Constitution d'un corpus de traduction de la parole : augmentation du corpus LibriSpeech

Ali Can Kocabiyikoğlu

Constitution d'un corpus de traduction de la parole : augmentation du corpus LibriSpeech

Ali Can Kocabıyıkoğlu

Sous la direction de Laurent Besacier

Tuteur : Olivier Kraif

Laboratoire d'Informatique de Grenoble — LIG

UFR LLASIC
Département Informatique intégrée en Langues, Lettres et Langages (I3L)

Mémoire de master 2 mention Sciences du langage - 20 crédits

Parcours : Industries de la langue

Année universitaire 2016-2017

Constitution d'un corpus de traduction de la parole : augmentation du corpus LibriSpeech

Ali Can Kocabıyıkoğlu

Sous la direction de Laurent Besacier

Tuteur : Olivier Kraif

Laboratoire d'informatique de Grenoble — LIG

UFR LLASIC
Département Informatique intégrée en Langues, Lettres et Langages (I3L)

Mémoire de master 2 mention Sciences du langage - 20 crédits

Parcours : Industries de la langue

Année universitaire 2016-2017

Remerciements

Je tiens tout d’abord à remercier M. Laurent Besacier de m’avoir encadré tout au long de ce mémoire de recherche, pour ses conseils précieux qu’il a apporté à mon travail, de m’avoir fait confiance tout au long du projet et de m’avoir guidé jusqu’au bout. De même, je tiens à remercier M. Olivier Kraif pour son encadrement continuel depuis le début de mon master.

Ensuite, je tiens à remercier les membres de jury Messieurs Georges Antoniadis et Benjamin Lecouteux qui ont accepté d’évaluer ce travail.

Je n’oublie pas tous mes collègues au LIG de l’équipe GETALP qui m’ont donné énormément de conseils, ont contribué à mon projet et m’ont encouragé quand j’en avais besoin.

Je tiens également à remercier toute l’équipe pédagogique du master IDL, pour ces deux années de master pleines d’apprentissages et de projets enrichissant qui m’ont servi tout au long de mes stages.

Je remercie de même tous mes camarades du master IDL, tout particulièrement William, Louise, Doriane, Anne-laure et Pauline qui ont été là tout au long la réalisation de ce mémoire.

Je remercie enfin Cécile Crépin et toute sa famille qui m’ont donné un foyer loin de chez moi, grâce à vous j'étais chez moi, chez vous.

UNIVERSITÉ Grenoble Alpes

Déclaration anti-plagiat
Document **à scanner** après signature
et **à intégrer** au mémoire électronique

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : KOCABIYIKOĞLU

PRENOM : Ali Can

DATE : 01/09/2017

SIGNATURE : [redacted]

Sommaire

Introduction

Ce mémoire de recherche, accompagné d'un stage de 6 mois au Laboratoire Informatique de Grenoble (LIG), s'inscrit dans le cadre de mon mémoire de fin d'études du Master sciences du langage, parcours industries de la langue (IDL) de l'Université Grenoble Alpes (UGA). Ce stage a été encadré par M. Laurent Besacier, professeur à l'UGA, et M. Olivier Kraif, maître de conférences à l'UGA, en tant que tuteur.

Au laboratoire informatique de Grenoble, le stage s'est déroulé au sein de l'équipe GETALP (Groupe d'Étude en Traduction Automatique des Langues et de la Parole). Cette équipe dynamique et pluridisciplinaire traite tout sujet impliquant la langue et l'informatique. Dans ce contexte, notre travail rentre dans le domaine de la traduction automatique, plus particulièrement dans la traduction automatique de la parole.

La traduction automatique (TA), ou *machine translation* (MT) en anglais, est un sous domaine du traitement automatique du langage (TAL) qui s'intéresse à la traduction d'une langue naturelle (texte ou parole) vers une autre langue à l'aide de logiciels. Cette tâche complexe, impliquant à la fois l'analyse linguistique fine d'une langue source et la génération du contenu linguistique dans une langue cible, nécessite diverses techniques de TAL, telle que la reconnaissance automatique de la parole (RAP). Étant l'un des sujets d'intérêt de l'Informatique depuis ses débuts, la TA représente aujourd'hui une industrie mondiale importante répondant à des besoins sociaux, gouvernementaux, commerciaux et militaires.

Depuis les premiers travaux conduits par l'un des pionniers du domaine, Yehosha Bar-Hillel, un long chemin a été parcouru en TA, suivant différentes approches méthodologiques dont les trois principales sont l'approche à base de règles, l'approche statistique et l'approche neuronale.

Ainsi, pour la traduction automatique de la parole (TAP), ou *Speech Translation* (ST) en anglais, la première approche consiste d'abord à transcrire automatiquement un signal de parole, puis à traduire automatiquement cette transcription dans une langue cible. La reconnaissance du flux de la parole étant un sujet de recherche à part entière, la TAP cumule les difficultés de ces deux domaines connexes.

Par ailleurs, en RAP, les approches statistiques s'appuyant sur un modèle du langage (ML) et un modèle acoustique permettent de déterminer les phrases les plus probables par calcul de probabilités. Dans cette approche, la décomposition du flux de parole dans une suite de phonèmes est nécessaire. Cependant, ces dernières années, l'augmentation

de la puissance de calcul et du stockage ainsi que les améliorations apportées aux technologies sous-jacentes ont mené à des approches neuronales.

En effet, au sein même de la TA et de la RAP, l'utilisation des réseaux de neurones profonds a réussi à faire ses preuves (Bérard et al., 2016 ; Bahdanau et al., 2015 ; Chorowski et al., 2015). Par l'intermédiaire d'un système à base de réseaux de neurones profonds avec un model *end-to-end,* il est possible de générer une séquence de sortie directement à partir d'un signal d'entrée sans passer par une transcription de la langue source. Dans les systèmes actuels de TAP, la méthodologie classique où la parole est d'abord transcrite puis traduite pourrait être changée avec une approche *end-to-end* (Bérard et al. 2016). Par exemple, pour les langues qui n'ont pas de système d'écriture ou qui ont un système d'écriture non-normalisé ou difficile, un tel système pourrait être utilisé pour traduire directement de la parole naturelle (Bérard et al. 2016). Cependant, un corpus parallèle (aligné et multimodal) comportant d'un côté l'enregistrement de la parole d'une langue source et de l'autre côté une traduction de cet enregistrement dans une langue cible, est nécessaire afin de réaliser des expériences sur de tels systèmes.

En vue de constituer un tel corpus, il existe de riches ressources contenant de nombreux livres numériques tel que le projet Gutenberg. Concernant l'enregistrement audio, le projet LibriSpeech propose un corpus de parole comportant 1000 heures de livres audio alignés avec leurs transcriptions.

Le travail effectué pendant ce stage a consisté à préparer un large corpus multimodal de données réelles (parole et texte) et aligné au niveau des phrases (utterances). Cette tâche nécessite un grand corpus composé de parole transcrite en grande quantité comme dans *LibriSpeech* dont on pourrait trouver la traduction dans la langue cible.

Ainsi donc, après ce premier état des lieux, mon travail débutera par une première étape de constitution d'un large corpus de données réelles (parole et texte), aligné au niveau des phrases (utterances), permettant de créer un corpus de traduction de la parole dans l'optique de réaliser une première expérience.

Pour ce faire, nous verrons dans un premier temps la traduction automatique au travers de deux de ses sous-branches : la traduction automatique statistique et la traduction automatique neuronale. Dans une optique de spécialisation de notre sujet, nous aborderons ensuite les ressources existantes, notamment les caractéristiques de *LibriSpeech* ainsi que ce qu'elle représente et sous quel angle nous allons l'étudier, ce qui nous mènera à la constitution même de notre corpus. Enfin, nous décrirons l'évaluation manuelle et les scores

d’alignement ajoutés entre les transcriptions et les traductions pour trier le corpus en fonction de ces scores.

Partie 1

-

Etat de l’art

Chapitre 1. Traduction Automatique

1.1 Traduction automatique statistique

La TA statistique part du constat que la création de règles nécessitant l'opinion experte des linguistes est trop coûteuse (Bérard et al., 2016). De plus, l'augmentation de la puissance de calcul des machines et la disponibilité de grands volumes de données comme Europarl[1] ont redirigé les chercheurs vers l'idée d'une TA empirique entrainée sur de gros volumes de données.

1.1.1 Principe général

La traduction automatique statistique (TAS) s'inscrit dans un contexte empirique où l'apprentissage est réalisé à partir des données. Ces données sont des collections de textes, c'est-à-dire des corpus. Les corpus parallèles comportant d'un côté des textes de la langue source et de l'autre côté les traductions dans la langue cible sont utilisés afin de créer des modèles de traduction. En outre, la vérification du contenu linguistique produit dans la langue cible est assurée par un modèle de langage extrait à partir d'un corpus monolingue.
L'approche statistique utilisée dans un modèle de TAS est fondée sur la distribution des probabilités. Dans cette approche, les probabilités sont utilisées pour l'attribution des scores aux événements par rapport aux fréquences rencontrées. (Koehn, 2010 : 9). Inspirée du modèle de la théorie de l'information de Shannon, un décodage est ensuite effectué afin de choisir la traduction la plus probable.

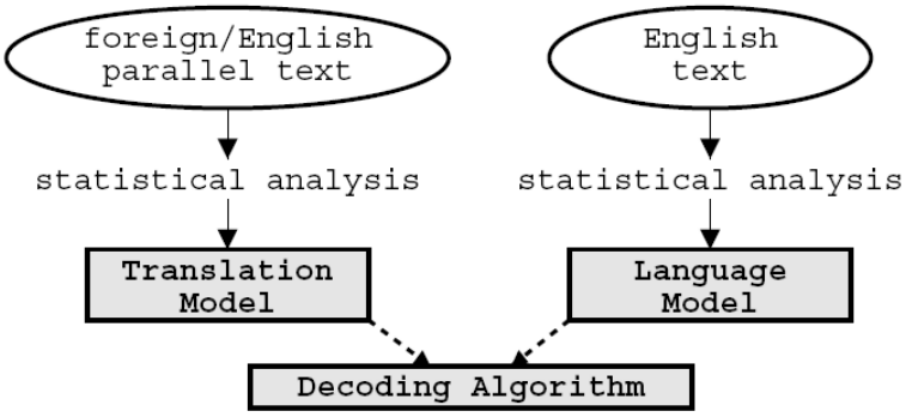


Figure 1 : Schéma d'un système de TAS (Besacier, Cours[1])

[1] Corpus parallèle des actes du parlement européen depuis 1996 jusqu'à nos jours.
Lien vers le site : http://www.statmt.org/europarl/

L'objectif d'un système de TAS pourrait être exprimé ainsi :

« Etant donné une phrase F dans une langue source (ex. français, espagnol), le but est de trouver une phrase Ê en langue cible (ex. anglais) qui maximise la probabilité conditionnelle d'avoir une phrase E en langue cible. » (Constant, 2009)

$$\hat{E} = argmax_E \, P(E|F) \qquad \textbf{Equation 1}$$

D'après la distribution de probabilité conditionnelle du théorème de Bayes :

$$argmax_{\hat{E}} \, P(E|F) = argmax_E \, \frac{P(F|E).P(E)}{P(F)} \qquad \textbf{Equation 2}$$

Le calcul de la fonction *argmax* ne dépend pas de la probabilité de la phrase F de la langue source et donc peut se simplifier en :

$$\hat{E} = argmax_E \, P(F|E).P(E) \qquad \textbf{Equation 3}$$

Les trois composants principaux d'un système de TAS sont :

- Un modèle de traduction
- Un modèle de langage
- Un décodeur

1.1.2 Modèle de traduction

L'un des composants principaux d'un système de TAS est le modèle de traduction extrait à partir d'un corpus parallèle aligné. Dans l'équation principale, celui-ci correspond au calcul de **P(F|E)**. L'apprentissage effectué sur le corpus aligné consiste à faire la correspondance entre la phrase en langue source et la traduction dans la langue cible. Différentes approches pourront être adoptées afin de réaliser cette tâche tels qu'un modèle à base de mots, de phrases et de syntaxe.

1.1.3 Modèle de langage

Les modèles de langages (ML) extraits à partir d'un corpus monolingue constituent une composante essentielle d'un système de TAS. Dans l'équation principale, le calcul de **P(E)** est réalisé par le ML. Il est utilisé pour assurer la génération des phrases acceptables en langue cible. Chaque hypothèse de traduction est associée à un score correspondant à la possibilité qu'une telle phrase soit produite dans telle langue (Koehn, 2010 : 10). Parmi les différentes approches adoptées, le modèle le plus utilisé est le modèle n-gramme.

1.1.4 Approches à base de mots

L'une des approches adoptées afin de réaliser la tâche de TAS est la décomposition des phrases en unités lexicales. Ce modèle provient de l'un des premiers travaux de *IBM Candide Project* par (Brown et al., 1980). Il s'agit d'une traduction de tous les mots isolés d'une phrase de la langue source vers les unités lexicales de la langue cible. L'estimation de la distribution de probabilités de la traduction lexicale est réalisée à partir d'un apprentissage sur la fréquence d'occurrences de mots alignés. Bien évidemment, un alignement simple des mots isolés ne suffit pas pour couvrir toutes les possibilités d'alignement. Autrement dit, un tel système est capable d'aligner un mot isolé à une ou plusieurs unités lexicales mais n'est pas capable du contraire.

1.1.5 Approches à base de segments

L'approche à base de segments, également appelée approche *phrase based* en anglais, est une approche plus récente utilisée dans le domaine. Beaucoup de systèmes actuels sont basés sur cette approche. Selon un travail conduit par (Koehn et al., 2003), les performances de ces systèmes sont supérieures à celles des systèmes utilisant l'approche à base de mots.

Dans un système à base de segments, les phrases sont d'abord segmentées en séquences. Ces dernières sont stockées dans des tableaux puis, selon leurs scores, mis en correspondance avec une ou plusieurs traductions possibles. Ces différentes associations sont ainsi traduites en séquence dans la langue cible. Elles pourront être réorganisées à l'aide d'un modèle de distorsion afin de trouver la meilleure configuration possible pour une phrase donnée. Cette dernière étape constitue l'intérêt principal de cette approche.

La figure ci-dessous démontre un exemple d'un système de TAS à base de segments :

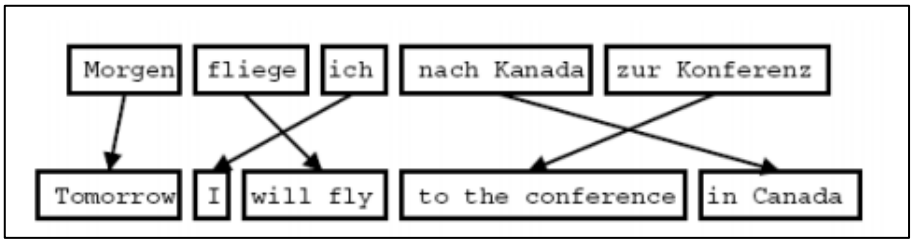


Figure 2 : Exemple de fonctionnement d'un système de TAS à base de segments (Besacier, Cours[1])

1.1.6 Approches syntaxiques

Un modèle de TAS utilisant une représentation arborescente semble naturel à imaginer, puisque la syntaxe à base de constituants utilise elle aussi une représentation sous forme d'arbre et de feuille. Dans cette représentation, les catégories syntaxiques sont représentées par les nœuds et les mots par les feuilles. Cette approche s'éloigne des premiers systèmes utilisés en proposant des techniques de *parsing* basées sur des probabilités.

L'un des avantages principaux d'un système à base de syntaxe porte sur la possibilité de réorganisation de la phrase en fonction de l'ordre syntaxique porté par la langue cible. Toutefois, ces systèmes ont besoin d'apprendre des règles à partir d'une grammaire hors-contexte bilingue. Ces règles sont extraites de l'alignement des mots et sont restreintes par les heuristiques pour la syntaxe (Specia, 2015).

1.2 Traduction automatique neuronale

Dernièrement, de nombreux travaux de recherches dans différents domaines du TAL, tel que RAP ou la synthèse vocale, ont été effectués sur la base des modèles fondés sur les réseaux de neurones appliqués au TAL : ce sont des traductions automatiques neuronales (TAN), ou en anglais *Neural Machine Translation* (NMT). Un réseau de neurones utilise différentes techniques d'apprentissage automatique afin de créer un ensemble de fonctions optimisées.

1.2.1 Principe Général

En TAN, les modèles de réseaux de neurones récurrents (RNN) capables de prendre en compte l'information contextuelle dans leur processus de décision sont utilisés. Dans un modèle de RNN, cette information contextuelle est assurée par une connexion en boucle permettant de prendre en compte à l'étape courante, une ou plusieurs informations prédites dans une étape précédente.

Les techniques d'apprentissage profond permettent la modélisation de données avec un haut niveau d'abstraction. L'intérêt des réseaux de neurones est leur capacité de classification et de généralisation. Autrement dit, un système de TAN modélise la globalité du processus de TA grâce aux réseaux de neurones profonds.

La plupart des systèmes actuels de TAN reposent sur le modèle d'*encoder-decoder* (Sutksever et al., 2014 ; Cho et al., 2014). Dans ces modèles, un réseau de neurones

récurrents d'encodage lit et encode une phrase en langue source dans un vecteur de longueur fixe (Bahdanau et al., 2014). Ensuite, la traduction est réalisée à partir de ce vecteur encodé.

Bien évidemment, dans un tel système de traduction neuronale de base, la génération de structures de longueurs variables à partir d'un simple vecteur de taille fixe n'est pas facile. Pour cela, différentes améliorations sur cette architecture de base sont réalisées par différents auteurs. L'une des extensions importantes faites par (Bahdanau et al., 2014) sur cette architecture, consiste en l'apprentissage automatique conjoint de l'alignement et de la traduction, permettant de capturer les informations d'alignement source-cible grâce à un modèle d'attention.

1.2.2 Architecture encoder-decoder

Un système de traduction neuronale utilise une architecture *encoder-decoder*, c'est-à-dire que la phrase source, par l'intermédiaire d'une séquence de vecteurs, est encodée, et grâce à cette séquence la phrase cible est créée mot par mot par le décodeur. Cette dernière production est produite de telle manière par le décodeur car ce dernier s'appuie sur la phrase source encodée mais aussi sur les mots qui ont précédemment été produits par le décodeur en langue cible.

La figure ci-dessous représente le processus de traduction neuronale de l'anglais vers le français basé sur un système d'*encoder-decoder :*

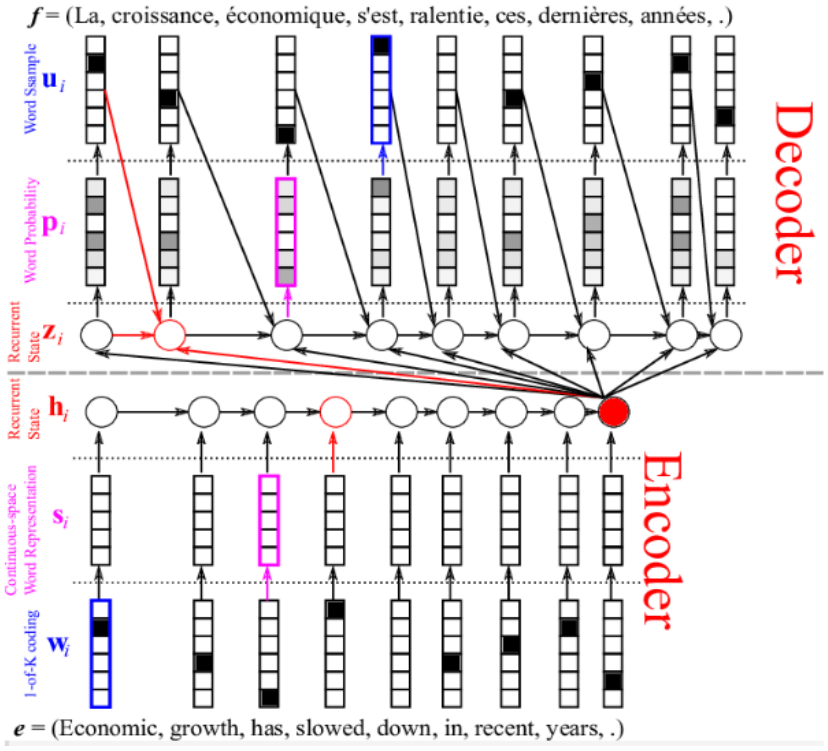


Figure 3: Architecture *encoder-decoder* d'un système de traduction neuronale (Cho, 2015)

Comme le montre la figure 3 par (Cho, 2015), l'encodeur, qui est un RNN, consomme une séquence d'entrée où chaque mot est représenté par un vecteur codé dit *one-hot* . Cette notation permet de représenter les caractéristiques principales dans un format plus adapté aux algorithmes de classification et de régression en apprentissage automatique. Ensuite, la représentation significative de chaque mot est projetée vers un espace continu de dimensionnalité plus faible grâce à une matrice. Cette dernière dispose d'autant de colonnes que la longueur du vecteur *one-hot* (Bérard et al., 2016). Dans la figure 3, la projection correspond à la deuxième partie de l'encodeur où les mots sont représentés par un espace continu. L'une des caractéristiques des RNN est leur capacité à réaliser un résumé séquentiel (Cho, 2015). Dans le décodeur, le résumé séquentiel du vecteur de l'espace continu est réalisé avec l'équation suivante :

$$h_i = \phi_\theta(h_{i_{-1}}, s_i) \qquad \textbf{Equation 4}$$

Le résultat obtenu par l'encodeur constitue la représentation de la phrase source sous forme de vecteur de taille fixe. Le dernier état de l'encodeur h_T permet d'initialiser le premier état du décodeur $z_{h..}$ Le décodeur utilise un autre RNN, dont (Cho, 2015) :

$$z_i = \phi_\theta(h_T, u_{i-1}, z_{i-1}) \qquad \textbf{Equation 5}$$

Cette partie correspond à celle entre le décodeur et l'encodeur dans la figure 3. Ensuite dans la partie décodeur, chaque mot cible se voit attribué un score correspondant à la possibilité qu'il suive tous les mots traduits précédemment (Cho, 2015). Pour chaque mot e = {economic,growth,has,slowed,down,in,recent,years}, sachant l'état caché, le score est calculé dans la première partie du décodeur de la figure 3 comme suivant :

$$e(k) = w_k^T z_i + b_k \qquad \textbf{Equation 6}$$

Enfin pour chaque mot, ces scores sont transformés en vraies probabilités par la fonction *softmax*. Ceci constitue la distribution de probabilités qui est utilisée dans la dernière étape de décodeur afin de générer la séquence de sortie.

1.2.3 Systèmes end-to-end

Dans un système de traduction neuronale *end-to-end*, l'intégralité de la chaine de traitement est effectuée en prenant pour entrée une phrase source et ensuite génère une phrase cible de sortie sans étape intermédiaire. L'optimisation simultanée de tous les paramètres nécessite un réseau de neurones ayant plusieurs couches.

Le fonctionnement d'un système end-to-end est expliqué par la société Systran comme suivant :

> […] comme dans le cerveau d'un humain, au sein de ce réseau de neurones unique, des sous-réseaux de neurones complémentaires s'activent au fur et à mesure de l'avancée de la traduction : un premier sous-réseau va traiter la phrase source pour en extraire le sens, un second, spécialisé dans la syntaxe (grammaire) ou la sémantique (sens des mots) va enrichir la compréhension, un troisième va contextualiser le contenu, un autre va attirer l'attention sur les mots clés.[2]

1.2.4 Alignements avec modèles d'attention

Comme expliqué dans la partie dédiée à l'architecture *encoder-decoder* (cf. 1.2.2), la phrase d'entrée est représentée par un vecteur de taille fixe. D'après les travaux effectués par (Cho, 2015), la performance de ces systèmes baisse rapidement au fur et à mesure que la longueur d'une phrase d'entrée augmente (Cho, 2015).

L'alignement dans un système de TAS classique représente un défi pour les systèmes de TA. Par conséquent, afin d'associer les mots de la langue source aux mots de la langue cible lorsqu'une phrase d'entrée est traduite vers une langue cible, il peut y avoir plusieurs possibilités comme présenté dans la figure 4.

En effet, la même approche utilisée dans un système statistique peut-être adoptée dans un système de traduction neuronale. Néanmoins, une extension à l'architecture *encoder-decoder* est proposée par (Bahdanau et al., 2014) à cet effet : les modèles d'attention.

---

[2] http://www.systran.fr/download/press-releases/fr/systran-pr-purely-neural-mt-engine-a-revolution-for-the-machine-translation-market-2016-08-30.pdf, consulté le 17/01/2017

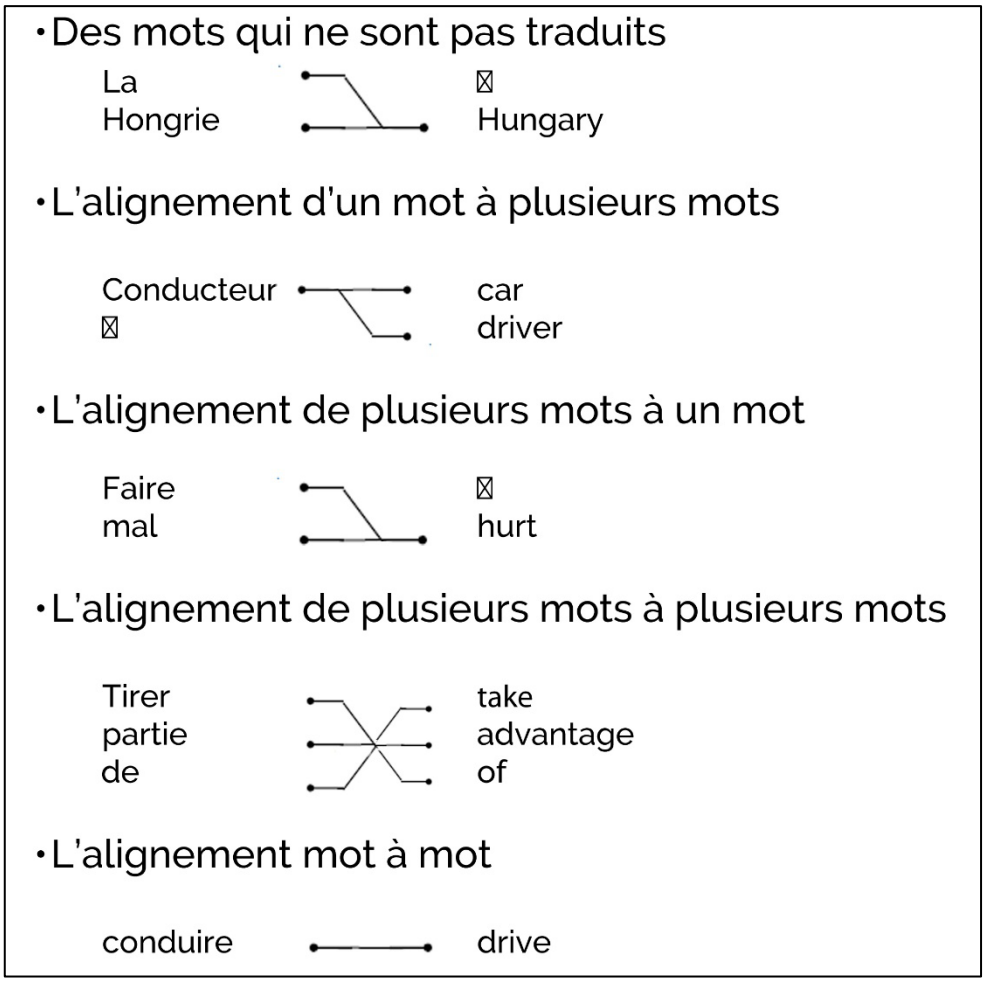


Figure 4: Processus d'alignement classique

Dans une architecture classique d'*encoder-decoder*, la phrase source (l'entrée) est une représentation d'un vecteur de taille fixe. Contrairement à un vecteur de taille fixe, les modèles d'attention permettent que la phrase source soit représentée par un vecteur de longueur variable. Les modèles d'attention, au lieu de prendre en compte seulement l'état caché précédent, prennent en compte l'ensemble des états cachés. A partir de cette représentation de longueur variable de la phrase d'entrée, le décodeur devrait être donc capable d'associer un ou plusieurs segments à un mot cible. La traduction est réalisée en fonction du poids attribué à chaque état caché. L'alignement est donc implicitement réalisé conjointement à la traduction.

Effectivement, l'apprentissage effectué en parallèle sur l'alignement et le modèle de traduction augmente les performances des systèmes, comme démontré dans le travail de (Bahdanau et al., 2015).

La figure ci-dessous illustre le mécanisme d'attention :

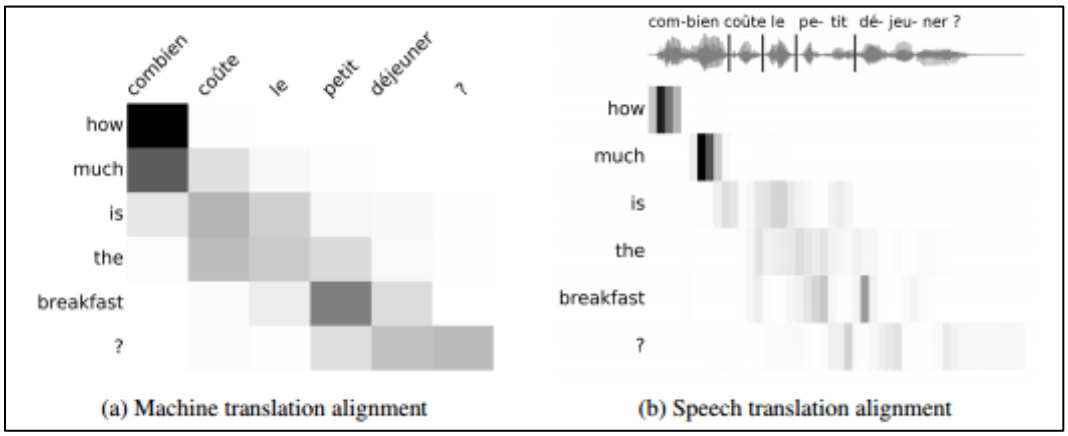


Figure 5 : Exemples de visualisation du modèle d'attention (Berard et al., 2016)

Le schéma ci-dessus (figure 5) représente le mécanisme d'attention utilisé pour aligner conjointement avec la traduction le texte et la parole.

1.2.5 Modèles de langue neuronaux

Pour s'assurer de la viabilité d'une séquence dans une langue donnée, la distribution de probabilités sur une séquence de mots est calculée par un modèle de langage. Dans un système de traduction statistique, la méthode la plus répandue consiste en l'utilisation du calcul des n-grammes.

L'estimation de la probabilité d'une séquence n mots, selon la règle de la chaîne la probabilité d'une phrase W est représentée en langage mathématique ainsi :

$$P(W) = \prod_i P(w_i \,|w_1 \dots w_{i-1)}) \qquad \textbf{Equation 7}$$

Toutefois, afin de prédire dans ces systèmes entre x différentes pour une de n mots, il est nécessaire de distinguer x^n configurations différentes. Le n dépassant les centaines, des milliers ou plus, les systèmes sont confrontés à un problème de *curse of dimensionality*, ou désigné en français comme fléau de dimensionnalité.

A cet effet, un modèle de langage neuronal améliore la qualité des modèles n-gram en utilisant le concept de *Word embedding*, ou embedding des mots proposés par (Bengio et al., 2003) qui pourront être utilisés dans tout système de TAL. Parmi les différents modèles proposés, deux architectures de modèles de langages neuronaux principales sont :

- Modèle de langage neuronal *feed-forward* par (Bengio et al., 2003)

– Modèle de langage neuronal récurrente par (Mikolov et al., 2010)

Le modèle *feed-forward* est le premier modèle de réseau de neurones, l'entrée avance dans une seule direction, c'est-à-dire vers l'avant, d'où la dénomination *feed-forward.* Par la suite, la couche d'entrée est projetée vers une couche intermédiaire contenant les vecteurs d'embeddings pour les mots ayant des représentations vectorielles similaires. Au final, ces vecteurs passent par une couche cachée non linéaire avant que la fonction *softmax (Bridle, 1990)* ne calcule la distribution de probabilités pour tout le vocabulaire dans la couche de sortie.

Ce processus est illustré par le schéma dans l'article publié par (Bengio et al., 2013) comme suivant :

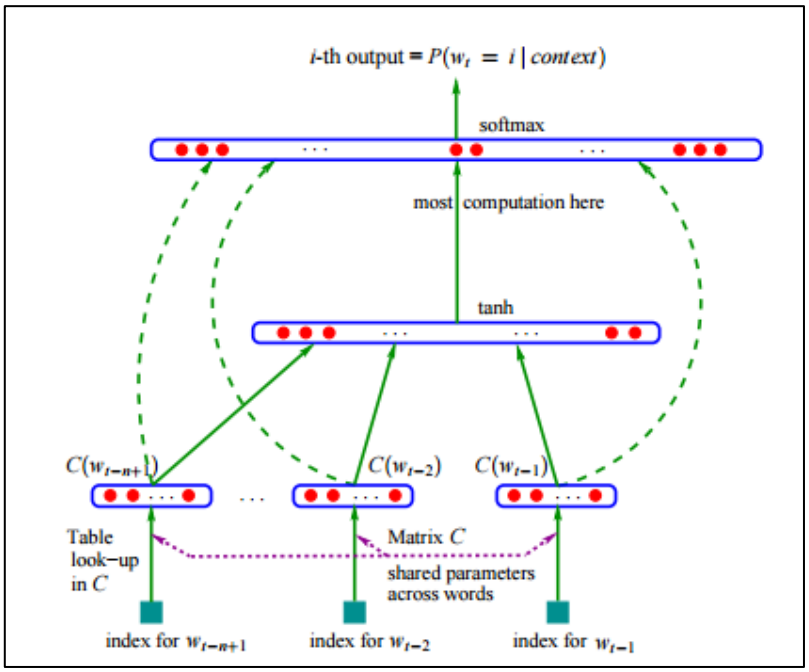


Figure 6: Processus du modèle feed-forward pour la création du modèle de langage neuronal (Bengio et al., 2003)

Le deuxième système proposé consiste à utiliser des réseaux de neurones récurrents pour la création de modèles de langage. Il est à noter que, comme dénoté dans l'équation 5, dans l'architecture *encoder-decoder* le ML cible est implicite.

Chapitre 2. Traduction automatique de la parole

Ces dernières années, en raison de l'utilisation croissante des appareils informatiques (téléphones, etc.), les systèmes reposant sur la parole ont progressé de façon constante. L'utilisation de programmes de visio-conférence, la publication de vidéos en direct sur les réseaux sociaux, *etc*., ont fait germer dans l'esprit des gens une idée de traduction automatique : ils s'exprimeraient dans leur langue maternelle et tout autre personne de langue étrangère recevrait le discours dans sa propre langue.

Dans les systèmes actuels, cette tâche est réalisée d'abord par la reconnaissance automatique du signal d'entrée dans une transcription textuelle et ensuite, la traduction est effectuée par des techniques de TA.

Contrairement à ces systèmes, comme expliqué dans l'article (Bérard et al., 2016), les systèmes de traduction neuronale avec un modèle « end-to-end », permettent d'effectuer une traduction neuronale directe de la parole. Un réseau de neurones artificiel unique comportant des couches cachées est capable de réaliser cette tâche complexe.

2.1 Reconnaissance automatique de la parole

La reconnaissance automatique de la parole (RAP), ou *automatic speech recognition* (ASR) en anglais, est la tâche dédiée à l'extraction automatique de paramètres d'un flux de parole composé des signaux acoustiques. Le signal brut étant impossible à traiter, les informations pertinentes sont d'abord échantillonnées en trames ou vecteurs afin d'extraire les paramètres acoustiques. L'approche principale utilisée dans ces systèmes est basée sur l'approche probabiliste avec des techniques d'apprentissage automatique. Toutefois, les approches neuronales ont démontré des résultats intéressants ces dernières années dans diverses tâches liées à la RAP.

2.1.1 Principe général

Les composants principaux d'un système de RAP statistique sont :

- Le modèle acoustique
- Le lexique phonétisé
- Le modèle de langage

Dans une approche empirique, le modèle acoustique est créé à partir de gros corpus de parole. Le modèle qui est utilisé dans la plupart des systèmes repose sur les modèles *Hidden Markov Models* (HMM), qui sont des transducteurs à états finis permettant de déterminer la probabilité d'une suite d'observations.

D'une manière formelle, pour trouver la phrase la plus probable W^*, l'équation qui maximise toutes les séquences de mots *W* sachant l'observation acoustique *X*, s'écrit comme suivant :

$$W^* = argmax_w \, P(W) \, . P(X|W) \qquad \textbf{Equation 8}$$

2.1.2 Création du lexique

Afin de partir du signal et d'arriver à une transcription, le modèle acoustique génère les hypothèses de phonèmes vérifiés par un lexique dit phonétisé qui génère les hypothèses de mots.

A partir du texte, une technique de transformation des graphèmes en phonèmes est utilisée : cette dernière consiste en l'estimation de la prononciation d'un mot à partir de l'orthographe. Cette tâche est plus ou moins complexe selon la transparence orthographique de la langue concernée. Par ailleurs, la création d'un lexique phonétisé nécessite le plus souvent un travail manuel.

2.2 La traduction directe de la parole

La tâche de traduction directe consiste à utiliser un système end-to-end : le modèle *seq2seq* avec l'architecture *encoder-decoder* permet justement une traduction directe grâce aux LSTM en projetant la séquence d'entrée vers un espace continu de dimensionnalité plus faible (Bérard et al., 2016). Cette représentation ensuite est transformée dans une séquence de sortie.

Un système fondé sur ce principe a été proposé par (Bérard et al., 2016). Dans ce système end-to-end, la séquence d'entrée étant le signal, il est donc possible de générer à partir du signal d'entrée la traduction textuelle correspondante. Dans le modèle *seq2seq*, la séquence d'entrée est d'abord transformée dans un vecteur de taille fixe. De plus, le système proposé par (Bérard et al., 2016), utilise un mécanisme d'attention où la phrase source est représentée par un vecteur de longueur variable.

Pour la parole, l'apprentissage peut être considéré comme la génération d'une séquence textuelle générée à partir d'une séquence du signal de la parole. Cependant, le

signal comporte des milliers de trames au lieu de quelques dizaines de mots (Chorowski et al., 2015). Le système proposé par (Bérard et al., 2016) utilise un modèle d'attention adapté pour le signal de parole proposé par (Chorowski et al., 2015). Vu le constat que le signal de parole peut être assez long, afin de ne pas traduire la même partie du signal deux fois, un modèle d'attention utilise un filtre évolutif capable de prendre en compte la pondération du cycle du temps précédent (Bérard et al., 2016).

D'après les expériences réalisées dans le cadre du travail de (Bérnard et al., 2016), malgré la taille du corpus et la constitution par synthèse vocale des données, les résultats obtenus sont proches d'un système de TAP baseline. De plus, le système montre une capacité à généraliser en ce qui concerne la variation inter-locuteurs. Néanmoins, la nature synthétique des données restreint la variabilité intra-locuteurs et rend la tâche un peu artificielle.

Par ailleurs, le système proposé par (Bérard et al., 2016) est implémenté avec la librairie *TensorFlow*[3] développée par Google. Dans cette librairie une implémentation du modèle séquence à séquence permettant de créer des systèmes end-to-end est disponible avec le nom *seq2seq.*

Dans le chapitre suivant, nous décrirons les sources existantes de corpus de parole, puis nous aborderons les caractéristiques de notre corpus de référence. Enfin, nous présenterons notre contribution à venir, c'est-à-dire l'augmentation du corpus avec l'ajout de traductions.

[3] https://www.tensorflow.org

Chapitre 3. Caractéristiques du Corpus de Référence : LibriSpeech

En vue d'expliquer au mieux le corpus développé dans le cadre de ce projet, il est essentiel d'exposer les caractéristiques et les particularités de notre corpus de référence. Les traitements effectués sur nos données sont étroitement liés aux particularités de LibriSpeech. Le schéma du fonctionnement général de notre projet fera l'objet d'un exposé plus détaillé ultérieurement dans le chapitre IV qui sera dédié à la constitution du corpus.

Dans ce chapitre, nous débuterons par une brève présentation des sources de corpus de parole existantes en abordant certaines des caractéristiques qui les composent. Ensuite, nous nous pencherons sur le projet LibriSpeech pour décrire en premier lieu le processus d'alignement réalisé en deux phases, puis aborderons, en second lieu la méthodologie de la segmentation des données.

3.1 Contexte général

Avant de citer quelques exemples de corpus de parole, il faut d'abord définir ce qu'est un corpus. De façon, purement littéraire, comme défini dans le dictionnaire Larousse, un corpus est le « Recueil de documents relatifs à une discipline, réunis en vue de leur conservation ». En informatique, plus particulièrement en TAL, le terme corpus est utilisé dans un sens plus étendu désignant « un ensemble de documents sélectionnés et assemblés à l'aide de critères explicites en vue d'un objectif clairement défini et stockés sous format électronique » (Kraif, Cours [2]). Afin de constituer une ressource qui sera utilisée dans le domaine de la traduction automatique de la parole, nous avons choisi d'utiliser par la suite ce terme de corpus dans un sens plus large comme une compilation systématique et structurée de données.

Les corpus d'enregistrements de la parole désigné comme *Speech Corpus* en anglais sont des collections d'enregistrements qui sont généralement accompagnées de fichiers de transcriptions correspondant à la parole enregistrée. Ils sont dotés des fichiers de métadonnées qui varient en fonction de la finalité du corpus. On distingue deux types de corpus de parole :

- Parole lue : les livres audio, diffusion de nouvelles, etc.
- Parole spontanée : dialogues, narrations, réunions, etc.

Les corpus de parole constituent une composante importante pour tout système ayant trait de près ou de loin à la parole naturelle. Contrairement aux corpus textuels, le nombre

de corpus disponibles librement et gratuitement est beaucoup plus limité pour la parole. Dans le domaine de la TA, c'est d'autant plus compliqué que les corpus doivent être alignés. Bien sûr, un tel corpus aligné peut être par nature multimodal, c'est-à-dire qu'un corpus de parole comporte d'un côté de la parole naturelle dans une langue source et de l'autre côté la traduction textuelle correspondante. Notons qu'avec les technologies de la synthèse vocale, il est possible de constituer des données dites *synthétiques*, constituées à partir de systèmes de synthèse vocale.

3.2 Sources de corpus de parole

Malgré la difficulté à trouver un corpus multimodal et aligné, il existe des projets libre de droit, comme *Gutenberg Project*[4] ou les émissions *Ted Talks*[5],qui peuvent être utilisés comme ressources dans un système de TAP. Ainsi, le projet *LibriSpeech*, basé sur le projet Librivox, propose 1000 heures de paroles alignées, segmentées et partitionnées.

3.2.1 Projet Gutenberg

Le projet Gutenberg est un projet d'accès libre dédié à la numérisation et à l'archivage des livres numériques. En 2016, cette archive propose au grand public plus de 53,000 livres au format électronique. La plupart des livres étant en anglais, il n'existe qu'un peu plus de 1500 manuscrits disponibles en français.

Les ouvrages sont en général issus de la littérature de genres textuels divers. En ce qui concerne la constitution du corpus, la plupart de ces ouvrages sont des ouvrages classiques. Ceci est important dans la constitution du corpus, notamment pour trouver l'ouvrage traduit dans la langue cible pour que le corpus soit parallèle. Aussi, dans l'archive de Gutenberg se trouvent plus de 2000 ressources audios : néanmoins ce ne sont pas que des livres lus par un humain, il s'agit également de livres sonorisés à l'aide de la synthèse vocale.

3.2.2 Ted Talks

Ted Talks sont des conférences internationales disponibles pour le grand public qui mettent principalement l'accent sur la technologie, le design et l'innovation. On trouve plus de 2400 conférences à disposition en ligne. La mission des *Ted Talks* est résumée par leur slogan : « partager les idées qui le méritent ». Il s'agit d'une prise de parole un peu

[4] https://www.gutenberg.org/
[5] http://www.ted.com/

particulière, car le but est de présenter une nouvelle idée, mais surtout d'intéresser le public en un temps limité : les conférenciers n'ont que 18 minutes pour développer leurs idées.

Les communautés locales ont également droit d'organiser leurs propres événements, désignés sous le nom « Tedx », tout en respectant l'esprit des conférences TED. Il est possible de trouver donc quelques conférences en français natif. Toutefois, l'intérêt principal des conférences repose sur le fait que les traductions sont faites manuellement par un comité de traducteurs volontaires. Par contre, l'alignement est nécessaire afin de constituer un corpus bilingue exploitable.

3.2.3 Librivox

Inspirée du projet Gutenberg, Librivox est une archive numérique sous forme de bibliothèque spécialisée sur les livres audio. Lues et enregistrées par des bénévoles, ce projet comporte plus de 10 000 ressources mises à disposition dans plus de 30 langues différentes. Comme le projet Gutenberg, la majorité des ressources sont en anglais mais plus de 500 livres audio sont également disponibles en français.

Quant à la qualité des enregistrements, leur site officiel présente des critères selon lesquels tout enregistrement compréhensible et fidèle à la traduction peut faire partie de cette ressource. La qualité des enregistrements dépend donc d'un locuteur (reader) à un autre.

Une autre caractéristique des enregistrements LibriVox est le fait d'avoir de multiples lecteurs pour un livre donné, voir pour un chapitre donné. Evidemment, un livre entier, notamment lorsqu'il s'agit d'un livre relativement long, n'est pas lu par un lecteur unique.

3.2.4 Corpus utilisés dans le domaine de la TAP

Les projets que nous venons de citer sont des ressources riches qui s'inscrivent dans un cadre plus large et qui ont donné suite à la création d'autres projets comme LibriSpeech. Nous nous intéresserons dans cette partie aux corpus qui sont constitués pour être utilisés dans les systèmes de traduction de la parole.

Comme cité dans l'article de (Bérard et al. 2016), *Fisher and Callhome Spanish--English Speech Translation Corpus* (Post, Matt, et al. 2013) est constitué de 38 heures de parole, de conversations téléphoniques fournies avec leurs transcriptions et traductions correspondantes en espagnol et en anglais. Nous pouvons aussi mentionner le corpus synthétique constitué à partir du corpus *BTEC* afin de réaliser des expérimentations sur le système end-to-end de TAP (Bérard et al. 2016). En outre, le corpus *Microsoft Speech Language Translation* (MSLT) (Federmann et al., 2016) est construit à partir de

conversations Skype autour des sujets simulés en anglais, français et allemand et comporte environ 4 heures de parole dans ces trois langues.

Les approches statistiques utilisées dans les systèmes de traduction nécessitent des données à grande échelle pour l'apprentissage. Pour la traduction automatique, les ressources telles que Europarl[6] et OpenSubtitles[7] comportent des milliers de phrases alignées permettant aux systèmes de faire un apprentissage. Alors que lorsqu'il s'agit de données de la parole, la disponibilité de ces ressources est beaucoup plus faible.

3.3 Présentation du projet LibriSpeech

Sous ensemble du projet LibriVox, le projet LibriSpeech correspond à des enregistrements de livres audio comportant 1000 heures de parole lue accompagné de leurs transcriptions. Celui-ci est constitué à partir des enregistrements provenant de LibriVox après avoir été soigneusement segmenté et aligné. Ce corpus s'inscrit bien dans le cadre des travaux réalisés sur la parole puisqu'il s'agit des données structurées qui sont bien documentées. Le travail préalable effectué sur les données a ciblé :

- L'alignement
- La création de données d'entraînement prétraitées
- La segmentation et le partitionnement de données

3.3.1 Processus d'alignement de Librispeech

D'après l'article publié par (Panayotov et al., 2015), les livres audio sont alignés en deux étapes. En premier lieu, l'algorithme de *Smith-Waterman*, algorithme utilisé initialement dans la bio-informatique pour l'alignement de séquences d'ADN, est appliqué au signal et au texte, permettant ainsi de trouver la meilleure région d'alignement entre l'audio et le texte. Celle-ci est trouvée à l'aide d'un calcul de mesure de distance entre une référence et une hypothèse. Contrairement à la distance Levenshtein ce calcul n'utilise pas d'office toute la référence ou l'hypothèse du début à la fin (Panayotov et al., 2015). De plus, avec ce processus différents poids sont associés aux différents types d'erreurs.

---

[6] Corpus parallèle des documents bilingues du parlement européen disponible en 21 langues. Plus d'informations sur ce corpus est disponible dans l'article (Koehn, 2005)

[7] Alignement des sous titres disponible sur le site www.opensubtitles.org. Ce corpus contient 22 millions de phrases alignés dans différentes langues. Plus d'informations sur cette ressource pourrait être trouvé dans l'article de (Tiedmann, 2009).

Ainsi, comme étape préalable, une étape de normalisation sur le texte est réalisée afin d'exclure les ponctuations, les abréviations et les mots non standards après que le texte ait été converti en majuscules. Contrairement à un système de RAP, pour un système de traduction, les ponctuations et la casse sont porteuses de sens. De ce fait, nous avons réalisé une étape d'association de la transcription avec le texte original du livre afin d'inclure les ponctuations lors de la constitution de notre corpus.

Dans le projet LibriSpeech, avant de passer à la deuxième étape d'alignement, l'audio est réparti en segments qui font au plus 35 secondes en utilisant un algorithme de programmation dynamique. Celui-ci présente, en effet, un deuxième challenge pour la constitution d'un corpus de traduction : les segments correspondent à la durée et non à la phrase. En effet, les transcriptions fournies pour un chapitre, qu'importe la donnée, comportent une segmentation qui peut correspondre à une, plusieurs ou même à une phrase incomplète. Etant donné que l'alignement bilingue des chapitres est réalisé au niveau des phrases, afin de trouver la correspondance entre une transcription d'un segment de parole et sa traduction, il serait idéal de se baser sur une segmentation au niveau des phrases.

Dans la deuxième phase d'alignement, chaque segment issu de la première étape d'alignement est confronté à un graphe de décodage qui se présente sous la forme d'un transducteur à états finis basé sur les modèles de langages bi-grammes au niveau des phonèmes. Celui-ci permet d'exclure des segments qui sont décalés par rapport à l'enregistrement audio. De plus, les insertions, délétions, substitutions, disfluences involontaires, *etc*. sont exclus également dans cette étape.

En outre, le modèle utilisé dans cette partie est un modèle adapté au lecteur (speaker adapted model) permettant d'exclure les décalages entre l'enregistrement et la transcription. Comme le projet LibriSpeech fait partie d'un corpus à grande échelle, la perte de certains segments n'est donc pas significative.

En revanche, lorsqu'il y a un segment exclu du LibriSpeech, il est exclu des enregistrements et donc des transcriptions. Les passages qui ne sont pas lus restent cependant intactes dans les chapitres originaux. Supposons que nous avons une phrase qui est correctement alignée, quand une partie de cette phrase n'est pas lu, dans la traduction nous aurons théoriquement une phrase qui ne correspondrait pas à la totalité de cette phrase.

Quelques exemples de segmentation de LibriSpeech :

1) **Une phrase coupée sur plusieurs segments**

a) Phrase anglais: Alice replied, rather shyly, 'I--I hardly know, sir, just at present--at least I know who I WAS when I got up this morning, but I think I must have been changed several times since then.'
b) Phrase français: Alice répondit, un peu confuse : « Je — je le sais à peine moi-même quant à présent. Je sais bien ce que j'étais en me levant ce matin, mais je crois avoir changé plusieurs fois depuis.
c) Transcription:
    a. <début> AND ADDRESSED HER IN A LANGUID SLEEPY VOICE WHO ARE YOU SAID THE CATERPILLAR THIS WAS NOT AN ENCOURAGING OPENING FOR A CONVERSATION **ALICE REPLIED RATHER SHYLY I** <fin>
    b. <début> **I HARDLY KNOW SIR JUST AT PRESENT AT LEAST I KNOW WHO I WAS WHEN I GOT UP THIS MORNING BUT I THINK I MUST HAVE BEEN CHANGED SEVERAL TIMES SINCE THEN** WHAT DO YOU MEAN BY THAT SAID THE CATERPILLAR STERNLY EXPLAIN YOURSELF <fin>

2) **Une phrase correspondant à un segment**

a) Phrase anglais: Poor Alice!
b) Phrase français: Pauvre Alice!
c) Transcription: <début> POOR ALICE <fin>

3) **Concaténation de plusieurs phrases dans un segment**

a) Phrases anglais:
    a. 'I can't explain MYSELF, I'm afraid, sir' said Alice, 'because I'm not myself, you see.'
    b. 'I don't see,' said the Caterpillar.
    c. 'I'm afraid I can't put it more clearly,' Alice replied very politely, 'for I can't understand it myself to begin with; and being so many different sizes in a day is very confusing.
b) Phrases français:
    a. « Je crains bien de ne pouvoir pas m'expliquer, » dit Alice, « car, voyez-vous, je ne suis plus moi-même.
    b. » « Je ne vois pas du tout, » répondit la Chenille.
    c. « J'ai bien peur de ne pouvoir pas dire les choses plus clairement, » répliqua Alice fort poliment ; « car d'abord je n'y comprends rien moi-même. Grandir et rapetisser si souvent en un seul jour, cela embrouille un peu les idées.
c) Transcription:
    <début> I CAN'T EXPLAIN MYSELF I'M AFRAID SIR SAID ALICE BECAUSE I'M NOT MYSELF YOU SEE **(a)** I DON'T SEE SAID THE CATERPILLAR **(b)** I'M AFRAID I CAN'T PUT IT MORE CLEARLY ALICE REPLIED VERY POLITELY <fin>
    <début> FOR I CAN'T UNDERSTAND IT MYSELF TO BEGIN WITH AND BEING SO MANY DIFFERENT SIZES IN A DAY IS VERY CONFUSING **(c)** IT ISN'T SAID THE CATERPILLAR WELL PERHAPS YOU HAVEN'T FOUND IT SO YET SAID ALICE <fin>

Figure 7: Exemples de segmentation des transcriptions de LibriSpeech

3.3.2 Organisation et segmentation des données de Librispeech

Après la première étape de segmentation, les données sont divisées en fragments plus petits en utilisant deux techniques : pour les données d'entraînement, un *split* à chaque intervalle de silence de plus de 0.3 secondes et pour les données de développement et de test avec un *sentence split*. Celui-ci est la raison pour laquelle nous avons adopté des approches différentes dans le traitement des données d'entraînement et de développement.

La restructuration du corpus est réalisée en vue de mettre en œuvre une tâche d'apprentissage automatique. Tout d'abord, ce corpus de 1000 heures (dont 40 heures de développement et de test) est divisé en 3 volumes pour l'entraînement : un premier de 100 heures, puis de 360 heures, et un dernier de 500 heures. L'attribution de segments audio par rapport à ces volumes est réalisée avec le WER[8], c'est-à-dire que pour les fichiers audio de meilleure qualité (qualité d'enregistrement, lecture par des locuteurs dont l'accent se rapproche de l'accent américain avec moins de WER) sont attribués au volume 'clean' de la parole. Les données représentant plus de WER sont choisies pour être classées dans la partie 'other'. (Panayotov et al., 2015). Dans ces deux catégories, 20 locuteurs homme et femmes sont ensuite choisis et classés dans un corpus de développement et de test.

En vue de s'assurer qu'il n'y ait pas de chevauchement de locuteurs entre les sous corpus d'entraînement, de développement et de test, plusieurs étapes de traitement sont effectuées. Par exemple, les techniques de suivi de locuteurs (*speaker diarization)* impliquant la segmentation et le regroupement automatique par locuteur sont utilisées afin de détecter les chapitres correspondant aux locuteurs.

Ainsi, pour différentes parties du corpus, pour des raisons d'équilibre, chaque locuteur est limité à une certaine durée d'enregistrement (Panayotov et al., 2015). Comme cité dans l'article (Panayotov et al., 2015), le projet similaire à LibriSpeech, VoxForge[9] issu d'un effort collectif de la collecte de la parole lue comporte des déséquilibres par locuteur et par sexe. En vue de constituer des modèles acoustiques plus représentatifs, LibriSpeech est fondé sur ce principe d'équilibre. Cependant, ceci a une incidence sur la constitution d'un

[8] Word Error Rate: Métrique calculé correspondant au taux d'erreur mots calculé par le métrique de la distance Levenshtein.

[9] Le projet VoxForge comporte la transcription d'environ 130 heures de parole distribuée avec une licence libre. Les transcriptions sont réalisées par des volontaires comme dans le projet LibriVox. Plus d'informations sur ce projet est disponible sur leur site officiel : http://www.voxforge.org

corpus de traduction. En effet, lorsque nous avons deux chapitres alignés, la transcription correspond seulement à une sous partie d'un chapitre. La limitation sur la durée de locution étant imposé pour chaque locuteur, la plupart des transcriptions correspondent donc à des sous parties des chapitres et non à l'intégralité.

Le répartition des données de Librispeech est résumée dans le tableau suivant :

Sous-ensemble	**Durée de lecture (h)**	**La durée maximale de lecture par personne (h)**	**Locuteurs femmes**	**Locuteurs hommes**	**Nombre de locuteurs total**
Dev-clean	5.4	8	20	20	40
Test-clean	5.4	8	20	20	40
Dev-other	5.3	10	16	17	33
Test-other	5.1	10	17	16	33
Train-clean-100	100.6	25	125	126	251
Train-clean-360	363.6	25	439	482	921
Train-other-500	496.7	30	564	602	1166

Tableau 1: Structure du corpus Librispeech

Comme schématisé dans le tableau 1, les données de LibriSpeech sont réparties selon deux critères : la qualité d'enregistrement et la durée des segments. Les quatre premiers sous-ensembles du corpus représentent les données de développement et de test qui font au plus 8 à 10 minutes d'enregistrement pour chaque locuteur. De la même manière, le corpus d'entraînement est partitionné en termes de qualité et durée d'enregistrement afin de faciliter son utilisation.

3.4 Bilan des corpus et contribution à venir

Notre projet est une augmentation de LibriSpeech pour les systèmes de traduction automatique. A part le projet Gutenberg, il existe d'autres collections de livres numériques telles que Wikisource, Gallica, beq, *etc*. Les enregistrements de LibriSpeech, issus du projet Librivox, se basent sur la lecture des documents du projet Gutenberg. Pour les segments de paroles de Librispeech, nous partons de l'hypothèse que nous pourrions aligner les transcriptions, la phrase dans le livre électronique anglais (avec les ponctuations), et les alignements dans la langue cible (les traductions) justement grâce au format électronique de cette traduction du même texte à partir d'une collection de livres numériques.

Le contexte de notre contribution et bilan des corpus est illustré dans la figure ci-dessous :

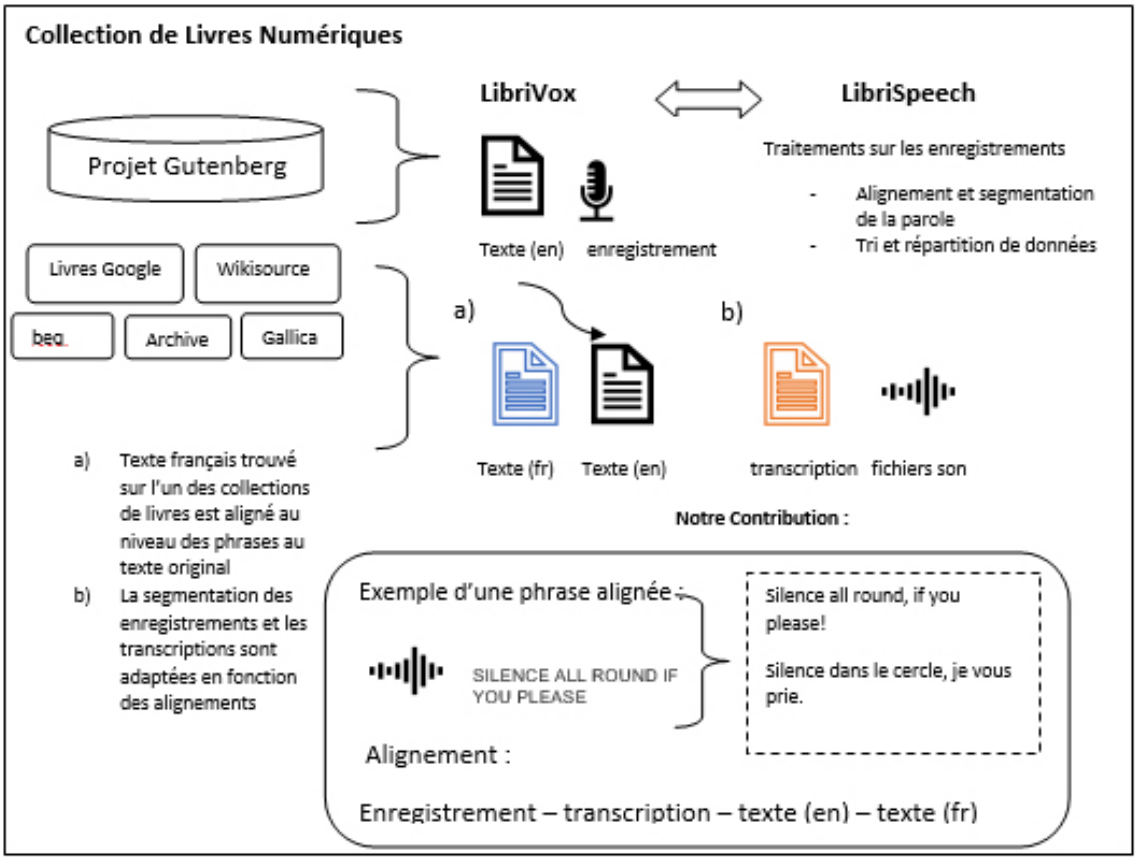


Figure 8: Schéma du bilan des sources existantes et notre contribution à venir

Partie 2

-

Constitution du corpus

Chapitre 4. Constitution du corpus

Une présentation globale des étapes de traitement ainsi que des outils utilisés dans la méthodologie est disponible en annexe 1.

Nous avons détaillé dans le chapitre précédent les caractéristiques des corpus de parole existants, notamment ceux de LibriSpeech. Dans ce chapitre, nous nous pencherons sur la constitution de notre corpus et sur la méthodologie que nous avons suivie afin d'exploiter les données de LibriSpeech, dans le but de constituer un corpus de traduction à partir de ces données.

Tout d'abord, nous décrirons la méthodologie utilisée pour le recueil des données à partir des métadonnées fournies par LibriSpeech. Nous détaillerons les sources qui nous ont servi à récolter les données sur le Web, ainsi que les techniques utilisées pour faciliter cette tâche. Nous aborderons également la structure de notre corpus et la technique d'extraction automatique du contenu des sites Web employée dans le but d'y récupérer des données.

Dans un second temps, nous décrirons l'étape de préparation des données pour l'alignement, c'est-à-dire les prétraitements appliqués sur les données. Nous aborderons d'abord l'étape d'extraction de chapitres, puis les traitements linguistiques effectués sur ces données.

En troisième partie, nous nous pencherons sur l'étape d'alignement. Pour cela, nous décrirons l'outil *hunalign* (Varga et al., 2005), et les étapes que nous avons suivies afin d'améliorer les alignements. Ensuite, nous présenterons l'étape de post traitement effectuée après l'utilisation de l'outil d'alignement.

Ensuite, nous aborderons les traitements réalisés sur la parole. La segmentation des données au départ, ainsi que les étapes d'alignement et de transcription forcés utilisés afin de resegmenter les fichiers et les transcriptions en fonction des chapitres alignés.

Finalement nous présenterons brièvement l'interface Web développée, qui permet d'une part la visualisation des alignements, et d'autre part l'écoute des segments de sons en parallèle.

4.1 Méthodologie pour le recueil du corpus

4.1.1 Exploitation des métadonnées fournies par Librispeech

LibriSpeech est fourni avec des métadonnées permettant aux utilisateurs d'accéder à la plupart des informations quant à la constitution de LibriSpeech. Ces métadonnées comprennent :

- Liens de téléchargement des fichiers sons, les checksum MD5 et la taille des fichiers téléchargés
- Annotations des fichier sons : bruité, mono/multi-speaker
- Annotations sur les chapitres : lecteur, durée, lien, numéro du projet, etc.
- Annotation sur les lecteurs : nom, natif/non-natif, sexe, WER, notes
- Auteurs, genres, détails sur la description de certains livres obtenus sur les fichiers RDF/XML de LibriVox

Ces métadonnées pourront servir lors de l'utilisation d'un système de RAP. Cependant, elles ne sont pas suffisantes pour trouver les œuvres cibles afin de constituer un corpus de traduction. De ce fait, la première partie de notre travail est consacrée à la complétion de ces métadonnées.

Les 1 000 heures de parole de LibriSpeech proviennent de 5 831 chapitres, soit 1 568 livres. Comme expliqué dans le chapitre III (cf. 3.3), LibriSpeech est une ressource basée sur les lecteurs (reader) et non sur les livres. Par conséquent, le nombre de chapitres est relativement réduit par rapport au nombre de livres disponibles.

En vue de trouver les œuvres traduites dans une langue étrangère, il faut d'abord trouver les titres traduits de ces ouvrages. Cette tâche qui peut paraître évidente, nécessite toutefois une étape de recherche et de vérification des différentes sources. Bien qu'il existe des bases de données telles que Wikipédia, BnF, WorldCat, *etc.*, la recherche bilingue des ouvrages implique la vérification de plusieurs sources afin de s'assurer que le nom traduit d'un tel ouvrage correspond à un autre dans une langue étrangère. Les traducteurs utilisent des stratégies diverses pour assurer la mission de médiateur interculturel, en traduisant parfois les titres de façon pragmatique en s'appuyant sur le sens plutôt que la traduction mot-à-mot.

Quelques exemples de traduction pragmatique de titres d'ouvrages :

Titre Anglais	Titre Français
In Search of the Castaways	Les enfants du Capitaine Grant
Clue of the Twisted Candle	Une lueur dans l'ombre
Off on a comet	Hector Servedac

Tableau 2: Exemples de traduction pragmatique de titres d'ouvrages

En plus des titres traduits, il est nécessaire de trouver les liens des livres électroniques des œuvres libres de droit. A cet effet, il existe des sources dans différentes langues. Les ressources que nous avons principalement utilisées sont décrites dans ce qui suit :

- **Wikisource :** une bibliothèque libre et gratuite disponible en plusieurs langues. Wikisource étant un *wiki*, tout individu peut modifier les pages. Il existe des statuts différents pour les livres disponibles sur le site. Les livres numérisés peuvent éventuellement comporter des pages manquantes ou abîmées. Certaines pages sont corrigées par les contributeurs de Wikisource et sont vérifiées au fur et à mesure par leur comité éditorial. Wikisource dispose d'un logiciel libre *WSexport* permettant aux utilisateurs d'exporter des livres en plusieurs formats. Il est donc possible d'extraire automatiquement un livre en format texte. Wikisource permet également un niveau de granularité supplémentaire ; il est effectivement possible de télécharger des chapitres, indépendamment des livres.
- **Gallica :** Bibliothèque numérique de la Bibliothèque Nationale de France et de ses partenaires. Elle comporte environ 3 millions de documents tels que des manuscrits, photographies, tirages, *etc*. Contrairement au Wikisource, les livres, dont le processus de correction des erreurs d'OCR n'est pas complété, ne sont téléchargeables qu'au format image. En ce qui concerne les livres téléchargeables au format texte, deux formats sont distribués : OCR brut et OCR corrigé.
- **Gutenberg :** le projet Gutenberg contient environ 1 500 œuvres en langue française et téléchargeables au format texte. Ces œuvres sont soigneusement numérisées et ne contiennent pas d'erreurs d'OCR.

- **Beq :** Bibliothèque électronique du Québec. Elle contient environ 2 700 volumes numérisés et soigneusement vérifiés. On y trouve beaucoup d'œuvres francophones, notamment celles issues de la littérature québécoise. Par contre, les livres ne sont téléchargeables qu'aux formats pdf et epub.
- **UQAC :** Bibliothèque numérique de l'Université de Québec à Chicoutimi, entièrement réalisée par des bénévoles et comportant environ 6 000 œuvres libres de droit. Cette ressource contient une bonne quantité des classiques des sciences sociales, disponibles dans les formats pdf et epub.
- **Livres Google :** contrairement aux ressources décrites auparavant, Google livres n'est pas une ressource complètement libre de droit. Cependant, dans le cadre du projet de préservation numérique de Google, les œuvres tombées dans le domaine public sont accessibles dans leur intégralité. Ces livres sont téléchargeables au format pdf. Ces livres comportent toutefois une particularité pour les outils de conversion de pdf : ils ne sont pas convertibles au format texte automatiquement ; c'est-à-dire que certaines pages sont conservées au format image, ce qui fait échouer la plupart des outils de conversion de formats.

Toutes ces ressources ainsi que d'autres non présentées ci-dessus comportent des particularités spécifiques. Etant donné que les textes des enregistrements sont basés sur les textes du projet Gutenberg, il serait idéal de préférer les textes français de Gutenberg. Cependant, parmi les 1 568 livres du LibriSpeech, nous n'avons trouvé que 47 livres, soit 516 chapitres, ce qui correspond à environ 84 heures de parole lue. Cette quantité ne couvre donc pas une grande partie des données de LibriSpeech.

Par ailleurs, les livres numériques nécessitent une étape de correction manuelle. En effet, les techniques basées sur un apprentissage à partir d'un corpus s'appuient sur des textes bien formés. Par exemple, l'étape de tokénisation en phrases (*sentence split)* part du principe que les ponctuations d'un texte sont correctes. Cependant, tous les textes ne sont pas au format texte et certains comportent des erreurs causées par la reconnaissance optique des caractères.

A part cela, certains textes font partie d'une traduction en ancien français. Nous rencontrons notamment ce phénomène pour des livres qui sont tombés dans le domaine public et qui sont disponibles sur Gallica et Google Livres. Lorsqu'il y a plusieurs sources

disponibles pour un même livre, nous avons donc ajouté les différentes sources disponibles dans la base de données au lieu de mettre un lien unique pour chaque livre.

Les métadonnées que nous avons ajoutées manuellement sont :

- Le titre traduit
- La source
- Le lien de téléchargement
- L'auteur : pour certains livres dont le nom de l'auteur pourrait aider à la recherche de sources différentes ou à la désambiguïsation, comme pour la République de Platon par exemple.
- Le fichier : le chemin relatif du fichier pour les fichiers téléchargés automatiquement

Ces métadonnées ajoutées ne sont pas exhaustives, mais elles sont toutefois suffisantes pour obtenir des données exploitables. Il serait envisageable d'ajouter d'autres informations par rapport aux livres comme le nom du traducteur et l'édition des textes. Il faudrait noter que d'une édition à une autre, les textes peuvent subir d'importantes modifications. Cependant, notre chaîne de traitement est appliquée sur une grande quantité de données et il est difficile de prendre en compte une granularité plus fine d'informations relatives aux livres. De plus, notre objectif étant d'aligner la parole à sa traduction textuelle, la prise en compte d'autres paramètres afin d'améliorer la qualité d'alignement textuel aurait nécessité plus de temps.

Ces 5 informations relatives aux livres ont fait l'objet d'une recherche manuelle et ont été obtenus pour 1 568 livres de LibriSpeech. Par contre, une caractéristique importante des livres bilingues est la différence entre les collections de textes. En effet, le même ouvrage en anglais au sein du projet Gutenberg peut faire partie d'une collection d'ouvrages qui serait différente en termes de volumes et de tomes en français.

Ci-dessous sont présentés quelques exemples des chapitres du livre *Les Misérables* par Victor Hugo présent en LibriSpeech, avec les tomes associés en français :

Nom du Chapitre (LibriSpeech)	Id du Livre sur Gutenberg (en)	Id du Livre sur Gutenberg (fr)	Tome-Volume
Bk 01 Ch 01-04	135	17494	Volume III, Livre I, Chapitres 1-4
Bk 08 Ch 15-16	135	17493	Volume III, Livre VIII, Chapitres 15-16
Bk 05 Ch 09: The Man with the Bell	135	17518	Volume II, Livre V, Chapitre 9

Tableau 3: Exemple d'association des chapitres aux tomes différents pour l'anglais et le français sur le projet Gutenberg

Le premier constat que nous pouvons faire à partir de ce tableau 3 est que les noms des chapitres fournis par LibriSpeech donnent une indication faussée sur la provenance d'un chapitre. Par exemple, l'indication donné pour le livre 8 et le chapitre 5 n'est pas suffisante pour savoir dans quel livre de Victor Hugo en français nous pourrons trouver le chapitre correspondant. Or, afin d'obtenir un alignement au niveau des phrases, il faut partir de textes qui sont à priori équivalents. Comme décrit dans le tableau 3, les cinq volumes de *Les Misérables* en anglais font partie d'une même collection. Le livre téléchargé correspond donc à l'intégralité des cinq livres alors qu'en français chaque volume est distribué indépendamment. Ce phénomène n'est pas spécifique au livre de Victor Hugo mais valable pour la plupart des livres composés de tomes et de volumes différents.

Les granularités des livres n'étant pas les mêmes, il est impératif de trouver les parties correspondantes des différents livres avant de réaliser l'étape d'alignement. Par conséquent, dans cette première partie, en vue de toutes ces différentes caractéristiques, nous avons choisi d'indiquer le plus d'informations concernant les différentes sources avant de nous pencher sur l'étape de traitement des données. Les métadonnées sont donc complétées dans un fichier *csv* pour l'intégralité des 5831 chapitres indiquant pour chacun, les cinq informations supplémentaires que nous avons précisées comme étant pertinentes.

Nous nous sommes servi des méthodes du Web sémantique afin d'obtenir automatiquement une partie de ces informations. Les métadonnées que nous avons ajoutées automatiquement feront l'objet du chapitre 4.1.2.

Parmi les 5831 chapitres de LibriSpeech, les métadonnées complétées correspondent à 1818 chapitres, soit 316 livres. Ces chapitres représentent 315 heures de parole lue en termes de quantité. Par contre, ces 315 heures de parole ne sont qu'une indication théorique de la durée maximale que nous pourrons obtenir, car tout au long du traitement, voire même avant cette étape, une perte de données est inévitable. Cette perte est due à diverses raisons telles que des problèmes d'encodage, des fichiers abîmés, ou encore des éditions différentes, *etc.*

La recherche de ces métadonnées supplémentaires est effectuée majoritairement sur le site Web : www.worldcat.org. Il s'agit d'une base de données des collections de bibliothèques à la fois locale et électronique. L'avantage de cette ressource est l'accessibilité aux œuvres bilingues à partir d'un livre. De plus, quand le format électronique d'un livre est disponible dans un projet de bibliothèque électronique comme Gutenberg, celui-ci est accessible par leur interface. Concernant les œuvres francophones, un deuxième site (www.noslivres.net) cité également dans le site de *beq* propose une base de données contenant uniquement les œuvres électroniques francophones.

Afin de faciliter les traitements, les œuvres dont nous avons ajouté un lien sont transférées dans un tableau annexe dans la base de données. Un extrait de ce tableau 'librispeech' contenant 1818 chapitres se trouve en annexe 2.

4.1.2 Utilisation du Web sémantique pour l'aide à la constitution du corpus

Comme décrit dans la partie 3.1.1, les titres des livres ainsi que les liens sont trouvés et ajoutés dans la base de données manuellement. En effet, les informations disponibles sur Wikipédia sont utilisées pour la création d'ontologies. Ce type de *wiki* sémantique considère les pages comme étant des concepts et les liens entre les pages comme des propriétés. Ces relations permettent l'extraction d'informations en formulant des requêtes SQL.

A cet effet, le projet DBPEDIA[10], disponible également en français, contient des données structurées permettant l'utilisation de Wikipédia comme une base de connaissances. En vue de récupérer les informations sur le même œuvre dans différentes langues, il est possible d'exploiter la relation d'équivalence avec la requête SPARQL[11] suivante :

[10] http://wiki.dbpedia.org/

[11] SPARQL est un langage de requêtes des graphs RDF : Un graphe RDF est un ensemble de triplets (?sujet ?predicat ?objet) que l'on peut interroger.

```
1.  prefix sameAs:<http://www.w3.org/2002/07/owl#sameAs>
2.  SELECT ?language  WHERE {
3.        <http://dbpedia.org/resource/$resource> sameAs:?language }
```

Par exemple, pour l'œuvre « *The man who laughs* » de Victor Hugo, le tableau suivant est extrait en utilisant la requête SPARQL ci-dessus :

http://it.dbpedia.org/resource/L'uomo_che_ride	http://es.dbpedia.org/resource/El_hombre_que_ríe
http://fr.dbpedia.org/resource/L'Homme_qui_rit	http://ko.dbpedia.org/resource/웃는_남자

Tableau 4 : Tableau des articles Wikipédia obtenus par la requête SPARQL utilisant la relation sameAs

Cette requête est ensuite complétée utilisant la relation titre[12] afin d'obtenir les titres d'ouvrages en français. Il faudra également noter que cette requête simple est basée sur des pages Wikipédia, et ne prend donc pas en compte les cas où la désambiguïsation est nécessaire. Par exemple, pour la page *Republic*, la précision *Republic_(Plato)* permettrait de désigner le livre et non d'autres concepts associés à cette ressource sur Wikipédia.

Etant donné que l'étape de vérification manuelle est primordiale afin de trouver la plupart des œuvres non-présentes sur Wikipédia, cette étape préalable a donc permis de trouver environ un tiers des titres français de LibriSpeech.

4.1.3 Récupération automatique des œuvres francophones

Après la récupération automatique des titres d'ouvrage, la deuxième partie automatique effectuée lors de la récupération des métadonnées est l'utilisation du framework *Selenium*. La base de données contenant les liens des œuvres francophones (www.noslivres.net) est un site dynamique dont le contenu est généré automatiquement via Javascript. Ce contenu n'est pas accessible par des méthodes de *scraping* classique car les données ne proviennent pas du code source des pages et sont interprétées par les navigateurs. A cet effet, ce framework est utilisé pour accéder aux informations dynamiques crées avec Javascript.

[12] http://dbpedia.org/property/title

Le tableau suivant représente les liens associés à l'œuvre *Candide* de Voltaire obtenus par le framework Selenium:

ID	Titre Original	Auteur	Source	Lien
302	Candide	Voltaire	Gutenberg	http://gutenberg.org/ebooks/4650
303	Candide	Voltaire	ELG	http://www.ebooksgratuits.com/details.php?book=637
304	Candide	Voltaire	BEQ	http://beq.ebooksgratuits.com/
305	Candide	Voltaire	BNR	https://ebooks-bnr.com/voltaire-candide-ou-loptimisme/

Tableau 5 : Extrait du tableau 'nosLivres' contenant les liens alternatifs pour l'œuvre Candide de Voltaire

4.1.4 Téléchargement automatique et structure des répertoires

La dernière étape du recueil de données (qui correspond à la première étape du traitement de l'annexe 1) est le téléchargement automatique des liens récupérés. Lors de cette dernière, et à partir des métadonnées de LibriSpeech, les livres en français sont téléchargés, renommés et organisés.

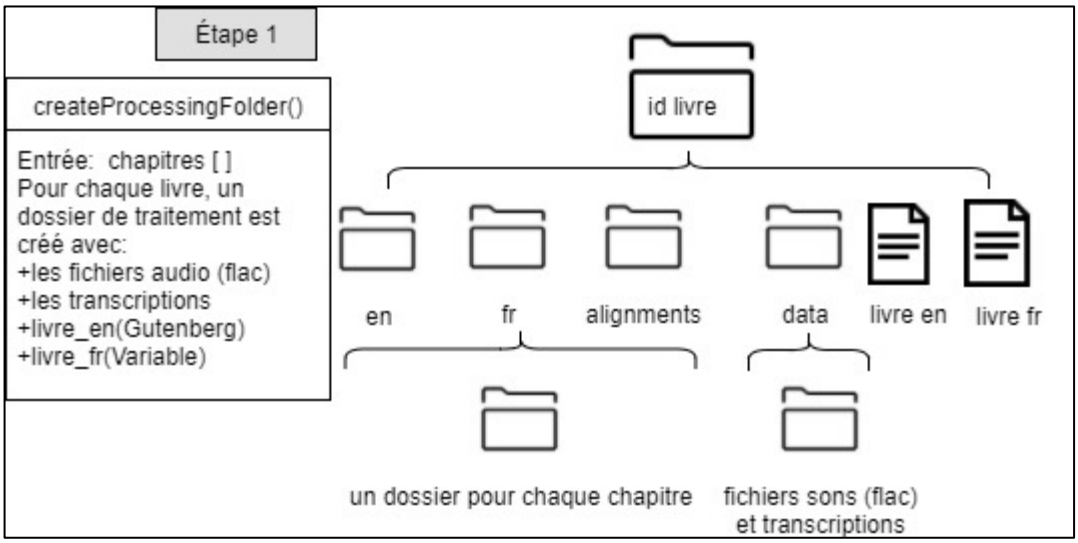


Figure 9 : création du dossier de traitement pour chaque livre – 1ère étape de la constitution du corpus

Concernant le téléchargement, les techniques de *scraping* Web sont utilisées. Tout d'abord, il nous faut différencier les notions de *crawling* et de *scraping*.

Tout d'abord l'indexation, aussi appelée *web crawling*, est une technique permettant la collection de ressources sur le Web. Celle-ci est réalisée en suivant les hyperliens d'une page principale récursivement. A l'inverse, le *scraping* est une technique qui vise à extraire le contenu à partir d'un site Web et est très répandue pour la collection de métadonnées.

Dans notre cas, nous avons besoin de récupérer des livres électroniques à partir de leurs liens, ce qui nous oblige donc à utiliser la méthode *scraping Web*.

Pour ce faire, la librairie BeautifulSoup[13] de Python est utilisée pour analyser (*parser*) les contenus de pages Web de différentes sources telles qu'Archive, Wikisource, Gallica, etc. Ensuite les fichiers sont téléchargés automatiquement utilisant des requêtes *urllib* de Python.

La convention utilisée tout au long du traitement correspond aux identifiants utilisés par LibriSpeech. Ces identifiants proviennent du projet Gutenberg et représentent le même projet (livre) dans LibriSpeech et Gutenberg. Notons également que pour certains livres, plusieurs éditions et enregistrements d'un même livre font partie du corpus avec des identifiants différents.

Comme décrit précédemment, différents liens sont récupérés pour un même livre. Lorsque plusieurs liens sont à télécharger, ils sont téléchargés sous le même identifiant séparés par un '_'. Le choix parmi différentes ressources ainsi que la vérification des livres téléchargés font partie du travail manuel effectué.

La dernière étape du recueil de données est la création d'un dossier de traitement pour chaque livre faisant partie du corpus.

[13] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Exemple d'un dossier de traitement pour le livre 11 (*Alice au Pays des Merveilles*) schématisé dans la figure ci-dessous :

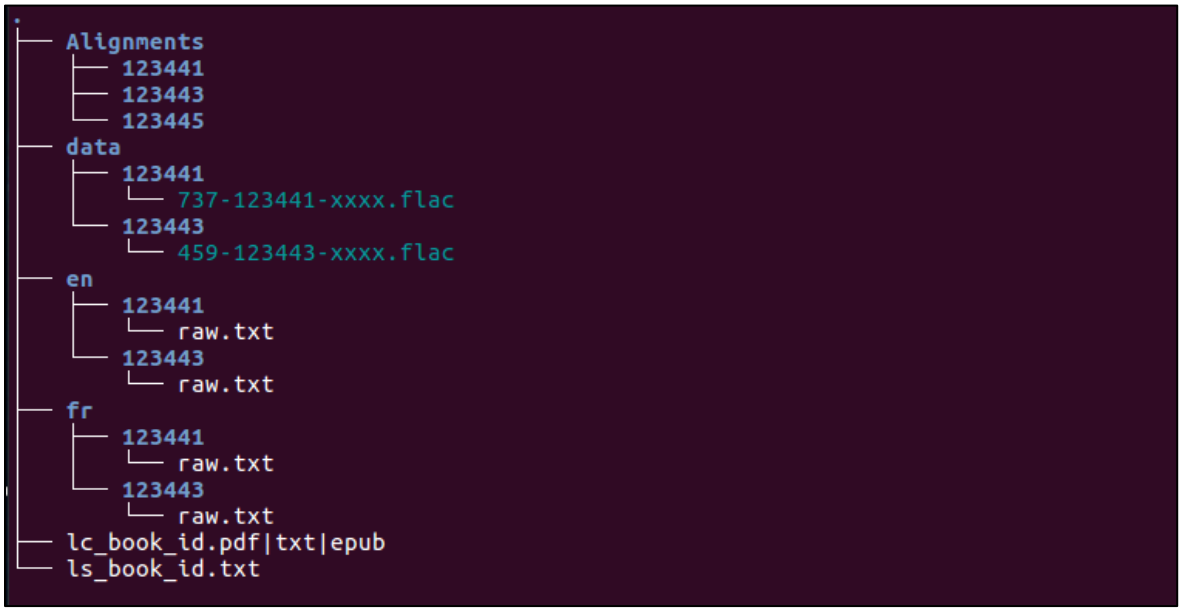


Figure 10 : L'organisation du dossier de traitement du livre id 11

Comme l'indique l'exemple ci-dessus, les noms des dossiers de traitements sont choisis en fonction de l'identifiant des livres. Le fichier de la langue source (ls_11.txt) et de la langue cible (lc_11.pdf) sont copiés dans ce répertoire. Tous les autres dossiers comportent un dossier pour chaque chapitre du livre. Par exemple, le dossier *data* contient pour chaque chapitre du livre 11 les fichiers sons de LibriSpeech ainsi que sa transcription. Les dossiers 'en' et 'fr' sont créés de la même manière afin de contenir les prétraitements effectués sur les données. Par ailleurs, le dossier 'alignments' est créé pour contenir les fichiers utilisés dans l'étape d'alignement.

4.2 Préparation des données pour l'alignement – Prétraitement des données

Une fois les documents téléchargés et les dossiers de traitements créés, nous nous intéresserons à l'étape consacrée à la préparation de données pour l'alignement. Dans cette partie, nous décrirons d'abord l'étape d'extraction des livres en chapitres, puis nous détaillerons les traitements linguistiques effectués en justifiant la motivation derrière chacun des traitements effectués.

4.2.1 Découpage des livres en chapitres

La deuxième étape du traitement effectuée est celle du découpage des livres en chapitres. Cette décomposition est réalisée pour chaque livre téléchargé pour les deux langues.

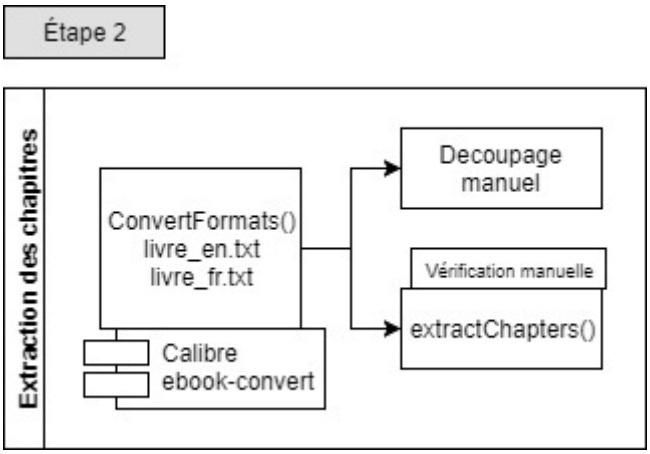


Figure 11: Conversion de formats et extraction des chapitres – 2ème étape de la constitution du corpus

Comme décrit dans le chapitre 3, LibriSpeech comporte des particularités, notamment au niveau de la segmentation de données. Les données de LibriSpeech sont décomposées par chapitres et sont dotées des fichiers de transcriptions.

Même si nous supposons que le chapitre en anglais est aligné parfaitement avec celui en français, l'association des transcriptions avec le chapitre en anglais n'est pas évidente. A cet effet, la délimitation du contenu permettrait d'une part la vérification de la correspondance et la qualité des livres téléchargés, et d'autre part l'obtention d'une granularité plus fine.

Par ailleurs, pour certains livres différents formats de différentes sources sont téléchargés. Le choix parmi différentes sources relève également d'un travail manuel. Ce travail manuel est notamment important pour les livres comportant des erreurs d'OCR, car il est parfois possible de trouver une version corrigée dans d'autres sources. La conversion des formats vers le fichier texte est réalisé en utilisant la fonctionnalité *ebook-convert* de la boite à outils et visionneuse des livres électroniques *Calibre*[14].

En vue d'extraire les chapitres des livres, un algorithme basé sur les expressions régulières est développé afin d'extraire tous les chapitres d'un livre lorsqu'il y a un motif récurrent. En effet, les chapitres commencent souvent de la même manière. Par exemple, les chapitres du livre *Alice au Pays de Merveilles* commencent toujours par le mot CHAPTER

[14] https://calibre-ebook.com/

suivi de la numération romaine indiquant le numéro du chapitre, et toujours suivi d'un point. Celui-ci pourrait-être exploité afin d'extraire automatiquement tout texte entre ce motif.

Les chapitres extraits sont vérifiés manuellement et mis dans les dossiers de traitement. Une expression régulière spécifique à chaque livre est créée manuellement pour lancer cette commande. Pour les livres dont le motif n'est pas récurrent, les chapitres sont extraits manuellement. Ce processus est présenté dans l'annexe 3.

Notons également que pour certains livres, la granularité des chapitres n'est pas équivalente dans les deux langues. Par exemple, pour le livre *Contes de Grimm*, le nombre de contes ainsi que leur contenu est très différent. Certains livres comme celui-ci ne sont pas traités et ont été exclus dans cette étape.

4.2.2 Traitements linguistiques effectués

Cette étape du traitement correspond à l'étape 3 dans le flux de données et de programmation : les données textuelles sont préparées pour l'étape d'alignement pour chaque chapitre et langue.

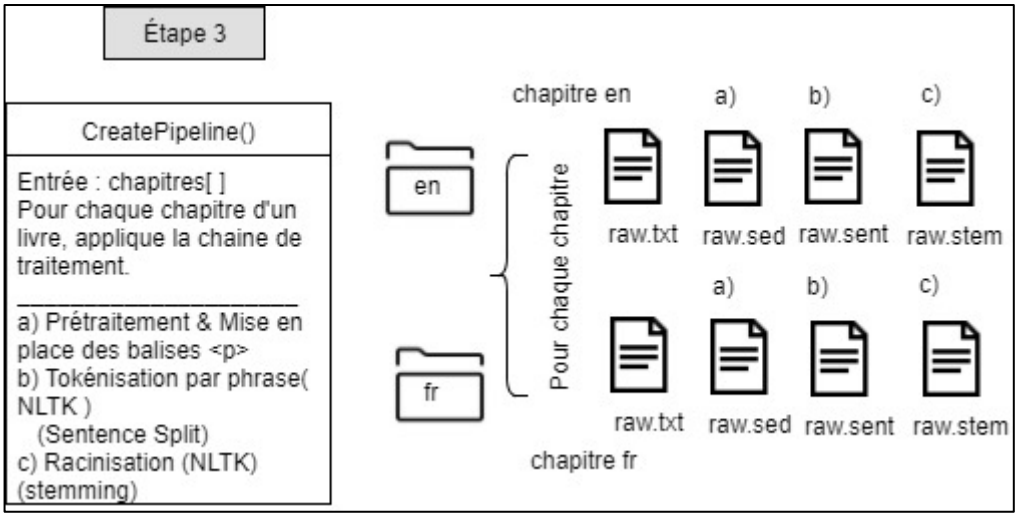


Figure 12: Application des pré-traitements pour chaque livre – 3ème étape de la constitution du corpus

Les traitements linguistiques effectués sur les données sont principalement liés aux particularités de l'outil d'alignement *hunAlign* (Varga et al., 2005). Tout d'abord un nettoyage par expressions régulières est effectué afin de normaliser les ponctuations et

d'effacer la numérotation de pages des PDF. Ensuite les fins de lignes séparées par un '-' sont concaténées aux suites de mots de la ligne suivante de la même manière.

La mise en place des frontières de paragraphes permet d'attribuer des scores différemment par rapport à un texte non-délimité pour le calcul de correspondance de *hunalign* (Varga et al., 2005). Toutefois, la mise en place des frontières nécessite que les textes soient bien formés. Un fichier extrait à partir d'un PDF peut comporter des déformations au niveau de la mise en page, ce qui pourrait fausser le calcul. Cependant, les scores de correspondances étaient globalement plus élevés lorsque les balises paragraphes étaient utilisées.

La suite des traitements linguistiques est effectuée en utilisant la boite à outils NLTK (Natural Language Toolkit). Facile d'utilisation, cet outil de TAL permet de réaliser des traitements de texte avec Python. Les traitements appliqués avec NLTK sont :

- Découpage par phrases (*sentence split*)
- Racinisation (*stemming*)

Ces prétraitements correspondent à ceux décrits dans l'article de (Varga et al.,2005). En effet, les expériences montrent que la racinisation augmente les performances de *hunalign* par rapport à d'autres méthodes utilisées sur des langues moyennement dotées.

4.2.2.1 Découpage par phrases

Le découpage par phrases répond à un double objectif : d'une part, il permet d'obtenir un nombre de lignes à peu près équivalent lorsque le document traité comporte des défauts de segmentation dus aux passages du format PDF au format texte, et d'autre part, il facilite le traitement par l'outil d'alignement. En effet, *hunalign* ne traite pas les fichiers comportant un nombre de lignes très différent.

Le découpage par phrase sur NLTK nécessite l'apprentissage à partir d'un corpus. Ce corpus d'apprentissage est simplement un corpus monolingue représentatif du français et de l'anglais. Nous nous sommes servi de tokeniseurs pré-entrainés de NLTK pour l'anglais et le français afin d'appliquer le découpage par phrase.

Le tableau suivant démontre la normalisation de la segmentation pour un paragraphe avec la technique tokenisation en phrases :

<table>
<tr><th>Paragraphe en anglais</th><th>Paragraphe en anglais avec sentence split</th></tr>
<tr><td rowspan="3">Then came the horses. Having four feet, these managed rather better than\n
the foot-soldiers: but even THEY stumbled now and then; and it seemed\n
to be a regular rule that, whenever a horse stumbled the rider fell off\n
instantly.\n</td><td><p> Then came the horses. \n
Having four feet, these managed rather better than the foot-soldiers: but even THEY stumbled now and then; and it seemed to be a regular rule that, whenever a horse stumbled the rider fell off instantly. \n</td></tr>
<tr><th>Paragraphe en français avec sentence split</th></tr>
<tr><td><p> Puis vinrent les chevaux. \n
Grâce à leurs quatre pattes, ils s'en <p> tiraient un peu mieux que les fantassins ; mais, malgré tout, eux <p> aussi trébuchaient de temps en temps ; et, chaque fois qu'un <p> cheval trébuchait, le cavalier ne manquait jamais de dégringoler. \n</td></tr>
</table>

Tableau 6: Tokenisation en phrases sur un paragraphe du livre *de l'autre côté du miroir*

4.2.2.2 Racinisation (Stemming)

La racinisation, parfois appelée *stemming*, consiste en un traitement heuristique permettant l'obtention des racines des mots. Contrairement au lemme qui forme une unité sémantique, le stem est obtenu à partir de la suppression des affixes dérivationnels, et ne constitue donc pas forcément une unité significative de la langue. Pour les livres comportant des erreurs d'OCR, la racinisation permet alors de récupérer des racines, même pour les unités non significatives. Ainsi, pour les livres en ancien français, la racinisation permettent une couverture lexicale plus large qui facilite le processus d'alignement des textes à priori difficiles.

Il existe différents algorithmes de racinisation. Celui employé dans notre cas est le *Snowball Stemmer* qui repose sur l'algorithme de racinisation de Porter (Porter, 1980).

L'application de l'algorithme de Snowball Stemmer sur une phrase comportant des erreurs d'OCR est présentée dans l'exemple ci-dessous :

a) Original : *Et voici, elle deviendra un arbre, bjaillissant en vous jusqu'à la vie éternelle.*

b) Racinisé : *et voic , elle deviendr un arbre , bjaill en vous jusqu'à la vi éternel.*

Dans l'exemple ci-dessus, nous pouvons constater que contrairement à la lemmatisation, la conjugaison du verbe est coupée comme s'il s'agissait d'un suffixe, celle-ci n'est pas transformée en sa forme canonique. En même temps, un mot inexistant comme '*bjallissant'* est considéré comme un mot du français et le suffixe est coupé.

Ce processus de racinisation correspond au dernier traitement avant l'utilisation de *hunalign*. Il faut noter que la racinisation du dictionnaire est nécessaire pour que ce processus de racinisation ait un effet sur l'alignement (Varga et al., 2005).

4.3 Alignement textuel

L'alignement textuel est la quatrième étape de la constitution de notre corpus. Dans cette étape les textes pré-traités sont soumis au système d'alignement afin d'obtenir un alignement parallèle pour chaque chapitre.

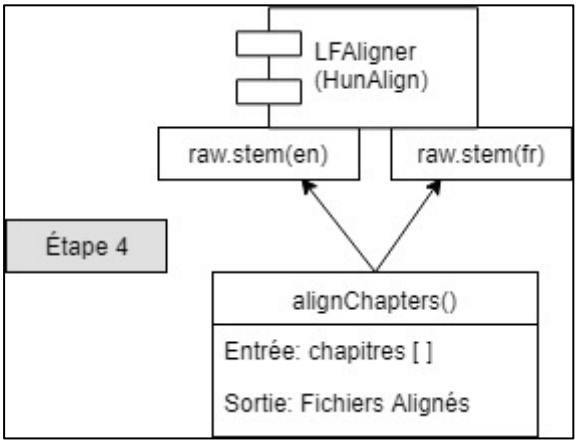


Figure 13 : alignement textuel des chapitres – 4ème étape de la constitution du corpus

Dans cette partie, nous verrons d'abord les particularités de *hunalign* ainsi que de celui du logiciel encapsuleur *LFAligner*. Puis, nous aborderons les post-traitements effectués sur les sorties de *hunalign* afin d'inverser le processus de racinisation tout en gardant l'ordre d'alignement des phrases.

4.3.1 Hunalign

Les aligneurs de phrases parallèles se basent principalement sur trois méthodes (D. Varga et al., 2005) :

- Méthode basée sur la longueur (Brown et al 1991, Gale and Church 1991)
- Méthode basée sur un dictionnaire ou de la traduction (Chen 1993 ; Melamed 1996 ; Moore 2002)
- Méthode basée sur une similarité partielle (Simard and Plamondon 1998)

L'algorithme non supervisé de hunalign repose sur une méthode hybride : basée à la fois sur la longueur et sur un dictionnaire. En effet, le calcul des scores de confiance de hunalign est un calcul issu de cette méthode mixte.

Par ailleurs, lorsqu'un dictionnaire bilingue est absent, un dictionnaire de *bootstrap* est créé automatiquement. Pour ce faire, en premier lieu, un alignement basé sur la longueur est réalisé, créant en même temps un dictionnaire automatique. Le texte est ensuite réaligné par rapport à ce dictionnaire. L'une des raisons pour lesquelles nous avons choisi hunalign comme aligneur repose sur la facilité d'utilisation ainsi que sur sa robustesse. Le logiciel encapsuleur *LFAligner*[15] développé par Andras Farkas propose encore d'autres fonctionnalités facilitant la tâche d'alignement. Par exemple, pour certaines langues, il dispose de dictionnaires bilingues et propose également certains prétraitements comme conversion de format, découpage par ligne, *etc*.

Le dictionnaire bilingue utilisé par hunalign est un fichier texte composé de couples comme le montre l'exemple suivant :

Phrase en langue cible @ phrase en langue source

Contrairement à un dictionnaire simple, ce dictionnaire contient des unités lexicales bilingues. Les entrées du dictionnaire peuvent être des expressions polylexicales, voire des phrases entières.

Le dictionnaire généré pour l'anglais et le français par *LFAligner* dispose de 40 000 entrées. La première tâche effectuée lors de cette étape d'alignement était d'augmenter la couverture lexicale. Pour ce faire, nous avons d'abord téléchargé d'autres dictionnaires disponibles pour l'anglais et le français. Les dictionnaires dont nous nous sommes servi font

[15] https://sourceforge.net/projects/aligner/

partie du projet Polyglotte[16]. Ces dictionnaires, initialement au format *stardict*, sont extraits vers un fichier texte utilisant *stardict2txt* disponible dans la boite à outils de *stardict*. Contrairement au format de dictionnaire de *hunalign*, la représentation de la polysémie était réalisée par des virgules en non comme des entrées à part. C'est pourquoi différents scripts pour différent types de dictionnaires sont développés. Ce dictionnaire est ensuite vérifié pour les doublons et trié par ordre alphabétique. Ce processus a permis d'obtenir un dictionnaire comportant 128,000 lignes d'entrées.

Notons également qu'au lieu d'utiliser le découpage par phrase et la conversion de format de *LFAligner*, nous avons préféré avoir recours à une étape de prétraitement. Ce choix est lié d'un part à la granularité des textes, d'autre part à d'autres traitements effectués tels que la racinisation, la mise en place des balises paragraphes, *etc*.

En termes de complexité, étant donné que les textes sont préalablement délimités, le processus d'alignement se réalise très rapidement. Lorsqu'il s'agit des textes longs, hunalign applique une sorte de réalignement permettant une meilleure cohésion du texte entier.

Différentes sorties sont proposées par hunalign et *LFAligner* :

- Représentation d'alignements dans un fichier texte comportant des tabulations : pour chaque ligne, ce fichier comporte la phrase de la langue source et de la langue cible ainsi que le score de confiance associé
- Sortie en forme de *ladder* : un fichier comportant le mapping (m : n) des phrases source-cible
- Fichier *tmx* sous forme de *xml* comportant de façon hiérarchique les trois informations représentées par le fichier tabulé

Même si le fichier tmx est plus confortable à la lecture, la sortie du fichier texte tabulé est utilisée dans la suite du traitement pour des raisons pratiques.

Il est à noter que les alignements ne comportent pas d'office le même nombre de lignes que d'entrées. Cela est compréhensible dans la mesure où les deux textes peuvent à priori contenir des différences au niveau du contenu. Lorsqu'un segment n'a pas de

[16]Un projet ayant pour but la facilitation d'apprentissage des langues grâce à l'usage de données et de logiciels libres. Lien : https://polyglotte.tuxfamily.org/

correspondance, il est associé à un segment vide. Ce segment vide peut être aussi bien du côté de la langue source que de la langue cible.

4.3.2 Post traitement sur les données : racinisation inverse

L'application du post-traitement sur la sortie du fichier aligné fourni par *hunalign* est la cinquième étape de la constitution de notre corpus. Dans cette étape, les alignements sont d'abord extraits dans des fichiers textes parallèles et sont ensuite comparés aux fichiers de tokénisation par phrase.

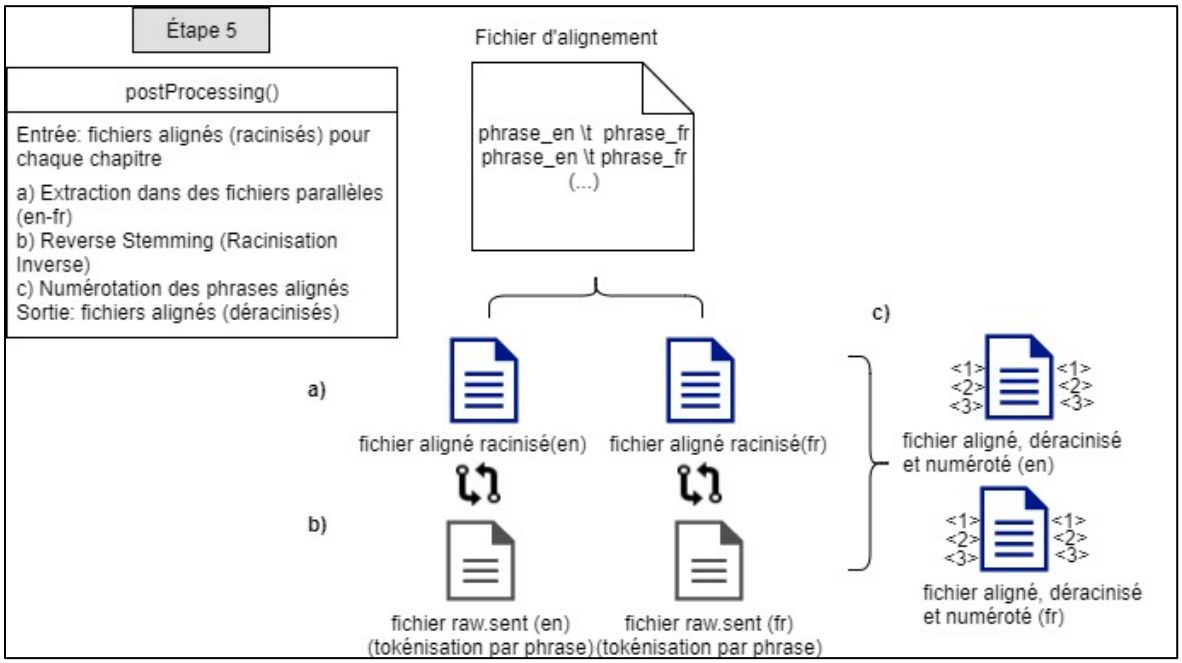


Figure 14: Post-traitements effectuées sur les alignements – 5ème étape de la constitution du corpus

Les fichiers d'entrées étant dans une forme racinisée, il faut inverser la racinisation afin d'obtenir les phrases dans leur forme originale. Or, il y a plusieurs étapes de prétraitement, et il y a donc plusieurs segmentations auxquelles on peut se référer. Nous avons choisi de nous référer aux fichiers comportant une phrase par ligne et non aux fichiers qui n'ont pas subi de traitement. Cependant, étant donné que l'alignement peut contenir des segments vides, la segmentation n'est pas équivalente à celle du fichier comportant une phrase par ligne. De ce fait, nous avons décidé d'utiliser une technique de comparaison de textes.

A cet effet, *Python* dispose de la librairie *difflib*, très populaire parmi les outils de comparaison des chaines de caractères et de textes. Prenons pour exemple la phrase suivante avec sa forme racinisée :

a) J'ai déjà consacré deux heures entières à mon habillage !

b) j'ai déjà consacr deux heur entier à mon habillag !

Le calcul de *difflib* retourne des deltas de différence au niveau des caractères. Ils sont représentés sous forme de liste permettant aux utilisateurs de les exploiter afin de gérer des changements entre deux séquences de textes. Pour chaque caractère un delta indique s'il est supprimé (-), s'il est ajouté (+) ou s'il n'a pas changé ( ).

L'exemple de la sortie de calcul de *difflib* est la suivante :

```
1. ['- J', '+ j', "  '", '  a', '  i', '   ', '  d', '  é', '  j', '  à', '   ', '  c', '
   o', '  n', '  s', '  a', '  c', '  r', '- é', '   ', '  d', '  e', '  u', '  x', '   ',
   '  h', '  e', '  u', '  r', '- e', '- s', '   ', '  e', '  n', '  t', '  i', '- è', '+ e
   ', '  r', '- e', '- s', '   ', '  à', '   ', '  m', '  o', '  n', '   ', '  h', '  a', '
     b', '  i', '  l', '  l', '  a', '  g', '- e', '   ', '  !']
```

Un autre moyen de faire la même chose aurait pu être d'introduire des balises phrase avec un identifiant pour chacune avant l'étape d'alignement. Par contre, nous avons constaté qu'il y avait des changements, notamment au niveau de la segmentation, entre les phrases de départ et la sortie des alignements.

Par ailleurs, la librairie de Google pour le calcul de différence des séquences, *Diff-patch-match*[17] développée par Neil Fraser, permet de comparer deux textes de façon plus robuste. Le même exemple confronté à l'algorithme de *diff* de Fraser donne le résultat suivant :

```
-J'ai déjà consacré deux heures entières à mon habillage !
```

Cette librairie est une implémentation de l'algorithme de différence de Myers (Myers, 1986) avec une couche de prétraitement pour l'optimisation et un post-nettoyage pour l'amélioration de la qualité des sorties. Comme démontré dans l'exemple ci-dessus, cette sortie regroupe les changements. Celui-ci est pratique dans notre cas, car nous ne nous intéressons pas qu'aux changements mais au contenu initial de ces phrases tout en gardant la segmentation des alignements. Il est à noter que les balises paragraphes qui étaient introduites en vue d'améliorer l'alignement sont exclues dans cette étape.

Un dernier traitement est effectué sur les sorties de *hunalign* : la numération des phrases. Le tableau suivant comporte les phrases alignées et numérotées :

[17] https://code.google.com/archive/p/google-diff-match-patch/

Alignement racinisé (en-fr)	
< p > chapter v. advic from a caterpillar < p > the caterpillar and alic look at each other for some time in silenc : at last the caterpillar took the hookah out of it mouth , and address her in a languid , sleepi voice.	< p > la chenill et alic se consider un instant en silence. enfin la chenill sort le houk de sa bouch , et lui adress la parol d ’ une voix endorm et traînante.
< p > who are you ? '	< p > « qui êtes-vous ?
said the caterpillar.	» dit la chenille.
< p > this was not an encourag open for a conversation.	ce n ’ était pas là une mani encourag d ’ entam la conversation
alic repli , rather shyli , 'i -- i hard know , sir , just at present -- at least i know who i was when i got up this morn , but i think i must have been chang sever time sinc then. '	alic répond , un peu confus : « je — je le sais à pein moi-mêm quant à présent. je sais bien ce que j ’ étais en me lev ce matin , mais je crois avoir chang plusieur fois depuis.
Application de racinisation inverse (en-fr)	
<0>CHAPTER V. Advice from a Caterpillar The Caterpillar and Alice looked at each other for some time in silence: at last the Caterpillar took the hookah out of its mouth, and addressed her in a languid, sleepy voice.<0>	<0>CHAPITRE V. CONSEILS D’UNE CHENILLE.La Chenille et Alice se considérèrent un instant en silence.Enfin la Chenille sortit le houka de sa bouche, et lui adressa la parole d’une voix endormie et traînante.<0>
<1>'Who are YOU?' <1>	<1>« Qui êtes-vous ?<1>
<2> said the Caterpillar.<2>	<2>» dit la Chenille.<2>
<3>This was not an encouraging opening for a conversation.<3>	<3>Ce n’était pas là une manière encourageante d’entamer la conversation.<3>
<4>Alice replied, rather shyly, 'I--I hardly know, sir, just at present--at least I know who I WAS when I got up this morning, but I think I must have been changed several times since then.'<4>	<4>Alice répondit, un peu confuse : « Je — je le sais à peine moi-même quant à présent.Je sais bien ce que j’étais en me levant ce matin, mais je crois avoir changé plusieurs fois depuis.<4>

Tableau 7 : Un extrait d’alignement du chapitre V d’Alice au Pays des Merveilles avec l’étape de racinisation inverse

4.3.3 Alignement des données de développement et de test

L’obtention de cet alignement n’est pas suffisante pour obtenir un alignement parallèle des phrases avec les transcriptions. La segmentation de la parole et donc des transcriptions est différente par rapport à cet alignement. Cependant, comme décrit dans la chapitre III (cf. 3.3), les données de test et développement de LibriSpeech sont segmentées du point de vue de la modélisation de la langue. Contrairement aux données d’entrainement de LibriSpeech, ces segments correspondent à des phrases. Pour cette partie des données, il

est possible d'aligner les transcriptions avec leur traduction sans avoir besoin de resegmenter la parole.

Par contre, ces transcriptions correspondent à des sous-ensembles des chapitres qui sont très courts. Par ailleurs, les transcriptions ne suivent pas l'ordre logique des phrases, c'est-à-dire que les transcriptions correspondent à des phrases quasi aléatoires dans le chapitre mais gardent l'ordre du texte. La figure suivante représente un extrait d'une transcription issue d'un chapitre de développement de LibriSpeech en contraste avec les phrases du chapitre :

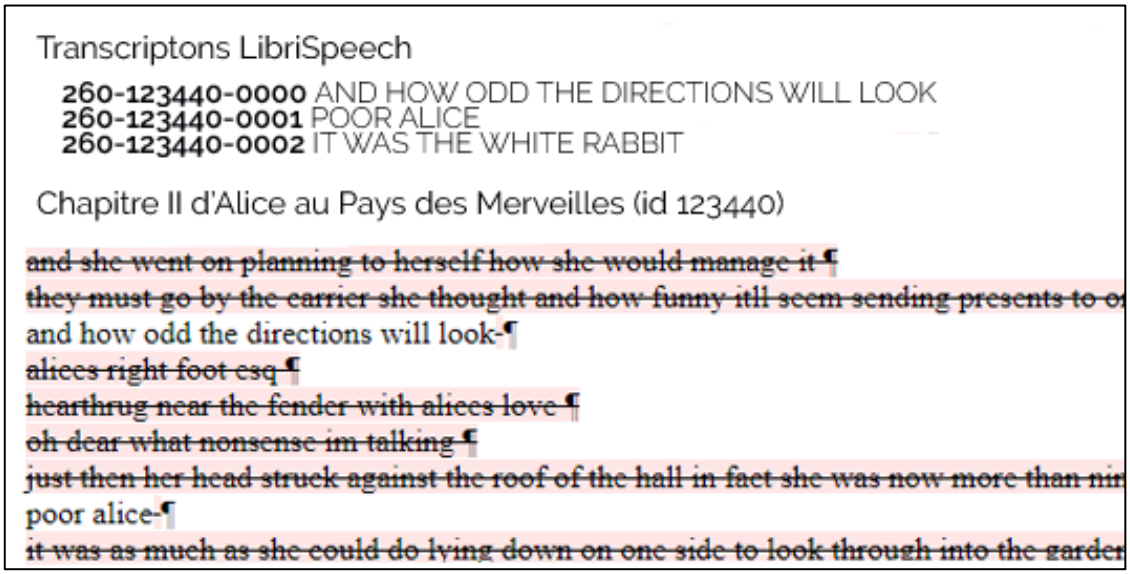


Figure 15: Exemple d'une transcription des données de développement de LibriSpeech en contraste avec le chapitre

Comme le montre l'exemple ci-dessus, il y a un écart considérable entre les transcriptions et le texte du chapitre. Les transcriptions correspondent souvent à des phrases mais ces phrases ne suivent pas l'ordre du texte. Pour cela nous avons réalisé une technique de recherche floue (*fuzzy-match)* entre la transcription et le chapitre afin de trouver les segments les plus semblables. De cette manière, il a été possible d'aligner la transcription, le chapitre et la traduction en parallèle. Evidemment, quand le découpage des transcriptions est différent de celui des phrases alignées, le segment de la parole est aligné à une phrase qui n'est pas correcte.

4.4 Alignements au niveau de la parole

Dans cette partie, les alignements obtenus, nous allons nous pencher sur l'alignement de la parole. Nous décrirons d'abord l'étape de la transcription forcée réalisée avec l'outil *mweralign* afin d'adapter la segmentation des chapitres à celui des transcriptions. Nous

expliquerons ensuite l'étape de l'alignement forcé utilisé entre le signal et les transcriptions. Nous finirons par l'explication de la resegmentation des chapitres par rapport au corpus parallèle aligné.

4.4.1 Transcription forcée

La transcription forcée est la sixième étape de la constitution du corpus en annexe 1. Le but ici est d'associer les phrases de transcription avec celles des alignements, nous obtenons en parallèle le flux de parole, la transcription, la phrase source et la phrase cible.

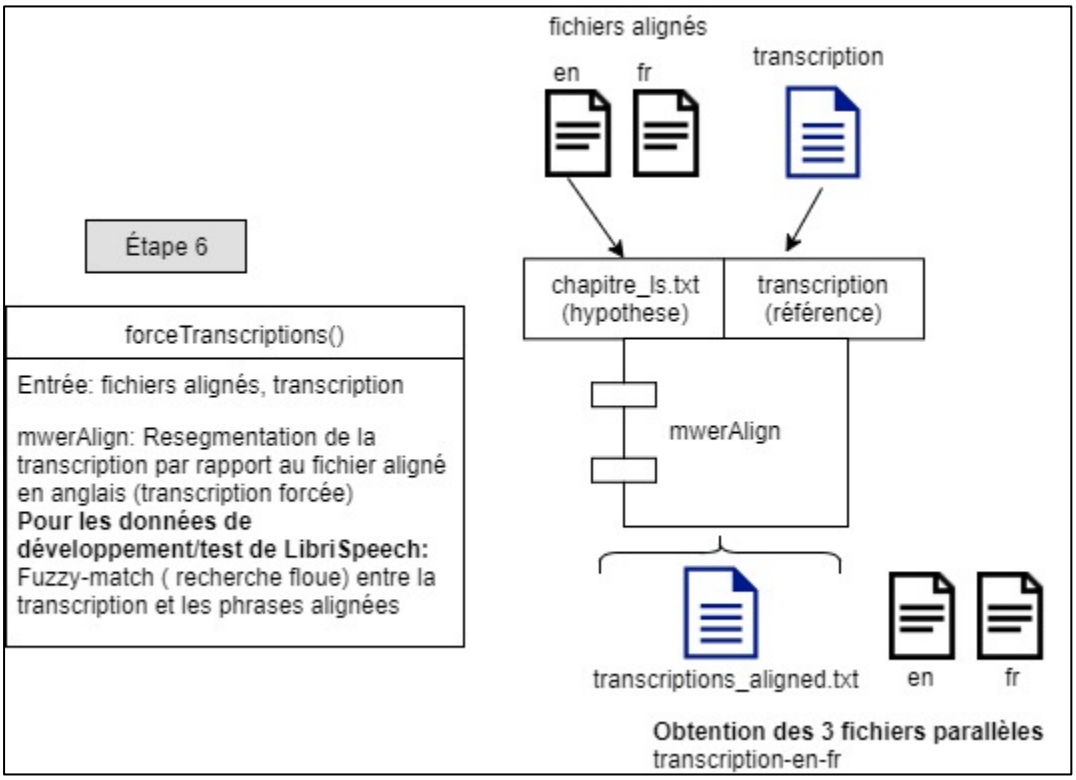


Figure 16 : Resegmentation des transcriptions par rapport aux fichiers alignés – 6ème étape de la constitution du corpus

Les transcriptions de LibriSpeech sont présentées dans des fichiers textes comportant pour chaque ligne la transcription d'un segment de la parole. Ces segments contiennent des bouts de parole qui peuvent correspondre ou non à une phrase entière.

Cette tâche pourrait ressembler à un alignement monolingue. Afin de segmenter les transcriptions par rapport aux phrases alignées, nous avons utilisé l'outil *mweralign* (Matusov et al. 2005). Cet outil permet d'obtenir une nouvelle segmentation à partir d'un fichier de référence et d'hypothèse. Dans notre cas, le fichier de référence sur lequel se base la segmentation est celui des transcriptions. Le fichier d'hypothèse est donc le texte du

chapitre comportant les numérations introduites dans l'étape de post-traitement. Dans ce processus, à l'aide de cet outil, la segmentation du texte de référence est 'forcée' au texte d'hypothèse, d'où la dénomination transcription forcée. *mweralign* aligne ces deux fichiers et adapte le fichier d'hypothèse selon la segmentation du fichier de référence.

Concernant l'utilisation de *mweralign*, une étape de pré-traitement est nécessaire pour que l'outil fonctionne correctement. Ce prétraitement consiste à la suppression des ponctuations et de la casse.

Le schéma suivant montre le processus de transcription forcée :

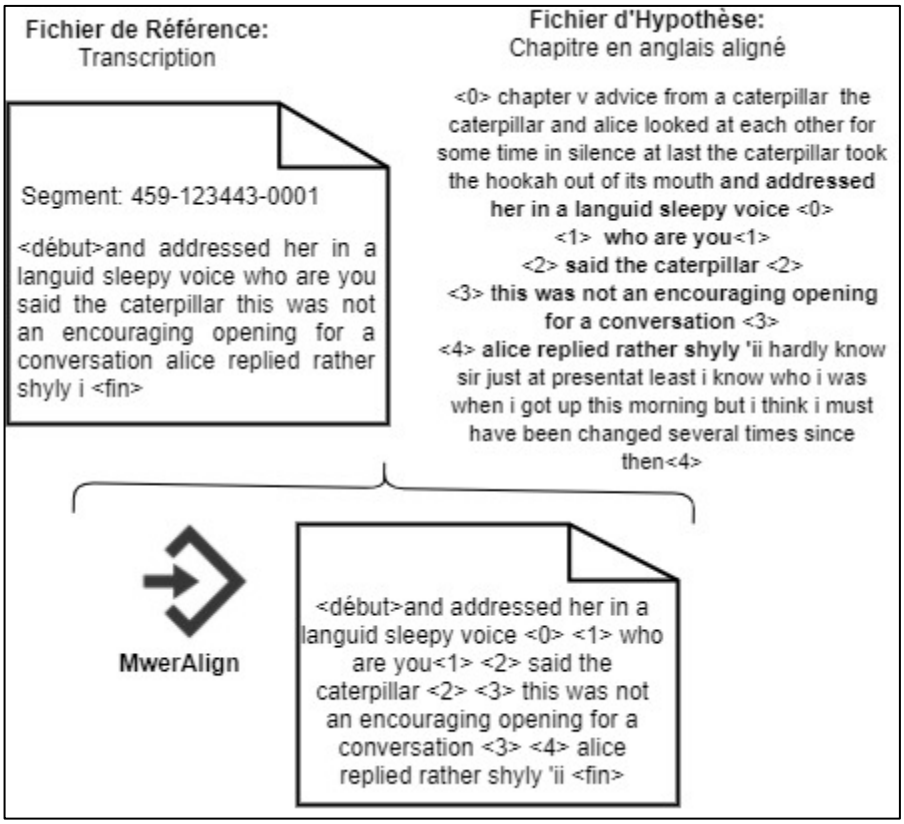


Figure 17: Processus de transcription forcée réalisé avec *mweralign*

Dans la figure ci-dessus, la technique d'alignement forcé est appliquée sur le texte du chapitre aligné en anglais. Comme les entrées de l'outil ont subi un prétraitement, la sortie ne comporte pas de ponctuations ni de caractères majuscules. De ce fait, le même post-traitement réalisé en vue d'inverser la racinisation est appliqué sur cette sortie.

Ce processus permet de forcer la segmentation des transcriptions dans le fichier aligné tout en gardant les numérotations qui seront utilisées plus tard pour l'association des traductions. Comme décrit dans l'exemple ci-dessus, les phrases complètes au sein des

transcriptions peuvent être facilement associées à leurs traductions. Or, comme les phrases numérotées 0 et 4 dans la figure 17, l'alignement de ces segments comporte aussi des phrases incomplètes.

Par ailleurs, la sortie de *mweralign* correspond à un fichier comportant le même nombre de lignes que la référence. Cependant, en termes de contenu les segments ne correspondent pas tout à fait aux transcriptions. En effet, le fichier obtenu est conforme au niveau de la segmentation mais pas au niveau du contenu. L'exclusion et le tri de ce dernier fait l'objet d'un traitement à part dans l'étape d'alignement forcé.

4.4.2 L'alignement forcé

La dernière étape de resegmentation appliquée sur les chapitres est l'étape d'alignement forcé. Cette étape correspond à l'étape 7 de l'annexe 1. En se basant sur les transcriptions et la parole, l'alignement forcé vise à obtenir les intervalles de temps des mots en indiquant pour chacun le début et la fin du mot. Cette information supplémentaire pourrait être utilisée afin de resegmenter la parole en fonction des phrases alignées.

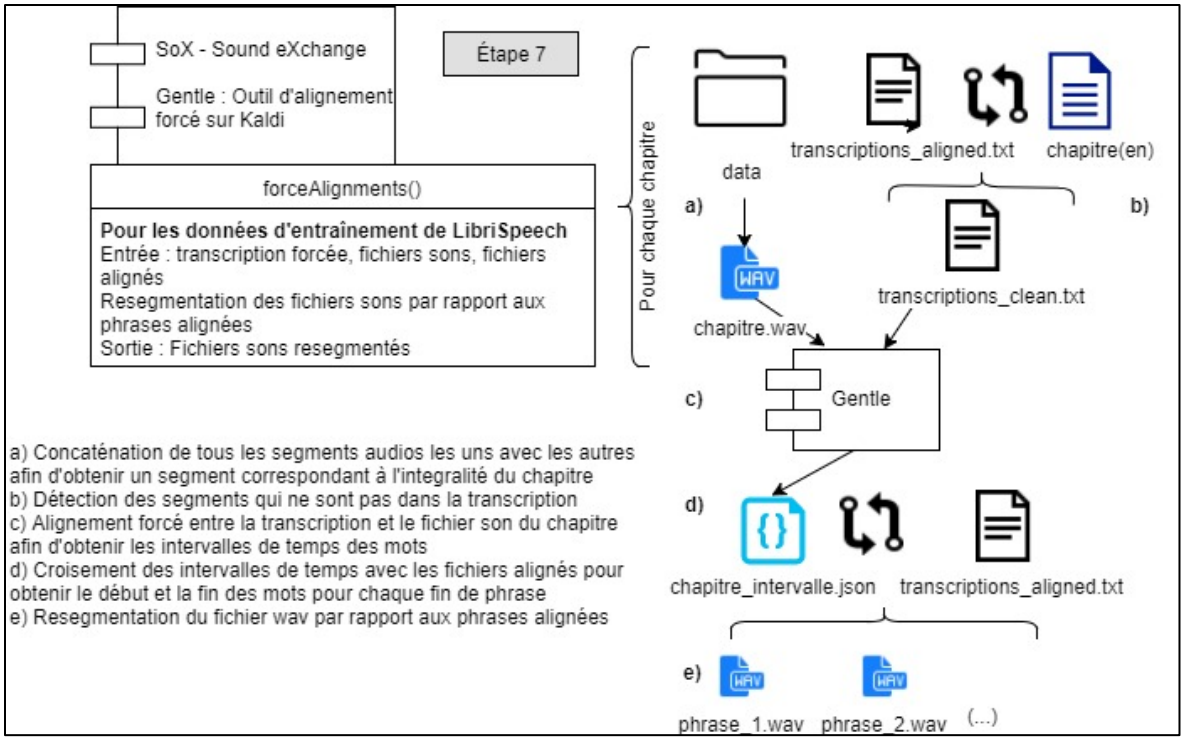


Figure 18 : Etape de l'alignement forcé utilisé pour resegmenter les fichiers sons par rapport aux phrases alignées – 7ème étape de la constitution du corpus

Pour ce faire, nous nous sommes servi d'un outil d'alignement forcé basé sur la boite à outils *Kaldi*. Développé par Robert M Ochshorn et Max Hawkins, l'outil *gentle*[18] prend un fichier audio et une transcription qui correspond à la parole, et fournit en sortie les intervalles de temps des mots dans un fichier json.

Les informations fournies pour l'alignement d'un mot sont représentées ainsi :

```
{
    "alignedWord": "five",
    "case": "success",
    "end": 1.43,
    "endOffset": 13,
    "phones": [{
        "duration": 0.12,
        "phone": "f_B"
 }, {
        "duration": 0.31,
        "phone": "ay_I"
    }, {
        "duration": 0.26,
        "phone": "v_E"
    }],
    "start": 0.74,
    "startOffset": 9,
    "word": "FIVE"
},
```

Après avoir aligné la parole, *gentle* fournit un fichier json comportant l'alignement de chaque mot comme dans l'exemple ci-dessus. Dans cet exemple extrait du chapitre III d'*Alice au Pays des Merveilles*, le début et la fin du mot 'five' est extrait sous forme de représentation d'objet de Javascript.

En vue d'appliquer le processus d'alignement forcé à grande échelle, la première étape consiste à appliquer un pré-traitement comme celui appliqué dans l'alignement forcé, à l'exception de la casse.

Pour chaque chapitre, le format des fichiers audio est *flac*. *Ils* sont décomposés en segments correspondant dans la transcription pour chaque ligne à un fichier son. Etant actuellement en développement, *gentle* ne prend que le format *.wav* comme entrée. Nous avons donc réalisé ensuite une conversion de format de *flac* vers *wav* en utilisant le logiciel *SoundConverter*.

Pour la suite du traitement des fichiers *wav*, nous nous sommes servi de *SoX-Sound Exchange*. Nous l'avons utilisé afin de concaténer tous les segments audios les uns avec les

[18] https://github.com/lowerquality/gentle

autres. Ce dernier a servi à obtenir un segment unique comportant l'intégralité des segments de parole d'un même chapitre.

Avant de fournir la transcription à l'outil d'alignement forcé, il faut préalablement établir une référence pour la resegmentation. Nous avons précédemment expliqué que le contenu des chapitres n'était pas équivalent à celui des transcriptions. Afin d'obtenir une référence qui permettra d'un part à resegmenter le flux de la parole, et qui d'autre part permettra d'exclure les alignements qui ne sont pas présents dans la parole, il faut comparer le chapitre et la transcription. Comme dans les étapes précédentes, cette comparaison est réalisée avec la librairie *diff-patch-match* de *Google*. La figure suivante montre quelques exemples de différences entre la transcription et le chapitre :

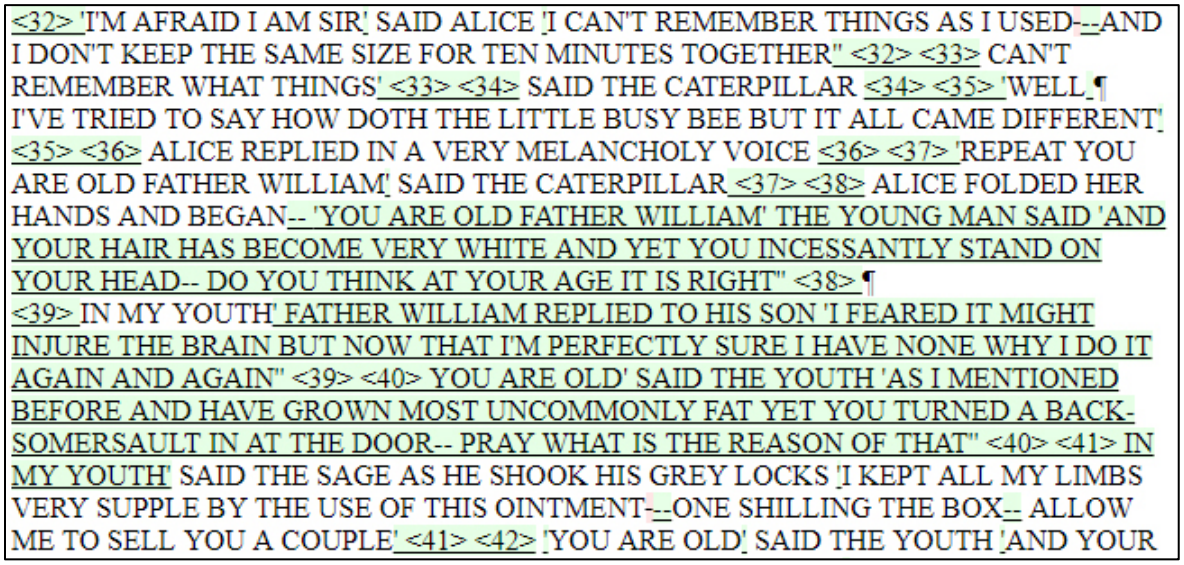


Figure 19: Calcul de diff entre la transcription et le chapitre extrait du livre d'*Alice au Pays des Merveilles*

La couleur verte dans le schéma ci-dessus montre qu'il y a une introduction d'un segment dans le chapitre qui n'est pas dans les transcriptions. Par exemple, la phrase numéro 40 de la figure *19*, représente une phrase complète qui est exclue des transcriptions. Ces phrases sont facilement identifiables grâce à des expressions régulières et ont été exclues. Cependant, les phrases comme la numéro 38 représentent un problème plus difficile à résoudre : ces segments correspondent aux transcriptions mais pas dans leur intégralité. L'alignement de la phrase 38 comporte toute la phrase et non la partie qui est lue.

Cette représentation intermédiaire entre le chapitre et la transcription est sauvegardée sans les numérotations des phrases et représente l'entrée (transcription) de l'outil

d'alignement forcé. A partir de cela, chaque mot correspondant à une fin de phrase est croisé avec le fichier *json* en vue d'extraire son intervalle de début et de fin dans le segment. Même si le taux de reconnaissance est relativement élevé, lorsqu'un mot correspondant à une fin de phrase n'est pas reconnu, le mot plus proche connu est recherché dans ses contextes gauche et droit. Les intervalles de temps des mots correspondant aux fins des phrases sont ensuite stockés dans une liste.

Le schéma suivant représente l'extraction des intervalles de temps pour chaque fin de phrase :

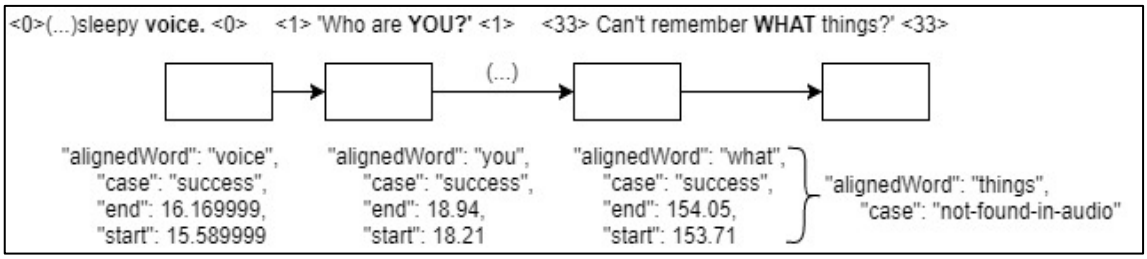


Figure 20: Liste contenant les intervalles du temps des mots correspondant aux fins de phrases

Cette liste est ensuite exploitée avec la fonctionnalité *trim* de Sox. Cette fonctionnalité permet de couper les fichiers sons en donnant le début et la durée de la locution qui sera coupée. Ce processus est appliqué itérativement pour chaque intervalle de la liste et est renommé suivant une identification unique qui permettra de l'associer à sa traduction.

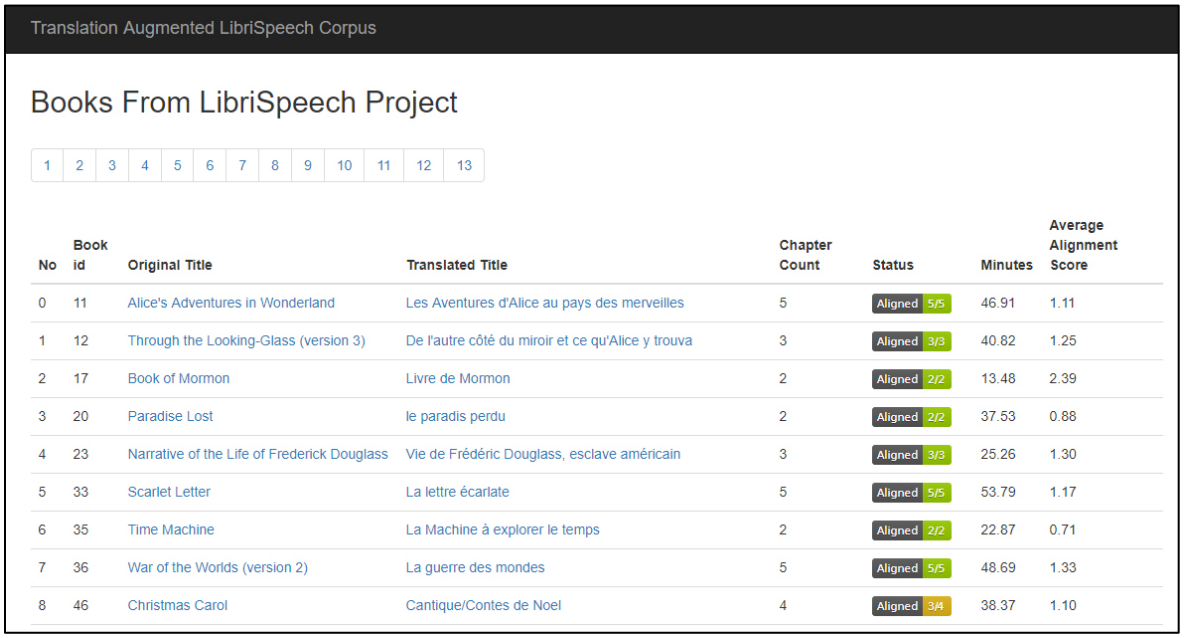
4.5 Visualisation des alignements – Interface Web

Il s'est avéré rapidement qu'en vue de vérifier les alignements aussi bien au niveau de la parole qu'à l'écrit, les fichiers séparés avec des tabulations ne sont pas très pratiques à lire. La visualisation est rendue plus lisible grâce aux fichiers *tmx* fournis par *LFAligner*. Par contre, la vérification des segments audio et des fichiers alignés tout au long du traitement a nécessité plusieurs étapes de vérification. C'est pour cette raison que nous avons développé une interface Web dynamique avec PHP. Le côté html du site est basé sur le *Twitter Bootstrap 3*[19].

Cette interface est liée avec la base de données et regroupe les livres faisant partie de notre corpus, indiquant pour chacun son état d'alignement. Les 315 livres du corpus sont

[19] http://getbootstrap.com/

représentés dans la page principale avec une pagination. La capture d'écran suivante représente la page principale :



No	Book id	Original Title	Translated Title	Chapter Count	Status	Minutes	Average Alignment Score
0	11	Alice's Adventures in Wonderland	Les Aventures d'Alice au pays des merveilles	5	Aligned 5/5	46.91	1.11
1	12	Through the Looking-Glass (version 3)	De l'autre côté du miroir et ce qu'Alice y trouva	3	Aligned 3/3	40.82	1.25
2	17	Book of Mormon	Livre de Mormon	2	Aligned 2/2	13.48	2.39
3	20	Paradise Lost	le paradis perdu	2	Aligned 2/2	37.53	0.88
4	23	Narrative of the Life of Frederick Douglass	Vie de Frédéric Douglass, esclave américain	3	Aligned 3/3	25.26	1.30
5	33	Scarlet Letter	La lettre écarlate	5	Aligned 5/5	53.79	1.17
6	35	Time Machine	La Machine à explorer le temps	2	Aligned 2/2	22.87	0.71
7	36	War of the Worlds (version 2)	La guerre des mondes	5	Aligned 5/5	48.69	1.33
8	46	Christmas Carol	Cantique/Contes de Noel	4	Aligned 3/4	38.37	1.10

Figure 21: Capture d'écran de la page principale de l'interface PHP

A partir de cette page principale, chaque livre comporte un page unique comportant un tableau de bord des chapitres disponibles sur LibriSpeech. Dans ce tableau, différentes métadonnées sur les chapitres et sur le livre sont récapitulées. La suite des pages des livres est divisée en deux parties pour contenir les données bilingues. Pour chaque livre, les étapes de pré-traitement et les alignements textuels sont consultables en parallèle.

Enfin, ces pages contiennent les segments de la parole alignés avec leurs traductions. Cette représentation regroupe les informations essentielles pour un alignement parallèle et multimodal. Elle comporte à la fois le score de confiance d'alignement et la durée du segment, la transcription et sa traduction et en même temps permet d'écouter le segment.

Les alignements finaux sont représentés comme il suit :

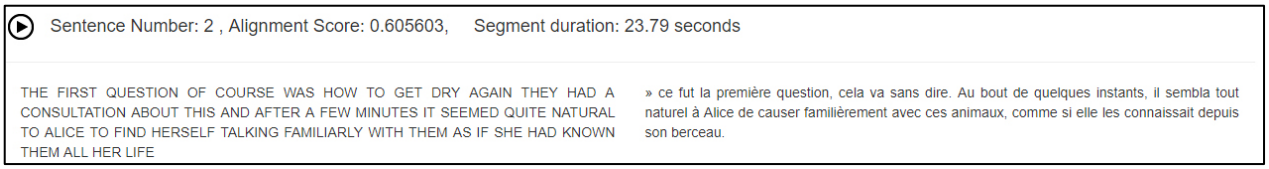


Figure 22 : Visualisation d'un alignement final comportant différentes informations

Il est à noter que les segments issus de développement et de tests de LibriSpeech qui ne sont pas resegmentés subissent une conversion de format (flac ⟶ wav) dans cette étape du traitement.

Des captures d'écran d'alignements textuels et de parole du livre *Alice au Pays des Merveilles* se trouvent dans les annexes 4 et 5.

Chapitre 5. Evaluation

Nous avons décrit la méthodologie que nous avons employée en vue d'augmenter les données de la parole de LibriSpeech avec les traductions récoltées sur internet. Dans cette partie nous nous pencherons sur l'évaluation de la qualité d'alignements. L'évaluation a un double objectif : d'une part celle-ci permettra d'obtenir un sous-ensemble de données qui sont vérifiés en termes de qualité et servir comme données de test aux systèmes de traduction. D'autre part, comme les alignements contiennent des décalages tant au niveau de l'écrit que de l'oral, le calcul des scores de correspondances permettra de trier le corpus en fonction de la qualité d'alignements.

Dans cette optique, nous présenterons en première partie l'évaluation manuelle réalisée sur 200 alignements du corpus. Nous décrirons ensuite les scores de correspondance des phrases alignées qui seront utilisés pour trier le corpus.

5.1 Evaluation manuelle sur 200 phrases

Comme présenté dans le chapitre IV, différents traitements sont effectués sur les données de LibriSpeech qui donne ensuite lieu à la création d'un corpus de grande taille de traduction de la parole. En termes de quantité, les œuvres constituées correspondent environ à 350 heures de parole. Toutefois, en réalité les segments de la parole ne correspondent pas à cette totalité mais correspondent à 236 heures de parole alignées au niveau de l'utterance avec leur traduction.

Nous avons choisi d'abord 4 chapitres représentatifs du corpus. Ces chapitres ont été choisis en fonction du score de confiance calculé par *hunalign*. Ces scores sont calculés pour chaque alignement et sont représentés dans le fichier de sortie d'alignements. Nous avons intégré ces scores sur l'interface Web, pour chaque chapitre et pour chaque livre en prenant la moyenne des scores d'alignement. Quand tous les livres sont pris en compte, le score global d'alignements est de *1,12*. Les 4 chapitres à évaluer ont été choisis en fonction de ce score.

Ces chapitres sont les suivants :

ID du Livre	ID du Chapitre	Nom du Livre	Chapitre	Score de Confiance
98	51758	Le conte de deux cités	Livre III Ch III	1.06
82	127083	Ivanhoé	Chapitre XXIII	1.43
11	123443	Alice au Pays des Merveilles	Chapitre V	1.19
76	163375	Aventures de Huck Finn	Chapitre VIII	0.75

Tableau 8: Métadonnées des chapitrés évalués manuellement

La somme des scores de confiance constitue une référence pour estimer la qualité d'alignement. Dans cette optique, nous avons choisi deux chapitres avec un score moyen pour l'un un peu plus haut que la moyenne, et pour l'autre un peu plus bas que la moyenne. Ensuite, un chapitre qui a un score relativement bas et un autre chapitre qui est relativement haut ont été choisis.

L'évaluation a porté sur l'évaluation des 50 premiers alignements des 4 chapitres de 4 livres différents, ce qui représente en termes de quantité de parole environ 27 minutes. Les alignements comportent d'un côté la parole et de l'autre côté l'alignement textuel. L'évaluation a donc été réalisée sur 2 axes : l'évaluation de la parole et l'évaluation de l'alignement entre transcriptions et traductions.

5.1.1 Les critères d'évaluation

Lorsqu'il s'agit de l'évaluation manuelle, il est primordial d'établir des critères clairs et explicites préalablement. Il est à noter que l'évaluation porte sur la qualité d'alignement et non de la traduction. Les traductions sont issues des livres électroniques où la traduction est réalisée par des traducteurs. Cependant, l'équivalence en termes de traduction comporte plusieurs dimensions et la granularité n'est pas toujours au niveau des phrases mais des fois au niveau des paragraphes, voire des chapitres. Cette évaluation porte sur l'équivalence de l'utterance avec son alignement et est réalisée du point de vue de traitement des données. Nous avons établi des critères d'évaluation pour la correspondance de la parole et pour la qualité d'alignement.

Comme décrit dans le chapitre III (cf. 3.3), les données de LibriSpeech sont traitées en fonction d'obtenir des données équitables et propres pour les systèmes de RAP. Concernant la parole, nous avons établi 3 critères pour évaluer les utterances présentées dans une échelle allant de 1 à 3. Cette échelle est la suivante :

- **1.** Le segment et la transcription ne correspondent pas du tout
- **2.** La transcription correspond au segment à l'exception d'un ou deux mots qui ne sont pas prononcés du tout ou qui sont coupés avant la fin de la prononciation
- **3.** La transcription et le segment correspondent parfaitement

Lorsqu'il y a un mot qui n'est pas bien aligné, le mot le plus proche est cherché dans le contexte à gauche et à droite. Cela implique un décalage dans les deux segments. Néanmoins, ce phénomène n'est pas très présent car le taux de réussite de l'alignement forcé est relativement haut.

Concernant l'évaluation des alignements au niveau de l'écrit, les phénomènes rencontrés sont plus subtils que par la parole. Nous avons établi 5 critères allant de 1 à 5 afin d'évaluer les alignements avec un exemple pour chacun. Ces critères et les exemples sont présentés dans le tableau suivant :

Score	Explication	Exemple
1	L'alignement se fait avec un segment vide « NA » ou la transcription et la traduction sont complètement différents	COMMIT TO ME I SHALL LET PASS NO ADVANTAGE
		Je sais, par exemple, que maintenant il souffre de.la faim (…)
2	L'alignement et la traduction correspondent légèrement : les phrases reflètent la même idée mais la traduction est soit littérale soit contextuelle	THAN IN SET TERMS AND IN COURTLY LANGUAGE
		Mais il paraît que tu préfères être courtisée avec l'arc et la hache, plutôt qu'avec des phrases polies et avec la langue de la courtoisie.
3	La transcription et la traduction correspondent en partie : la traduction contient la transcription avec plus ou moins d'informations	SO AT LAST BEGAN THE EVENING PAPER AT LA FORCE
		» C'est ainsi qu'enfin débuta le journal du soir à la Force, le jour où la pauvre Lucie avait vu danser la carmagnole.
4	La transcription et la traduction correspondent globalement à l'exception des quelques mots proportionnellement à la longueur du segment	THE NIGHT WAS DARK AND A COLD WIND BLEW
		La nuit était sombre ; le vent âpre et froid chassait devant lui avec rage les nuages rapides.

5	La transcription et la traduction correspondent parfaitement	WHAT IS A CAUCUS RACE
		» « Qu'est-ce qu'une course cocasse?

Tableau 9: Critères d'évaluation manuelle des 200 phrases

A partir des critères définis dans le tableau 9, et les critères d'évaluation de la parole, les 50 premières phrases des 4 chapitres ont été évaluées par 3 annotateurs sur les échelles présentées ci-dessus. Cette évaluation a été réalisée sur un formulaire créé sur l'interface. Ces scores sont ensuite mis dans la base de données.

5.1.2 Résultats

Chapitre	Score de confiance en moyenne	Score moyen de l'évaluation de la parole (max 3)	Score moyen de l'évaluation des alignements (max 5)
Ivanhoé Chapitre XXIII	1.34	2.82	4.64
Alice au Pays des Merveilles Chapitre V	1.14	2.98	4.28
Le conte de deux cités Livre III,Chapitre III	0.96	2.86	3.86
Aventures de Huck Finn Chapitre VIII	0.66	2.9	2.58
Moyenne	1.25	**2.89**	**3.84**

Tableau 10: Résultats d'évaluation manuelle

Le premier constat que nous pourrons faire sur les résultats est la corrélation entre les scores de confiance de *hunalign* et les évaluations des alignements. Cela signifie que dans le cas où le score de confiance moyen pour un chapitre est élevé, alors la moyenne des évaluations manuelles se trouve être aussi élevée.

Deuxième constat à faire sur les scores : l'alignement de la parole au niveau de l'utterance est très élevé. En vue donc d'évaluer la qualité des données, l'alignement textuel a un impact plus fort que l'alignement de la parole. Cela n'est pas surprenant vu que nous disposons des transcriptions exactes de la parole avec une technique d'alignement forcé. Nous obtenons un score d'évaluation en moyenne de 3.84/5 pour les alignements au niveau de l'écrit et de 2.89/3 pour la parole.

Lorsqu'une évaluation manuelle est effectuée par l'arbitrage de plusieurs annotateurs, il est important de vérifier l'accord entre les annotateurs. Un moyen de calculer

ce score est le calcul du Kappa de Cohen. Kappa est un coefficient utilisé pour mesurer l'accord entre deux variables qualitatives attribuées par deux annotateurs. Dans Kappa, tout désaccord est considéré comme un désaccord total. A l'inverse, lorsqu'il s'agit d'une échelle linéaire, par exemple de 1 à 5, une variation de calcul, Kappa modéré (Cohen, 1968), est plus adapté. Ce calcul prend en compte différents niveaux d'accords et pondère le calcul, ce qui fait que tout désaccord n'est pas un désaccord total.

Le calcul de Kappa modéré de nos évaluations sur l'alignement textuel est le suivant : 0.76: ce calcul correspond à la moyenne de kappa modéré des 3 annotateurs entre eux. Le score 0.76 montre qu'il y a un accord fort entre annotateurs et que les évaluations sont pertinentes avec les critères d'évaluation.

5.2 Calcul des scores d'alignement

Le calcul du score de confiance de *hunalign* est calculé en fonction de deux composants majeurs : l'un basé sur la longueur et l'autre basé sur les tokens (Varga et al, 2005). Un premier score est calculé en fonction de la correspondance des tokens et est normalisé par la phrase comportant le plus de tokens. La deuxième méthode consiste au calcul d'un score basé sur le ratio du nombre de caractères du plus élevé au plus bas (Varga et al., 2005). Dans cet article, le score de confiance est ensuite calculé en attribuant un poids relatif à ces deux méthodes en vue d'optimiser le résultat sur le corpus d'entraînement d'hongrois-anglais.

Le score d'alignement de *hunalign* peut être trompeur lorsqu'il y a une correspondance entre les mots sur deux textes distincts. Dans notre cas, les chapitres fournis à *hunalign* sont préalablement délimités et sont plus ou moins vérifiés (cf. 4.2). En revanche lorsqu'il s'agit de deux documents qui couvrent le même contenu dans les deux langues, les scores sont pertinents pour obtenir la qualité d'alignement.

Le score de corrélation entre les évaluations manuelles et les scores de confiance de *hunalign* est de *0,41*. Cela nous montre qu'il y a une corrélation linéaire faible – modérée. Dans cette situation, il ne faut pas oublier que les évaluations manuelles sont faites entre les *utterances* et leurs traductions, ce qui peut justifier une baisse dans la corrélation puisque les scores de confiance de *hunalign* eux portent sur l'alignement textuel.

En effet, comme décrit dans la section 4.4, les segments de la parole peuvent comporter des différences en termes de contenu par rapport au texte du chapitre. Notre corpus s'inscrit dans le cadre d'un corpus de grande taille créé de manière automatique et

comporte des erreurs d'alignements ainsi que d'erreurs de segmentation. Il faut donc faire une évaluation entre les transcriptions et la traduction pour avoir une estimation de qualité plus adéquate.

Par ailleurs au-delà des scores de *hunalign*, nous avons ajouté des scores issus de l'approche par similarité interlangue utilisée dans le domaine de détection de plagiat (Ferrero et al., 2016). Le premier score calculé est un score mixte entre le calcul de n-grammes au niveau des caractères (CL-CNG) et une méthode de similarité à base de thésaurus multilingue (CL-CTS). Comme le score de confiance de *hunalign*, ces scores sont mis dans la base de données dans un tableau séparé.

CL-CNG est une méthode utilisée dans l'extraction d'informations qui consiste à comparer deux textes dans une représentation vecteur de n-grammes. Cette méthode est efficace lorsqu'il s'agit de comparer deux textes qui partagent la même origine (Ferrero et al., 2016). Dans notre cas, les vecteurs utilisés sont de longueur de 3 caractères.

La deuxième méthode est une mesure de similarité sémantique entre deux vecteurs de concepts : les documents sont représentés en tant que vecteurs et sont comparés en utilisant des ontologies. (Ferrero et al., 2016). Pour cette méthode, l'ontologie utilisée est *Dbnary*. Un deuxième calcul basé sur le modèle de taille de Pouliquen (Pouliquen et al., 2003) et la méthode CL-CTS est ajouté dans la base de données. Ce modèle est différent du modèle de taille utilisé par *hunalign*.

Les scores de corrélation sont les suivantes :

Scores / Evaluations	CL-CNG & CL-CTS	Longueur & CL-CTS	hunalign
Evaluation manuelle	**0.41**	-0.11	**0.41**
CL-CNG & CL-CTS			0.34
Longueur & CL-CTS			-0.26

Tableau 11: Tableau des scores de corrélation

Les scores de corrélation bas pourront être expliqués du même fait que la corrélation entre l'évaluation manuelle et les scores de confiance, il s'agit d'une corrélation faible – modérée entre la méthode mixte CL-CNG & CL-CTS. Ainsi, nous pouvons observer que le score de corrélation de cette méthode n'est pas supérieur à celui de *hunalign*, il lui est égal.

De plus, nous avons réalisé un traitement qui consiste à exclure les segments qui sont incohérents en termes de longueur en se basant sur la longueur des transcriptions et leurs traductions : ceux qui comportent 3 fois plus de tokens que leurs transcriptions ont été exclus.

Par ailleurs, un script a été développé pour manipuler et extraire les données pour l'entraînement et le développement en se basant sur ces critères d'évaluation. La requête SQL suivante permet donc de repérer les segments non-exclus triés par rapport aux scores d'évaluation choisis par l'utilisateur (hunalign, CL-CNG & CL-CTS, *etc)* :

```
1. SELECT * FROM alignments
2. JOIN(alignments_excluded JOIN alignments_scores USING(audio_filename))
3. USING(audio_filename) WHERE excluded != "True"
4. ORDER BY alignment_score DESC
```

Le script python basé sur cette requête SQL prend en compte les paramètres tels que le score d'alignement à utiliser pour trier, la durée maximale et minimale des segments, *etc.*
Les paramètres du script *TA-LibriSpeech.py* sont les suivants :

```
usage: TA-LibriSpeech.py [-h] [--size SIZE] [--listTrain LISTTRAIN]
                         [--useEvaluated USEEVALUATED]
                         [--sort {None,hunAlign,CNG}] [-v]
                         [--maxSegDuration MAXSEGDURATION]
                         [--minSegDuration MINSEGDURATION] [--extract]
                         action output
```

Les segments de la parole ainsi que la transcription et traduction sont extraits dans le dossier entré en paramètre. Lorsqu'il s'agit d'extraire les données de test, les données évaluées manuellement sont privilégiées. Si la durée d'extraction dépasse les données évaluées, les meilleurs alignements triés par le score choisi par l'utilisateur sont pris en compte pour compléter cette dernière.

Des exemples d'alignements de sortie triés par le score mixte CL-CNG & CL-CTS et par le score de *hunalign* obtenus sont présentés en annexe 6 et 7.

Conclusion et Perspectives

Conclusion

Après avoir passé en revue les avancées dans les domaines de TA, TAP et la RAP, nous avons présenté les ressources existantes qui servent à entraîner ces systèmes. Il existe des ressources en grande quantité pour la TA comme *Europarl, OpenSubtitles, etc.* Pour la RAP, des données de parole existent mais elles sont plus restreintes. Toutefois, pour la TAP, les corpus existants sont très limités. Dans ce cadre-là, en se basant sur le corpus de RAP de LibriSpeech, nous avons abouti à un travail d'augmentation de ce projet en ajoutant les traductions. Cela nous a donné un corpus aligné au niveau de l'utterance avec les traductions correspondantes.

Nous avons tout d'abord expliqué comment nous avons exploité les métadonnées de LibriSpeech afin de trouver les œuvres françaises disponibles parmi celles disponibles sur le projet. Nous avons détaillé cela en expliquant notre méthodologie pour le recueil de données.

Ensuite, nous nous sommes penchés sur le sujet d'alignement des œuvres bilingues avec *hunalign* ainsi que le pré-traitement et les post-traitements que nous avons effectués sur les données. Nous avons présenté les techniques de transcription et d'alignement forcé afin de resegmenter le flux de la parole en fonctions des phrases alignées.

Au final, nous avons obtenu un corpus de traduction de la parole de 236 heures. Les scores d'associations entre les transcriptions et les traductions sont ajoutés pour trier le corpus en fonction de la qualité d'alignements. Nous avons complété ceci avec un script permettant d'extraire les données pour l'entraînement, le développement et le test.

Perspectives

Dans ce travail, nous avons constitué un corpus de grande taille comportant des données de parole en anglais avec les traductions associées en français. Il serait intéressant de mener une expérience afin d'évaluer l'approche de traduction directe présentée dans le chapitre 2 (cf. 2.2). Ce corpus pourrait tout aussi bien être utilisé afin entraîner ce système de TAP.

Indépendamment de la langue, la même méthodologie pourrait être utilisée avec quelques changements afin de l'adapter à d'autres langues. Celle-ci nécessiterait la récolte

des livres électroniques dans une autre langue, ainsi qu'un dictionnaire bilingue (idéalement) pour l'alignement. Les prétraitements réalisés avec NLTK sont réalisables pour d'autres langues ou peuvent être adaptés facilement.

D'autre part, il pourrait être envisageable de modifier la langue des enregistrements afin de créer d'autres données dont les transcriptions sont fournies. Notons que s'il y a un besoin de resegmentation de la parole comme décrit dans le chapitre 4 (cf. 4.4.2), les modèles acoustiques devront être entraînés dans cette langue-ci afin d'obtenir un alignement forcé.

Par ailleurs, il est envisageable dans un futur proche d'augmenter la quantité des données, d'une part avec des livres électroniques qui ne font pas partie de nos données, et d'autre part via d'autres sources de corpus de parole. En effet, les projets comme Ted, OpenSubtitles, *etc*. comportent des transcriptions de la parole et pourront être ajoutés en vue d'augmenter la quantité des données.

Bibliographie

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*.

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, *3*, 1137–1155.

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

Besacier, L. (-). *Cours[1] - Introduction à la traduction automatique statistique*.

Bridle, J. S. (1990). Advances in Neural Information Processing Systems 2. In D. S. Touretzky (Éd.) (p. 211–217). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Brown, P. F., Cocke, J., Peitra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., … Roossin, P. S. (1990). A statistical approach to machine translation. *Journal Computational Linguistics*, *16*(2), 79-85.

Brown, P. F., Lai, J. C., & Mercer, R. L. (1991). Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics* (p. 169–176). Stroudsburg, PA, USA: Association for Computational Linguistics.

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information (p. 9-16). Présenté à Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Ohio, USA: Association for Computational Linguistics.

Cho, K. (2015, juin 14). Introduction to Neural Machine Translation with GPUs (Part 2). Consulté 17 janvier 2017, à l'adresse https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-2/

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv:1406.1078 [cs, stat]*.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-Based Models for Speech Recognition. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Éd.), *Advances in Neural Information Processing Systems 28* (p. 577–585). Curran Associates, Inc.

Cieri, C., Miller, D., & Walker, K. (2004). The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. In *LREC* (Vol. 4, p. 69–71).

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, *70*(4), 213.

Constant, M. (2009), *Cours[2]*. *Traduction automatique*. Consulté à l'adresse http://igm.univ-mlv.fr/~mconstan/enseignement/m1/tal/C6.pdf

Federmann, C., & D. Lewis, W. (2016). Microsoft speech language translation (mslt) corpus: The iwslt 2016 release for english, french and german. In *The IWSLT 2016 Evaluation Campaign*. Seattle, USA.

Ferrero, J., Agnes, F., Besacier, L., & Schwab, D. (2016). A Multilingual, Multi-Style and Multi-Granularity Dataset for Cross-Language Textual Similarity Detection. In *10th edition of the Language Resources and Evaluation Conference*.

Gale, W. A., Church, K. W., & others. (1991). Identifying Word Correspondences in Parallel Texts. In *HLT* (Vol. 91, p. 152–157).

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, *9*(8), 1735–1780

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1* (p. 48–54). Stroudsburg, PA, USA: Association for Computational Linguistics.

Kraif, O. (2016). *Cours[3] - Corpus électroniques écrits*.

Lecouteux, B., Linarès, G., Nocera, P., & Bonastre, J. (2006). Reconnaissance de la parole guidée par des transcriptions approchées. *Journées d'Etudes sur la Parole (JEP 2006), Dinard, France*, 53–56.

Matusov, E., Leusch, G., Bender, O., Ney, H., & Vi, L. F. I. (2005). Evaluating Machine Translation Output with Automatic Sentence Segmentation. In *in Proc. of IWSLT* (p. 138-144).

Melamed, I. D. (1996). Automatic Construction of Clean Broad-Coverage Translation Lexicons. In *arXiv:cmp-lg/9607037* (p. 125–134). Montreal, Canada.

M.F. Porter. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130-137.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Interspeech* (Vol. 2, p. 3).

Moore, R. (2002). Fast and accurate sentence alignment of bilingual corpora. *Machine Translation: From Research to Real Users*, 135–144.

Myers, E. W. (1986). AnO(ND) difference algorithm and its variations. *Algorithmica*, *1*(1-4), 251-266.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books (p. 5206-5210). IEEE.

Post, M., Kumar, G., Lopez, A., Karakos, D., Callison-Burch, C., & Khudanpur, S. (2013). International Workshop on Spoken Language Translation (IWSLT 2013). *International Workshop on Spoken Language Translation*, 7.

Pouliquen, B., Steinberger, R., & Ignat, C. (2003). Automatic Identification of Document Translations in Large Multilingual Document Collections. In *RANLP'03* (p. 401-408). Borovets, Bulgaria.

Specia, L. (2015). *Statistical Machine Translation*. Consulté à l'adresse http://lxmls.it.pt/2015/Talk_LuciaSpecia_LxMLS.pdf

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., & Trón, V. (2007). Parallel corpora for medium density languages. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Éd.), *Current Issues in Linguistic Theory* (Vol. 292, p. 247-258). Amsterdam: John Benjamins Publishing Company.

Sigles et abréviations utilisés

ASR: automatic speech recognition

CL-CNG: cross language character N-gram

CL-CTS: cross language conceptual thesaurus-based similarity

LPC: linear predictive coefficients

MFCC: mel-frequency cepstral coefficients

ML : modèle de langage

MT : machine translation

RAP : reconnaissance automatique de la parole

ST : speech translation

TAN : traduction automatique neuronale

TAP : traduction automatique de la parole

TAS : traduction automatique statistique

WER: word error rate

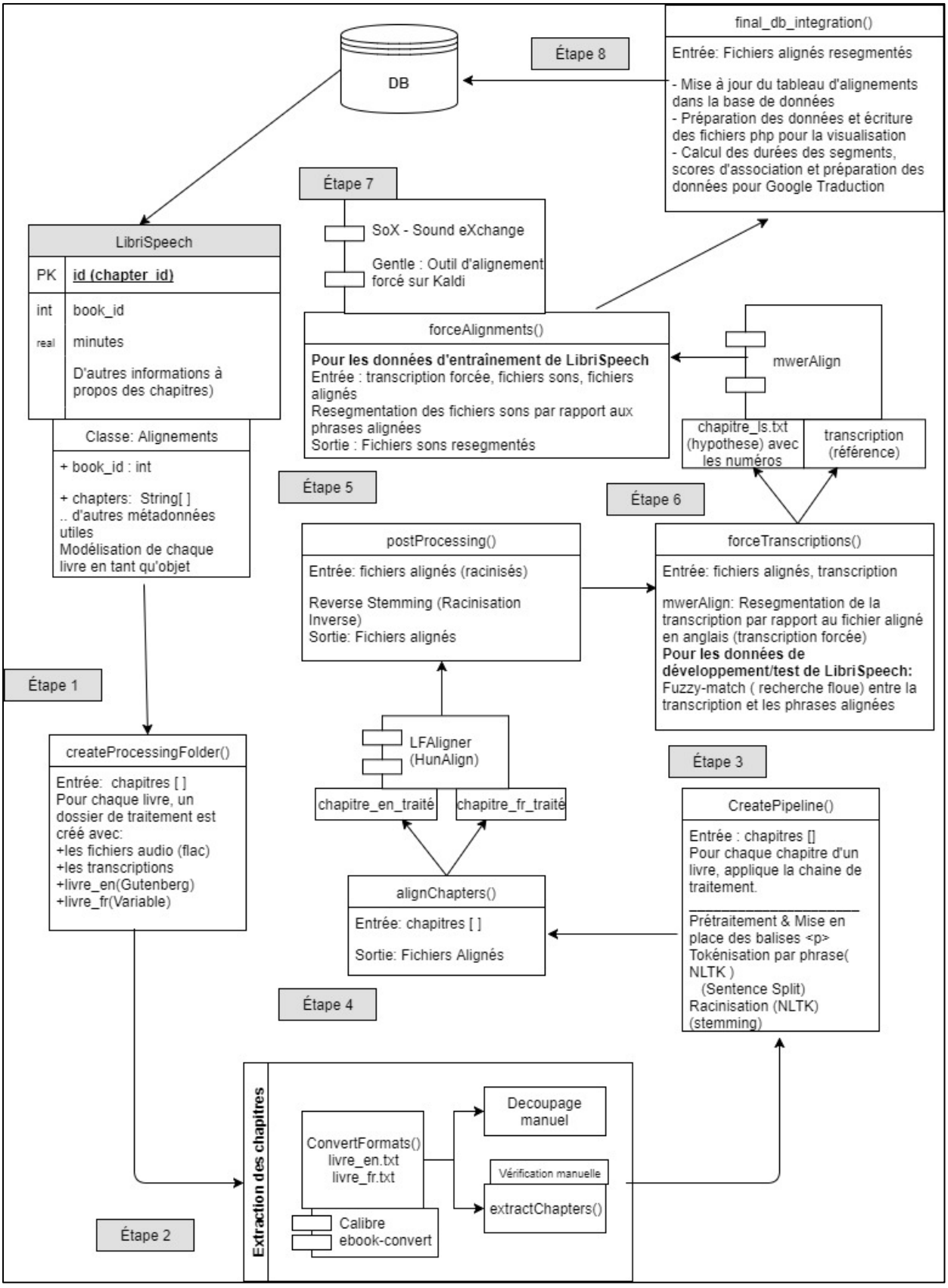
Table des illustrations

Liste des tableaux

Table des équations

Table des annexes

Annexe 1
Flux de données et programmation (Flowchart)



Annexe 2

Extrait du tableau des métadonnées complétés de LibriSpeech

librispeech [TA-LibriSpeechCorpus.db] | TA-LibriSpeechCorpus.db.sql

1-500 of 1818 | Tx: Auto | Tab-se...d (TSV) | DDL | Viev

<Filter criteria>

	book_id	id	minute	reader_id	corpus_name	chapter	original_title	translated_title	link	source	author	file
1	1023	1	19.77	110	train-other-500	In Chancery	Bleak House	Bleak House	https://fr.wik…	Wikisource	Dickens	/corpus_data/…
2	1023	2	10.3	110	train-other-500	In Fashion	Bleak House	Bleak House	https://fr.wik…	Wikisource	Dickens	/corpus_data/…
3	121	198	8.42	19	train-clean-100	Chapter 01	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
4	121	199	11.68	98	train-clean-360	Chapter 02	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
5	121	200	11.25	173	train-other-500	Chapter 03	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
6	121	201	7.57	44	train-other-500	Chapter 04	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
7	121	204	12.76	92	train-other-500	Chapter 07	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
8	121	205	12.82	20	train-other-500	Chapter 08	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
9	121	207	18.33	44	train-other-500	Chapter 10	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
10	121	208	12.95	14	train-clean-360	Chapter 11	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
11	121	209	8.2	198	train-clean-100	Chapter 12	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
12	121	212	12.09	14	train-clean-360	Chapter 15	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
13	121	214	6.19	37	train-other-500	Chapter 17	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
14	121	215	8.28	37	train-other-500	Chapter 18	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
15	121	216	6.3	403	train-clean-100	Chapter 19	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
16	121	218	12.78	89	train-clean-100	Chapter 21	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
17	121	219	12.3	89	train-clean-100	Chapter 22	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
18	121	222	15.59	40	train-clean-100	Chapter 25	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
19	121	224	7.56	246	train-clean-360	Chapter 27	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
20	121	225	15.58	37	train-other-500	Chapter 28	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
21	121	227	16.77	19	train-clean-100	Chapter 30	Northanger Abbey	L'Abbaye de Northanger …	https://beq.eb…	beq	Jane Austen	/corpus_data/…
22	84	348	30.21	728	train-other-500	00 - Letters	Frankenstein, or …	Frankenstein ou le Prom…	http://www.ac-…	direct_l…	Mary Shelley	/corpus_data/…
23	84	352	10.65	166	train-clean-360	04 - Chapter 5	Frankenstein, or …	Frankenstein ou le Prom…	http://www.ac-…	direct_l…	Mary Shelley	/corpus_data/…
24	84	354	13.84	66	train-other-500	06 - Chapter 7	Frankenstein, or …	Frankenstein ou le Prom…	http://www.ac-…	direct_l…	Mary Shelley	/corpus_data/…
25	84	355	16.38	66	train-other-500	07 - Chapter 8	Frankenstein, or …	Frankenstein ou le Prom…	http://www.ac-…	direct_l…	Mary Shelley	/corpus_data/…
26	84	358	9.82	730	train-clean-100	10 - Chapter 11	Frankenstein, or …	Frankenstein ou le Prom…	http://www.ac-…	direct_l…	Mary Shelley	/corpus_data/…
27	84	359	7.9	730	train-clean-100	11 - Chapter 12	Frankenstein, or …	Frankenstein ou le Prom…	http://www.ac-…	direct_l…	Mary Shelley	/corpus_data/…
28	84	360	7.48	730	train-clean-100	12 - Chapter 13	Frankenstein, or …	Frankenstein ou le Prom…	http://www.ac-…	direct_l…	Mary Shelley	/corpus_data/…
29	84	362	15.5	17	train-clean-360	14 - Chapter 16	Frankenstein, or …	Frankenstein ou le Prom…	http://www.ac-…	direct_l…	Mary Shelley	/corpus_data/…
30	84	363	9.54	17	train-clean-360	15 - Chapter 17	Frankenstein, or …	Frankenstein ou le Prom…	http://www.ac-…	direct_l…	Mary Shelley	/corpus_data/…

Annexe 3

Algorithme d'extraction automatique des chapitres

Afin d'extraire toute les lignes entre le motif donné pour le livre *Alice au Pays des Merveilles* dans des fichiers textes différents, la méthode d'extraction automatique est lancée par la commande suivante :

```
1. Alice.extractChapters("CHAPTER [IVXCLD]+\. .+ ?\n", "CHAPITRE [IVXCLD]\..+ ?\n")
```

Algorithm 1 Extract Chapters

```
procedure EXTRACTCHAPTERS(motif(en),motif(fr)) ▷ Motif réccurent(exp. rég.) pour en,fr
    for all language ∈ languages do ▷ Pour chaque langue
        chapitreslanguages ← regexp.findAll(motif(langue),book id)
        for all chapter ∈ chapters do ▷ Pour chaque chapitre détecté
            while ≠ EOF do ▷ Tant que la fin du fichier n'est pas atteinte
                for line ∈ book do ▷ Passer jusqu'au début du chapitre
                    if match(line, chapter) then ▷ Si le chapitre est trouvé
                        print(line) ▷ Ecrire le titre du chapitre
                        break ▷ Sortir de la boucle
                    end if
                end for
                for line ∈ book do ▷ Lire jusqu'à la fin du chapitre
                    if match(line, chapter + 1) then ▷ Si le titre du chapitre suiv. est trouvé
                        print(text) ▷ Ecrire le contenu du chapitre
                        break ▷ Sortir de la boucle
                    end if
                    text += line ▷ Concaténation du contenu du chapitre
                end for
                break ▷ Répéter pour chaque chapitre jusqu'à la fin du fichier
            end while
        end for
    end for
```

Annexe 4

Alignement textuel du chapitre V d'*Alice au Pays des Merveilles*

Translation Augmented LibriSpeech Corpus

CHAPTER V. Advice from a Caterpillar The Caterpillar and Alice looked at each other for some time in silence: at last the Caterpillar took the hookah out of its mouth, and addressed her in a languid, sleepy voice.	CHAPITRE V. CONSEILS D'UNE CHENILLE.La Chenille et Alice se considérèrent un instant en silence.Enfin la Chenille sortit le houka de sa bouche, et lui adressa la parole d'une voix endormie et traînante.
'Who are YOU?'	« Qui êtes-vous ?
said the Caterpillar.	» dit la Chenille.
This was not an encouraging opening for a conversation.	Ce n'était pas là une manière encourageante d'entamer la conversation.
Alice replied, rather shyly, 'I--I hardly know, sir, just at present--at least I know who I WAS when I got up this morning, but I think I must have been changed several times since then.'	Alice répondit, un peu confuse : « Je — je le sais à peine moi-même quant à présent.Je sais bien ce que j'étais en me levant ce matin, mais je crois avoir changé plusieurs fois depuis.
'What do you mean by that?'	» « Qu'entendez-vous par là ?
said the Caterpillar sternly.	» dit la Chenille d'un ton sévère.
'Explain yourself!'	« Expliquez-vous.
'I can't explain MYSELF, I'm afraid, sir' said Alice, 'because I'm not myself, you see.'	» « Je crains bien de ne pouvoir pas m'expliquer, » dit Alice, « car, voyez-vous, je ne suis plus moi-même.
'I don't see,' said the Caterpillar.	» « Je ne vois pas du tout, » répondit la Chenille.
'I'm afraid I can't put it more clearly,' Alice replied very politely, 'for I can't understand it myself to begin with; and being so many different sizes in a day is very confusing.'	« J'ai bien peur de ne pouvoir pas dire les choses plus clairement, » répliqua Alice fort poliment ; « car d'abord je n'y comprends rien moi-même.Grandir et rapetisser si souvent en un seul jour, cela embrouille un peu les idées.
'It isn't,' said the Caterpillar.	» « Pas du tout, » dit la Chenille.
'Well, perhaps you haven't found it so yet,' said Alice; 'but when you have to turn into a chrysalis--you will some day, you know--and then after that into a butterfly, I should think you'll feel it a little queer, won't you?"	« Peut-être ne vous en êtes-vous pas encore aperçue, » dit Alice.« Mais quand vous deviendrez chrysalide, car c'est ce qui vous arrivera, sachez-le bien, et ensuite papillon, je crois bien que vous vous sentirez un peu drôle, qu'en dites-vous ?
Not a bit,' said the Caterpillar.	» « Pas du tout, » dit la Chenille.
'Well, perhaps your feelings may be different,' said Alice; 'all I know is, it would feel very queer to ME.'	« Vos sensations sont peut-être différentes des miennes, » dit Alice.« Tout ce que je sais, c'est que cela me semblerait bien drôle à moi.
'You!'	» « À vous !
said the Caterpillar contemptuously.	» dit la Chenille d'un ton de mépris.
'Who are YOU?'	« Qui êtes-vous ?
Which brought them back again to the beginning of the conversation.	» Cette question les ramena au commencement de la conversation.
Alice felt a little irritated at the Caterpillar's making such VERY short remarks, and she drew herself up and said, very gravely, 'I think, you ought to tell me who YOU are, first.'	Alice, un peu irritée du parler bref de la Chenille, se redressa de toute sa hauteur et répondit bien gravement : « Il me semble que vous devriez d'abord me dire qui vous êtes vous-même.
'Why?'	» « Pourquoi ?

Annexe 5

Alignement des utterances du chapitre V d'*Alice au Pays des Merveilles*

Translation Augmented LibriSpeech Corpus

Chapter 5

Sentence Number: 0 , Alignment Score: 0.699618, Segment duration: 16.17 seconds

CHAPTER FIVE ADVICE FROM A CATERPILLAR THE CATERPILLAR AND ALICE LOOKED AT EACH OTHER FOR SOME TIME IN SILENCE AT LAST THE CATERPILLAR TOOK THE HOOKAH OUT OF ITS MOUTH AND ADDRESSED HER IN A LANGUID SLEEPY VOICE	CHAPITRE V. CONSEILS D'UNE CHENILLE. La Chenille et Alice se considérèrent un instant en silence. Enfin la Chenille sortit le houka de sa bouche, et lui adressa la parole d'une voix endormie et traînante.

Sentence Number: 1 , Alignment Score: 1.7, Segment duration: 2.77 seconds

WHO ARE YOU	« Qui êtes-vous?

Sentence Number: 2 , Alignment Score: 0.997059, Segment duration: 1.16 seconds

SAID THE CATERPILLAR	» dit la Chenille.

Sentence Number: 3 , Alignment Score: 1.33091, Segment duration: 4.28 seconds

THIS WAS NOT AN ENCOURAGING OPENING FOR A CONVERSATION	Ce n'était pas là une manière encourageante d'entamer la conversation.

Sentence Number: 4 , Alignment Score: 0.377273, Segment duration: 13.36 seconds

ALICE REPLIED RATHER SHYLY I I HARDLY KNOW SIR JUST AT PRESENT AT LEAST I KNOW WHO I WAS WHEN I GOT UP THIS MORNING BUT I THINK I MUST HAVE BEEN CHANGED SEVERAL TIMES SINCE THEN	Alice répondit, un peu confuse: « Je — je le sais à peine moi-même quant à présent. Je sais bien ce que j'étais en me levant ce matin, mais je crois avoir changé plusieurs fois depuis.

Sentence Number: 5 , Alignment Score: 1.404, Segment duration: 3.14 seconds

WHAT DO YOU MEAN BY THAT	» « Qu'entendez-vous par là?

Annexe 6

Les alignements triés par rapport aux scores de CL-CNG & CL-CTS

Transcription	Traduction	Score CLT-CNG	Durée de Locution(s)
E SENIORATU ERIPIMUS	E senioratu eripimus.	1.0	3.17
HE CALLED NICLESS GOVICUM FIBI VINOS URSUS HOMO	Il appela Nicless, Govicum, Fibi, Vinos, Ursus, Homo.	0.9231	5.96
HE WAS INEXCUSABLE INCOMPREHENSIBLE	Il était inexcusable et incompréhensible.	0.9191	3.48
WHAT MODULATIONS POSSIBLE	Que de modulations possibles!	0.9063	3.15
THE INVISIBLE INEXORABLE WHAT AN OBSESSION	L'invisible inexorable, quelle obsession!	0.9057	3.62
FAR LA RIRA TOUR LA RIBAUD RICANDEAU SANS REPOS REPIT REPOS	Far la rira, Tour tala rire, Tour la Ribaud, Ricandeau, Sans repos, répit, répit r	0.9002	10.7
WHO IS PINOCCHIO	– Pinocchio?	0.8746	3.19
GWYNPLAINE WAS THE RELIGION OF DEA	Gwynplaine était la religion de Dea.	0.8696	3.61
OH MONSIEUR LE DUC	monsieur le duc!	0.858	3.18
VIOLENT RESISTANCE CONCLUSION A REFUSAL	Résistance violente; conclusion, refus.	0.8553	3.66
THE CHIEF SIGNED GAIZDORRA CAPTAL	Le chef signa Gaïzdorra, captal.	0.8539	3.51
THE MURMURS REDOUBLED	Les murmures redoublèrent.	0.8529	3.4
EUROPE WILL HAVE HER AMPHICTYONS THE GLOBE WILL HAVE ITS AI	L'Europe aura ses amphictyons; le globe aura ses amphictyons.	0.8504	6.12
THE PRESENCE OF A SPECTRE IN THE HORIZON IS AN AGGRAVATION	La présence d'un spectre dans un horizon est une aggravation a la solitude.	0.8433	5.75
ENNUI CONSOLED BY THE PREMEDITATION OF EXPLOSION	Ennui consolé par la préméditation de l'explosion.	0.8382	4.01
CACAMBO APPLAUDED THIS WISE RESOLUTION	Cacambo applaudit, à cette sage résolution.	0.804	3.29
CRIED THE DUC DANJOU	s'écria le duc d'Anjou.	0.8024	3.04
YES JOHN BUNSBY MASTER OF THE TANKADERE	—Oui, John Bunsby, patron de la_Tankadère_.	0.8018	4.06
COSETTE HAD MARIUS MARIUS POSSESSED COSETTE	Cosette avait Marius, Marius possédait Cosette.	0.8002	3.02
THE CONJECTURE OF MAJOR HEYWARD WAS TRUE	Le major Heyward ne se trompait pas dans sa conjecture.	0.7986	3.71
I SUPPOSE SO	— Je le suppose.	0.7947	3.4
SUBSTANCE ABSOLUTELY INFINITE IS INDIVISIBLE	La substance absolument infinie est indivisible.	0.7903	4.54
YOUR AFFECTIONATE AND AFFLICTED FATHER ALPHONSE FRANKENS	Ton père affectionné et affligé, Alphonse Frankenstein.	0.7858	4.47
IT IS A DECLARATION OF INDIFFERENCE	C'est une déclaration d'indifférence.	0.7797	3.1
EVEN AT WOKING STATION AND HORSELL AND CHOBHAM THAT WA	Même à Woking, à Horsell et à Chobham, tel était le cas.	0.7784	5.68
BARKILPHEDRO ASSUMED AN ATTITUDE OF DEFERENTIAL GRAVITY	Barkilphedro prit l'attitude de la gravité déférente.	0.7772	4.75
OF A VERY DIFFERENT TYPE WAS HIS COMPANION JULIUS BURGER	Son compagnon Julius Burger était d'un type très différent.	0.769	4.2
DEMOISELLE I AM MONSIEUR MALICORNE	-- Demoiselle je suis, monsieur Malicorne.	0.7687	3.12
THEY VISITED THE ISLANDS OF MONTREAL MACONOCHIE AND OGLE	Ils visitèrent les îles de Montréal, Maconochie, pointe Ogle.	0.7674	4.97
THEN A FORMIDABLE SPECTACLE WAS SEEN	Alors on vit un spectacle formidable.	0.767	3.89
MISTER PICKWICK AND MISS BOLO AGAINST LADY SNUPHANUPH AN	On tira les places, et M. Pickwick se trouva avec miss Bolo, contre lady Snupha	0.7669	5.07
THE NORTHERN BASQUE SIGNED HIMSELF GALDEAZUN	Le basque du nord signa Galdeazun.	0.7648	3.46

Annexe 7

Les alignements triés par rapport aux scores de confiance de *hunalign*

Transcription	Traduction	Score	Durée
LOUIS PHILIPPE IS EIGHTEEN THIRTY MADE MAN	Louis-Philippe, c'est 1830 fait homme.	1.0	3.17
ADMISSION FIFTY CENTS CHILDREN HALF PRICE	50, moitié pour les enfants.	0.9231	5.96
THE BRUJON OF EIGHTEEN ELEVEN WAS THE FATHER OF THE BRUJON OF EIGHTEEN THIRT	Le Brujon de 1811 était le père du Brujon de 1832.	0.9191	3.48
THE GREAT EASTERN PUT BACK TO SEA ON JULY	Le_Great-Eastern_ reprit la mer le 13 juillet 1866.	0.9063	3.15
IN GENEVA FIVE HUNDRED PERSONS WERE BURNED DURING FIFTEEN FIFTEEN AND FIFTE	A Genève, cinq cents personnes furent brûlées de 1515 a 1516.	0.9057	3.62
AH SAID A GALLANT OLD GENERAL WHO IN EIGHTEEN O NINE HAD SUNG PARTANT POUF	en 1809, nous n'irons pas seuls au jardin.	0.9002	10.7
FORTY THOUSAND WERE EXECUTED IN ENGLAND FROM SIXTEEN HUNDRED TO SIXTEEN	Quarante mille furent exécutées en Angle— terre dc l'année 1600 à 1680.	0.8746	3.19
BY THE NEXT DAY MARCH TWENTY SEVENTH SIX METERS OF ICE HAD BEEN TORN FROM T	Le lendemain, 27 mars, six mètres de glace avaient été arrachés de l'alvéole.	0.8696	3.61
O FIVE THE TWENTY SIX BISHOPS WERE REDUCED TO TWENTY FIVE THE SEE OF CHESTER E	En 1705, les vingt-six évêques n'étaient que vingt-cinq, le siège de Chester étant vacant.	0.858	3.18
INCENSE DECLARES THE SOSHI RYAKU IS THE MESSENGER OF EARNEST DESIRE	« L'encens, déclare le Soshi-Ryaku(1) , est le Messager du Désir Sincère.	0.8553	3.66
WE WERE OFF THE SOUTHERN TIP OF THE GRAND BANKS OF NEWFOUNDLAND	Le 15 mai, nous étions sur l'extrémité méridionale du banc de Terre-Neuve.	0.8539	3.51
EIGHT THERE WERE IN FULL BLAST IN NEW YORK AND BROOKLYN SIXTEEN WITCHES AND	En 1858, ils étaient en grande faveur à New York et a Brooklyn; on y comptait seize sor— cier	0.8529	3.4
BUT THE EVENING OF NOVEMBER FOURTH ARRIVED WITH THIS UNDERWATER MYSTERY S	Mais le soir du 4 novembre arriva sans que se fût dévoilé ce mystère sous-marin.	0.8504	6.12
WHY DO YE GO ABOUT PERVERTING THE WAYS OF THE LORD	Et il arriva que le grand prêtre lui dit: Pourquoi vas-tu partout pervertir les voies du Seigneur	0.8433	5.75
THUS OF EIGHTEEN TWENTY SEVEN THIRTY SIX WHICH ARE PRODUCTS OF NINE YOU MAK	Ainsi, pour 18, 27, 36, qui sont des multiples de 9, vous obtenez 9 en additionnant 1 à 8, 2 à 7,	0.8382	4.01
ON AUGUST THIRTEENTH SEVENTEEN SEVENTY EIGHT COMMANDED BY LA POYPE VERTRI	En 1778, le 13 août, commandé par La Poype-Vertrieux, il se battait audacieusement contre l	0.804	3.29
AT SIX OCLOCK IN THE MORNING JANUARY EIGHTH I CLIMBED ONTO THE PLATFORM	A six heures du matin- 8 janvier je remontai sur la plate-forme.	0.8024	3.04
THE NEXT DAY MARCH TWENTY SIXTH I RETURNED TO MY MINERS TRADE WORKING TO R	Le lendemain, 26 mars, je repris mon travail de mineur en entamant le cinquième mètre.	0.8018	4.06
AND JUNE EIGHTEEN FORTY EIGHT KNEW A GREAT DEAL MORE ABOUT IT THAN JUNE EIGI	Chapitre II Que faire dans l'abîme à moins que l'on ne cause?Seize ans comptent dans la sou	0.8002	3.02
THOU HAST HAD SIGNS ENOUGH WILL YE TEMPT YOUR GOD	Mais Alma lui dit: Tu as eu assez de signes; tenteras-tu ton Dieu?	0.7986	3.71
NINETY FOUR THE NEW REPUBLIC OF FRANCE CHANGED THE NAME OF THIS SHIP	En 1794, la république française lui changeait son nom.	0.7947	3.4
THE MANNER OF MARRIAGE IN EIGHTEEN THIRTY THREE WAS NOT THE SAME AS IT IS TO	La mode du mariage n'était pas en 1833 ce qu'elle est aujourd'hui.	0.7903	4.54
DURING THE GREAT CHOLERA SCARE OF EIGHTEEN SEVENTY ONE OUR NEIGHBOURHOOD	Lors de la grande épidémie de choléra de 1871, notre voisinage en fut curieusement épargn	0.7858	4.47
HE ABDICATED AT FONTAINEBLEAU IN EIGHTEEN FOURTEEN AND WAS SENT TO THE ISLA	--Il a abdiqué à Fontainebleau en 1814 et a été relégué à l'île d'Elbe.	0.7797	3.1
THIRTY THREE A HUNDRED YEARS AGO MARRIAGE WAS NOT CONDUCTED AT A FULL TRO	En 1833, il y a cent ans, on ne pratiquait pas le mariage au grand trot.	0.7784	5.68
DURING THE DAY OF DECEMBER ELEVENTH I WAS BUSY READING IN THE MAIN LOUNGE	Pendant la journée du 11 décembre, j'étais occupé à lire dans le grand salon.	0.7772	4.75
BUT IT IS SO COUNTED A HUNDRED AND FIFTY IN HER FLEET	En 1705, l'Angleterre, qui n'avait que treize vaisseaux de guerre sous Élisabeth et trente-six	0.769	4.2
WILL YE DENY AGAIN THAT THERE IS A GOD AND ALSO DENY THE CHRIST	Alors Alma lui dit: Nieras-tu encore qu'il y a unDieu, et nieras-tu aussi le Christ?	0.7687	3.12
AT NOON ON THIS DAY OF NOVEMBER EIGHTH WE HEREBY BEGIN OUR VOYAGE OF EXPL	C'est aujourd'hui 8 novembre, à midi, que commence notre voyage d'exploration sous les ea	0.7674	4.97
GENERAL JAMES CLINTON THE BROTHER OF GEORGE CLINTON	Le général James Clinton, frère de George Clinton, alors gouverneur de New-York, et le père	0.767	3.89
IN THE SUMMER OF EIGHTEEN FORTY ONE I FOUND MYSELF AT LITTLEMORE WITHOUT AN	» Dans l'été de 1841, je me trouvai à Littlemore l'esprit libre de toute angoisse et de toute ir	0.7669	5.07
IN OCTOBER EIGHTEEN FIFTEEN HE WAS RELEASED HE HAD ENTERED THERE IN SEVENTEE	En octobre 1815 il fut libéré; il était entré là en 1796 pour avoir cassé un carreau et pris un pa	0.7648	3.46

Table des matières

MOTS-CLÉS : traitement automatique des langues, alignement, corpus parallèle, traduction automatique neuronale, traduction automatique de la parole

RÉSUMÉ

Il existe des corpus parallèles en grande quantité tels que *Europarl*, *OpenSubtitles*, *etc*. pour les systèmes de traduction automatique. Toutefois, dans le domaine de la traduction automatique de la parole, le nombre de corpus disponibles est très restreint. Dans ce mémoire de recherche, nous nous sommes intéressés à la constitution d'un corpus réel de grande taille (236h) pour les systèmes de traduction automatique de la parole à partir du corpus LibriSpeech. Tout d'abord, nous avons récupéré les livres électroniques français correspondant aux livres audios présents dans LibriSpeech en anglais. Ensuite, après avoir aligné les œuvres en français avec les œuvres en anglais correspondantes, nous avons également aligné les segments de paroles des livres audio avec les livres électroniques en anglais, au niveau de la phrase (*utterance*). Ceci nous a permis d'obtenir l'alignement des segments de parole avec la traduction du texte original en français. Finalement, nous avons évalué manuellement 200 alignements et avons ajouté des scores de correspondance entre les transcriptions et les traductions pour trier le corpus en fonction de ces scores.

KEYWORDS : natural language processing, alignment, parallel corpus, speech translation, machine translation

ABSTRACT

Large quantities of parallel corpora such as *Europarl, OpenSubitles, etc.* is available for machine translation systems. However, for speech translation systems, number of available corpora is very limited. In this research paper, we investigate building a large scale (236h) non-synthetic corpus for speech translation systems from prepared speech recordings of LibriSpeech project. First, we gathered available French e-books corresponding to the English audio-books from LibriSpeech. Then, after aligning English and French texts at the sentence level, we also aligned speech segments at the sentence level with their respective translations. This allowed us to obtain an alignment at the utterance level aligned with their translations. Lastly, we manually evaluated 200 alignments and added alignment scores between transcriptions and their respective translations to sort the corpus according to those scores.