vecteurs de mots issus de modèles de traitement automatique des langues. Plus particulièrement, notre intérêt se porte sur leur valeur linguis-*

tique et la relation qu'ils entretiennent avec la sémantique distributionnelle, le champ d'études fondé sur l'hypothèse que le contexte est corrélé au sens. L'objet de notre recherche est d'établir si ces plongements lexicaux peuvent être considérés comme une implémentation concrète de la sémantique distributionnelle.

Notre première approche dans cette étude consiste à comparer les plongements lexicaux à d'autres représentations du sens, en particulier aux définitions telles qu'on en trouve dans des dictionnaires. Cette démarche se fonde sur l'hypothèse que des représentations sémantiques de deux formalismes distincts devraient être équivalentes, et que par conséquent l'information encodée dans les représentations sémantiques distributionnelles devrait être équivalente à celle encodée dans les définitions. Nous mettons cette idée à l'épreuve à travers deux protocoles expérimentaux distincts : le premier est basé sur la similarité globale des espaces métrisables décrits par les vecteurs de mots et les définitions, le second repose sur des réseaux de neurones profonds. Dans les deux cas, nous n'obtenons qu'un succès limité, ce qui suggère soit que la sémantique distributionnelle et les dictionnaires encodent des informations différentes, soit que les plongements lexicaux ne sont pas motivés d'un point de vue linguistique.

Le second angle que nous adoptons ici pour étudier le rapport entre sémantique distributionnelle et plongements lexicaux consiste à formellement définir ce que nous attendons des représentations sémantiques distributionnelles, puis à comparer nos attentes à ce que nous observons effectivement dans les plongements lexicaux. Nous construisons un jeu de données de jugements humains sur l'hypothèse distributionnelle. Nous utilisons ensuite ce jeu pour obtenir des prédictions sur une tâche de substituabilité distributionnelle à partir de modèles de plongements lexicaux. Bien que nous observions un certain degré de performance en utilisant les modèles en question, leur comportement se démarque très clairement de celui de nos annotateurs humains. Venant renforcer ces résultats, nous remarquons qu'une large famille de modèles de plongements qui ont rencontré un franc succès, ceux basés sur l'architecture Transformer*, présente des artéfacts directement imputables à l'architecture qu'elle emploie plutôt qu'à des facteurs d'ordre sémantique.*

Nos expériences suggèrent que la validité linguistique des plongements lexicaux n'est aujourd'hui pas un problème résolu. Trois grandes conclusions se dégagent de nos expériences. Premièrement, la diversité des approches en sémantique distributionnelle n'implique pas que ce champ d'études est voué aux approches informelles : nous avons vu que le linguiste peut s'appuyer sur la substituabilité distributionnelle. Deuxièmement, comme on ne peut pas aisément comparer la sémantique distributionnelle à une autre théorie lexicale, il devient nécessaire d'étudier si la sémantique distributionnelle s'intéresse bien au sens, ou bien si elle porte sur une série de faits entièrement distincte. Troisièmement, bien que l'on puisse souligner une différence entre la qualité des plongements lexicaux et ce qu'on attend qu'ils puissent faire, la

possibilité d'étudier cette différence sous un angle quantitatif est de très bon augure pour les travaux à venir.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-03723503

Pedro ORTIZ SUAREZ : pedro@portizs.eu

Titre : Une approche basée sur les données pour le traitement automatique du langage naturel en français contemporain et historique

Mots-clés : modèle de langue, corpus de pré-entraînement, traitement automatique des langues, français contemporain, français historique, apprentissage par transfert.

Title: *A Data-driven Approach to Natural Language Processing for Contemporary and Historical French*

Keywords: *language model, pre-training corpora, natural language processing, contemporary French, historical French, transfer learning.*

Thèse de doctorat en informatique, ALMAnach, Centre de Recherche Inria de Paris, Sorbonne Université, sous la direction de Laurent Romary (DR, Inria) et Benoît Sagot (DR, Inria). Thèse soutenue le 27/06/2022.

Jury : M. Laurent Romary (DR, Inria, codirecteur), M. Benoît Sagot (DR, Inria, codirecteur), M. Francis Bach (DR, Inria, président), Mme Maud Ehrmann (MC, École polytechnique fédérale de Lausanne, Suisse, examinatrice), M. Alexander Geyken (DR, Berlin-Brandenburgischen Akademie der Wissenschaften, Allemagne, examinateur), Mme Anna Korhonen (DR, University of Cambridge, Royaume-Uni, rapporteuse), M. Holger (DR, Meta AI Research, rapporteur).

Résumé : *Depuis plusieurs années, les approches neuronales ont régulièrement amélioré l'état de l'art du traitement automatique des langues (TAL) sur une grande variété de tâches. L'un des principaux facteurs ayant permis ces progrès continus est l'utilisation de techniques d'apprentissage par transfert. Ces méthodes consistent à partir d'un modèle pré-entraîné et à le réutiliser, avec peu ou pas d'entraînements supplémentaires, pour traiter d'autres tâches. Même si ces modèles présentent des avantages évidents, leur principal inconvénient est la quantité de données nécessaire pour les pré-entraîner. Ainsi, le manque de données disponibles à grande échelle a freiné le développement de tels modèles pour le français contemporain et a fortiori pour ses états de langue plus anciens.*

Cette thèse met l'accent sur le développement de corpus pour le pré-entraînement de telles architectures. Cette approche s'avère extrêmement efficace, car nous sommes en mesure d'améliorer l'état de l'art pour un large éventail de tâches de TAL pour le français contemporain et historique, ainsi que pour six autres langues contemporaines. De plus, nous montrons que ces modèles sont extrêmement sensibles à la qualité, à l'hé-

térogénéité et à l'équilibre des données de pré-entraînement et montrons que ces trois caractéristiques sont de meilleurs prédicteurs de la performance des modèles que la taille des données de pré-entraînement. Nous montrons également que l'importance de la taille des données de pré-entraînement a été surestimée en démontrant à plusieurs reprises que l'on peut pré-entraîner de tels modèles avec des corpus de taille assez modeste.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-03770337

Léon-Paul SCHAUB : lp.schaub@gmail.com

Titre : Dimensions mémorielles de l'interaction écrite humain-machine : une approche cognitive par les modèles mnémoniques pour la détection et la correction des incohérences du système dans les dialogues orientés-tâche

Mots-clés : système de dialogue orienté tâche, modèle de mémoire, réseaux de neurones, détection d'incohérences, interaction homme-machine.

Title: *Memory Dimensions of Human-Machine Written Interaction: a Cognitive Approach Using Mnemonic Models for the Detection and Correction of System Inconsistencies in Task-oriented Dialogues*

Keywords: *task-oriented dialogue system, memory model, neural networks, inconsistency detection.*

Thèse de doctorat en informatique, Laboratoire Indisciplinaire des Sciences du Numérique, Université Paris-Saclay, sous la direction de Patrick Paroubek (IR, CNRS). Thèse soutenue le 23/03/2022.

Jury : M. Patrick Paroubek (IR, CNRS, directeur), M. Frédéric Landragin (DR, CNRS, rapporteur, président), Mme Chloé Clavel (Pr, Télécom ParisTech, rapporteuse), M. Yves Lepage (Pr, Waseda Université, Tokyo, Japon, examinateur), Mme Magalie Ochs (MC, Université Aix-Marseille, examinatrice), M. Frédéric Béchet (Pr, Université Aix-Marseille, examinateur).

Résumé : *Dans ce travail, nous nous intéressons aux systèmes de dialogue orientés tâche. Nous nous concentrons sur la différence de traitement de l'information et de l'utilisation de la mémoire, d'un tour de parole à l'autre, par l'humain et la machine, pendant une conversation écrite. Après avoir étudié les mécanismes de rétention et de rappel chez l'humain en dialogue, nous émettons l'hypothèse qu'un des éléments susceptibles d'expliquer les moindres performances des machines par rapport aux humains est leur incapacité à posséder une image de l'utilisateur, mais également une image de soi, explicitement convoquée pendant les inférences liées à la poursuite du dialogue. Améliorer la machine se traduit par trois axes. Tout d'abord, par l'anticipation, à un tour de parole, de l'énoncé suivant de l'utilisateur. Ensuite, par la détection d'une incohérence dans son propre énoncé, facilitée par l'anticipation du tour suivant*

de l'utilisateur en tant qu'indice supplémentaire. Enfin, par la prévision du nombre de tours de paroles restant dans le dialogue afin d'avoir une meilleure vision de la progression du dialogue, en prenant en compte la présence d'une incohérence dans son propre énoncé. Pour les mettre en place, nous exploitons les réseaux de mémoire de bout en bout, un modèle de réseau de neurones récurrent qui possède la spécificité de créer des sauts de réflexion, permettant de filtrer l'information contenue à la fois dans l'énoncé de l'utilisateur et dans celui de l'historique de dialogue. De plus, ces trois sauts de réflexion servent de mécanisme d'attention « naturel » pour le réseau de mémoire, à la manière d'un décodeur de transformeur. Pour notre étude, nous améliorons, en y ajoutant nos trois fonctionnalités, un type de réseau de mémoire appelé WMM2Seq. *Ce modèle s'inspire des modèles cognitifs de la mémoire, en présentant les concepts de mémoire épisodique, sémantique et de travail. Il obtient des résultats performants sur des tâches de génération de réponses de dialogue sur les corpus* DSTC2 *et* MultiWOZ *qui sont les corpus que nous utilisons pour nos expériences. Les trois axes mentionnés apportent deux contributions principales à l'existant. En premier lieu, ils complexifient l'intelligence du système de dialogue en le dotant d'un garde-fou. En second lieu, ils optimisent à la fois le traitement des informations dans le dialogue et la durée de celui-ci. Les résultats obtenus avec nos différentes mesures d'évaluation montrent l'intérêt d'orienter les recherches vers des modèles de gestion de la mémoire plus cognitifs afin de réduire l'écart de performance dans un dialogue entre l'humain et la machine.*

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03647756

Antoine SIMOULIN : antoine.simoulin@gmail.com

Titre : Plongements de phrases et leurs relations avec les structures de phrases

Mots-clés : traitement automatique des langues naturelles, plongements de phrases, apprentissage profond, réseaux de neurones structurés.

Title: *Sentence Embeddings and Their Relation with Sentence Structures*

Keywords: *natural language processing, sentence embeddings, deep learning, structured neural networks.*

Thèse de doctorat en informatique, Laboratoire de Linguistique Formelle, école doctorale Sciences Mathématiques de Paris Centre, Université Paris Cité, sous la direction de Benoît Crabbé (Pr, Université de Paris). Thèse soutenue le 07/07/2022.

Jury : M. Benoît Crabbé (Pr, Université de Paris, directeur), Mme Claire Gardent (DR, CNRS, rapporteuse), M. Éric Gaussier (Pr, Université Grenoble Alpes, rapporteur), Mme Rachel Bawden (CR, Inria, examinatrice), M. Loïc Barrault (MC, Le Mans Université, examinateur), M. Nicolas Brunel (Pr, ENSIEE, Laboratoire de Mathématiques et Modélisation d'Évry, examinateur).

Résumé : *Historiquement, la modélisation du langage humain suppose que les phrases ont une structure symbolique et que cette structure permet d'en calculer le sens par composition. Ces dernières années, les modèles d'apprentissage profond sont parvenus à traiter automatiquement des tâches sans s'appuyer sur une structure explicite du langage, remettant ainsi en question cette hypothèse fondamentale. Cette thèse cherche ainsi à mieux identifier le rôle de la structure lors de la modélisation du langage par des modèles d'apprentissage profond. Elle se place dans le cadre spécifique de la construction de plongements de phrases — des représentations sémantiques basées sur des vecteurs — par des réseaux de neurones profonds.*

Dans un premier temps, on étudie l'intégration de biais linguistiques dans les architectures de réseaux neuronaux, pour contraindre leur séquence de composition selon une structure traditionnelle, en arbres. Dans un second temps, on relâche ces contraintes pour analyser les structures latentes induites par ces réseaux neuronaux. Dans les deux cas, on analyse les propriétés de composition des modèles ainsi que les propriétés sémantiques des plongements.

La thèse s'ouvre sur un état de l'art présentant les principales méthodes de représentation du sens des phrases, qu'elles soient symboliques ou basées sur des méthodes d'apprentissage profond. La deuxième partie propose plusieurs expériences introduisant des biais linguistiques dans les architectures des réseaux de neurones pour construire des plongements de phrases. Le premier chapitre combine explicitement plusieurs structures de phrases pour construire des représentations sémantiques. Le deuxième chapitre apprend conjointement des structures symboliques et des représentations vectorielles. Le troisième chapitre introduit un cadre formel pour les transformer selon une structure de graphes. Finalement, le quatrième chapitre étudie l'impact de la structure vis-à-vis de la capacité de généralisation et de composition des modèles.

La thèse se termine par une mise en concurrence de ces approches avec des méthodes de passage à l'échelle. On cherche à y discuter les tendances actuelles qui privilégient des modèles plus gros, plus facilement parallélisables et entraînés sur plus de données, aux dépens de modélisations plus fines. Les deux chapitres de cette partie relatent l'entraînement de larges modèles de traitement automatique du langage et comparent ces approches avec celles développées dans la deuxième partie d'un point de vue qualitatif et quantitatif.

URL où le mémoire peut être téléchargé :
https://hal.archives-ouvertes.fr/tel-03791935

Chunxiao YAN : yanchunxiao5597@gmail.com

Titre : Complexité syntaxique et flux de dépendance. Études quantitatives dans les *treebanks* Universal Dependencies

Mots-clés : flux de dépendance, syntaxe de dépendance, complexité syntaxique, métrique, mémoire de travail, *Universal Dependencies treebanks*.

Title: *Syntactic Complexity and Dependency Flux. Quantitative Studies in Universal Dependencies Treebanks*

Keywords: *dependency flux, dependency syntax, syntactic complexity, metrics, working memory, Universal Dependencies treebanks.*

Thèse de doctorat en sciences du langage, MoDyCo, Université Paris Nanterre, sous la direction de Sylvain Kahane (Pr, Université Paris Nanterre). Thèse soutenue le 01/12/2021.

Jury : M. Sylvain Kahane (Pr, Université Paris Nanterre, directeur), M. François Lareau (Pr, Université de Montréal, Canada, rapporteur), M. Philippe Blache (DR, CNRS, rapporteur), Mme Marie Candito (MC, Université Paris-Diderot, examinatrice), Mme Marie-Catherine de Marneffe (Pr, The Ohio State University, Columbus, États-Unis, examinatrice), M. Kim Gerdes (Pr, Université Paris-Saclay, examinateur).

Résumé : *Nous nous intéressons à la complexité syntaxique et aux contraintes liées à la mémoire de travail chez l'humain. La mémoire de travail concerne non seulement la capacité de retenir des informations, mais aussi la capacité de les manipuler temporairement. Elle a été montrée limitée à 7 ± 2 éléments et est aujourd'hui actualisée autour de 4 selon Cowan. La limitation de la mémoire de travail peut rendre le traitement de certaines structures de phrase difficile, voire impossible. Dans cette thèse, nous nous penchons sur trois pistes d'étude : étudier et mesurer la complexité syntaxique sous différentes hypothèses cognitives, savoir s'il existe des limites à la complexité syntaxique dans les langues naturelles, et comprendre les phénomènes impliqués par les contraintes sur la complexité syntaxique.*

De ce fait, nous mesurons la complexité syntaxique en utilisant des métriques basées sur le flux de dépendance dans le corpus (les treebanks Universal Dependencies*). Ces métriques incluent non seulement des métriques devenues classiques comme la longueur de dépendance, des métriques proposées dans des travaux plus récents, mais aussi de nouvelles métriques également basées sur le flux de dépendance.*

En nous basant sur les résultats donnés par ces différentes métriques dans les plus de 100 langues appartenant à la collection des treebanks Universal Dependencies*, nous pouvons déterminer celles qui sont les plus appropriées pour étudier la complexité syntaxique. Nous montrons qu'il existe pour certaines métriques du flux des contraintes universelles, dont nous postulons qu'elles sont liées à la mémoire de tra-*

vail. Enfin, nous essayons également d'expliquer certains des phénomènes linguistiques observés dans nos données qui impliquent la complexité syntaxique.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-03649621
