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

François Buet, François Yvon

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.

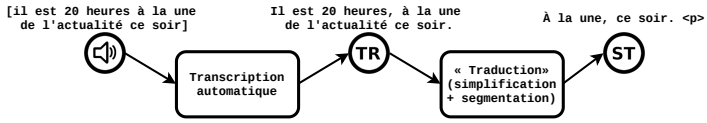

Figure 1. *Architecture globale pour le sous-titrage automatique. Nos expériences se concentrent sur la tâche de « traduction » de la transcription vers les sous-titres segmentés.*

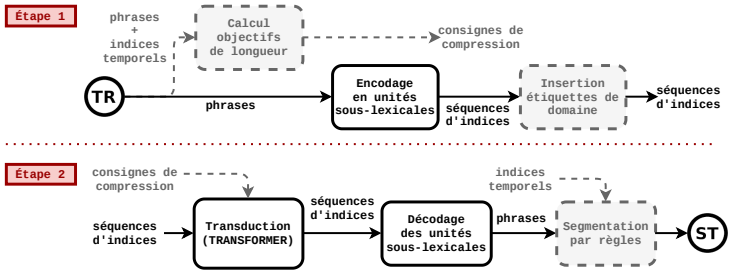

Figure 2. *Architecture détaillée pour la transduction de la transcription vers les sous-titres segmentés. En noir sont représentées les étapes partagées par tous les systèmes ; en gris et en pointillé celles réalisées seulement par certains systèmes : le calcul de consignes de compression (introduites 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 + LRPE$_{CpR = 0,75}$	28,6	35,3	50,7	0,34	3,1	10,0
+ BS + LRPE$_{CPS = 14,5}$	31,6	39,2	52,2	0,30	3,3	0,5
+ BS + LDPE$_{CpR = 0,75}$	27,8	34,8	50,6	0,34	3,1	9,3
+ BS + LDPE$_{CPS = 14,5}$	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 + LRPE$_{CpR^*}$	29,8	36,8	51,3	0,32	3,2	11,1
+ BS + LRPE$_{CPS^*}$	32,1	40,1	52,6	0,30	3,2	7,7
+ BS + LRPE$_{CPS^*}$ + 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.