La Traduction automatique statistique dans un contexte multimodal

Haithem Afli

UNIVERSITÉ DU MAINE
LABORATOIRE D'INFORMATIQUE DE L'UNIVERSITÉ DU MAINE

T H È S E

pour obtenir le titre de

Docteur en Science

de l'Université du Maine
Mention : INFORMATIQUE

Présentée et soutenue par

Haithem AFLI

La traduction automatique statistique dans un contexte multimodal

Thèse dirigée par Loïc BARRAULT & Holger SCHWENK

soutenue le 7 juillet 2014

Jury :

Rapporteurs :			
	M. Kamel SMAÏLI	-	LORIA, Université du Lorraine
	M. Philippe LANGLAIS	-	RALI, Université de Montréal
Directeur :	M. Holger SCHWENK	-	LIUM, Université du Maine
Co-directeur :	M. Loïc BARRAULT	-	LIUM, Université du Maine
Examinateurs :			
	M. Alexandre ALLAUZEN	-	LIMSI-CNRS, Université de Paris XI
	M. Emmanuel MORIN	-	LINA, Université de Nantes

To the three hundred brave men and women killed during the Tunisian Revolution between 17/12/2010 and 14/01/2011.

To the thousands Egyptian killed during the revolution of 25/01/2011 in Tahrir square and after the coup of 03/07/2013 in Rabiaa square.

To the hundred of thousands killed in Lybia, Syria, Yaman, Borma and Ukraine during these three years of my thesis defending their choice and freedom.

To those who defend freedom anywhere.

Remerciements

Merci à ...

Laurent Besacier et Loïc Barrault,

mes parents, Rebeh et Abdelhafid, pour tout,

mon directeur de thèse Holger Schwenk pour ses conseils, et surtout sa confiance,

tous mes collègues du LIUM.

Cette thèse a été financée par la région des Pays de la Loire sous le projet DEPART.

Résumé

Les performances des systèmes de traduction automatique statistique dépendent de la disponibilité de textes parallèles bilingues, appelés aussi bitextes. Cependant, les textes parallèles librement disponibles sont aussi des ressources rares : la taille est souvent limitée, la couverture linguistique insuffisante ou le domaine des textes n'est pas approprié. Il y a relativement peu de paires de langues pour lesquelles des corpus parallèles de tailles raisonnables sont disponibles pour certains domaines. L'une des façons pour pallier au manque de données parallèles est d'exploiter les corpus comparables qui sont plus abondants.

Les travaux précédents dans ce domaine n'ont été appliqués que pour la modalité texte. La question que nous nous sommes posée durant cette thèse est de savoir si un corpus comparable multimodal permet d'apporter des solutions au manque de données parallèles dans le domaine de la traduction automatique.

Dans cette thèse, nous avons étudié comment utiliser des ressources provenant de différentes modalités (texte ou parole) pour le développement d'un système de traduction automatique statistique. Une première partie des contributions consiste à proposer une technique pour l'extraction des données parallèles à partir d'un corpus comparable multimodal (audio et texte). Les enregistrements sont transcrits avec un système de reconnaissance automatique de la parole et traduits avec un système de traduction automatique. Ces traductions sont ensuite utilisées comme requêtes d'un système de recherche d'information pour sélectionner des phrases parallèles sans erreur et générer un bitexte.

Dans la deuxième partie des contributions, nous visons l'amélioration de notre méthode en exploitant les entités sous-phrastiques créant ainsi une extension à notre système en vue de générer des segments parallèles. Nous améliorons aussi le module de filtrage. Enfin, nous présentons plusieurs manières d'aborder l'adaptation des systèmes de traduction avec les données extraites.

Nos expériences ont été menées sur les données des sites web TED et Euronews qui montrent la faisabilité de nos approches.

Table des matières

Liste des tableaux

Table des figures

Introduction

Dans notre monde de communication à l'échelle mondiale, la traduction automatique est devenue une technologie clef. Actuellement, deux types d'approches coexistent : la première est basée sur l'utilisation de règles syntaxiques et linguistiques (analyse, transfert et génération) tandis que l'approche statistique extrait automatiquement toutes les connaissances à partir de corpus existants [Brown *et al.*, 1990]. Habituellement, on a besoin pour cette dernière approche de corpus bilingues alignés afin d'apprendre un modèle de traduction, et des corpus dans la langue cible pour entraîner le modèle de langue. Cependant, les textes parallèles librement disponibles sont des ressources rares : leur taille est souvent limitée, la couverture linguistique insuffisante ou le domaine des textes n'est pas approprié. Il y a relativement peu de paires de langues pour lesquelles des corpus parallèles de tailles raisonnables sont disponibles pour certains domaines.

Le projet DEPART

Les travaux de cette thèse s'inscrivent dans le cadre du projet DEPART [1] (Documents Écrits et PAroles – Reconnaissance et Traduction) dont l'un des objectifs est l'exploitation de données multimodales et multilingues pour la traduction automatique. Nous considérons le cas, assez fréquent pour des domaines spécifiques, où un manque de données textuelles peut être pallié par l'exploitation de données audio, et où le manque de données parallèles peut être amendé par des corpus comparables.

La question que nous nous sommes posée est de savoir si un corpus comparable multimodal permet d'apporter des solutions au problème du manque de données parallèles dans le domaine de la traduction automatique.

Problématique

Une façon de pallier au manque de données parallèles est d'exploiter les corpus comparables, qui sont plus abondants. Un corpus comparable est un ensemble de textes dans deux langues différentes, qui ne sont pas parallèles au sens strict du terme, mais qui contiennent les mêmes informations. Les travaux précédents dans ce domaine n'ont été appliqués que pour la modalité texte.

Dans cette thèse, nous avons étudié comment utiliser des ressources provenant de différentes modalités (texte ou parole) pour le développement d'un système de traduction statistique. Nous pouvons distinguer plusieurs cas :

– des ressources multimodales : la source est disponible dans une modalité, par exemple sous forme de textes, sa traduction dans une autre modalité (ici audio) ;

1. http://www.projet-depart.org/

- des ressources approximatives : on dispose de données audio ainsi que leur transcriptions ou traductions approximatives (exemple : sous-titres) ;
- des ressources complémentaires : la même information existe simultanément dans plusieurs modalités ;
- des ressources monolingues spécialisées : données en format texte et audio sans leur traduction.

Dans notre contexte de travail, nous nous intéressons à l'exploitation de corpus comparables multimodaux avec différents niveaux de similitude. La multimodalité, dans notre cas, concernera l'utilisation de documents textuels et audio.

Notre but est de trouver une méthode pour exploiter les données comparables multimodales, afin d'en extraire des données nécessaires pour construire, adapter et améliorer nos systèmes de traduction automatique statistique. L'objectif du processus d'extraction de traductions est de permettre la génération automatiquement ou la complétion des ressources multilingues.

Structure du document

Cette thèse est organisée comme suit :

Dans une première partie, nous introduirons d'abord les principes de la traduction automatique et l'état de l'art de l'approche statistique. Nous détaillerons, ensuite, les corpus comparables multilingues et les travaux existant pour l'exploitation de ces types de corpus.

Dans la deuxième partie, nous présenterons les outils et les données que nous avons utilisés dans nos travaux. Puis, nous décrirons les directions de recherche que nous avons suivies dans ce travail.

Dans la troisième partie, nous proposons une approche d'extraction des données parallèles à partir des données comparables multimodales. D'abord, nous examinerons le système d'extraction des phrases parallèles et les expériences de faisabilité de la méthode proposée. Ensuite, nous étendrons cette approche en travaillant sur les segments parallèles. Enfin, nous présentons nos améliorations de la méthode dans les différents modules.

Pour finir, nous conclurons ce manuscrit par un développement sur un ensemble de perspectives envisagées.

Première partie

État de l'art

Chapitre 1

Introduction à la traduction automatique statistique

Sommaire

Ce chapitre est un survol de l'état de l'art en traduction automatique statistique. Après un bref historique de la traduction automatique, nous présenterons l'architecture générale des systèmes de traduction automatique. Les notions et les travaux de référence pour la traduction automatique statistique feront l'objet d'une troisième section. Nous y détaillerons les fondements théoriques, les différents modèles et la notion d'alignement. Nous présenterons aussi les modèles la traduction mot-à-mot et ceux à base de segments, puis nous terminerons par les modèles log-linéaires. Nous conclurons ce premier chapitre par une description des différentes métriques d'évaluation de la traduction automatique.

1.1 Bref historique de la traduction automatique

L'idée de mécaniser la traduction des textes fut émise dès la sortie des premiers ordinateurs, en 1947 par Warren Weaver dans une conversation avec Andrew Booth [Hutchins, 2004]. En 1948, Booth collabora avec Richard Richens sur les premières expériences à l'aide de *traduction mécanique* (mechanical translation) de cartes perforées en 1948. C'est lors de la seconde guerre mondiale que les premiers essais d'élaboration de machines à traduire automatiquement des messages voient le jour, lorsque l'armée américaine tente de mettre au point des ordinateurs susceptibles de déchiffrer les messages codés de l'armée japonaise.

En 1949, il a été rapporté dans un journal que Harry Huskey envisage la traduction avec les ordinateurs du SWAC [1] à Los Angeles. Puis, en juillet de la même année, Warren Weaver a écrit le fameux *memorandum* qui a stimulé les débuts de la recherche en traduction automatique : à l'université de Washington (Erwin Reifler), université de Californie à Los Angeles (Victor Oswald, Stuart Fletcher) ainsi qu'à l'institut technologique de Massachusetts (MIT). Ensuite IBM, en collaboration avec l'Université de Georgetown, réalisèrent en 1954 à New York, la première démonstration publique d'un système de traduction automatique russe-anglais. C'était une expérience à petite échelle de 250 mots seulement et six règles « grammaticales » selon [Hutchins, 2004]. Il fallu attendre la fin des années 90 et l'avènement de la société de l'information pour voir apparaître de véritables débouchés aux logiciels de traduction automatique, par exemple : traduction à la volée de pages WEB, traduction de dépêches d'actualités ...

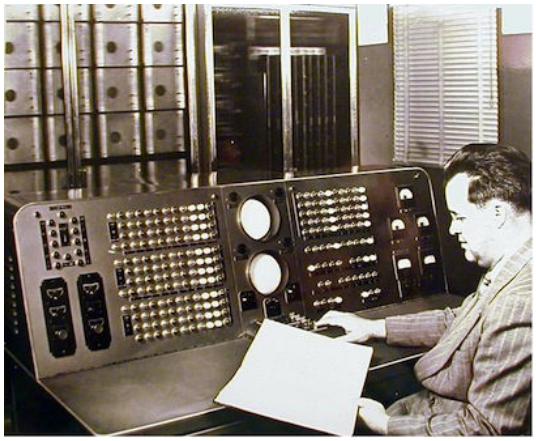

FIGURE 1.1 – Les ordinateurs de type SWAC en 1950.

1. Standards Western Automatic Computer

La traduction automatique s'est enrichie, par la suite, d'une dizaine de paires de langues supplémentaires et d'une analyse statistique. Bien que ces programmes aient pu bénéficier de progrès théoriques et techniques considérables, les résultats obtenus sont restés quelque peu décevants. Aucune machine n'était en mesure de traduire, seule et de façon pleinement satisfaisante, un texte, quel qu'il soit, même si la traduction automatique rendait de grands services aux traducteurs humains, et permettait de réaliser des gains de productivité assez significatifs.

Depuis quelques années, la traduction automatique connaît un essor considérable grâce au développement d'internet, puisque divers systèmes en ligne permettent de traduire automatiquement des pages web et des textes plus ou moins brefs. À l'heure actuelle, cette technologie est appréciée du grand public, car elle permet de déchiffrer de façon grossière le thème d'un texte dans une langue totalement inconnue, et les principaux faits ou éléments d'informations qu'elle contient. Toutefois, la plupart des outils disponibles se limitent à un moteur sans vocabulaire « métier » ou à des dictionnaires traduisant mot-à-mot ; ils ne traitent qu'un nombre limité de mots, et offrent des prestations généralistes, inadaptées aux documents trop techniques ou écrits dans une langue soutenue.

De tels outils présentent plusieurs autres avantages, pour les utilisateurs : ils sont gratuits (pour la plupart), « instantanés », acceptent des entrées de différentes natures (mots, textes, mail, pages web, etc.) et proposent la traduction de plusieurs langues (exemple Google traduction [2] : qui gère 145 langues). D'autre part, pour les entreprises développant ces outils en ligne (Reverso, Systran ...), ils représentent une version de démonstration qui incite les utilisateurs à acheter des logiciels payants, plus avancés.

1.2 Architecture linguistique des systèmes de traduction automatique

Selon le triangle de Vauquois [Vauquois et Boitet, 1985] présenté dans la figure 1.2, il existe trois types de base d'architectures linguistiques :

Les systèmes directs : ils se basent sur des équivalences de termes, et à partir de la consultation d'un dictionnaire, ils traduisent mot à mot sans faire aucune analyse. Ce sont des systèmes qui traitent une seule paire de langues et unidirectionnelle. Ils opèrent directement au niveau du texte d'entrée (source) et du texte de sortie (cible) sans utiliser de représentation intermédiaire. Ces systèmes peuvent être utiles dans certains cas d'application restreinte, mais ils sont limités. Ils représentent les systèmes de première génération des années 1950 (russe-anglais & anglais-russe).

2. https ://translate.google.com

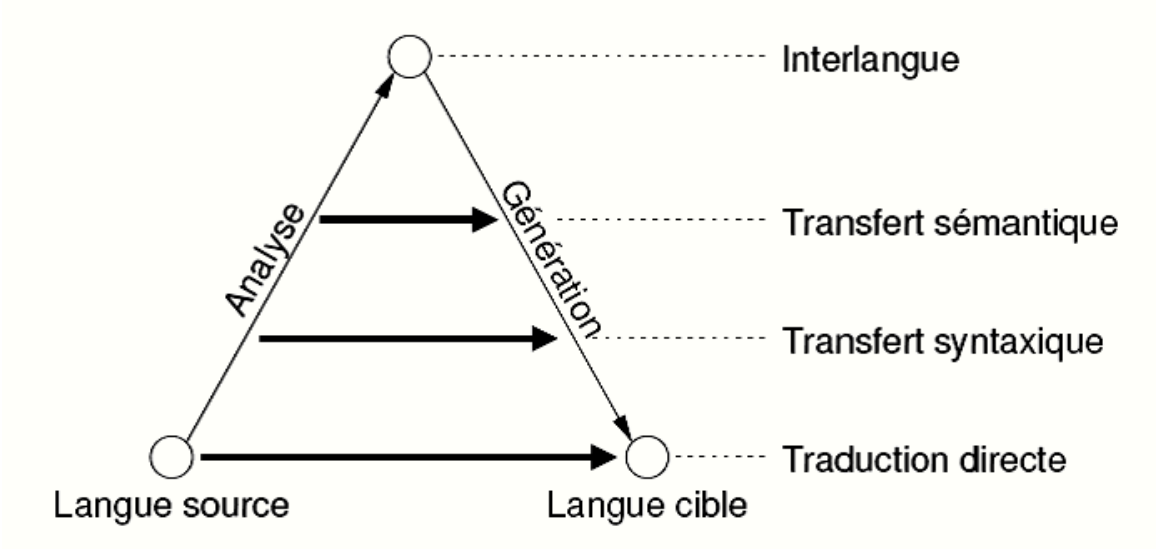

FIGURE 1.2 – Triangle de Vauquois : représentation des différentes architectures linguistiques.

Les systèmes à transfert : ces systèmes de deuxième génération ont un principe plus complexe que celui des systèmes de première génération (directs). Actuellement, les systèmes à architecture basés sur le transfert utilisent trois modules : l'analyse du texte en langue source, le transfert, et la génération dans la langue cible. Ils sont les plus couramment utilisés puisqu'ils facilitent l'intégration d'une nouvelle langue, contrairement aux systèmes directs où l'ajout d'une langue revient à créer un nouveau système. Pour passer d'une phrase source à sa traduction, ils utilisent des représentations intermédiaires basées sur une analyse syntaxique ou sémantique plus ou moins profonde.

Les systèmes à pivot : cette approche est basée sur une représentation dite *pivot* qui réduit le problème de traduction en deux grandes étapes : construire une représentation pivot à partir d'une phrase source et générer une phrase cible à partir de cette représentation. L'importance de cette approche vient du fait que les modules d'analyse et de génération sont réutilisables pour la création d'un système pour un nouveau couple de langues. Cette approche reste toutefois un domaine de recherche actif.

1.3 Les principes de la traduction automatique

Les tendances actuelles de la recherche en traduction automatique répartissent les chercheurs du domaine entre deux approches dominantes. Une première utilise des méthodes expertes et vise à formaliser toutes les connaissances nécessaires à la traduction. La seconde, basée sur des méthodes empiriques, faite en sorte que toute

connaissance linguistique est acquise de manière empirique et automatique à partir de corpus.

Historiquement, la première approche utilisée pour traduire des textes fut la méthodologie experte basée sur des règles linguistiques. L'ensemble des règles définit les possibilités d'association des mots selon leurs catégories lexicales et permet de modéliser la structure d'une phrase donnée. Celle-ci nécessite beaucoup de travail de la part des linguistes pour définir le vocabulaire et la grammaire. Cette méthode peut fournir de bons résultats. Par ailleurs, l'un des inconvénients des modèles construits sur une approche structurelle est la conception des grammaires. Celles-ci doivent être suffisamment robustes pour prendre en compte tous les phénomènes syntaxiques d'une langue.

D'autres types d'approches furent développées : l'approche à base d'exemples (mémoire de traduction) et l'approche statistique qui font en sorte que toute connaissance linguistique soit acquise de manière empirique et automatique à partir de grandes quantités de texte (les distributions de probabilités remplacent ici les règles). L'un des avantages des approches statistiques est leur indépendance vis-à-vis de l'expertise des linguistes. Les systèmes apprennent automatiquement à partir de millions d'exemples des données parallèles collectées. Nous détaillerons, par la suite, cette approche de traduction statistique.

1.4 L'approche par canal bruité

La traduction automatique a été profondément renouvelée par le développement de modèles statistiques. En utilisant ces modèles, il est possible, à partir de corpus parallèles annotés, d'apprendre automatiquement les associations les plus statistiquement significatives entre deux langues, donnant lieu à des systèmes de traduction entièrement statistiques. L'approche la plus courante de la traduction statistique s'inscrit dans le cadre des approches dites « canal bruité » [3] présenté dans la figure 1.3 introduite par [Shannon, 1948] que l'on explique ici de manière intuitive. Deux personnes, un émetteur E et un récepteur R, souhaitent communiquer via un canal bruité. Ce canal est « tellement bruité » qu'une phrase S déposée par E à l'entrée du canal est reçue par R comme une autre phrase T, traduction de S.

Le but du récepteur R est de retrouver la phrase source à partir de la phrase reçue et ses connaissances du canal bruité. Chaque phrase (S) de la langue source est une origine possible pour la phrase reçue T. On assigne une probabilité $P(S|T)$ à chaque paire de phrases (S, T).

3. Terme venant du domaine de reconnaissance de la parole (*Noisy Channel Model*)

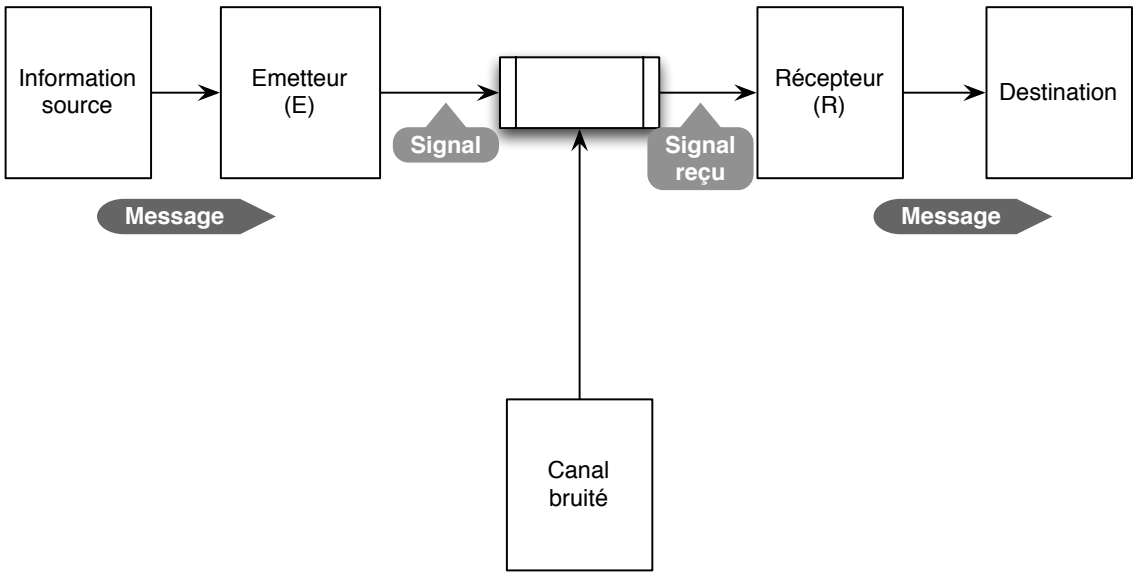

FIGURE 1.3 – llustration de la théorie du *canal bruité*.

1.5 Modélisation statistique de la traduction automatique

La modélisation de la traduction automatique statistique repose sur la théorie mathématique de distribution et d'estimation probabiliste développée en 1990 par Peter F. Brown et ses collègues à *IBM* au *Watson* décrite dans [Brown *et al.*, 1990].

L'hypothèse initiale est que toute phrase d'une langue est une traduction possible d'une phrase dans une autre langue. Si on traduit depuis une langue source s vers une langue cible t, le but est de trouver la phrase cible t la plus appropriée pour traduire la phrase source s.

Pour chaque paire de phrases possible (s, t), on attribue une probabilité $P(t|s)$ qui peut être interprétée comme la probabilité qu' « un traducteur » produise t dans la langue cible, lorsque la phrase s a été énoncée dans une langue source, ou autrement dit, la probabilité que la traduction de s soit t. Ceci s'explique par l'équation fondamentale que nous présenterons dans la section 1.5.2.

En traduction automatique statistique, les modèles probabilistes sont utilisés pour trouver la meilleure traduction possible t^* d'une phrase source donnée s, parmi toutes les traductions t possibles dans la langue cible. Il s'agit alors d'appliquer des méthodes d'apprentissage statistiques afin d'entraîner le système avec des millions de mots, dont des textes monolingues en langue cible et des textes alignés composés d'exemple de traduction entre les deux langues.

La figure 1.4 [4] présente une comparaison entre le système statistique basé sur l'apprentissage et le système expert de traduction automatique basé sur les règles codées manuellement.

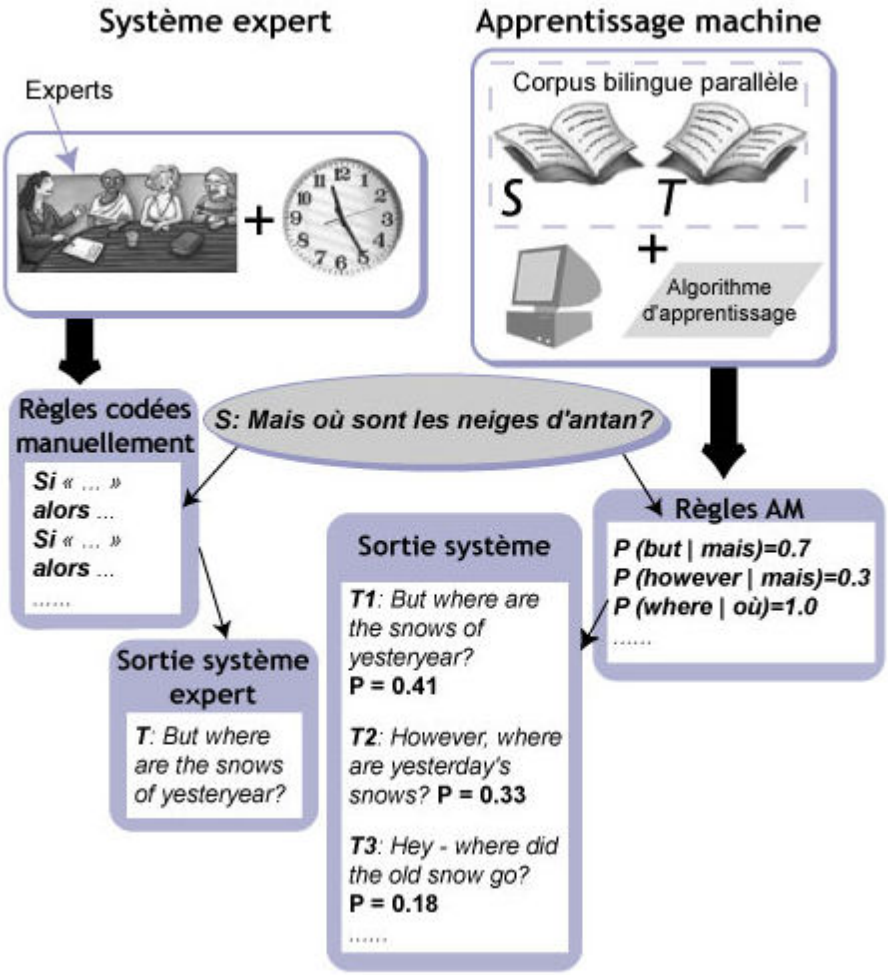

FIGURE 1.4 – Comparaison entre les systèmes de traduction automatique experts et statistiques.

4. extrait du rapport technique du projet PORTAGE (http ://archive.nrc-cnrc.gc.ca/eng/projects/iit/machine-learning/portage-technical.html)

1.5.1 Notion de corpus

Les paramètres des modèles statistiques sont estimés à partir de l'analyse d'une grande quantité de données d'apprentissage monolingue ou bilingue : appelé *corpus*.

Les corpus sont indispensables et précieux en traitement automatique du langage naturel. Ils permettent en effet d'extraire un ensemble d'informations utiles pour les traitements statistiques.

Dans le cas de la traduction automatique statistique, nous avons besoin de textes composés d'exemples de traduction entre les deux langues, ou plus précisément d'un ensemble de phrases traduites en langue source et en langue cible et alignées par paire. Ce que l'on appelle *bitexte* représente un corpus bilingue parallèle (un texte dans une langue source et sa traduction) où les liens de traduction entre les phrases sont explicites. Un bitexte est obtenu à partir d'un corpus bilingue en alignant le corpus au niveau des phrases. Il existe deux types d'informations exploitées dans les algorithmes d'alignement des phrases :

- Les informations métriques : qui utilisent la longueur des phrases (en terme de nombre de caractères ou mots) comme critère de mise en correspondance [Gale et Church, 1991].

- Les informations à caractère linguistique : où on propose d'aligner des corpus bilingues en exploitant le fait que deux phrases en relation de traduction dans deux langues proches partagent souvent des mots communs ou proches : comme des noms propres, des données chiffrées, ou encore des mots partageant la même racine [Simard *et al.*, 1993] . (Exemple : accès/access, activité/activity, parlement/parliament. . .).

Le bitexte est utilisé pour l'entraînement, le développement et l'évaluation du système de traduction automatique statistique. Le corpus d'apprentissage a pour but d'entraîner et de construire le modèle à l'aide des méthodes d'apprentissage statistique. Le corpus de développement servira à l'ajustement et l'amélioration des modèles apprit tandis que le corpus de test permet de vérifier et tester la qualité du modèle appris.

1.5.2 Équation fondamentale

La traduction statistique se définit par la recherche de la phrase cible ayant la plus grande probabilité d'être la traduction d'une phrase source. En appliquant le théorème de Bayes sur la paire de phrases (s, t), où la phrase t en langue cible est la traduction de la phrase s en langue source, on obtient pour chacune des paires une probabilité $Pr(t|s)$ que la machine produise le mot t comme traduction de la phrase s :

$$Pr(t|s) = \frac{Pr(s|t)Pr(t)}{Pr(s)} \tag{1.1}$$

Puisque nous calculons l' $\arg\max_t$ et s est indépendante de t, en utilisant seulement le produit $Pr(s|t)Pr(s)$, on arrive à l'équation fondamentale 1.2 en traduction automatique :

$$t^* = \arg\max_t Pr(t|s) = \arg\max_t Pr(s|t)Pr(t) \tag{1.2}$$

Dans cette formule, $Pr(t)$ est appelée le modèle sur la langue *cible* et $Pr(s|t)$ est appelée le modèle de traduction. Ces deux modèles sont appris empiriquement à partir de corpus. $Pr(t)$ est la probabilité de la phrase t dans la langue cible et $Pr(s|t)$ a pour fonction de vérifier que la phrase source s est bien une traduction de la phrase cible t. La phrase t^* retenue pour la traduction de la phrase s sera celle qui maximise le produit des deux modèles probabilistes de l'équation 1.2. Le problème de la traduction peut être divisé en trois sous problèmes :

1. calculer les paramètres du modèle de langue *Pr(t)* ;
2. calculer les paramètres du modèle de traduction *Pr(s|t)* ;
3. réaliser un mécanisme capable d'effectuer l'opération de maximisation de l'équation 1.2 en un temps acceptable (c'est ce que l'on appelle le décodeur).

Cette décomposition a pour intérêt de donner un double contrôle sur la qualité de la traduction en apprenant les deux modèles (modèle de langue en langue cible et modèle de traduction) indépendamment, de telles sortes que les erreurs d'un des modèles peuvent être compensées par l'autre, ce qui rend le modèle final plus robuste.

1.5.3 Modèle de langue

Selon l'équation 1.2, $Pr(t)$ représente le modèle de langue, c'est-à-dire, la composante du système de traduction qui est en charge d'introduire les contraintes imposées par la syntaxe et la sémantique de la langue cible. Le rôle d'un modèle de langue est d'estimer la probabilité d'une séquence de mots ou d'une phrase. Ainsi, plus une séquence de mots (ou phrase) sera vraisemblable et conforme au modèle de langue, plus sa probabilité, $Pr(t)$, sera élevée [Brown *et al.*, 1990]. L'objectif final est de guider la recherche des séquences de mots les plus probables en

se basant sur des connaissances extraites d'un corpus monolingue de la langue cible [Koehn *et al.*, 2003].

Sans perte d'information, si l'on considère que t^I est une suite de I mots, $t^I = w_1, w_2, ..., w_I$, alors :

$$Pr(t^I) = \prod_{i=1}^{I} P(w_i|w_1...w_{i-1}) \tag{1.3}$$

Cette formule peut être approximée par l'hypothèse que le mot w_i de t^I ne dépend que des $n-1$ mots précédents, ce qui permet de réécrire la probabilité $Pr(t^I)$ comme suit :

$$Pr(t^I) = Pr(w_1)Pr(w_2|w_1)P(w_3|w_1w_2) \cdot ... \cdot Pr(w_I|w_1w_2.....w_{I-2}w_{I-1}) \tag{1.4}$$

Par la suite, on appellera ce modèle de langue *modèle n-grammes*. Le traitement de modèle de langue en traduction automatique était sujet de plusieurs travaux comme les modèles d'espace continue [Schwenk, 2007].

1.5.4 Modèle de traduction

Suite à l'application du modèle du canal bruité [Shannon, 1948, Shannon, 2001] et la formulation avec le théorème de Bayes, le modèle de traduction modélise le processus de génération d'une phrase source à partir d'une phrase cible dans le sens inverse de la traduction. Nous nous intéressons ici au problème du calcul de $Pr(s|t)$, la probabilité qu'une phrase s soit la traduction de la phrase t. Ce modèle est appris à partir d'un corpus bilingue aligné en phrases.

Étant donné que, généralement, les données du corpus ne sont pas assez suffisantes pour apprendre directement $Pr(s|t)$, nous pouvons décomposer les phrases t et s en des unités plus petites. Alors, chaque couple de phrases se décompose en $t = t_1t_2 \dots t_M$ et $s = s_1s_2 \dots s_N$ où M et N sont le nombre de subdivisions des phrases. Selon une technique d'alignement, les éléments de t sont ensuite mis en correspondance avec les éléments de s. Il est donc nécessaire d'apprendre des alignements entre les phrases du corpus parallèle utilisé.

Exemple (figure 1.5) : (Jean aime Marie| John loves Mary), il est raisonnable de supposer que *John* produit *Jean*, *loves* produit *aime*, et *Mary* produit *Marie*. En pratique, une telle relation entre les phrases de la langue cible et de la langue source n'est pas le cas le plus fréquent, puisque généralement les deux langues ne

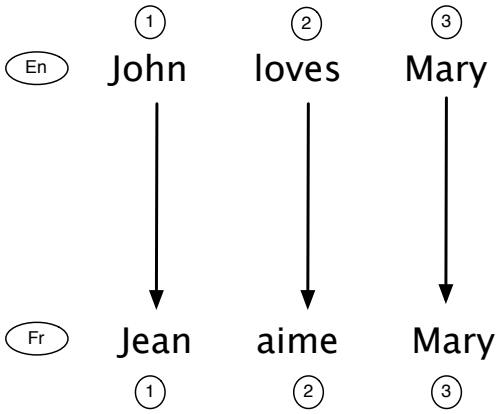

FIGURE 1.5 – Exemple d'alignement 1 : 1 (anglais/français).

possèdent pas la même longueur de phrases (c'est notamment le cas pour les phrases françaises qui sont habituellement plus longues que leur équivalent en anglais.). De plus, même à longueur égale, souvent il n'a y pas de relation 1 :1.

1.5.5 Alignement

Un alignement entre les mots d'une phrase et sa traduction consiste à associer à chaque mot ou groupe de mots de la phrase traduite le mot ou le groupe de mots de la phrase source correspondant à sa traduction. La prédiction de ces relations de traduction joue un rôle central dans les systèmes de traduction. Ceux-ci reposent en effet sur une table de traduction, construite à partir de ces alignements, qui décrit la probabilité de traduire un mot ou un groupe de mots de la langue source en un groupe de mots de la langue cible.

Les modèles basés sur les mots proposés dès les années 90 par [Brown *et al.*, 1990] ont ouvert la porte à de nombreuses recherches en alignement et en traduction en général. Afin d'estimer les probabilités de traduction, il est nécessaire d'établir des correspondances entre les mots de la langue source et de la langue cible.

Il existe deux types d'alignement : les alignements type 1 : 1 (exemple figure 1.5) si pour chaque mot dans la langue source correspond un et un seul mots dans la langue cible, et les alignements type $n : m$ (exemple figure 1.6) [5], dans le cas où un groupe de mots dans la langue source correspond à une traduction utilisant un ou plusieurs mots dans la langue cible et réciproquement.

La probabilité de traduction est estimée par la somme des alignements possibles entre les éléments de s et ceux de t, nous considérons $P(s|t) = \sum_{a \in A} P(s, a|t)$, avec a un alignement possible entre la phrase source et la phrase cible. Cette somme est néanmoins trop grande pour être calculée directement, car le nombre d'alignements croît très rapidement avec le nombre d'unités de phrases. À titre indicatif, pour une

5. Exemples empruntés de [Brown *et al.*, 1993]

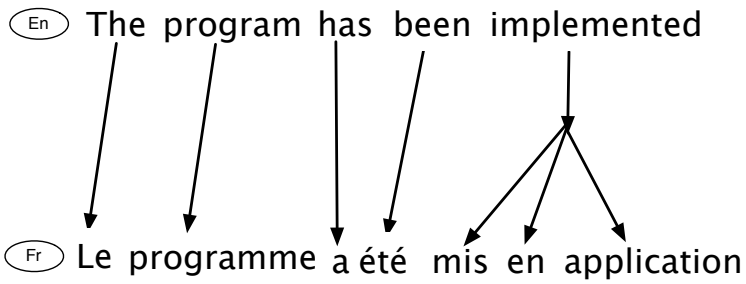

FIGURE 1.6 – Exemple d'alignement 1 : N (anglais/français).

phrase source de N mots et une phrase cible de M mots, nous avons $(N+1)^M$ alignements possibles entre les mots. Dans notre exemple 1.5, la fonction d'alignement a est telle que :

$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3\}$.

Les mots en langue cible qui ne s'alignent à aucun mot en langue source sont alignés à un mot spécial vide (NULL), afin que tous les mots de la phrase cible possèdent un alignement.

1.5.6 Modèles de traduction à base de mots

Dans ces modèles, les mots sont les unités fondamentales de traduction. Les éléments traduits sont les mots, le but est donc d'aligner les corpus parallèles en mots. Le système d'alignement le plus utilisé est Giza++ qui met en œuvre plusieurs modèles génératifs, les modèles IBM.

Les modèles IBM

Les articles de référence des méthodes de traduction probabilistes à base de mots d'IBM font mention de 5 modèles de traduction dont le but est d'évaluer la probabilité $P(s|t)$. Le premier modèle d'IBM, décrit dans [Brown *et al.*, 1990], considère la distribution des mots comme uniforme, donc tous les alignements comme équiprobables (il ne prend pas en compte l'ordre des mots). Il faut attendre le deuxième modèle d'IBM, décrit dans [Brown *et al.*, 1991], pour que l'ordre des mots soit pris en compte. Celui-ci intègre un modèle de distorsion, appelé aussi modèle de réordonnancement, qui représente la distance entre un mot de la phrase source s et le mot de la phrase cible t qui l'a produit. Ce modèle permet l'alignement à un mot spécial appelé *null* utilisé lorsqu'un ou plusieurs mots d'une phrase n'ont pas de correspondance dans l'autre phrase (formellement, il y a un mot *null* dans chacune des langues).

Le modèle IBM 3 introduit la notion de *fertilité* qui autorise un mot source à

générer plusieurs mots cibles. Les principaux paramètres du modèle IBM 3 sont définis par :

- $P(n_a(s)|s)$ est le modèle de fertilité où $n_a(t)$ est le nombre de mots de t alignés avec s dans l'alignement a ;
- P_{null} est la probabilité d'insertion du mot $null$;
- $P(t|s_a(t))$ est le dictionnaire des mots où $s_a(t)$ est le mot de s aligné avec t dans l'alignement a ;
- $P_{distorsion}(t, a, s)$ est le modèle de distorsion qui est calculé avec s_i, la i-ème position de s et t_j est le j-ème position de la traduction dans t et N, M , respectivement le nombre de mots des phrases source et cible.

Ce modèle est donné par la formule suivante :

$$Pr(s|t) = \prod_{i \in T} P(t|s_a(t)) \cdot \prod_{i \in S} P(n_a(s)|s) \cdot P_{distorsion}(t, a, s) \tag{1.5}$$

Les modèles 4 et 5 d'IBM utilisent toutefois une modélisation plus complexe que celle du modèle 3 pour le réordonnancement. Dans ces deux modèles, la phrase cible est développée par étape. Tout d'abord, pour chaque mot dans la phrase source s, on regarde le nombre de mots dans la phrase cible t qui lui correspond, puis on essaye de définir la structure de ces mots. Enfin, après avoir connecté les mots dans les deux phrases on cherche les bonnes positions des mots dans la phrase cible, c'est le travail de l'algorithme de réordonnancement. Dans la phase de réordonnancement, on définit réellement les connexions entre les mots.

Pour le modèle 3, comme dans le modèle 2, la probabilité de connexion $Pr(t|s)$ dépend de la position des mots et de la longueur des deux phrases source et cible. Par contre, dans le modèle 4, la probabilité de connexion dépend des structures des mots liés et aussi des positions des autres mots cibles connectés avec le même mot source. Quant au modèle 5, reste toujours le plus utilisé.

Entraînement des Paramètres (GIZA)

L'équipe d'IBM [Brown *et al.*, 1993] a fourni l'algorithme d'alignement de mots le plus populaire à ce jour, celui qui a été mis en œuvre dans les logiciels *GIZA* [Al-Onaizan *et al.*, 1999] et son extension Giza++ [Och et Ney, 2003] qui sont adoptés par la plupart des systèmes de traduction automatique comme Pharaoh [Koehn, 2004] et Moses [Koehn *et al.*, 2007]. Ces programmes réalisent l'entraînement des paramètres des modèles que nous avons présentés ci-dessus, et sont disponibles publiquement [6]. L'entraînement est réalisé à partir d'un corpus parallèle, aligné par phrases. Aucune information n'est nécessaire, même un lexique. Giza++ entraîne successivement des modèles de complexité croissante, comme

6. Giza++ : http ://www.fjoch.com/GIZA++.html

proposé par [Brown *et al.*, 1993].

Limitations des modèles de traduction à base de mots

Cette méthode d'alignement en mots trouve ses limites au niveau des phrases dont la traduction de certains mots dépend de la traduction d'autres mots de la phrase. La notion de *fertilité*, en effet, autorise un mot cible à être aligné avec plusieurs mots sources $(m : 1)$ mais n'autorise pas que plusieurs mots cibles soient alignés avec un seul mot source $(1 : n)$, comme présenté dans la figure 1.7.

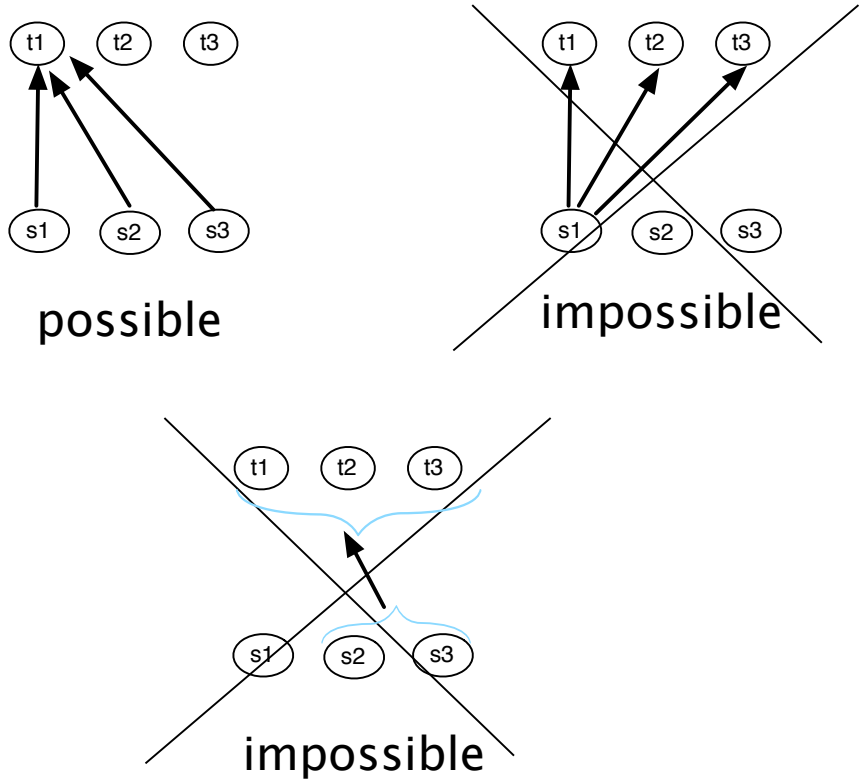

FIGURE 1.7 – Possibilités d'alignement en mots avec les modèles IBM.

Pour prendre en compte toutes les complexités du langage, il est nécessaire de pouvoir générer des correspondances $(n : m)$. Et pour cela, il est indispensable de considérer, non plus les mots, mais plutôt des groupes de mots.

1.5.7 Modèles de traduction à base de segments de mots

La notion d'alignement en segments de mots es introduite par [Och, 1999] et [Koehn *et al.*, 2003], où les auteurs proposent de pallier aux difficultés que rencontre l'alignement en mots des modèles IBM, pour pouvoir construire des alignements $(n : m)$. Comme leur nom l'indique, l'unité de traduction de ces modèles est le segment de mots. Un segment de mots peut compter un ou plusieurs mots, par

exemple, $e = e_i, ..., e_{i+l-1}$ regroupe L mots, avec $l >= 1$. Cette notion est élaborée dans [Koehn, 2004] sous le terme de *table de traductions*. Très vite, ces modèles se révèlent plus performants que ceux basés sur les mots. Plusieurs heuristiques sont alors imaginées pour effectuer l'extraction de segments possibles et créer la table de traduction. La plus utilisée est celle qui consiste à aligner les phrases en segments à partir de l'alignement en mots du corpus [Och, 1999]. Les principaux composants de l'équation 1.6 des modèles basés sur les segments sont :

- le modèle de traduction $P(s|t)$;
- le modèle de langage $P(t)$;
- le modèle de distorsion ou réordonnancent $\Omega(s|t)$.

$$Pr(t|s) = P(t) \times P(s|t) \times \Omega(s|t) \tag{1.6}$$

Lors de la traduction, la phrase source s est alors décomposée en segments. Tous ces segments sont traduits en langue cible à l'aide du modèle de traduction et les segments sont ensuite réordonnancés avec le modèle de distorsion. Une des limites des modèles basés sur les segments concerne leurs applications pour des segments non contigus. D'autres parts, il est nécessaire de souligner que l'espace nécessaire pour stoker la table de traduction et le nombre d'hypothèses à explorer augmente considérablement avec la taille du corpus.

1.5.8 Le modèle log-linéaire

Pour combiner les différents modèles utilisés dans la traduction statistique, [Och et Ney, 2003] ont proposé d'utiliser un modèle linéaire discriminant, avec la log-probabilité comme fonction caractéristique. D'une façon générale, le modèle log-linéaire usuel a pour fonction de prédire la valeur d'une variable attendue à partir d'un ensemble de variables descriptives non nécessairement indépendantes. Le principal avantage de cette méthode est qu'il permet facilement l'ajout de variables descriptives.

Dans le cas de la traduction automatique statistique, un modèle de combinaison log-linaire est utilisé à la place d'une simple combinaison linéaire, car les différentes valeurs des probabilités de $P(t|s)$ différent souvent en ordre de grandeur. Dans un tel cas, la combinaison log-linéaire est meilleure que la combinaison linéaire, car celle-ci présuppose que chaque fonction fournit une quantité d'information proportionnelle à sa probabilité correspondante. Le modèle initial tel que décrit par IBM [Brown *et al.*, 1993] a évolué, par la suite, en un modèle log-linéaire. Ce modèle est pour la première fois utilisé en traduction automatique dans [Och et Ney, 2003], où les auteurs ont décrit la manière dont le modèle combine différentes composantes

(modèle de langage, modèle de traduction, modèle de réordonnancement, etc.) en utilisant l'estimation des paramètres par maximisation de vraisemblance.

1.6 Évaluation de la qualité des traductions

Une fois qu'une traduction est réalisée, il s'agit d'en évaluer la qualité. Prenons l'exemple de la phrase source *« La pratique, c'est quand tout fonctionne et que personne ne sait pourquoi. »*[7] ; les traductions suivantes obtenues par deux différents traducteurs automatiques ; Googlequi se base sur l'approche statistique et Systran qui se base sur l'approche des règles ; ne sont pas de la même qualité :

- **Google** : The practice is when everything works and nobody knows why.
- **Systran** : The practice, it is when all functions and that nobody knows why.

Pour évaluer la qualité des systèmes de traduction automatique, plusieurs approches sont utilisées.

1.6.1 Évaluation humaine

Lors d'une évaluation humaine de la traduction automatique, plusieurs participants évaluent chaque traduction en fonction de critères précis. Les critères de qualité peuvent être multiples et inclure, par exemple, des critères de correction grammaticale et du sens du texte. HTER [Snover *et al.*, 2006] présente un des critères humains habituellement utilisés, qui est une mesure manuelle avec laquelle les humains ne classent pas directement les traductions, mais plutôt génèrent une nouvelle traduction de référence qui est plus proche de la sortie du système en conservant la fluidité du sens de la référence. Cette référence permet de calculer les erreurs produites par le système, en la comparant avec la traduction automatique.

Ces critères de qualité constituent la vraie mesure de la qualité du système, mais requièrent une coûteuse intervention humaine. Par ailleurs, toute évaluation subjective souffre des problèmes de non-reproductibilité et de variabilité inter-annotateur. C'est pourquoi plusieurs mesures automatiques ont été développées au fil des années. Leur objectif est d'être corrélé avec les scores que produirait une évaluation humaine, tout en étant reproductible, beaucoup moins coûteuse et rapide pour pouvoir optimiser et comparer les systèmes.

1.6.2 Évaluation automatique

L'évaluation automatique est l'un des défis majeurs dans le développement d'un système de traduction automatique. Elle présente en soi un domaine de recherche actif [Snover *et al.*, 2009], alors qu'elle est souvent considérée comme un problème résolu et trivial dans de nombreux autres domaines. Par exemple, en reconnaissance

7. d'Albert Einstein

de la parole, pour toute entrée i, il n' y a qu'une seule sortie correcte σ. Tandis que, dans la traduction automatique, il y a un ensemble de traductions correctes $\sigma = \{\sigma_1, \sigma_2, .., \sigma_n\}$.

Diverses méthodes ont été proposées dans le but d'automatiser l'évaluation de systèmes de traduction automatique. Cette évaluation nécessitait auparavant une intervention humaine, requérant beaucoup de temps et de ressources. Une des solutions retenues actuellement consiste à produire un ou des scores reflétant la similarité ou la distance entre les sorties d'un système et une ou plusieurs références.

Le score BLEU (Bilingual Evaluation Understudy) est proposé par Papineni dans [Papineni *et al.*, 2002]. L'idée principale est la comparaison de la sortie du traducteur avec une ou plusieurs traduction(s) de référence. Les statistiques de co-occurrence et de *n-grammes*, basées sur les ensembles de $n-grammes$ (une séquence de n mots) pour les segments de traduction et de référence appelés *précision modifiée* (p_n), sont calculées pour chacun de ces segments et sommées sur tous les segments. Cette moyenne est multipliée par une pénalité de brièveté (BP), destinée à pénaliser les systèmes qui essaieraient d'augmenter artificiellement leurs scores en produisant des phrases délibérément courtes.

L'expression générale pour BLEU est définie par l'équation 1.7, où le paramètre w_n donne la possibilité de pondérer des $n-grammes$ de tailles différentes. dans la version standard de BLEU, $w_n = \frac{1}{N}$, et $N = 4$. Il varie de 0 à 1 et il est d'autant meilleur que grand.

$$BLEU = BP \cdot \exp(\sum_{n=1}^{N} w_n \cdot \log p_n) \tag{1.7}$$

La mesure BLEU a gagné le statut de mesure automatique de référence au sein de la communauté de traduction automatique.

Le score METEOR [Lavie et Agarwal, 2007] introduit plusieurs concepts intéressants. Il utilise un algorithme itératif qui aligne, lors d'une première étape, les mots strictement identiques et tente dans une seconde étape, d'aligner les mots restants. Au cours de cette dernière, *joli* et *jolie* pourront ainsi être alignés. Le score METEOR intègre, d'autre part, une pénalité dont le but est de favoriser une traduction qui a de longs segments consécutifs alignés avec la référence.

Le score TER[8] [Snover *et al.*, 2006] et représente une évolution du score WER[9] en lui ajoutant l'opération de « décalage » (dite shift). Un décalage permet de déplacer un groupe de mots contigus vers la gauche ou la droite ; tout décalage

8. Translation Edit Rate
9. Word Error Rate : score utilisé pour l'évaluation des systèmes de reconnaissance automatique de la parole.

compte comme une seule édition, quels que soient le nombre de mots déplacés et l'amplitude du déplacement. Lorsque plusieurs références sont disponibles, le score TER est déterminé par le nombre d'éditions entre la phrase candidate et la référence la plus proche.

La variante *TERp* [Snover *et al.*, 2009] introduit trois nouvelles opérations : l'accord sur la racine du mot (stem match), l'accord sur les synonymes (synonym match) et la paraphrase.

Autres mesures

La liste des mesures automatiques que nous avons présentée ici est loin d'être exhaustive. Nous avons présenté les mesures les plus utilisées dans le développement des systèmes de traduction automatique statistique.

1.6.3 Utilité des mesures automatiques et humaines

Nous avons choisi dans notre travail le score BLEU comme mesure automatique pour nos systèmes de TA puisqu'il est toujours une mesure officielle dans la plupart des campagnes d'évaluation et qu'il est également le plus utilisée dans des différents travaux de recherche en traduction automatique. Mais BLEU, comme plusieurs mesures automatiques est de plus en plus critiqué pour sa faible corrélation avec les jugements humains dans certains cas, car il est basé sur la précision.

Donc, les mesures automatiques s'avèrent très utiles au cours du développement d'un système de TA pour comparer rapidement deux versions successives de ce système sans intervention humaine. Des exemples d'évaluation automatique des systèmes de traduction automatique avec les scores TER et BLEU sont présentés dans les chapitres 3 et 4 et 6. Nous avons également effectué des évaluations manuelles sur des exemples aléatoires de corpus de test, en plus de l'évaluation automatique des systèmes de traduction utilisés dans nos expériences.

1.7 Conclusion

La traduction automatique a opéré un virage important par le développement de modèles statistiques. En utilisant ces modèles, il est possible, à partir de corpus parallèles annotés, d'apprendre automatiquement les associations les plus statistiquement significatives entre deux langues, donnant lieu à des systèmes de traduction statistique.

Nous avons présenté dans ce chapitre un survol sur la traduction automatique statistique (TAS) ainsi que les enlèves des différentes théories et les composantes d'un système de TAS. Ce que présente le cadre général de nos travaux de recherche durant cette thèse. Dans le chapitre suivant, qui s'intéresse aux corpus comparables

multilingues, nous présenterons les travaux précédents d'extraction de textes parallèles qui constitue l'objectif de nos recherches.

Chapitre 2

Corpus comparables multilingues

Sommaire

Dans ce chapitre, nous présenterons les travaux existants qui sont pertinents pour cette thèse. Nous décrirons, tout d'abord, les types de corpus multilingues abordés dans nos travaux et introduirons leurs différents propriétés et usages. Puis, nous présenterons les différentes approches d'exploitation des corpus comparables existant dans la littérature.

2.1 Corpus multilingues

Les documents textes et audio sont de plus en plus nombreux et volumineux, notamment en raison de l'augmentation des publications sous forme électronique ou audiovisuelle, rendant leurs traitements et leur stockage plus aisé. Selon « Le Trésor de la Langue française [1] », un corpus est le recueil réunissant, en vue de leur étude scientifique, des documents disponibles d'un genre donné. Plus précisément en traitement automatique des langues naturelles, Sinclair définit un corpus comme :

« *une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques et extralinguistiques explicites pour servir d'échantillon d'emplois déterminés d'une langue.* »[Sinclair, 1996] [2].

1. Le Trésor de la langue française informatisé : http ://atilf.atilf.fr/
2. « *A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language* »

Les corpus multilingues sont des corpus composés de documents dans des langues différentes. Ces corpus sont utilisés pour mettre en évidence des applications langagières. L'une des principales utilisations est la traduction automatique. Nous distinguons et décrivons dans ce domaine deux types de corpus multilingues : les corpus parallèles et les corpus comparables.

2.1.1 Corpus parallèles

Un corpus parallèle est un ensemble de paires de documents bilingues qui sont la traduction l'un de l'autre. Généralement, les relations de traduction (alignement) sont au niveau de la ligne. [Harris, 1988] a inventé le terme « bitexte » pour identifier ce genre de corpus. Mais historiquement, les premières utilisations des corpus parallèles dans l'histoire étaient depuis 1822, lorsque Champollion a pu décrypter les écritures des hiéroglyphes à l'aide de la Pierre de Rosette qui était un exemple canonique de corpus parallèle.

FIGURE 2.1 – Pierre de Rosette.

Dans notre cadre de travail, nous définissons un « bitexte » par une collection de textes bilingues alignés au niveau de la phrase, c'est-à-dire, des textes en langue source avec leurs traductions.

Ce type de corpus est précieux pour la traduction automatique statistique, et aussi pour d'autres tâches, comme la recherche d'information cross-lingue [Davis et Dunning, 1995], la projection d'annotation [Diab et Resnik, 2002] ou encore, l'acquisition lexicale automatique [Gale et Church, 1993].

Les textes parallèles librement disponibles sont des ressources rares : la taille est souvent limitée, la couverture linguistique insuffisante ou le domaine des textes n'est pas approprié. Il y a relativement peu de paires de langues pour lesquelles des corpus parallèles de taille raisonnable sont disponibles. Nous pouvons citer, entre autres, l'anglais, le français, l'espagnol, l'arabe, le chinois et quelques langues européennes [Hewavitharana et Vogel, 2011]. De plus, ces corpus disponibles proviennent principalement de sources gouvernementales, comme le parlement canadien *Hansard* ou Européen *Europarl*, ou de l'Organisation des Nations Unies (UN). Ceci est problématique en TAS parce que les systèmes de traduction appris sur des données provenant, par exemple, d'un domaine politique ne donnent pas de bons résultats lorsqu'ils sont utilisés pour traduire des articles scientifiques.

Face aux insuffisances des corpus parallèles, les recherches se sont tournées vers d'autres ressources et méthodes comme l'exploitation des corpus comparables.

2.1.2 Corpus comparables

Depuis quelques années, les corpus comparables font l'objet d'une attention particulière [Fung et Yee, 1998, Koehn et Knight, 2000, Vogel, 2003, Gaussier *et al.*, 2004, Li et Gaussier, 2010]. Ils sont constitués de documents dans des langues différentes n'étant pas en relation de traduction, contrairement au corpus parallèle qui est clairement défini par des textes et leurs traductions. L'acquisition de ce type des corpus comparable est moins difficile que celle des corpus parallèles. Puisque, dans de nombreuses langues et pour plusieurs thématiques, les corpus comparables sont généralement plus disponibles et plus faciles à constituer. [Déjean et Gaussier, 2002] les définissent de la manière suivante :

« *Deux corpus de deux langues L1 et L2 sont dits comparables s'il existe une sous-partie non négligeable du vocabulaire du corpus de langue L1, respectivement L2, dont la traduction se trouve dans le corpus de langue L2, respectivement L1.* »

Nous pouvons, par exemple, citer les actualités multilingues comme celles de la figure 2.2 produites par des organismes de presse telles que l'Agence France Presse (AFP), Xinhua, l'agence Reuters, CNN, BBC, etc. Ces textes sont largement disponibles sur le Web pour de nombreuses paires de langues [Resnik et Smith, 2003].

2.1.3 Comparabilité

D'après ces définitions des corpus comparables et parallèles, nous pouvons en conclure que la notion de comparabilité est relative à la quantité d'informations partagées par les documents du corpus, comme illustré dans la figure 2.3. En effet, il en ressort de cette dernière que la comparabilité varie entre les corpus très faiblement reliés, connus sous le nom de corpus indépendants et corpus parallèles.

FIGURE 2.2 – Exemple de ressource de données comparables, extrait de [Resnik et Smith, 2003].

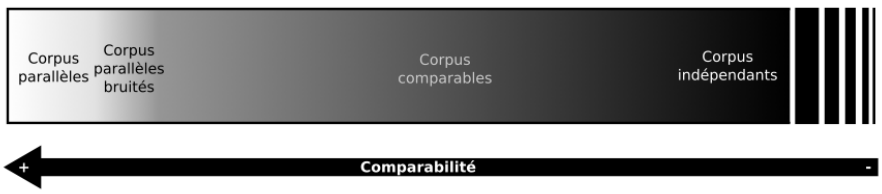

FIGURE 2.3 – Comparabilité des corpus multilingues, extrait de [Prochasson, 2009].

Nous présenterons dans la section 2.2 différents usages des corpus comparables liés à notre contexte de travail.

2.1.4 Corpus multimodal

Le terme « multimodalité » peut être utilisé de plusieurs façons, mais la définition que nous allons adopter est que « l'information multimodale » se rapporte plus à une « multimodalité de production » des documents qui contiennent les informations : texte, audio, vidéo ou image. La modalité dans ce contexte est liée au type des données.

Définition. Nous pouvons définir alors un *corpus multimodal* par :
« le recueil informatisé des documents contenant de plus qu'une seule modalité pour servir le traitement des langues. »

Des exemples de corpus multimodaux pourraient donc être une collection numérisée de textes illustrés avec des photos et/ou schémas, ou une collection numérisée de films avec des transcriptions de conversations dans les films. Nous nous sommes intéressés dans nos travaux aux modalités texte et audio, c.-à-d., nous considérons les cas où le corpus est constitué par une partie en texte et une partie en audio. Dans ce contexte, nous définissons les différents types de corpus multimodaux comme suit.

Définition. Un *corpus multilingue multimodal* est une collection de corpus composés de différentes modalités dans deux langues ou plus.

Nous considérons dans ce cas que chaque partie du corpus est dans une modalité et langue différente que l'autre.

Définition. Un *corpus multimodal comparable* est une collection de données de différentes modalités qui contiennent les mêmes informations, mais ne présentent pas des traductions les uns des autres.

Nous citons les exemples des films et leurs sous-titrages ou les audio et leurs transcriptions manuelles ou automatiques. Dans les travaux de cette thèse, nous nous sommes intéressés par les modalités texte et audio. Un corpus multimodal dans notre contexte est composé par une partie audio dans une langue et une partie texte dans une autre langue.

2.2 Exploitation des corpus comparables multilingues

L'exploitation des corpus comparables a marqué un tournant en traduction automatique dans la tâche d'extraction des données parallèles. Nous présentons dans ce qui suit de cette section les principaux travaux dans l'extraction des phrases et segments parallèles.

2.2.1 Extraction de textes et phrases parallèles

De nombreux travaux ont été réalisés sur l'extraction des phrases parallèles à partir d'un corpus comparable bilingue. Un critère de maximum de vraisemblance a été proposé par [Zhao et Vogel, 2002], combinant des modèles de longueur de phrases avec un lexique extrait d'un corpus parallèle aligné existant. Comme présenté dans la figure 2.4, le lexique est itérativement adapté au travers d'un processus de réapprentissage en utilisant les données extraites. Les données utilisées dans ces expériences ont été collectées du site de l'organisme de presse Xinhua. Les auteurs ont obtenu des améliorations significatives dans les modèles de leurs systèmes de traduction automatique.

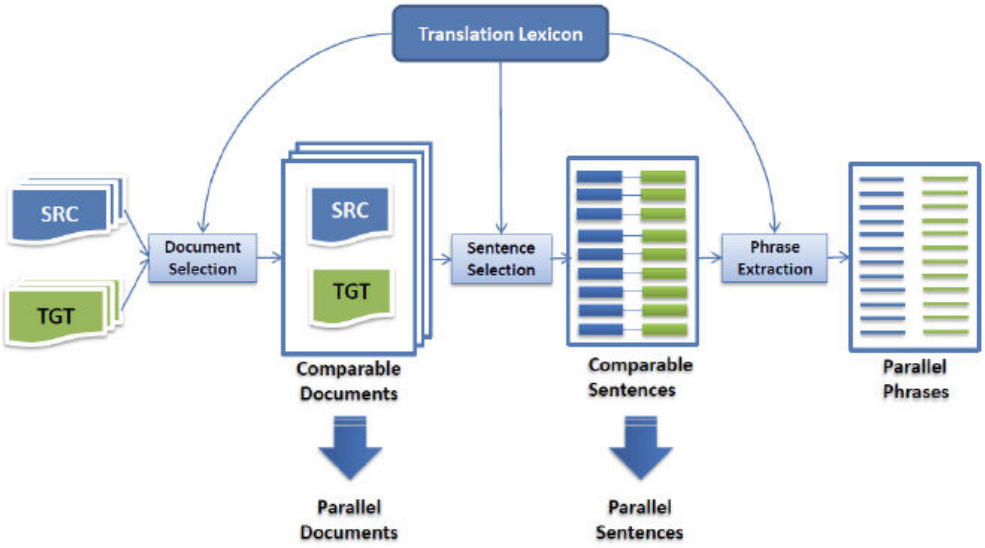

FIGURE 2.4 – Architecture du système d'extraction proposé par [Zhao et Vogel, 2002].

[Yang et Li, 2003] utilisent une approche basée sur la programmation dynamique pour identifier le parallélisme des pages web anglais/chinois au niveau des titres, mots et caractères. L'utilisation de la plus longue séquence commune (LCS), ainsi qu'un score de confiance est appliqué pour identifier la bonne traduction chinoise d'une séquence de mots anglais. Une fonction est utilisée pour déterminer les paires de titres optimales. Cette méthode est testée sur les articles collectés des sites Web de *Hong Kong SAR government* et *Hongkong & Shanghai Banking Corporation Limited*.

[Resnik et Smith, 2003] ont montré qu'il est possible de générer un grand nombre de documents parallèles à partir du WEB en utilisant leur système d'extraction de

textes parallèles, « STRAND »[3]. Leur méthode se résume en trois étapes :

- localisation des pages qui pourraient avoir des traductions ;
- génération des paires de documents qui pourraient être des traductions les uns des autres ;
- filtrage des paires de documents résultants.

L'évaluation de cette méthode est réalisée à l'aide de calcul des mesures de rappel et précision, ainsi que le jugement humain. Dans le cas de l'évaluation humaine, une k mesure[4] (mesure de Cohen) [Cohen, 1960] est calculée suivant les réponses à la question « Est-ce que ces deux phrases ont le même sens dans les deux langues ? ». C'est pourquoi ils ont eu de bons résultats en terme de score calculé, car la vérification n'était pas sur la qualité de traduction, mais sur la possibilité d'avoir une traduction dans les paires de phrases sélectionnées. Ainsi, les données extraites n'étaient pas utilisées dans des systèmes de traduction pour tester leurs impacts.

Afin de construire un corpus parallèle anglais/Japonais, [Utiyama et Isahara, 2003] utilisent la recherche d'information cross-lingue et la programmation dynamique pour l'extraction de phrases parallèles à partir d'un corpus comparable du domaine des actualités. Les paires d'articles similaires sont identifiées et traitées comme des textes parallèles afin d'aligner leurs phrases. La procédure d'alignement, utilisant un dictionnaire bilingue, commence par la traduction mot à mot des textes japonais des textes japonais pris comme requêtes de recherche d'informations dans la partie des textes rédigée en anglais.

L'approche de [Fung et Cheung, 2004] utilise la mesure « cosinus » pour calculer le degré de similarité des phrases. Toutes les paires de phrases possibles d'un corpus « non-parallèle » ont été considérées, et celles ayant un niveau de similarité supérieur à un certain seuil sont utilisées pour construire un dictionnaire qui sera réappris itérativement.

[Munteanu et Marcu, 2005] ont présenté l'une des méthodes d'extraction des phrases parallèles. Comme nous pouvons le voir dans la figure 2.5 un dictionnaire bilingue existant est utilisé pour traduire chaque document en langue source vers la langue cible afin d'extraire le document cible qui correspond à cette traduction. Pour chaque paire de documents, des paires de phrases et de segments parallèles sont extraites en utilisant un lexique de traduction, et un classifieur pour le choix final des phrases parallèles.

[AbduI-Rauf et Schwenk, 2009, Abdul-Rauf et Schwenk, 2011] présentent une nouvelle technique basée sur la traduction automatique.

3. Structural Translation Recognition, Acquiring Natural Data
4. K (kappa) mesure l'accord entre observateurs lors d'un codage qualitatif en catégories.

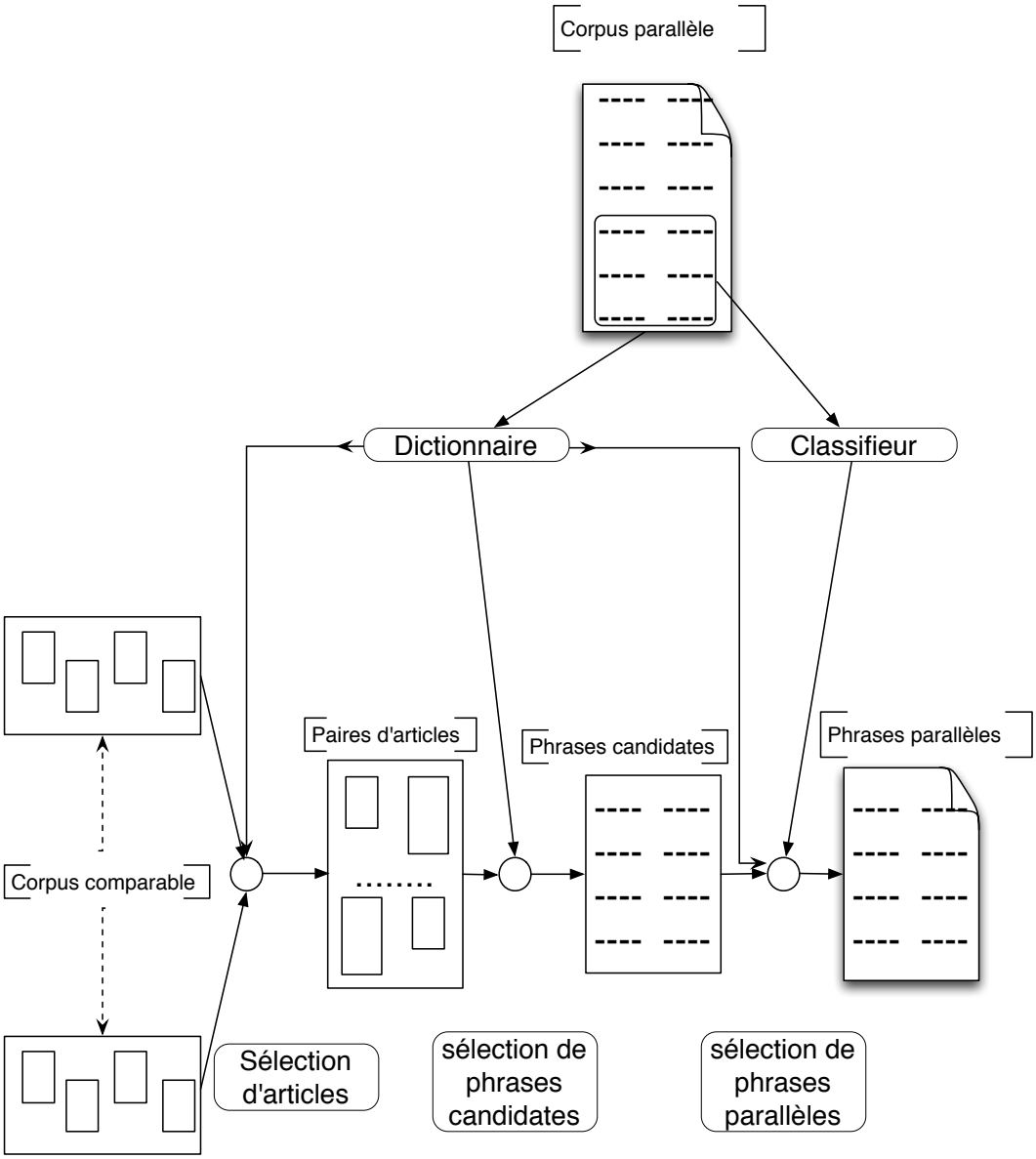

FIGURE 2.5 – Architecture du système d'extraction proposé par [Munteanu et Marcu, 2005].

L'idée principale de cette approche réside dans l'utilisation d'un système de TAS pour traduire toutes les phrases en langue source du corpus comparable. Chacune de ces traductions est ensuite utilisée en tant que requête afin de trouver des phrases potentiellement parallèles. Les phrases obtenues sont comparées aux traductions automatiques afin de déterminer si elles sont effectivement parallèles à la phrase correspondante en langue source. Les corpus comparables utilisés dans ces travaux se situent dans le domaine des actualités, plus précisément, des dépêches d'actualités des agences de presse telles que «Agence France Press (AFP)», «Associate press» ou «Xinua News». Les textes parallèles extraits ont permis d'améliorer significativement les systèmes de traduction statistique.

Les différences majeures avec celle de [Munteanu et Marcu, 2005], comme nous pouvons les distinguer dans la figure 2.6 résident dans l'utilisation d'un système de TA statistique à la place du dictionnaire bilingue et dans l'utilisation de métriques d'évaluation, comme le taux d'erreur mot (WER) ou le taux d'édition de la traduction (TER),afin d'évaluer le degré de parallélisme des phrases extraites.

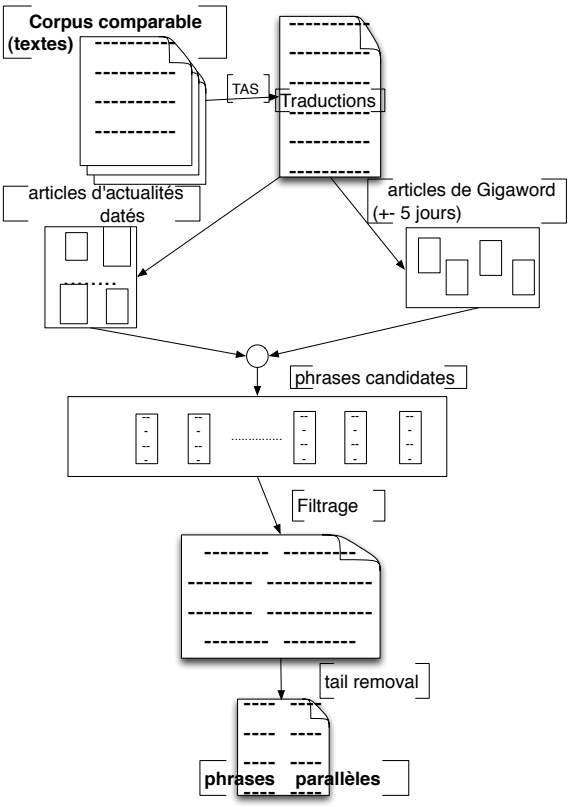

FIGURE 2.6 – Architecture du système d'extraction proposé par [Abdul-Rauf et Schwenk, 2011].

[Do *et al.*, 2010] ont utilisé une méthode non-supervisée pour extraire des paires de phrases parallèles à partir d'un corpus comparable et ont montré que cette

approche est intéressante surtout pour les langues peu dotées. La détection des paires de phrases parallèles dans le corpus comparable est réalisée via un système de traduction automatique de base amélioré à l'aide d'un processus itératif.

[Gahbiche-Braham *et al.*, 2011] proposent d'adapter un système de traduction en utilisant un corpus parallèle extrait d'un corpus comparable. Ce dernier est collecté à partir des dépêches arabes et françaises produites par l'*AFP*, contenant environ un million de phrases en arabe (150 millions de mots) et 5M de phrases en français. Le système de base est appris sur 7.7 millions de phrases parallèles. L'extraction des données parallèles se fait en trois étapes :

- traduction : traduire le côté source (arabe) en utilisant un système de base hors domaine.
- sélection des documents : sélectionner les paires de documents parallèles comme dans l'exemple 2.7 en calculant la similarité avec les documents traduits. Pour faciliter la tâche, seuls les documents du même jour sont comparés.
- sélection des phrases parallèles : réalisé à partir des documents parallèles en utilisant l'outil *hunalign*[5] qui calcule un score de similarité entre chaque paire de phrases.

Arabic:
واضاف نحن في حماس لا مانع لدينا من استئناف المفاوضات غير المباشرة حول الصفقة من النقطة التي انتهت اليها والتي حاول ان يفشلها نتانياهو.
And he added, we in Hamas don't have a problem to resume indirect negotiations about the deal from the point at which it ended and at which Netanyahu tried to fail.

French: Le porte-parole a réaffirmé que le Hamas était prêt à reprendre les tractations au point où elles s'étaient arrêtées.
The spokesman reaffirmed that Hamas was ready to resume negotiations at the point where they stopped.

FIGURE 2.7 – Exemple de documents similaires arabe/français, extrait de [Gahbiche-Braham *et al.*, 2011].

L'évaluation de la méthode est faite, comme la méthode de [AbduI-Rauf et Schwenk, 2009], en fonction du score BLEU après l'injection des données extraites dans le système de base de traduction appris avec des données

5. https ://github.com/herrnici/HunAlign

hors domaine.

2.2.2 Extraction des segments parallèles

L'identification des segments sous phrastiques parallèles à partir des corpus comparables est un problème qui a fait l'objet de plusieurs travaux de recherche ces dernières années [Su et Babych, 2012, Hewavitharana et Vogel, 2011, Riesa et Marcu, 2012]. Nous pouvons grossièrement classer ces méthodes en deux catégories.

La première, l'approche associative, qui repose sur des mesures comme l'information mutuelle ou le rapport de vraisemblance. [Munteanu et Marcu, 2006] ont suivi cette méthode en utilisant un lexique qui s'améliore avec un réapprentissage après chaque passe d'extraction. Ils ont obtenu de bons résultats dans l'amélioration d'un système de TA avec des données de segments parallèles extraits à partir des phrases comparables. Néanmoins, leur méthode reste imparfaite du fait que les fragments sources et cibles sont filtrés séparément ce qui ne garantit pas l'obtention de bonnes traductions dans tous les cas. Suivant la même approche, [Hewavitharana et Vogel, 2011] ont obtenu de bons résultats par leur méthode basée sur le calcul des corrélations entre les paires de segments.

La seconde approche, se base sur l'alignement. Le but est de déterminer le meilleur ensemble de liens d'alignement entre des groupes de mots sources et cibles de chaque paire de phrases ou documents. Cette méthode a été introduite par [Brown *et al.*, 1990] sur les corpus parallèles. Des travaux récents ont essayé de l'appliquer sur les données comparables [Quirk *et al.*, 2007, Riesa et Marcu, 2012]. Ces travaux sont prometteurs, mais il est difficile jusqu'à maintenant de dire s'ils sont efficaces surtout concernant l'amélioration des systèmes de TA avec les données extraites.

2.2.3 Vers l'exploitation des corpus multimodaux

Toutes ces méthodes précédemment décrites sont présentées comme des techniques efficaces pour extraire des données parallèles à partir d'un corpus comparable de modalité texte. Au cours de ces dernières années, des travaux ont été réalisés en vue d'initier des essais d'exploitation de la modalité audio et extraire des données parallèles.

[Paulik et Waibel, 2009, Paulik et Waibel, 2013] ont montré que les modèles de traductions statistiques peuvent être appris automatiquement d'une manière non-supervisée à partir des données parallèles audio. Ils ont proposé un système d'extraction (présenté dans la figure 2.8) des données d'apprentissage du système de traduction automatique à partir de l'audio parallèle.

Leurs expériences étaient basées sur les enregistrements audio des sessions du parlement européen et leurs traductions directes par les différents interprètes.Ils

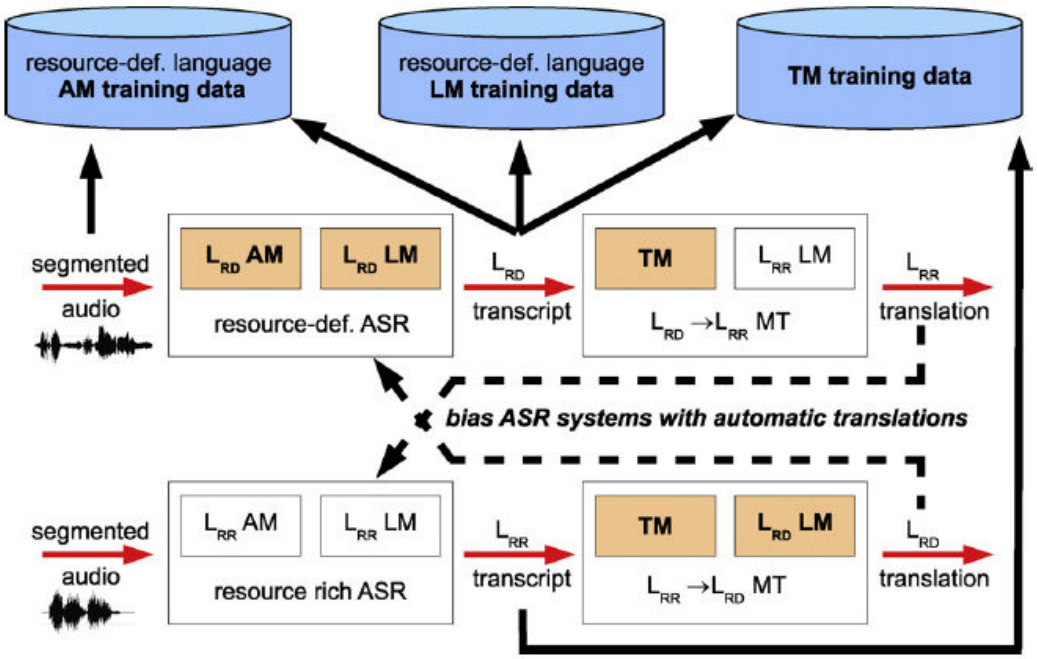

FIGURE 2.8 – Architecture du système d'extraction proposé par [Paulik et Waibel, 2013].

ont aussi appliqué leurs méthodes pour le couple de langue anglais/pashto [6] où les ressources de textes parallèles sont très rares. Les résultats présentés ont montré que les données extraites à partir des ressources audio parallèles peuvent :

- améliorer un système de TA existant en ajoutant du vocabulaire ;
- apprendre un système de TA pour des paires de langues dites *peu doté*.

L'idée d'utilisation des ressources audio parallèles pour apprendre des systèmes de traduction automatique a été proposée aussi par [Besacier *et al.*, 2006] en utilisant des séquences téléphoniques en *Iraqi* [7] et leurs traductions orales en anglais. Cette méthode appelée *Phone-Based* a permis de fournir des ressources pour construire des systèmes de TA d'une langue non écrite. Les résultats en terme de score BLEU étaient comparables au système appris avec des données textes parallèles construites manuellement.

[Sarikaya *et al.*, 2009] ont exploité les transcriptions des films sous-titrés et les émissions de TV traduites anglaises/espagnoles qui présentent un corpus bilingue bruité. Leurs méthodes semi-supervisées consistent à utiliser un système de traduction appris sur des données parallèles de base et faire des itérations lors desquelles

6. La langue parlée en Afghanistan.
7. Un des dialectes de la langue arabe parlée en Iraq.

les données extraites sont ajoutées au corpus initial parallèle pour reconstruire le système de traduction. Cette idée a déjà été proposé par [Caroline *et al.*, 2007] afin de produire un corpus parallèle anglais/français à partir de 40 filmes sous-titrées.

Selon l'ensemble des connaissances dans la revue de littérature, l'exploitation des ressources comparables de modalité différentes pour la traduction automatique n'a pas encore été abordée à ce jour. C'est la raison pour laquelle nous avons décidé de proposer d'une part, dans le chapitre 4, notre méthode permettant l'exploitation de ce type de données pour l'extraction des données parallèles, et d'autre part, nos expériences du test de faisabilité et d'amélioration de l'approche.

2.3 Conclusion

Dans ce chapitre, nous nous sommes intéressés aux principaux travaux d'exploitation des corpus comparables existant dans la littérature. Après une définition du cadre linguistique des différents corpus multilingues, nous avons proposé un tour d'horizon de l'exploitation des corpus comparables de modalité texte passant par l'extraction des phrases et segments parallèles se terminant par les récents travaux d'utilisations des corpus multimodaux. Ces bases définissent le contexte théorique de cette thèse. Dans le chapitre suivant, nous décrirons le contexte des corpus et outils existant au sein du LIUM.

Deuxième partie

Contexte de travail et proposition du problème

Chapitre 3

Ressources et systèmes utilisés

Sommaire

L'orientation de nos recherches a été inspirée par les ressources et technologies disponibles au sein du LIUM : un système de reconnaissance automatique de la parole, un système de traduction automatique, et les différents corpus parallèles et comparables dans les modalités texte et audio. Nous avons aussi développé d'autres ressources durant cette thèse, afin de tester ou valider nos approches proposées. Dans ce chapitre, nous allons présenter ces ressources avant de passer à la description des axes de recherches suivies.

3.1 Ressources linguistiques

Les ressources linguistiques sont les données relatives aux langues, accessibles dans un format électronique, et utilisées pour le développement de systèmes de traitement automatique des langues comme les systèmes de TA.

Nous présenterons les ressources de corpus parallèles et comparables utilisées dans cette thèse.

3.1.1 Corpus parallèles

L'apprentissage de notre système de TA statistique de base demande un corpus parallèle de taille suffisante pour construire le modèle de traduction, et un corpus monolingue pour le modèle de langage. Afin de construire nos systèmes de TAS de base, nous avons utilisé dans nos expériences les corpus parallèles *Europarl* et *News Commentary*.

Europarl

Dans la majorité de nos expériences, nous avons utilisé le corpus Europarl [1] [Koehn, 2005], construit à partir des procès-verbaux des réunions du parlement européen, traduit par des professionnels et disponible en 21 langues. Nous avons utilisé dans nos expériences la partie anglais-français de ce corpus dans sa version 7. Nous appellerons par la suite ce corpus *Eparl7*.

News Commentary

Nous avons également utilisé pour certaines de nos expériences, le corpus de News Commentary dans sa version 7 fourni au cours de la campagne d'évaluation WMT-12 [2]. Ce corpus parallèle a été créé dans le cadre du projet européen *EuroMatrix* à partir de différents éditoriaux en ligne appelé par la suite ce corpus *NC7*.

3.1.2 Corpus comparables multimodaux

De nos jours, les données comparables de niveau de comparabilité variable existent en quantité très importante. Mais dans une étude préliminaire sur les ressources linguistiques existantes, nous avons constaté un manque de corpus multimodaux librement disponibles dédiés à la traduction automatique ou autres applications du traitement automatique des langues. D'après l'ensemble de ces observations, nous avons décidé d'utiliser les données *TED* et de construire le corpus *Euronews*.

FIGURE 3.1 – Exemple de ressources de données comparables multimodales du site TED.

Corpus TED-LIUM

Nous avons exploité les données de la campagne d'évaluation *IWSLT'11*[3] au sein de laquelle des données bilingues multimodales sont disponibles avec le corpus *TED-LIUM*. Cette tâche, détaillée dans [Rousseau *et al.*, 2011], consiste à traduire des discours de *TED*[4] de l'anglais vers le français. Comme présenté dans la figure 3.1, ces données sont basées sur des présentations scientifiques orales en anglais disponibles avec leurs traductions en plusieurs langues dont le français. Le tableau 3.1 décrit l'ensemble des caractéristiques de départ des données du corpus TED.

Nous avons considéré la partie audio en anglais et les traductions textes correspondantes en français comme un corpus multimodal. Nous avons enrichi ce corpus avec des textes parallèles extraits de ces présentations scientifiques fournis par IWSLT'11 appelé par la suite *TEDbi*.

Dans les travaux de cette thèse, ce corpus est utilisé pour la tâche d'adaptation d'un système SMT à un domaine différent de celui des données d'entrainement.

1. http ://www.statmt.org/europarl/
2. http ://www.statmt.org/wmt12/
3. http://www.iwslt2011.org/
4. http://www.ted.com/

Nombre de shows	779
Avec locuteur masculin	526
Avec locuteur féminin	253
Durée totale de l'audio	205h, 49m
Moyenne par show	15m
Durée totale de parole	117h, 45m
Dont locuteur masculin	82h, 26m
Dont locuteur féminin	35h, 19m
Moyenne par show	9m
Nombre de segments de parole	56 803
Dont locuteur masculin	39 389
Dont locuteur féminin	17 414
Durée moyenne d'un segment	7,46 secondes
Nombre de locuteurs uniques	666
Masculins	452
Féminins	214
Shows par locuteur unique	1,16
Nombre de mots dans les transcriptions	1 690 775
Nombre de mots moyen par show	2 184

TABLE 3.1 – Caractéristiques du corpus TED-LIUM.

Corpus Euronews-LIUM

Afin de disposer d'une quantité exploitable de données multimodale d'un autre domaine, nous avons exploité le site web *Euronews*[5] en deux langues : français et anglais. Nous avons également développé un outil écrit en *Perl* nous permettant de télécharger de façon automatique tous les articles datant de la période comprise entre 2010-2012. La plupart bénéficiaient à la fois d'une vidéo en anglais (ou en français) mais aussi d'une version texte de l'article comme présenté dans la figure 3.2. La version texte n'est pas la transcription exacte de la vidéo, et les articles qui parlent du même sujet en français et en anglais ne sont pas la traduction exacte.

L'outil fonctionne en deux phases. La première ayant pour but de connaître la liste des articles en anglais à extraire par jour et à chercher les correspondances en français, la deuxième servant effectivement à récupérer les données.

La première phase a lieu comme suit :

1. une liste complète des articles est téléchargée quotidiennement dans les deux langues ;

5. http://www.euronews.com/

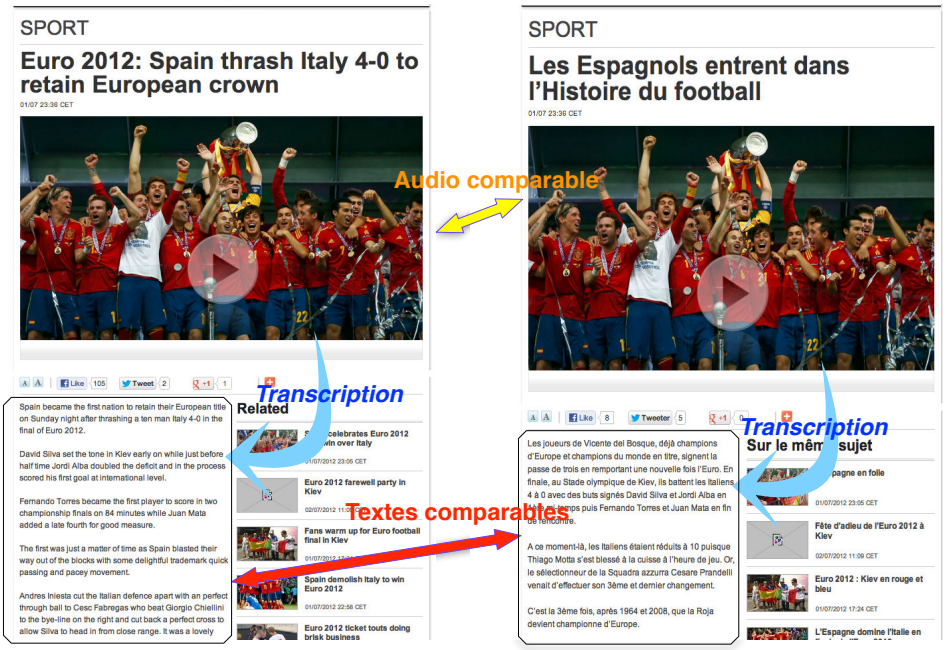

FIGURE 3.2 – Exemple de ressource de données comparables multimodales dans le domaine du sport à partir du site Euronews.

2. puis ces listes sont comparées afin d'en conserver ; l'intersection (*i.e.* la liste des articles qui ont des correspondances comparables en français),
3. enfin, le code HTML de chaque page retenue dans la liste sera extrait du site.

Pour la seconde phase, nous procédons de cette manière :

1. le nom et le domaine de la vidéo sont repérés à partir du code HTML des pages téléchargées,
2. puis, le lien des vidéos est extrait et récupéré localement dans le dossier portant le nom de chaque domaine,
3. ensuite, les textes des articles sont extraits dans les mêmes dossiers,
4. enfin, le flux audio de la vidéo est extrait puis converti dans le format utilisé pour les systèmes de reconnaissance de la parole (format *NIST Sphere*), grâce à une suite d'outils libres (respectivement *mplayer* et *sox*).

À la fin de la collecte des données, celles-ci sont traitées par un script spécifique, dont le but est de nettoyer les données textuelles téléchargées. Les vidéos téléchargées sont automatiquement transcrites en utilisant le système de reconnaissance automatique de la parole du LIUM. Nous avons ainsi automatiquement collecté environ 2,2 millions de mots transcrits et environ 6,2 millions de mots des textes liés aux transcriptions. L'ensemble des documents collectés relève de sept domaines

Domaine	nb de mots	nb de phrases
Economie	289 k	7 k
Sport	81 k	2 k
Culture	388 k	12 k
Euro	398 k	12 k
Style de vie	28 k	1 k
Politique	806 k	26 k
Science	231 k	9 k
Total	2.2 M	76 K

TABLE 3.2 – Nombre de mots et de phrases de la transcription automatique du corpus audio anglais Euronews.

Domaine	nb de mots Fr	nb de mots En
Économie	425 K	613k
Sport	112 k	102 k
Culture	262 k	274 k
Euro	302 k	287 k
Style de vie	18 k	19 k
Politique	4 M	4 M
Science	147 k	141 k
Total	6.2 M	6.1 M

TABLE 3.3 – Quantité en termes de mots de la partie texte anglais/français du corpus Euronews.

de dépêches. Les détails en terme de quantité de données texte et audio après la transcription sont présentés dans les tableaux 3.2 et 3.3.

Le corpus *Euronews-LIUM* présente, d'après notre connaissance, le premier corpus comparable multimodale construit dans le but de l'adaptation et l'amélioration d'un système de traduction automatique. Nous proposons dans [Afli *et al.*, 2014] de le distribuer librement afin qu'il soit utilisable par l'ensemble de la communauté scientifique.

Dans les travaux de cette thèse, ce corpus est utilisé pour l'amélioration des performances d'un système de TAS appris avec des données du même domaine. Les expériences sur *Euronews-LIUM* ont permis la validation de nos approches sur des données réelles.

3.2 Systèmes de traduction

Les modèles basés sur les segments ont donné lieu à des améliorations significatives des traductions. Nous avons testé une autre approche dite *hiérarchique* pour la comparaison avec notre approche existante au LIUM. Pour cela nous avons construit deux systèmes SMT basés sur ces deux approches lors de la participation à la compagne d'évaluation WMT'11.Cette participation nous a permis de décider de l'approche adéquate pour le reste de nos expériences.

3.2.1 Le système à base de segments (Phrase-based System)

Nous avons développé un système de TAS basé sur le décodeur libre *Moses* et l'approche à segments (*Phrase Based*). Dans cette section nous allons décrire l'architecture de ce système, et les différentes techniques développées.

L'architecture générale de notre système de TAS est présentée dans la figure 3.3. Nous avons trois étapes principales : apprentissage, modélisation (ressources) et décodage (traduction).

3.2.1.1 Apprentissage

Cette étape commence par l'alignement des textes parallèles. Chaque paire de phrases est alignée mot-à-mot à l'aide de l'outil *GIZA++*. Cet alignement est réalisé dans les deux sens de traduction, afin de pouvoir en extraire les paires de séquences nécessaires à l'estimation du modèle de traduction. Ensuite, la construction de la table de traduction se fait en extrayant les paires de séquences de mots. Un algorithme est utilisé pour l'extraction, où toutes les séquences en langue source d'une phrase donnée sont passées en revue pour déterminer la séquence minimale en langue cible correspondant à chacune d'entre elles. La détermination se fait en identifiant tous les points d'alignement de la séquence source puis en trouvant la séquence cible la plus courte qui inclut toutes les traductions des mots de la séquence source, tel qu'illustré dans le figure 3.4.

3.2.1.2 Ressources

Le résultat de ce processus d'apprentissage est une table de traduction contenant toutes les traductions connues (mots seuls ou segments de quelques mots). Elle contient aussi les scores exprimant la confiance du processus d'apprentissage dans la qualité de la traduction automatiquement extraite. Cette table est présentée dans la partie *ressource* de la figure 3.3 par le *modèle de traduction*. Nous avons aussi dans cette partie un autre modèle très important pour la construction du système SMT, qui est le *modèle de langage*.

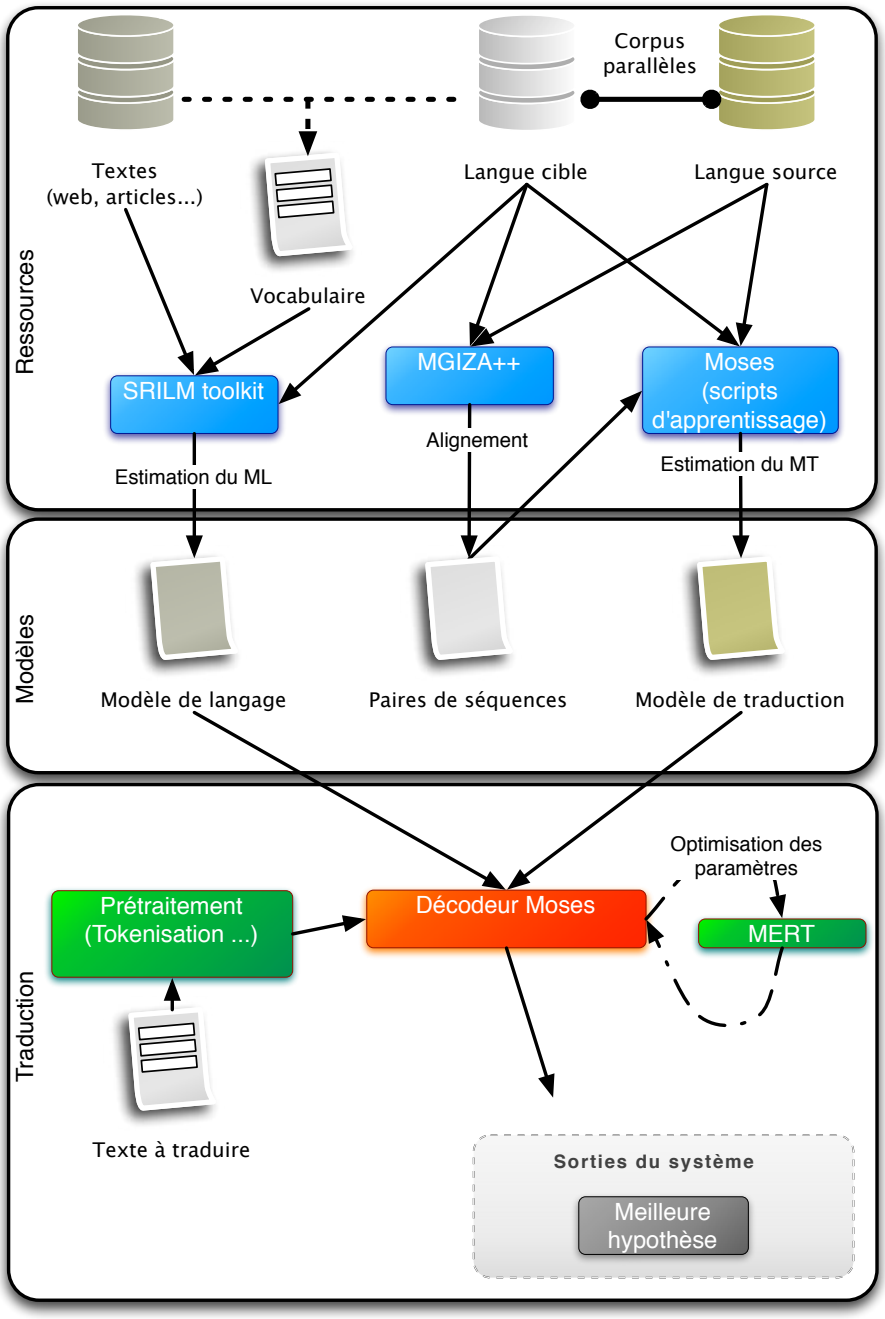

FIGURE 3.3 – Architecture générale du système de traduction automatique à base de segments.

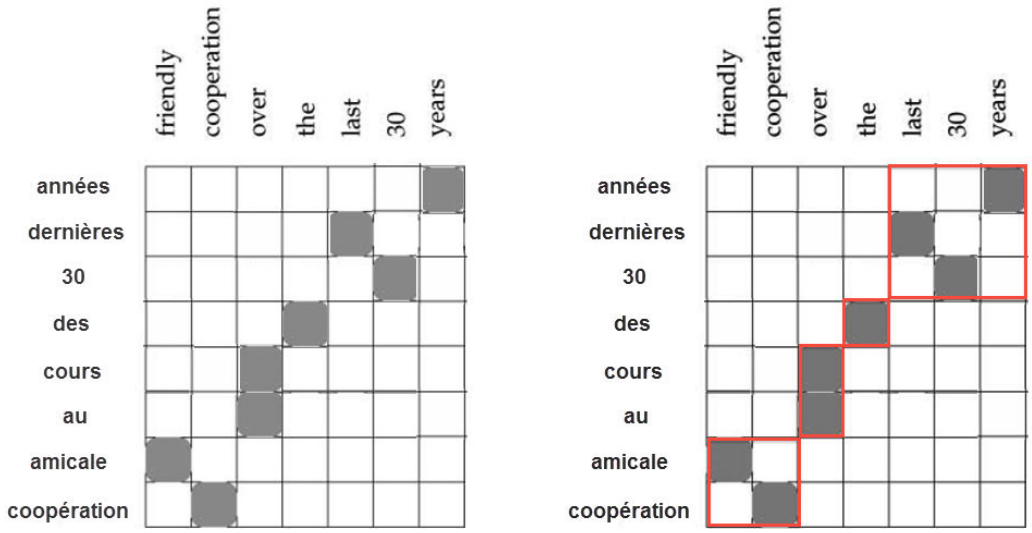

FIGURE 3.4 – Exemples d'alignements en mots (à gauche) et en séquences de mots (à droite).

L'apprentissage du modèle de langage nécessite des textes monolingues en langue cible. Pour le couple de langue anglais/français qu'on a traité dans nos expériences, le LIUM dispose de beaucoup de données monolingues pour ces deux langues, notamment une grande collection de textes journalistiques disponibles auprès du LDC [6]. Ces corpus sont connus sous le nom *Gigaword*. Le corpus *Gigaword* anglais contient plus de quatre milliards de mots et le corpus *Gigaword* français environ un milliard de mots. Les deux corpus couvrent une longue période (environ 1995 à 2012). En plus de ces données génériques, nous utilisons également le côté langue cible des textes parallèles qui sont de l'ordre de quelques dizaines de milliers de mots.

3.2.2 Le système hiérarchique (Hierarchical Phrase-based System)

Les modèles statistiques à base des séquences ont permis de passer du mot comme unité de traduction aux séquences de mots, mais pas nécessairement toute une expression complète. L'idée de l'approche hiérarchique a permis d'améliorer les systèmes SMT pour certaines langues morphologiquement riches comme le français et l'arabe [Besacier *et al.*, 2010]. Nous avons décidé de tester cette approche pour le couple de langues anglais-français.

3.2.2.1 Apprentissage et ressources

Notre système de TAS hiérarchique utilise la branche *moses_chart* du logiciel libre Moses [Koehn *et al.*, 2007], qui est développé en suivant le même principe du

6. https ://www.ldc.upenn.edu/

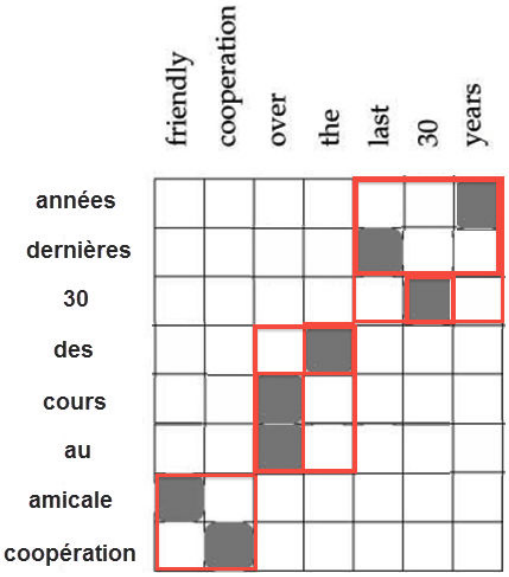

FIGURE 3.5 – Exemple d'alignement de sous séquences de mots.

système HIERO [Chiang, 2007].

Comme les systèmes habituels basés sur les segments, les systèmes hiérarchiques sont construits de la façon suivante : d'abord le logiciel *GIZA++* est utilisé afin d'obtenir les alignements mot à mot dans les deux directions. Ensuite les groupes de mots sont extraits, avec les valeurs de défaut de l'outil Moses.

Nous utilisons les mêmes modèles de langages qui sont construits pour les systèmes à base de segments. Les groupes de mots et les règles sont extraits, avec les valeurs de défaut de l'outil Moses tel que présenté dans la section suivante.

3.2.2.2 Extraction des règles

La table des règles qui remplace la table de traduction du système à base de segments se compose de règles extraites automatiquement en exécutant *GIZA++* sur le corpus pour construire un alignement en mots dans les deux sens (figure 3.4). Nous extrayons alors à partir de chaque paire de mots alignés un ensemble de règles qui sont compatibles aux alignements de mots, et pouvant être fait en deux étapes : d'abord, nous extrayons des paires de séquences selon la même méthode que pour les systèmes à base de segments, à savoir, qu'aucun mot à l'intérieur d'une expression ne peut être aligné qu'à un mot en dehors de l'autre expression parallèle (figure **??**). Puis, afin d'obtenir des règles sur ces expressions, nous recherchons celles qui contiennent d'autres expressions (figure 3.5) et remplaçons ces derniers par des symboles non terminaux (figure 3.6).

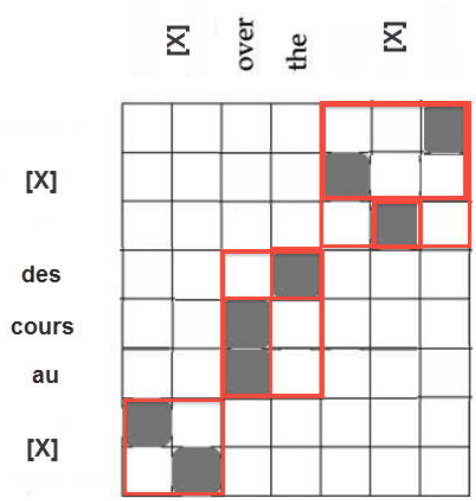

FIGURE 3.6 – Exemple d'extraction d'une règle.

3.2.3 Traduction

Cette partie est commune aux deux approches utilisant les mêmes modules, à savoir, la tokénisation et le décodage les données à traduire.

Tokénisation

La théorie de la traduction statistique considère que la phrase à traduire est constituée de plusieurs unités élémentaires, communément appelées des mots. Cependant, il convient plutôt d'utiliser le terme *token* puisque ces unités ne ressemblent pas nécessairement aux mots dans le sens linguistique ou grand public du terme. Ce ne sont au final que des suites de caractères entourés par des espaces.

Le processus de segmentation d'une phrase en plusieurs *tokens* et appelée *tokénisation*. Les systèmes de traduction développés utilisent des outils de tokénisation fournis avec Moses pour l'anglais et le français afin de tokéniser les données d'apprentissage, de développement et de test.

Optimisation des systèmes

L'optimisation des modules de traduction est un point très important du modèle log-linéaire dans la construction du système de TAS. Ce réglage des fonctions caractéristiques est réalisé à l'aide l'un des algorithmes les plus utilisés, *MERT* [Och, 2003]. Ces optimisations sont faites pour chaque système sur un corpus de développement.

Certains résultats des systèmes construits via les méthodes précédemment mentionnées sont présentés dans le tableau 3.4.

Données utilisées

Pour l'apprentissage de nos modèles, nous avons utilisé les données distribuées par les organisateurs de WMT. Ces données sont réparties de la manière suivante :

- deux corpus bilingues sont utilisés pour l'apprentissage des paramètres des modèles de traduction et ils sont constitués par les dernières versions de :
  - News-Commentary (NC) ;
  - Europarl (Eparl) ;
- d'un corpus pour le développement utilisé en vue d'optimiser les paramètres de nos modèles (newstest2009)
- d'un corpus de test pour évaluer les performances du système (newstest2010).

Résultats

Bitext	Système à base de segments		Système hiérarchique	
	newstest2009	newstest2010	newstest2009	newstest2010
Eparl+NC	26.20	28.06	26.12	27.92

TABLE 3.4 – Résultats des systèmes de TA anglais/français en terme du score BLEU sur les données de développement (newstest2009) et les données de test (newstest2010) construit lors de la participation du LIUM à WMT11.

Pour évaluer nos systèmes, nous avons utilisé le score BLEU. Les performances du système anglais/français sont résumées dans le tableau 3.4. Nous remarquons qu'il n'y a pas de différences significatives entre les résultats des deux approches.

3.2.4 Système de base (*Baseline*)

Après avoir comparé les deux approches présentées lors de la compagne d'évaluation WMT'11,nous avons décidé d'utiliser un système de TA à base de segments pour le reste de nos expériences. Nous avons fait le choix de cette approche étant donné les bons résultats qu'elle a fournie pour le couple de langue anglais-français.Ainsi que les performances des deux approches (à base de segment et hiérarchique) sont très proches sur ce couple de langues selon nos expériences. Cela nous permet également de choisir l'approche la plus simple des deux en terme de temps de calcul et de complexité.

3.3 Système de reconnaissance automatique de la parole

Dans ce paragraphe, nous allons nous intéresser au système de reconnaissance automatique de la parole (RAP) du LIUM utilisé lors de nos expériences [Deléglise *et al.*, 2009]. Ce système utilise comme base le décodeur CMU *Sphinx* [Lee *et al.*, 1989], diffusé sous licence libre depuis 2001. Nous verrons que le LIUM

lui a apporté de nombreuses modifications afin de rendre le système plus performant. Deux versions du décodeur *CMU Sphinx* sont utilisées pour réaliser le système du LIUM :

- *Sphinx 3* : cette version a pour objectif de permettre la meilleure précision possible dans le processus de décodage. Il s'agit d'un décodeur qui se base sur les modèles de Markov caché. Il est entièrement codé en langage C [Ravishankar *et al.*, 2000].
- *Sphinx 4* : cette version est une réécriture complète du décodeur en Java décrit dans [Walker *et al.*, 2004] avec une nouvelle conception. *Sphinx 3* et *Sphinx 4* utilisent les mêmes modèles acoustiques et modèles de langage.

La figure 3.7 présente l'architecture générale du système de reconnaissance automatique de la parole du LIUM, reprise d'après [Estève, 2009]. Nous pouvons y voir l'apprentissage nécessaire à la création des modèles du système, ainsi que le processus de décodage (transcription) , qui seront développés un peu plus loin.

3.3.1 Apprentissage et ressources

Nous avons utilisé dans nos expériences le système de RAP anglais du LIUM, appris sur les données TED-LIUM. Les modèles acoustiques de ce système sont basés sur les modèles de Markov cachés emploient un ensemble de 39 phonèmes de l'anglais. Les modèles de langage utilisés dans ce système sont réalisés à l'aide de l'outil *SRILM*.

L'apprentissage des différents modèles (modèles acoustiques, dictionnaire phonétisé, modèles de langage) est indispensable pour obtenir de bonnes performances. Le réglage des poids et des modèles est réalisé à l'aide de l'optimiseur mathématique *CONDOR* [Vanden Berghen et Bersini, 2005].

3.3.2 Transcription

Le système de RAP du LIUM est un système multi-passes, qui consiste, en l'utilisation d'un algorithme de recherche manipulant les données de la passe précédente. Elles sont au nombre de cinq :

- la première passe consiste en l'utilisation d'un modèle de langage trigramme et des modèles acoustiques adaptés au genre du locuteur et aux conditions acoustiques ;
- dans la deuxième passe, les meilleures hypothèses produites dans la passe 1 sont utilisées pour le calcul de la matrice de transformation *CMLLR* ;
- la troisième passe permet de re-scorer les graphes produits lors de la passe précédentepar analyse en composantes principales des paramètres de sortie ;
- lors de la quatrième passe, les scores linguistiques obtenus des graphes de mots de la troisième passe sont recalculés à l'aide d'un modèle de langage quadrigramme ;

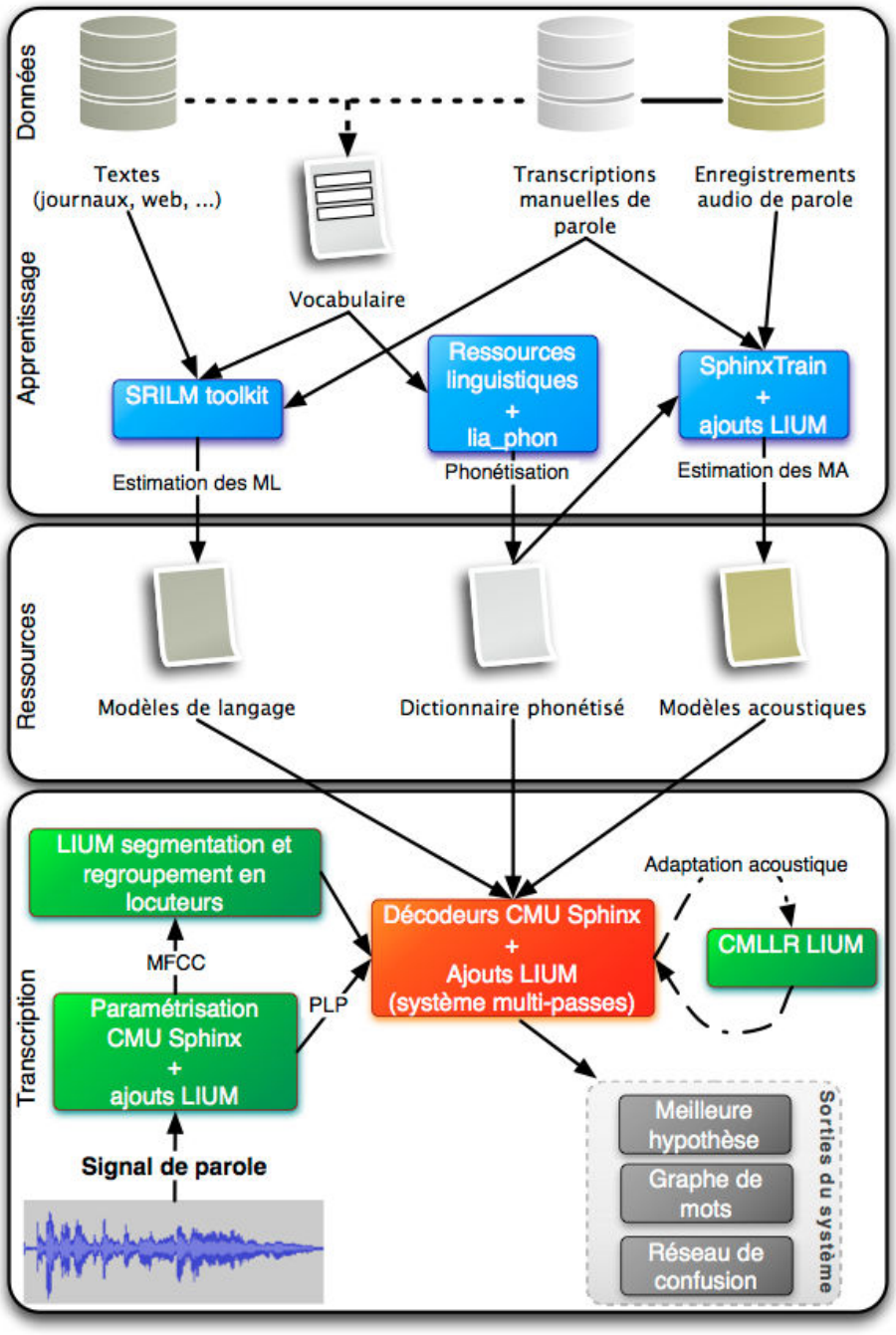

FIGURE 3.7 – Architecture générale du système de transcription automatique du LIUM, extrait de [Estève, 2009].

– finalement, la dernière passe transforme le graphe de mots issu de la passe précédente en un réseau de confusion qui permet l'obtention de l'hypothèse de reconnaissance finale.

3.3.3 Évaluation du système de RAP

Afin de pouvoir évaluer le système de RAP, la méthode la plus utilisée est le taux d'erreurs de mots (*Word Error Rate*). Le WER correspond à la distance d'éditions ou distance Levenshtein [Levenshtein, 1966] fondée sur la distance entre les mots de la sortie du système et les mots de la référence divisée par la longueur de la référence. Le score WER est utilisé dans de nombreux domaines en traitement automatique des langues. Mais en traduction, le score WER peut pénaliser injustement des traductions correctes si elles organisent les mots différemment des traductions de référence. Il est particulièrement adapté à la reconnaissance automatique de la parole du fait de la monotonicité entre le signal audio et la transcription. Cette métrique prend en compte les erreurs de :

– *Substitution* : mot reconnu à la place d'un mot de transcription manuelle.
– *Insersion* : mot inconnu inséré par rapport à la transcription de référence.
– *Suppression* : mot de référence oublié dans l'hypothèse fournie par le système de RAP.

Le WER s'écrit selon la formule mathématique suivante :

$$WER = \frac{nombre\ de\ substitutions + nombre\ d'insertion + nombre\ de\ suppressions}{nombre\ de\ mots\ de\ la\ reference} \tag{3.1}$$

3.4 Le système de recherche d'information *Lemur*

Lemur[7] est un système de recherche d'information (RI) libre, développé par *CIIR* à l'université du *Massachusetts* et le *LTI*. Il contient différentes fonctionnalités pour l'indexation et la recherche. Nous utilisons le moteur de recherche, , *Indri*[8], qui permet de mener des recherches précises sur une collection de textes indexées selon des différents paramètres. Cette indexation peut être vue comme un modèle bayésien représentant les variables sous forme d'un graphe orienté acyclique.

Nous avons utilisé dans nos expériences la technique de *Sac de mots* (Bag-of-words) comme dans l'exemple suivant :

$$\#combine(abraham\ lincoln\ gettysburg)$$

de manière à donner des poids équivalents pour tous les mots de la requête, après l'exclusion des mots fréquents (stop words). Les mots de la requête sont équipondérés et combinés avec un OU logique.

7. http ://www.lemurproject.org/
8. http ://lemur.sourceforge.net/indri

3.5 Conclusion

C'est donc dans ce contexte que s'inscrivent mes travaux de thèse. Nous nous sommes intéressés à l'exploitation des données multimodales pour la traduction automatique. Donc, afin de structurer notre approche, nous avons préparé les outils et les ressources utiles pour le traitement de ce sujet. Nous avons présenté dans ce chapitre les ressources linguistiques disponibles au LIUM, notamment les corpus parallèles et comparables, ainsi que les outils et les technologies développés au sein du laboratoire qui ont permis le démarrage de nos travaux. Toutes les approches présentées sont évaluées afin de choisir celles qui sont les mieux adaptées à notre contexte de travail. Dans le chapitre suivant, nous présenterons le système développé pour appliquer notre approche d'extraction des données parallèles à partir d'un corpus multimodal.

Troisième partie

Contributions

Chapitre 4

Mise en oeuvre d'un système d'extraction de données parallèles

Sommaire

Comme nous avons présenté dans le chapitre 2, une façon de pallier le manque au données parallèles est d'exploiter les corpus comparables qui sont plus abondants.

Dans ce chapitre nous présentons une méthode pour l'utilisation des corpus comparables multimodaux, en nous limitant aux modalités texte et audio, pour l'extraction de données parallèles.

Dans un premier temps, nous présentons le contexte de nos travaux et notre système proposé. Cette approche soulève plusieurs problématiques, qui sont présentées dans un second temps. Nous présentons dans la section 4.4 une série d'expériences qui permet de répondre à nos questions, et nous analysons nos résultats dans un dernier temps.

4.1 Contexte

Dans notre contexte de travail, nous nous intéressons à l'exploitation de corpus comparables multimodaux avec différents niveaux de comparabilité.

Notre but est de trouver une méthode pour exploiter les données comparables multimodales, afin d'en extraire des données parallèles nécessaires pour construire, adapter et améliorer nos systèmes de traduction automatique statistique.

4.2 Architecture générale

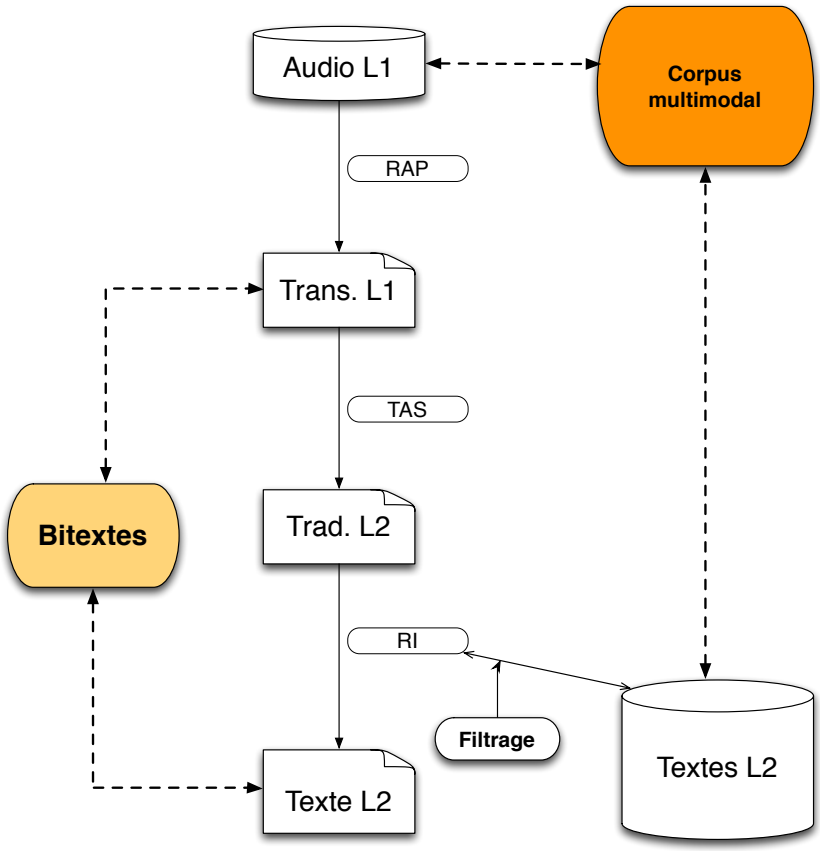

FIGURE 4.1 – Architecture générale du système d'extraction des données parallèles à partir d'un corpus multimodal multilingue

L'architecture générale de notre approche présentée dans la figure 4.1 se résume en trois étapes.

Notre corpus comparable multimodal est constitué de données audio en langue source L1 et de données textuelles en langue cible L2.

Dans une première étape, les données audio sont tout d'abord transcrites par un système de RAP. Ce système produit une hypothèse de transcription qui est ensuite traduite par le système TAS, en second étape. La troisième étape consiste à utiliser la meilleure hypothèse de traduction comme requête dans le système de recherche d'information (RI), dont le corpus indexé correspond à la partie textuelle en langue cible du corpus comparable multimodal.

Dans cette approche, qui se base sur les travaux de [Abdul-Rauf et Schwenk, 2011], nous utilisons le logiciel libre *Lemur* comme décrit dans la section 3.4 du chapitre 3 pour effectuer la RI. Au final, nous obtenons un bitexte constitué d'une part, de la transcription automatique et d'autre part, du résultat de la RI. Ces données pourront être exploitées de différentes manières

pour améliorer le système de base.

4.3 Problématiques

Ce cadre de travail soulève plusieurs problèmes. Nous détaillons dans cette section les différentes problématiques rencontrées et nos propositions pour répondre aux questions scientifiques posées lors de l'élaboration de notre méthode.

4.3.1 Enchaînement des modules

Chaque module mis en jeu pour la traduction de la parole introduit un certain nombre d'erreurs. Il est important de mettre en évidence la faisabilité de l'approche ainsi que l'impact de chaque module sur la qualité des données générées. Pour cela, nous avons effectué trois types d'expérience, décrits dans la figure 4.2.

Le premier type d'expérience (*Exp 1*) consiste à utiliser la référence de traduction comme requête pour la RI. Ce cas est le plus favorable, cela simule le fait que les modules de RAP et de TAS ne commettent aucune erreur. Le second type d'expérience (*Exp 2*) utilise la référence de transcription pour alimenter le système de traduction automatique. Cela permet de mettre en évidence l'impact des erreurs de traduction. Enfin, le troisième type d'expérience (*Exp 3*) met en œuvre l'architecture complète décrite dans la section 4.2. Cela correspond au cas réel auquel nous sommes confrontés.

4.3.2 Degré de comparabilité

Une autre problématique concerne l'importance du degré de similitude (*comparabilité*) des corpus comparables utilisés. Nous avons donc artificiellement créé des corpus comparables plus ou moins ressemblants en intégrant une quantité plus ou moins grande (*25%, 50%, 75% et 100%*) de données du domaine dans le corpus indexé par la RI. Ainsi, nous pouvons tester notre approche dans différents scénarios de comparabilité qui varient entre les corpus très faiblement reliés comme celui qui contient que *25%* des données du domaine, et les corpus très similaires comme celui qui contient *100%* des données du domaine.

Les données que nous injectons dans la partie texte en langue L2 (qui est le français dans notre cas) correspondent aux phrases que le système doit retourner à partir des requêtes anglaises traduites. Nous les appelons *phrases de références*. Nous pouvons mentionner ici que cette expérience ne correspond pas aux cas réelles des corpus comparables, mais son intérêt est de tester la faisabilité de notre approche.

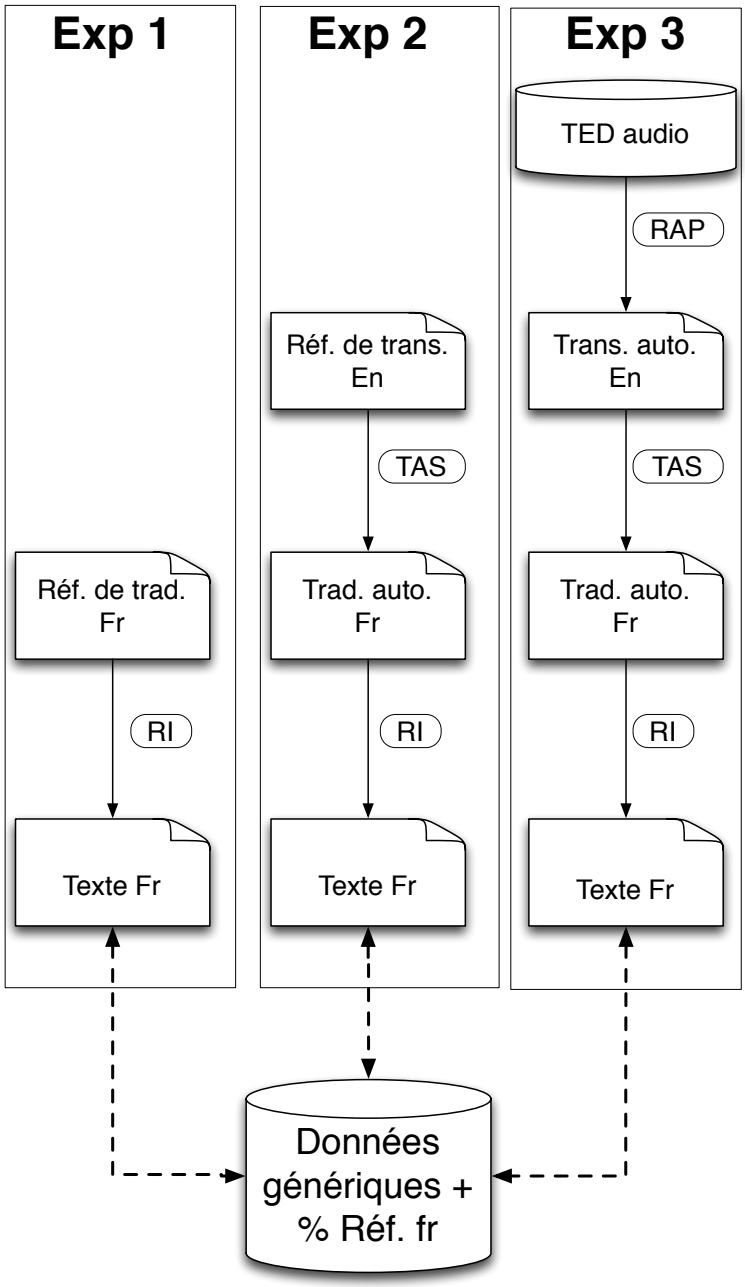

FIGURE 4.2 – Expériences permettant de mesurer l'impact des différents modules mis en jeu sur le corpus bilingue extrait.

Les différentes conditions de comparabilité permettent aussi d'évaluer le module de recherche d'information. En effet, en injectant des données de références de différents quantités dans les données génériques, nous pouvons mesurer la capacité du système à trouver les phrases de références correspondant à ce qu'on doit trouver pour chaque requête.

4.3.3 Filtrage

Les résultats de la RI ne sont pas toujours satisfaisants, il est donc nécessaire de filtrer ces résultats afin de ne pas ajouter des phrases non parallèles dans le bitexte final. [Cettolo *et al.*, 2010] ont montré que l'ajout des phrases extraites sans filtrage n'améliore pas le système de traduction de base, même lorsque les corpus sont bien similaires.

Nous nous intéressons dans ce module à exclure les couples de phrases générées par le système et qui sont fortement bruités en terme de parallélisme. Nous avons exploité la méthode utilisée par [Abdul-Rauf et Schwenk, 2011] dans le développement de notre système d'extraction. Nous proposons une amélioration de ce module dans la section 5.2 du chapitre suivant.

Le Taux d'Edition de la Traduction (*Translation Edit Rate* - TER) calculé entre les phrases retournées par la RI et la requête est utilisée comme métrique de filtrage des phrases trouvées. Les phrases ayant un TER supérieur à un certain seuil (déterminé empiriquement) sont exclues.

4.4 Expériences et résultats

Dans cette partie nous présentons les données et le cadre des expériences réalisées pour répondre aux problématiques évoquées par la proposition de notre méthode d'extraction. Nous analysons nos résultats après la description du déroulement de nos expériences.

4.4.1 Cadre expérimental

Afin d'expérimenter la méthode d'extraction, un corpus comparable a été simulé en assemblant des paires de phrases parallèles et non parallèles avec différents nivaux de similitude. Nous présentons ce corpus par la suite.

Le corpus de développement, issu de TED, est composé de 19 discours représentant un peu plus de 4 heures de parole. Ce corpus présenté dans le tableau 4.1 est transcrit également avec les systèmes de reconnaissance automatique de la parole.

Les corpus bilingues présentés dans le tableau 4.2 sont utilisés pour l'apprentissage des modèles de traduction : *News-Commentary* version 7(*nc7*), le corpus des actes du parlement européen (*eparl7*).

Le système de reconnaissance de la parole utilisé est basé sur le système libre CMU Sphinx que nous avons déjà évoqué lors du chapitre précédent dans la section 3.3. Les performances du système sont présentées dans le tableau 4.3. Ces résultats ont permis de classer ce système en première place lors de la campagne d'évaluation IWSLT'11 [Rousseau *et al.*, 2011]. Nous utilisons ce système pour transcrire les données audio TED qui vont être traduites et utilisées comme requêtes de RI.

Données	# de mots anglais	# de mots français
DevTED	36k	38k
TestTED	8.7k	9.1k

TABLE 4.1 – Données de développement et de test. DevTED (anglais) sont les données de développements en anglais issu de la transcription du système ASR. DevTED (français) sont les références de traduction en français des données de développement. TestTED (anglais) sont les données de test en anglais transcrites par le système ASR. TestTED (français) sont les références de traductions des données de test.

données	# de mots (en)	# de mots (fr)	du domaine TED ?
nc7	3.4 M	4 M	non
eparl7	55.7 M	61.7 M	non
TEDasr	1.8 M	-	oui
TEDbi	1.8 M	1.9 M	oui

TABLE 4.2 – Données utilisées pour l'apprentissage des systèmes de traduction automatique. nc7 et eparl7 sont utilisés comme données génériques pour l'apprentissage du système de traduction de base. TEDasr sont les données TED (anglais) transcrites par le système de RAP. TEDbi sont les données TED (français) de référence (manuelle) de traduction qui sont injectées dans le corpus de RI.

Données	WER
DevTED	19.2%
TestTED	18.2%

TABLE 4.3 – Performance du système de RAP en terme de WER sur les données de développement et de test.

Nous appelons ces données comme présentées dans le tableau 4.2 *TEDasr*. Les données *TEDbi* sont les références de traductions (en français) des phrases de *TEDasr*, qui sont injectées dans la partie texte de notre corpus comparable avec de différents pourcentages afin de tester l'efficacité de notre système à trouver les bonnes phrases, si elles existent.

Nous utilisons un système de traduction de base appris avec la méthode décrite dans la section 3.2.4, avec les données hors domaine (génériques) présentées dans le tableau 4.2. Ce système est optimisé et testé respectivement avec les corpus de développement (DevTED) et de test (TestTED) détaillé dans le tableau 4.1.

4.4.2 Déroulement des expériences

Dans tous les cas, l'évaluation de l'approche est nécessaire. Ainsi, les données parallèles extraites sont réinjectées dans le système de traduction de base, qui est

ensuite utilisé pour traduire, de nouveau, les données de test. L'évaluation peut ensuite se faire avec une mesure automatique, comme BLEU.

4.4.3 Synthèse des résultats

Comme mentionné précédemment, le score *TER* est utilisé comme métrique de filtrage des phrases résultantes de la RI, c'est-à-dire que les phrases ayant un TER supérieur à un certain seuil ne sont pas conservées. Ce seuil est déterminé empiriquement. Pour cela, nous avons filtré les corpus extraits dans les différentes conditions d'expérimentation avec différents seuils *TER* (de 0 à 100). Pour chaque seuil TER nous obtenons un nombre de phrases supposées parallèles. Le corpus obtenu est ajouté aux données d'entraînement du système de base (*eparl7* et *nc7*) pour obtenir le système adapté. Les résultats, en terme de score *BLEU* sur le corpus de développement obtenu avec les différents systèmes adaptés, sont présentés dans les figures 4.3 et 4.4 avec le corpus de développement.

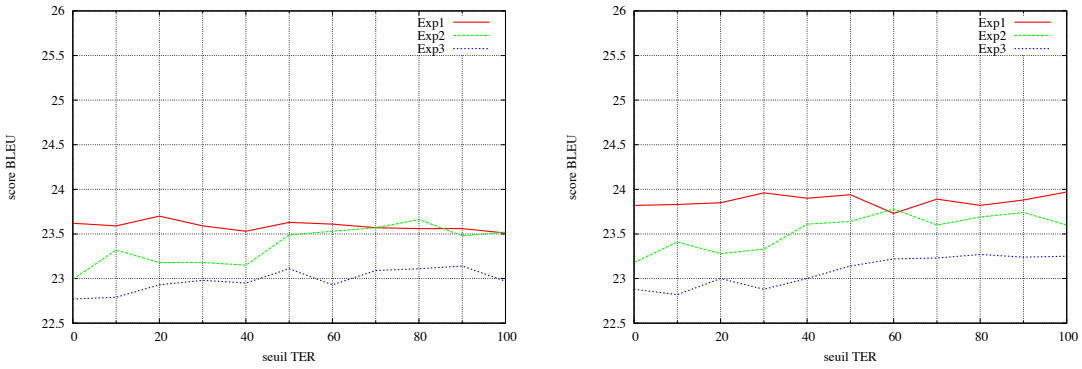

FIGURE 4.3 – Score BLEU de la traduction du Dev en utilisant les systèmes adaptés avec les bitextes correspondant à différents seuils TER, extraits d'un corpus d'index constitué par *des données génériques + 25% TEDbi* (à gauche) et *des données génériques + 50% TEDbi* (à droite).

Ces résultats montrent que le choix du seuil de *TER* adéquat dépend de la nature des données. En effet, dans la condition de l'*Exp1* où les requêtes de la RI sont sans erreur, nous remarquons que les résultats sont stables. Dans les deux autres conditions (*Exp2* et *Exp3*), le meilleur seuil est dans l'intervalle [80-90]. Dans nos expériences, nous retiendrons le seuil de 80 pour le filtrage des résultats de la RI.

Notre choix est confirmé par les résultats obtenus sur TestTED présenté dans le tableau 4.4 pour les expériences avec 100% TEDbi injectées dans les données génériques, et dans les figures 8.1, 8.2 et 8.3 en annexe.

Nous avons aussi remarqué dans d'autres expériences avec des seuils TER qui dépassent 100, les performances des systèmes dans toutes les conditions commencent à se dégrader, surtout lorsque le corpus indexé n'est pas très comparable (*le cas de*

TER	Exp1	Exp2	Exp3
0	25.62	24.54	24.24
10	25.40	24.75	24.06
20	25.41	24.67	23.88
30	25.58	24.70	23.87
40	25.62	25.21	24.05
50	25.50	25.20	24.41
60	25.51	25.15	24.22
70	25.75	25.09	24.68
80	25.14	25.15	24.69
90	25.48	25.11	24.72
100	25.22	24.99	24.54

TABLE 4.4 – Score BLEU obtenu sur TestTED dans les conditions *Exp1*, *Exp1* et *Exp1* pour chaque seuil TER et avec les données génériques + 100% TEDbi.

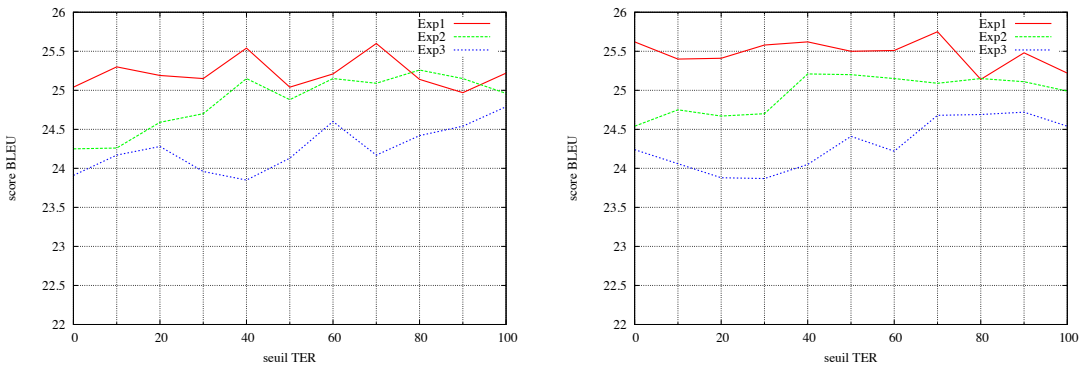

FIGURE 4.4 – Score bleu des systèmes adaptés avec les bitextes de différents seuils TER, extraite d'un corpus d'indexe constitué par des données génériques + 100% TEDbi (à droite) et des données génériques + 75% TEDbi (à gauche).

25% TEDbi). Ce résultat est dû à l'insertion de fausses traductions dans le bitexte, d'où l'importance du filtrage.

Les résultats obtenus après adaptation du système de base sont présentés dans le tableau 4.5. Dans ce cas, le corpus indexé par la RI est constitué par l'injection de *TEDbi* (100%).

Expérience	DevTED	TestTED
Système de base	22.93	23.96
Exp1	24.14	25.14
Exp2	23.90	25.15
Exp3	23.40	24.69

TABLE 4.5 – Scores BLEU obtenus sur le Dev et Test après l'ajout des bitextes extraits au système de base, dans les conditions *Exp1, Exp2 et Exp3*.

Nous remarquons également que les résultats obtenus dans les 3 conditions sont relativement proches, et que même lorsque la requête fournie au système de recherche d'informations est dégradée, les performances des systèmes adaptés avec les bitextes résultants sont proches. Nous pouvons donc en déduire que l'enchaînement des systèmes de RAP et de TAS n'impactent que légèrement les résultats de la RI.

Le tableau 4.6 présente les résultats des systèmes adaptés en fonction du degré de similitude du corpus comparable, dans les conditions d'expérimentation *Exp3*.Nous pouvons remarquer que le degré de similitude est un facteur important.

Expérience	DevTED	TestTED	# mots
Système de base	22.93	23.96	-
25% TEDbi	23.11	24.40	∼110k
50% TEDbi	23.27	24.58	∼215k
75% TEDbi	23.43	24.42	∼293k
100% TEDbi	23.40	24.69	∼393k

TABLE 4.6 – Scores BLEU et quantités de données obtenus dans la condition d'expérience *Exp3* avec les systèmes adaptés lorsque le degré de similitude du corpus comparable varie.

Un résultat attendu est que lorsque nous augmentons la proportion de corpus du domaine dans le corpus indexé, les performances sont meilleures. Il est important de noter que lorsque les corpus sont moins similaires, le nombre de phrases conservées est réduit drastiquement par le filtrage, et donc l'impact de l'adaptation est plus faible. Sans filtrage, les performances du système de base peuvent être dégradées.

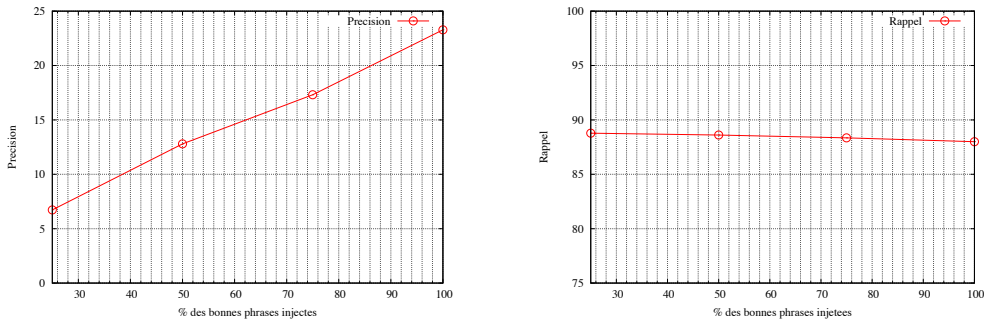

FIGURE 4.5 – Courbes de la précision (à gauche) et du rappel (à droite) du système d'extraction.

Pour faire une évaluation complète du système, nous avons aussi mesuré la performance du processus d'extraction en calculant le rappel de la quantité des phrases renvoyées dans les différents cas de comparabilité et la précision des bonnes phrases renvoyées. Nous définissons la précision, dans ce contexte, par le quotient du nombre de paires de phrases identifiées comme parallèles par le nombre total des paires de phrases extraites (pour chaque seuil TER). Le rappel est défini par le quotient du nombre de paires de phrases extraites par le système sur le nombre total des phrases, c-à-d, les phrases du domaine (*TEDbi*) et hors domaine. Les deux métriques sont exprimées en pourcentage comme suit :

$$Precision = \frac{100 * nb\ de\ phrases\ paralleles\ trouvees}{total\ nb\ de\ phrases\ extraites} \tag{4.1}$$

$$Rappel = \frac{100 * nb\ de\ phrases\ paralleles\ extraites}{total\ nb\ de\ requetes} \tag{4.2}$$

La combinaison des deux mesures avec des poids équivalents est présentée par l'équation suivante :

$$Fmesure = 2\ \frac{Rappel \cdot Precision}{Rappel + Precision} \tag{4.3}$$

Comme nous pouvons remarquer dans la figure 4.5, la valeur du rappel est stable puisque le nombre de requêtes initiales reste le même dans toutes les expériences. Aussi, nous pouvons voir clairement dans la figure 4.6 que la performance du système d'extraction en terme de F-mesure dépend du degré de parallélisme du corpus comparable. Ces courbes valident les résultats précédents obtenus en terme de score BLEU.

Le tableau 4.7 présente un exemple d'adaptation des systèmes de TA avec les données extraites par notre système. Nous extrayons à partir du corpus comparable du domaine les deux phrases comme dans la première case du tableau.

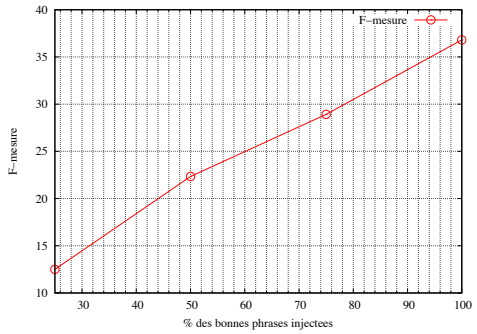

FIGURE 4.6 – Courbe du F-mesure du système d'extraction.

La première correspond à la version originale en anglais traduite en requête présentée dans la deuxième phrase. La troisième phrase correspond à ce que le système a trouvé dans le corpus de RI. Ces phrases sont utilisées, après leur injection aux données d'apprentissage de base, pour enrichir la traduction proposée par le mot *superordinateur* qui n'était pas dans le vocabulaire du système de base. Nous pouvons remarquer aussi dans cet exemple que l'erreur de transcription dans la phrase de test (*humans and only* au lieu de *humans have only*) a un impact sur la traduction finale proposée par les deux systèmes de traduction (*les humains et seulement* au lieu de *les humains n' avaient plus que*).

4.5 Conclusion

Dans cette partie nous avons proposé une méthode permettant d'extraire des textes parallèles à partir d'un corpus comparable multimodal (audio et texte) pour améliorer un système de traduction automatique statistique. Plusieurs modules sont utilisés pour extraire du texte parallèle : la reconnaissance automatique de la parole, la traduction automatique et la recherche d'information. Nous validons notre méthode en injectant les données produites dans l'apprentissage de nouveaux systèmes de TAS. Des améliorations en termes de BLEU sont obtenues dans différents cadres expérimentaux. Il en ressort que l'enchaînement des modules ne dégrade que faiblement les résultats. Nous nous intéressons dans le chapitre suivant au module de filtrage, ainsi que l'augmentation de la quantité des données extraites en exploitant les segments sous phrastiques.

	Extraction
Phrase en anglais	so you need ten thousand laptops so would you go you a good idea and you get a supercomputer because they know how to take ten thousand laptops and put into the size of a refrigerator
Requête en français	si vous avez besoin de dix mille ordinateurs portables alors iriez vous vous avez une bonne idée et vous obtenez un supercomputer parce qu' ils savent comment prendredix mille ordinateurs portables et mis en de la taille d' un réfrigérateur
Phrases trouvée	où aller vous allez chez ibm et vous prenez un superordinateur car ils savent comment concentrer dix mille ordinateurs portables dans les dimensions d' un réfrigérateur
	Test audio
Sortie ASR Référence	a supercomputer has calculated that humans and only ... a supercomputer has calculated that humans have only ...
	Traductions de la sortie ASR
Système de base Système adapté Référence	un **supercomputer** a calculé que les humains et seulement ... un **superordinateur** a calculé que les humains et seulement ... un **superordinateur** a calculé que les humains n' avaient plus que ...

TABLE 4.7 – Exemple d'amélioration du système de base en utilisant un vocabulaire enrichi à partir les phrases parallèles extraites dans la condition *Exp3*.

Chapitre 5

Extraction d'entités sous-phrastiques

Sommaire

Nous poursuivons dans ce chapitre avec l'exploitation des documents multimodaux en nous intéressant à la génération des données parallèles composées de segments parallèles.

La plupart des études existantes traitent de l'extraction des données parallèles au niveau des phrases [Zhao et Vogel, 2002, Utiyama et Isahara, 2003, Munteanu et Marcu, 2005, Abdul-Rauf et Schwenk, 2011]. Le degré de parallélisme peut varier considérablement du parallèle bruité au quasi parallèle [Fung et Cheung, 2004]. Les corpus de cette dernière catégorie ne contiennent pas une grande quantité de phrases parallèles. Cependant, les phrases non parallèles peuvent contenir aussi des segments parallèles qui pourront aider à l'amélioration des systèmes de TA [Munteanu et Marcu, 2006]. Dans l'exemple des figures 5.1 et 5.2, nous avons deux articles d'actualité en modalité texte avec leurs sources vidéo de l'édition anglaise et française du site Euronews [1]. Le même évènement est abordé dans ces articles à travers l'utilisation de phrases contenant des traductions au niveau des segments ou groupes de mots. La plupart des phrases dans ces deux articles ne sont pas des traductions exactes les unes des autres, donc les techniques d'extraction des phrases parallèles ne vont pas donner de bons résultats. Nous avons besoin dans ce cas d'une méthode d'extraction qui agit au niveau des segments de phrases, pour en extraire les groupes de mots traduits.

Nous avons choisi de travailler sur les segments à la suite d'une étude menée sur les types d'erreurs obtenues par l'utilisation de l'approche d'extraction des phrases

1. www.euronews.com/

xx

FIGURE 5.1 – Exemple d'articles d'actualité en modalité texte avec leurs sources vidéo en langues anglaise et française. Les deux paragraphes entourés parlent des mêmes informations, mais ne contiennent pas de phrases parallèles en terme de traduction exacte.

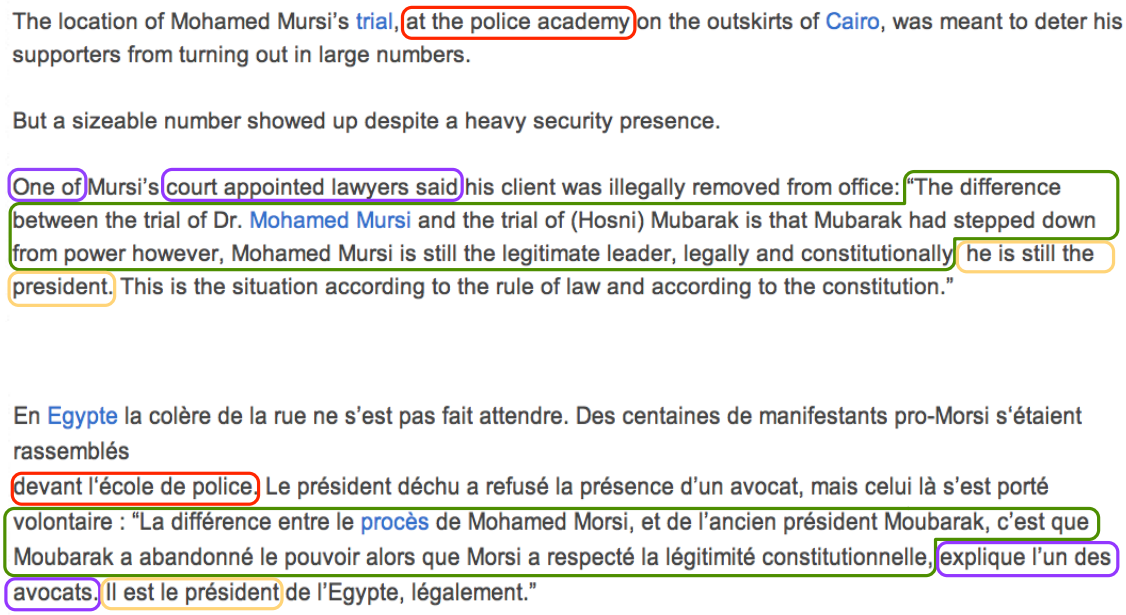

FIGURE 5.2 – Exemples de traductions continus dans deux paragraphes comparables.

parallèles, notamment sur le nombre des mots à éditer, à insérer ou à supprimer pour chaque phrase extraite. Nous pouvons remarquer à partir des figures 5.4 et 8.4 qu'en général, les phrases qui sont constituées d'un plus grand nombre de mots ont un score TER plus élevé que les petites phrases. Les figures 5.5, 5.6, 8.5, et 8.6 des statistiques sur le nombre de mots à supprimer ou insérer. Elles montrent aussi que les phrases longues contiennent plus d'erreurs que les petites, en terme de nombre de mots.

5.1 Systèmes d'extraction des phrases et segments parallèles

Nous allons décrire dans cette partie notre méthode d'extraction des segments sous-phrastiques parallèles appelée *PhrExtract* [Afli *et al.*, 2013]. Nous comparons cette méthode avec celle de l'extraction des phrases parallèles [Afli *et al.*, 2012] décrite dans le chapitre 4, appelée par la suite *SentExtract*.

L'architecture générale de la méthode d'extraction des segments parallèles proposée est présentée dans la figure 5.3. La différence par rapport à l'extraction des phrases réside en la segmentation des données transcrites en groupes de mots avant la traduction. Afin de conserver une cohérence lors de la recherche d'information, nous avons également segmenté les phrases de la langue cible par la même méthode. La taille des segments varie de deux à dix mots. Cette segmenta-

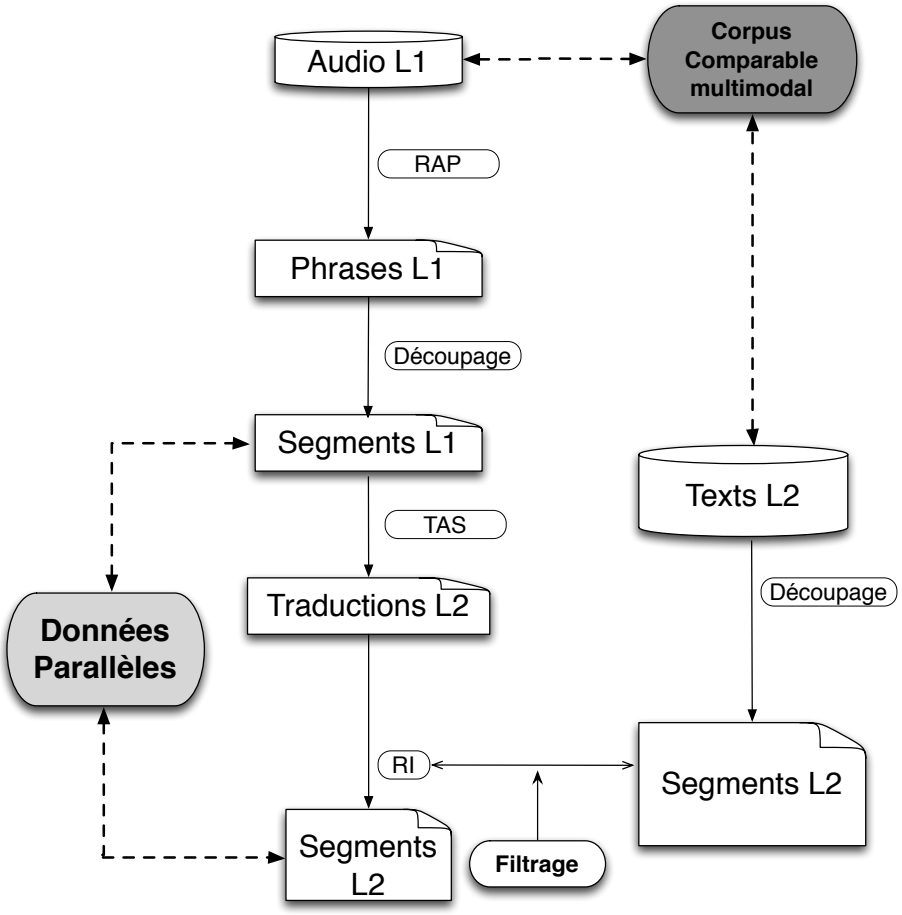

FIGURE 5.3 – Principe du système d'extraction des segments parallèles avec la méthode *PhrExtract*.

tion consiste à générer tous les groupes des mots successifs possibles dans une phrase.

5.1.1 Cadre expérimental

Dans cette partie, nous avons utilisé les deux corpus multimodaux TED et Euronews. Le système de TA de base est le même pour toutes les expériences. Il est appris à l'aide des bitextes Europarl (Eparl7) et News Commentary (nc7) du tableau 5.1. Puisque les domaines de nos deux corpus multimodaux sont différents, nous utilisons deux différents corpus de développement et de test du même domaine de chaque corpus. *devTED* et *tstTED* sont respectivement les données de développement et de test utilisées dans les expériences sur le corpus TED-LIUM, présenté dans le chapitre précédent. Concernant les expériences sur les données Euronews-LIUM, nous avons utilisé les données de développement et de test *devEuronews* et *tstEuronews* qui sont des données du domaine d'actualité (news). Les statistiques de ces données sont présentées dans le tableau 5.1.

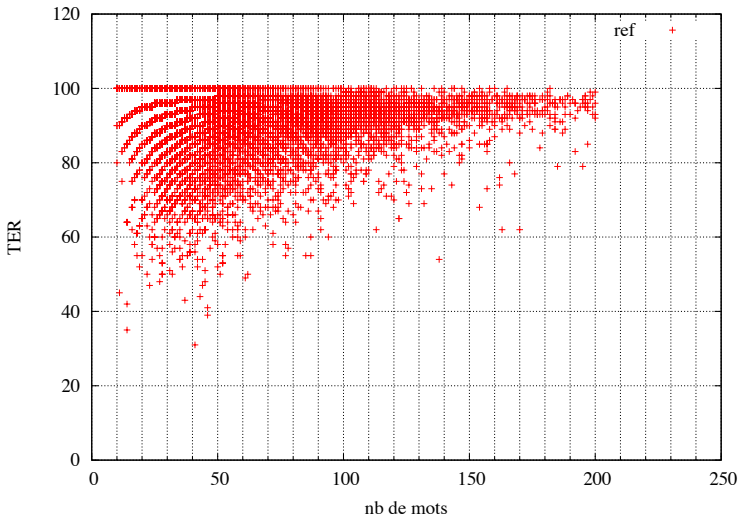

FIGURE 5.4 – Distribution du seuil TER par rapport au nombre de mots de chaque phrase (Corpus Euronews).

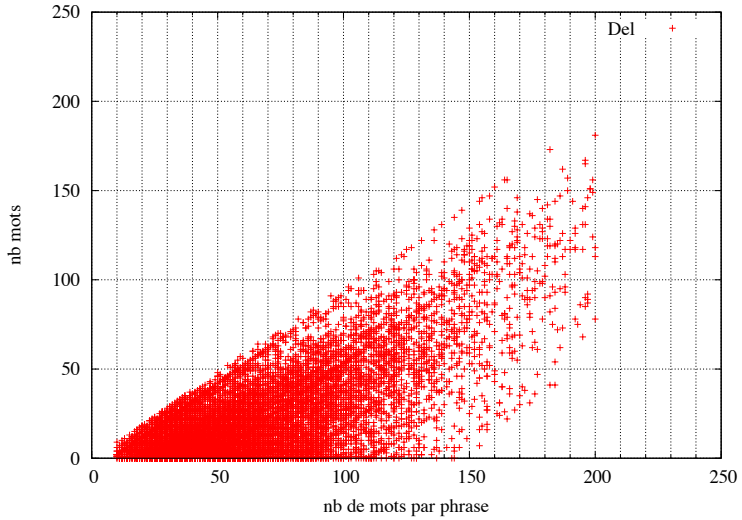

FIGURE 5.5 – Distribution de nombre de mots supprimés pour chaque phrase (Corpus Euronews).

Pour comparer les deux méthodes *SentExtract* et *PhrExtract*, nous ajoutons les données extraites par les deux méthodes aux données du système de base. Ce qui nous permet de mesurer la qualité de ces données en regardant leurs impacts sur les performances du système de traduction de base. Les différents systèmes de TA sont évalués à l'aide de la métrique BLEU.

Comme mentionné dans l'architecture du système d'extraction, dans cette partie

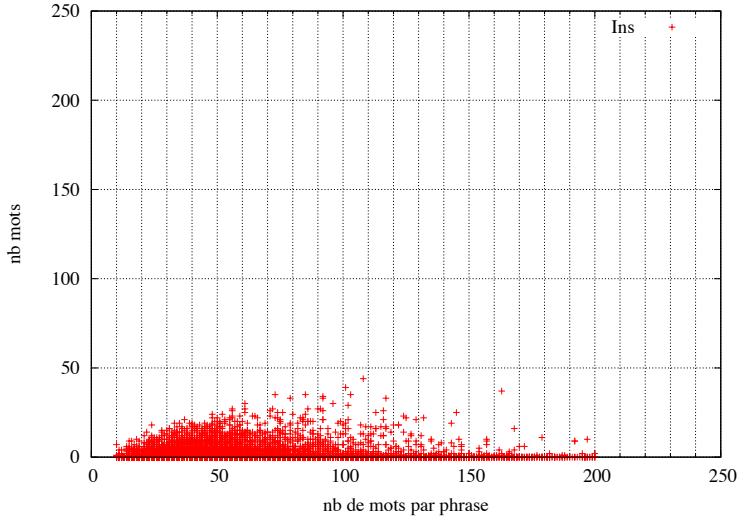

FIGURE 5.6 – Distribution de nombre de mots insérés pour chaque phrase (Corpus Euronews).

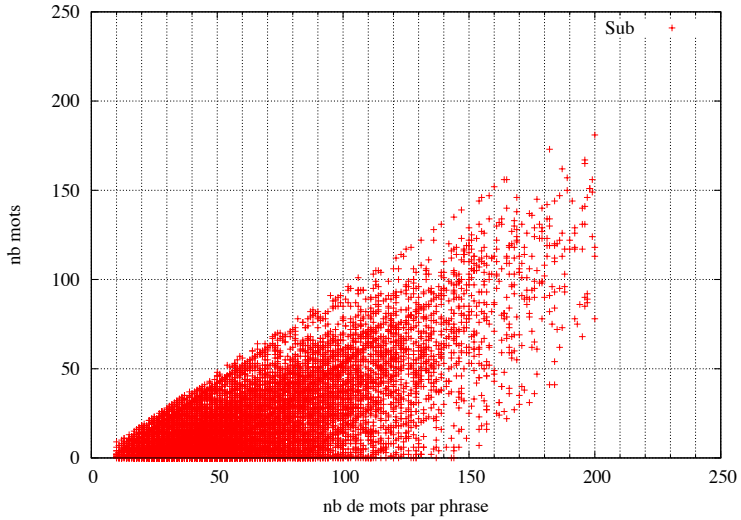

FIGURE 5.7 – Distribution de nombre de mots de substitution pour chaque phrase (Corpus Euronews).

Corpus	# mots anglais	# mots français
devEuronews	74k	84k
tstEuronews	61k	70k

TABLE 5.1 – Données de développement et de test utilisées pour les expériences de corpus Euronews pour les méthodes *SentExtract* et *PhrExtract*.

nous utilisons le score TER comme métrique de filtrage des résultats de la recherche d'information. Cette méthode de filtrage nous permet de générer différents corpus avec différents seuils TER allant 0 et 100 par pas de 10.

5.1.2 Résultats

Les résultats de ces expériences montrent, comme présentés dans le tableau 5.2 pour les données de développement et dans le tableau 5.3 pour les données de test, l'intérêt de notre méthode pour exploiter des documents multimodaux dans un contexte de traduction automatique. Notre système de TA de base est appris avec des données génériques et atteint un score BLEU de 22.93. Nous pouvons remarquer une amélioration de ce système avec les deux méthodes d'extraction, notamment avec *PhrExtract*. Ce dernier dépasse la méthode d'extraction de phrases parallèles *SentExtract* dans la plupart des expériences présentées dans le tableau 5.2 avec les différents seuils TER. Plus précisément, *PhrExtract* donne toujours de bons résultats d'amélioration jusqu'à le seuil TER de 80, où les deux systèmes adaptés atteignent le même score BLEU de 23.40 avec les deux méthodes. Comme nous pouvons le voir dans la dans la figure 5.8, cette valeur de TER présente un point d'inflexion pour la courbe de *SentExtract*. Les performances des systèmes adaptés avec la méthode *PhrExtract* commencent à baisser par rapport aux systèmes adaptés avec la méthode *SentExtract* à ce même point aussi. Nous pouvons également voir dans la figure 5.8 que le meilleur système qu'on a pu l'avoir est celui qui est adapté avec les données de TER 60 extraites en utilisant la méthode *PhrExtract*, souligne l'importance de cette méthode, et relève une relation étroite entre le seuil TER et le choix de la méthode utilisée.

Ces résultats nous ont motivés à combiner ces deux méthodes d'extraction. Cette combinaison consiste en l'injection des phrases et des segments extraits par les deux méthodes pour chaque seuil TER aux données d'apprentissage du système de TA de base. Nous appelons par la suite cette méthode *CombExtract*.

La figure 5.9 présente les résultats de cette combinaison en terme de score BLEU. Nous remarquons dans cette figure que la courbe de combinaison suit en général celle de la méthode *PhrExtract*. Nous pouvons expliquer ce fait à partir du tableau 8.3 où nous pouvons remarquer la grande différence de quantité de données extraites entre les deux méthodes *PhrExtract* et *SentExtract*. Cette grande différence influence la combinaison à ne pas trop considérer les données de *SentExtract* qui sont statistiquement moins importantes dans les données globales d'apprentissage. Nous traitons la problématique de la manière d'utilisation des données extraites pour l'adaptation du système de traduction dans le chapitre 6.

Nous avons appliqué la meilleure méthode *PhrExtract*, sur les données Euronews, où nous avons eu un gain plus important en qualité de traduction et en score d'évaluation BLEU. Nous pouvons voir dans le tableau 5.4 qu'on pu atteindre un

TER	# mots (fr) SentExtract	# mots (fr) PhrExtract	BLEU SentExtract	BLEU PhrExtract
0	55	1.06M	22.86	23.39
10	313	1.4M	22.97	23.35
20	1.7k	2.5M	23.06	23.53
30	6.9k	4.3M	22.95	23.39
40	23.5k	7.02M	22.92	23.45
50	62.4k	11.4M	23.26	23.54
60	13.8k	**13.8M**	23.10	**23.70**
70	25.1k	18.04M	23.29	23.41
80	**39.3k**	**25.3M**	**23.40**	**23.40**
90	57.5k	35.9M	23.39	23.18
100	83.6k	45.3M	23.34	23.26
Baseline	60.1M			

TABLE 5.2 – Score BLEU obtenu sur devTED et quantités de mots extraites avec la méthode *PhrExtract* pour chaque seuil TER (données TED).

TER	SentExtract	PhrExtract	CombExtract
0	24.24	24.68	24.45
10	24.06	24.38	24.26
20	23.88	24.75	24.51
30	23.87	24.64	24.77
40	24.05	24.29	24.41
50	24.41	23.92	23.97
60	24.22	24.84	24.25
70	24.68	24.39	24.26
80	24.69	24.60	24.27
90	24.72	24.50	24.05
100	24.54	23.95	24.28
Baseline	23.96		

TABLE 5.3 – Score BLEU obtenu sur tstTED avec les méthodes *CombExtract*, *PhrExtract* et *SentExtract* pour chaque seuil TER.

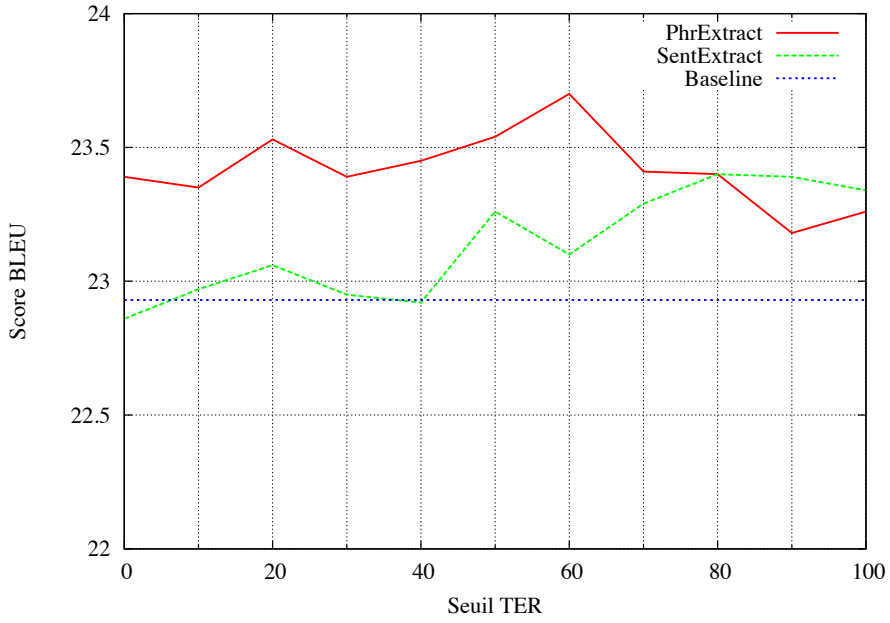

FIGURE 5.8 – Évolution de la performance des systèmes de traduction adaptés avec les données extraites à l'aide des méthodes *PhrExtract* et *SentExtract* en terme de score BLEU pour chaque seuil TER sur le corpus de développement devTED.

score BLEU de 30.04 en ajoutant les données extraites du corpus Euronews avec la méthode *PhrExtract* qui ont un seuil 30 de TER. Ce qui nous fait un gain de 4.85 points de BLEU sur le corpus de développement et 5.45 points sur le corpus de test.

TER	# mots (fr)	devEuronews	tstEuronews
0	90 k	29.95	27.13
20	168 k	30.01	27.33
30	322 k	**30.04**	**27.59**
40	769 k	29.92	27.45
50	1.9 M	30.03	27.27
60	3.1M	29.92	27.30
70	5.7M	29.83	27.11
80	12.39 M	29.73	27.10
90	25.7 M	29.73	26.88
100	40.3 M	29.52	26.53
Baseline	60.1M	25.19	22.12

TABLE 5.4 – Score BLEU obtenu sur devEuronews et tstEuronews et quantités de mots extraites avec la méthode *PhrExtract* pour chaque seuil TER (données Euronews).

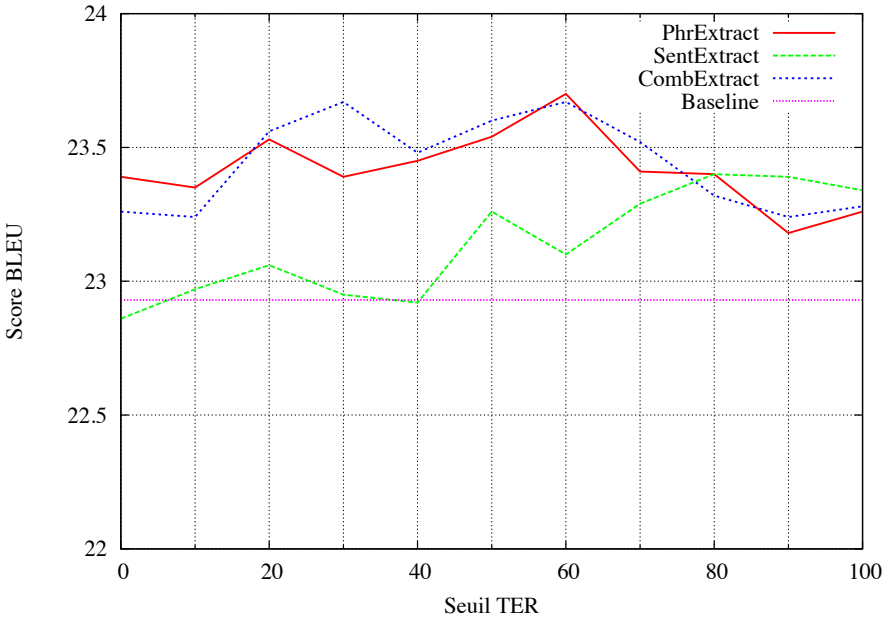

FIGURE 5.9 – Évolution de la performance des systèmes de traduction adaptés avec les données extraites à l'aide des méthodes *CombExtract*, *PhrExtract* et *SentExtract* en terme de score BLEU pour chaque seuil TER sur le corpus de développement devTED.

Nous pouvons constater aussi que le filtrage TER est très important en analysant la quantité de données ajoutées et le score BLEU obtenu pour chaque seuil. D'après le tableau 5.4, 322.5K de mots ajoutés pour le seuil 30 donnent de meilleures améliorations que 40.3M de mots pour le seuil 100. Ce résultat est attendu, car la plupart des phrases de seuil TER 100 sont très bruitées. Ce qui nous donne une idée sur la quantité de données qui a un impact sur l'adaptation du système de base. Mais cette méthode reste un peu couteuse en temps, puisque nous sommes obligés pour chaque expérience d'extraction de générer plusieurs sous-corpus avec différents seuils TER. Ces corpus sont utilisés pour chercher le meilleur seuil TER. Pour améliorer notre système d'extraction, nous avons proposé de modifier ce module de filtrage, afin de chercher une solution qui permet de minimiser le nombre d'expériences et le temps de cette étape en préservant les mêmes performances du système. Nous présenterons par la suite notre proposition.

5.2 Amélioration du module de filtrage

Dans le système d'extraction des fragments parallèles proposé par [Munteanu et Marcu, 2006], un lexique est utilisé pour le module de filtrage. Les probabilités de traduction des mots de ce lexique sont estimées à l'aide de la méthode *LLR* (Log-Likelihood-Ratio) de [Dunning, 1993]. Nous avons proposé

d'utiliser cette méthode pour remplacer notre module de filtrage basé sur le TER. Cette proposition consiste à utiliser le système d'extraction des phrases, considérées quasi parallèles, comme première étape. Dans une deuxième étape le module de filtrage basé sur la *LLR*, qu'on appelle par la suite *LLR_ lex*, permet de détecter, à l'aide du lexique, les fragments parallèles dans les phrases générées. Ce lexique est construit dans notre cas, à partir des bitextes du système de TA de base.

Nous l'utilisons pour estimer l'indépendance des paires de mots qui co-occurent dans notre corpus parallèle. Le score LLR d'une paire de mots est faible lorsque les mots sont indépendants, et augmente lorsque les mots sont fortement associés.

La figure 5.10 présente l'architecture générale du nouveau système d'extraction des segments parallèles en utilisant le filtrage avec le lexique LLR. Nous appelons par la suite ce système *SentExtract +LLR_ lex*, puisqu'il est basé sur la combinaison entre notre méthode *SentExtract* et la méthode *LLR_ lex*.

5.2.1 Cadre expérimental

Afin d'expérimenter notre nouvelle méthode, nous avons utilisé les mêmes corpus multimodaux comparables *TED* et *Euronews* pour en extraire des données parallèles. Nous avons utilisé aussi le même système de TA de base appris sur *Eparl7* et *nc7*.

5.2.2 Résultats

Les tableaux 5.5 et 5.6 présentent les résultats en terme de score BLEU sur les corpus de développement et de test, des systèmes de TA adaptés avec les données parallèles générées par les méthodes *PhrExtract +TER_ filter* et *SentExtract +LLR_ lex* à partir des corpus *TED* et *Euronews*. Dans les tableaux 5.7 et 5.8 nous présentons les quantités de données extraites en utilisant les différentes méthodes.

Nous pouvons remarquer que les données extraites et filtrées par la méthode de lexique LLR améliorent les résultats de systèmes de base de façon presque équivalente à notre ancienne méthode. En effet, notre méthode proposée avec le filtrage LLR atteint pour les données TED 23.63 comme score BLEU sur *devTED* et 24.88 sur *tstTED* d'après le tableau 5.5. Ces résultats sont comparables aux meilleurs résultats de la méthode *PhrExtract +TER_ filter* où on atteint 23.70 de score BLEU sur *devTED* et 24.84 sur *devTED*. Ce qui prouve l'efficacité en terme de qualité de données extraites de la nouvelle méthode qui ne demande pas beaucoup de temps d'expérimentation pour chercher le meilleur seuil comme pour la technique *PhrExtract +TER_ filter*.

Une autre différence qu'on a pu remarquer entre les deux méthodes réside dans la quantité de données extraites. D'après le tableau 5.7, avec la méthode *SenExtract*

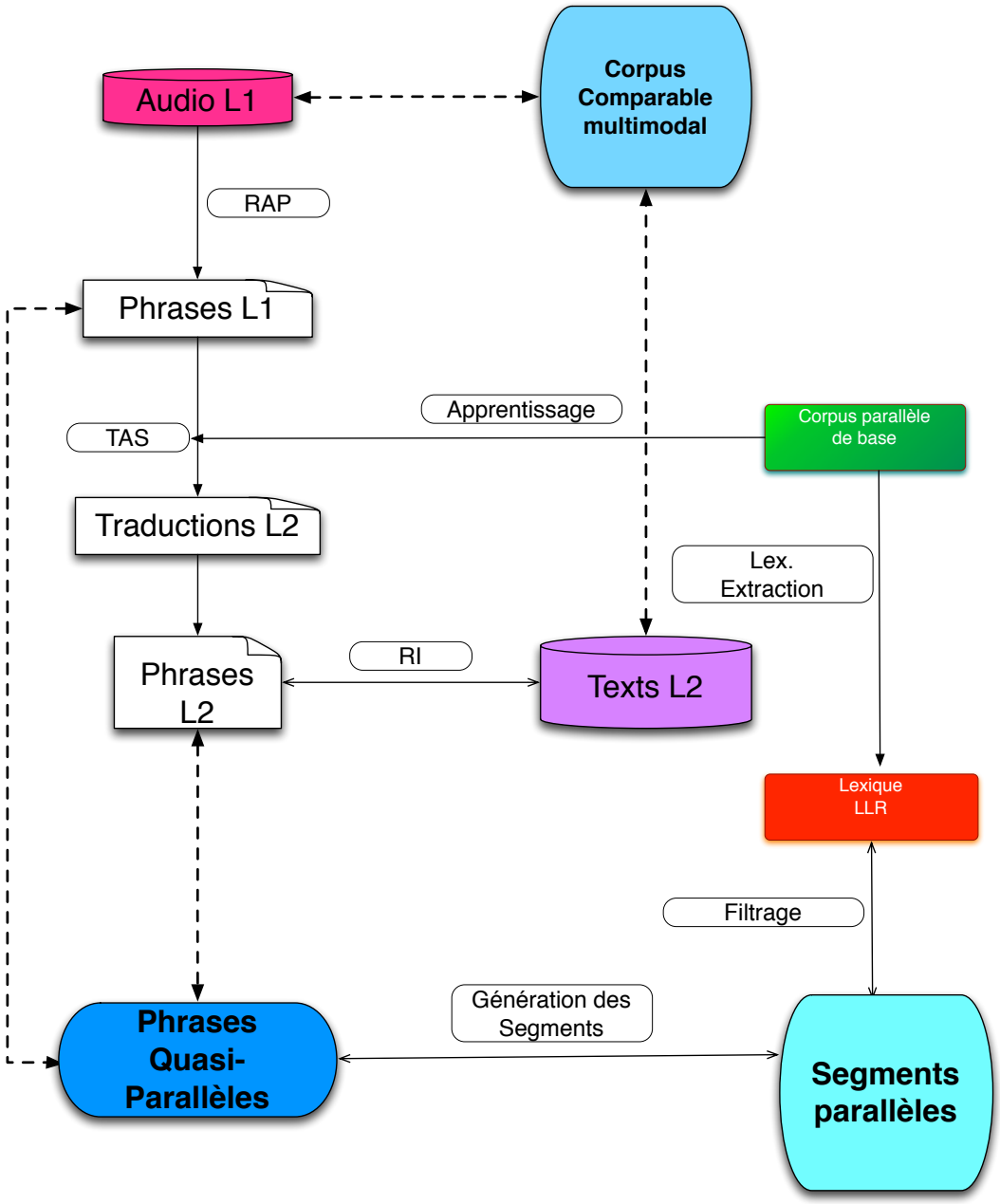

FIGURE 5.10 – Principe du système d'extraction des segments parallèles avec la méthode *SentExtract +LLR_lex*.

+LLR_lex nous avons pu extraire une quantité de données parallèles égale à 16% de la quantité extraite par la méthode *PhrExtract +TER_filter* de seuil 30 TER. Ce qui prouve que ce nouveau filtrage avec lexique sert à bien sélectionner les données les plus importantes dans l'adaptation du système de TA.

En changeant le domaine et le type de données, les résultats de tableaux 5.5 et 5.6 montre la stabilité de notre système d'extraction avec les différentes méthodes en terme de qualité de données extraites.

Systèmes	devTED	tstTED
Baseline	22.93	23.96
PhrExtract +TER_filter (TER 60)	23.70	24.84
SenExtract +LLR_lex	23.63	24.88

TABLE 5.5 – Score BLEU obtenu sur devTED et tstTED avec les méthodes d'extraction des entités sous phrastiques *PhrExtract +TER_filter* et *SentExtract +LLR_lex* (données TED).

Systèmes	devEuronews	tstEuronews
Baseline	25.19	22.12
PhrExtract +TER_filter (TER 30)	30.04	27.59
SenExtract +LLR_lex	30.00	27.47

TABLE 5.6 – Score BLEU obtenu sur devEuronews et tstEuronews avec les méthodes d'extraction des entités sous phrastiques *PhrExtract +TER_filter*et *SentExtract +LLR_lex* (données Euronews).

Méthodes	# mots (en)	# mots (fr)
PhrExtract +TER_filter (TER 60)	16.6 M	13.8 M
SenExtract +LLR_lex	1.6 M	2.2 M

TABLE 5.7 – Quantités de données extraites avec les méthodes *PhrExtract +TER_filter* et *SentExtract +LLR_lex* (données TED).

Méthodes	# mots (en)	# mots (fr)
PhrExtract +TER_filter (TER 30)	321 k	322 k
SenExtract +LLR_lex	636 k	224 k

TABLE 5.8 – Quantités de données extraites avec les méthodes *PhrExtract +TER_filter* et *SentExtract +LLR_lex* (données Euronews).

5.3 Conclusion

Nous avons présenté dans ce chapitre nos contributions en extraction des entités sous-phrastiques. Nous avons comparé le système d'extraction des segments parallèles à partir des données comparables multimodales avec celui des phrases parallèles. Cette comparaison faite sur deux différents corpus a montré l'efficacité de notre méthode pour l'amélioration de la qualité de traduction du système de TA de base avec l'injection des données extraites. Nous avons aussi proposé une extension du module de filtrage où nous avons remplacé le filtrage TER par un module basé sur un lexique. Cette extension a amélioré le système d'extraction en terme de minimisation de nombre d'expériences (pour trouver le seuil optimal) et de la quantité de données extraites, en préservant les mêmes performances de l'amélioration du système de TA. Nous nous intéresserons dans le chapitre suivant aux méthodes d'adaptation du système de TA avec ces données extraites.

Chapitre 6

Adaptation des systèmes de traduction automatique statistique

Sommaire

Dans ce chapitre, nous abordons l'adaptation des modèles de traduction génériques à un domaine particulier en utilisant des données bilingues extraites à partir d'un corpus comparable multimodal.

Nous avons initié l'utilisation des données multimodales pour l'extraction des phrases et segments parallèles. Nous visons dans ce qui suit de ce chapitre de s'intéresser plus à la manière d'utilisation de ces données dans l'adaptation des systèmes de traduction automatique statistique.

6.1 Adaptation d'un système de TA

Plusieurs techniques ont été proposées dans la littérature pour aborder le problème d'adaptation le modèle de traduction. On peut distinguer deux façons d'effectuer cette adaptation : premièrement, on ajoute de nouveaux mots en langue source ou de nouvelles traductions ; et deuxièmement, on modifie les distributions de probabilité du modèle existant pour qu'elles conviennent mieux au domaine. Une technique classique pour adapter un modèle statistique consiste à utiliser un mélange de plusieurs modèles et à optimiser les coefficients d'interpolation à la tâche. Ceci a été étudié par plusieurs auteurs dans le cadre de la traduction statistique, par exemple pour l'alignement des mots [Civera et Juan, 2007], pour la

modélisation linguistique [Zhao *et al.*, 2004, Koehn et Schroeder, 2007], et pour le modèle de traduction [Foster et Kuhn, 2007]. Une autre direction consiste à pondérer automatiquement les corpus d'apprentissage en fonction de leur importance dans le domaine de la tâche de traduction [Shah *et al.*, 2011, Sennrich, 2012].

Notre but final de l'extraction des données parallèles est l'adaptation des systèmes de TAS. Ainsi, l'approche classique que nous l'avons utilisé dans les deux chapitres précédents est de réinjecter les données parallèles extraites dans le système de traduction de base, qui est ensuite utilisé pour traduire les données de test à nouveau. L'évaluation peut ensuite se faire avec une mesure automatique comme BLEU. Nous avons testé d'autres méthodes d'adaptation : la méthode « fill-up », la méthode à plusieurs tables et la méthode d'adaptation non supervisée. Nous présenterons ces méthodes par la suite, afin de comparer et dechoisir la meilleure manière pour chaque cas.

6.1.1 Méthode de combinaison par remplissage « fill-up »

La méthode « fill-up » [Bisazza *et al.*, 2011], est une technique de combinaison de plusieurs tables de traductions.

D'abord, deux modèles séparés sont appris sur deux différents types de données : les données du système de base et les données extraites. Ce qui implique de faire séparément l'alignement des mots, l'extraction des segments et le calcul des probabilités. La différence par rapport aux méthodes classiques d'ajout des données est que au lieu d'augmenter simplement la quantité de données du système de base, la méthode « fill-up » permet d'ajouter seulement les segments de mots qui n'existent pas dans la table prioritaire. Considérons T_1 et T_2 les tables de traductions apprises sur les données de base et les données extraites respectivement. Le modèle de traduction final donne un vecteur caractéristique $\phi(\tilde{s}, \tilde{t})$ pour chaque paire de segments, où $\tilde{s}$ et $\tilde{t}$ sont respectivement le segment source et cible. Donc $\forall(\tilde{s}, \tilde{t}) \in T_1 \cup T_2$:

$$\phi(\tilde{s}, \tilde{t}) = \begin{cases} (\phi_1(\tilde{s}, \tilde{t}), exp(0)) & \text{si} \quad \phi(\tilde{s}, \tilde{t}) \in T_1 \\ (\phi_2(\tilde{s}, \tilde{t}), exp(1)) & \text{sinon} \end{cases}$$

Ce qui implique que les entrées du modèle final du « fill-up » correspondent à l'union des deux tables de traduction. Les segments et les scores sont repris de la table principale, et des nouvelles entrées qui n'existaient pas avant sont ajoutées.

6.1.2 Méthode à plusieurs tables de traduction

Cette méthode consiste à utiliser deux tables de traductions ou plus au même temps (sans les combiner au préalable), où une est principale et la deuxième est utilisée pour les traductions non trouvées. La différence par rapport à la méthode « fill-up » est que les deux tables sont toujours utilisées lors du décodage et ne sont pas combinées dans une seule table comme pour le « fill-up ». Nous appelons par la suite cette méthode « *MT* ». Cette méthode est connue sous le nom de *Backoff* dans l'outil *Moses*.

6.2 Adaptation non supervisée

L'idée de l'adaptation non supervisée proposée pour la première fois par [Ueffing, 2006] consiste à traduire les données de test, à filtrer les traductions avec une mesure de confiance et à utiliser les meilleures traductions pour entraîner un nouveau (petit) modèle de traduction qui est utilisé conjointement avec la table de traduction générique.

Nous utilisons une amélioration de cette approche proposée par [Schwenk et Senellart, 2009] qui se résume dans le protocole suivant :

- Traduction automatique des données côté source avec le système de base
- Choix des traductions à conserver :
  - Selon un seuil sur le score de traduction,
- injection des bitextes dans les données d'apprentissage du système de base.

Les différentes étapes sont présentées dans la figure 6.1 et détaillées dans ce qui suit.

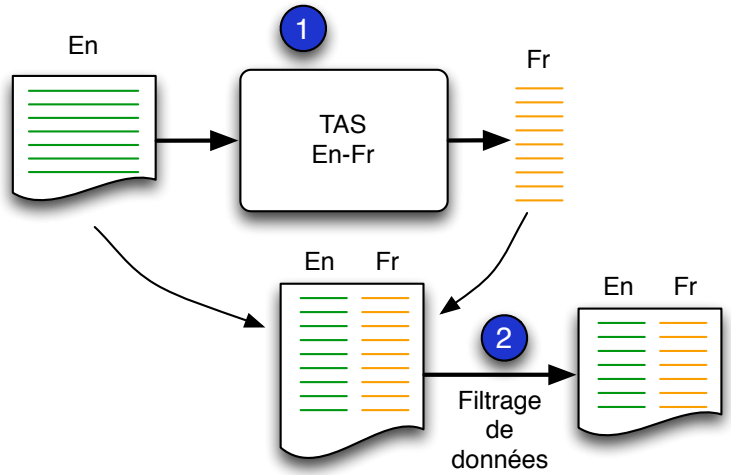

FIGURE 6.1 – Utilisation des ressources monolingues pour générer des bitextes à laide d'un système de TA de base.

Dans notre contexte, nous considérons les données anglaises transcrites comme données monolingues du domaine. Pour traduire ces données, nous avons utilisé le système de base avec toutes les données bilingues parallèles disponibles *Europarl* et *News Commentary*. Une fois les traductions automatiques produites, nous utilisons un filtrage basé sur le score de traduction de l'outil Moses. Ce score donne une indication sur la confiance du système qu'une traduction produite est bonne (la meilleure ayant le plus haut score), cependant ce score n'est pas le mieux corrélé à la qualité effective de la traduction comme nous remarquons dans des expériences de la section 6.3.2. Nous appelons par la suite cette méthode *Unsup*.

6.3 Expériences et résultats

Nos expériences réalisées dans cette étude ont pour but d'adapter et améliorer un système de TAS de l'anglais vers le français partant des données comparables multimodales.

6.3.1 Adaptation des tables de traduction

Dans le chapitre précédent, nous avons présenté les résultats d'adaptation avec l'approche classique d'ajout de données extraites aux données de base. Nous comparons dans cette section cette méthode avec les deux techniques présentés dans la section 6.1.

Dans nos expériences sur les données TED et Euronews, nous avons utilisé les corpus suivants :

- Baseline : corresponds aux données de base.
- TED
  - TED_ter60 : correspond à toutes les données extraites du corpus TED avec la méthode *PhrExtract* de seuil TER 60.
  - TED_ter80 : correspond à toutes les données extraites du corpus TED avec la méthode *SentExtract* de seuil TER 80.
- Euronews
  - Euronews_ter30 : correspond à toutes les données extraites du corpus Euronews avec la méthode *PhrExtract* de seuil TER 30.

Nous avons utilisé ces données pour améliorer le système de base appelé dans cette partie *Baseline*.

Systèmes	DevTED	tstTED
Baseline	22.93	23.96
TED_ter60	19.32	20.71
Baseline + TED_ter60 (ajout)	23.70	24.84
Baseline + TED_ter60 (MT)	22.93	23.85
TED_ter60 + Baseline (MT)	19.66	20.27
Baseline_pt1 + TED_ter60_pt2 (fill-up)	23.23	24.44
TED_ter60_pt1 + Baseline_pt2 (fill-up)	23.16	24.66

TABLE 6.1 – Scores BLEU obtenus des systèmes adaptés avec différentes méthodes d'adaptation (Segments parallèles avec les données TED). pt1 présente la table de traduction principale et pt2 la secondaire.

Pour avoir une idée sur la qualité du corpus extrait, nous avons appris des systèmes TAS en utilisant que les données extraites. Les résultats des tableaux 6.1, 6.2 et 6.3 montrent que ces données peuvent aider à produire de bonnes traductions. En effet, les résultats du système appris seulement avec les données *TED_ter80* ou *TED_ter60* ne sont pas très mauvais surtout en les comparant avec le système de

Systèmes	DevEuronews	TestEuronews
Baseline	25.19	22.12
Euronews_ter30	15.76	14.07
Baseline + Euronews_ter30 (ajout)	30.04	27.59
Baseline + Euronews_ter30 (MT)	25.19	22.13
Euronews_ter30 + Baseline (MT)	18.93	16.56
Baseline_pt1 + Euronews_ter30_pt2 (fill-up)	25.29	22.14
Euronews_ter30_pt1 + Baseline_pt2 (fill-up)	25.22	21.99

TABLE 6.2 – Scores BLEU obtenus des systèmes adaptés avec différentes méthodes d'adaptation (Segments parallèles avec les données Euronews). pt1 présente la table de traduction principale et pt2 la secondaire.

Systèmes	DevTED	TestTED
Baseline	22.93	23.96
TED_ter80	20.97	21.90
Baseline + TED_ter80 (ajout)	23.40	24.69
Baseline + TED_ter80 (MT)	22.86	24.21
TED_ter80 + Baseline (MT)	22.04	22.62
Baseline + TED_ter80 (fill-up)	22.94	23.96
TED_ter80 + Baseline (fill-up)	22.83	23.62

TABLE 6.3 – Scores BLEU obtenus des systèmes adaptés avec différentes méthodes d'adaptation (phrases parallèles des données TED).

base qui est un système de domaine et très performant vu la grande quantité de données utilisées lors de son apprentissage, même qu'il est générique. Pour le cas des données Euronews, la différence est plus grande entre le système de base et le système appris sur les données extraites *Euronews_ ter30* qui sont du même domaine. Ce qui montre que le système d'extraction est plus fiable dans la tâche d'adaptation du système TAS dans certains cas de domaines comme celui de Eronews. Notre explication est que le système de TA apprend de nouvelles bonne traductions et vocabulaire des rapports et articles transcrits de ce site. Ce résultat est spécifique pour les données extraites des corpus multimodaux, car ces méthodes ont été bien appliquées dans la littérature avec d'autres modalités de corpus comme le texte.

Comparant les différentes méthodes d'adaptation, nous pouvons remarquer que la méthode d'ajout des données reste la meilleure technique d'adaptation pour ce type de données. Les tableaux 6.1 et 6.3 montrent que le meilleur score BLEU atteint est 23.70 dans le cas d'ajout des segments parallèle lors de l'apprentissage. Avec les différentes combinaisons des méthodes *fill-up* et *MT*, nous n'avons pas dépassé le score du système de base. Le même résultat pour les données Euronews présentées

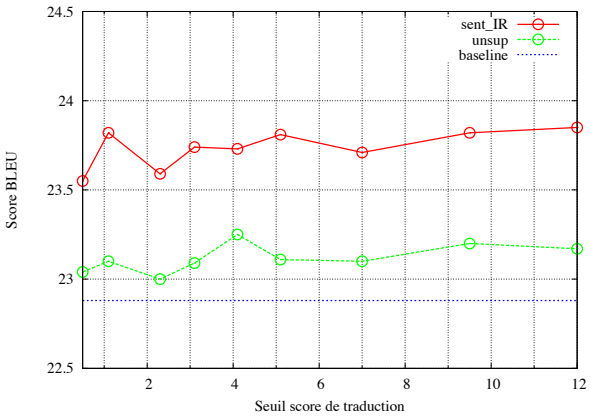

FIGURE 6.2 – Evolution du score BLEU calculé sur les données DevTED après l'adaptation d'un système de base appris sur eparl7nc7.

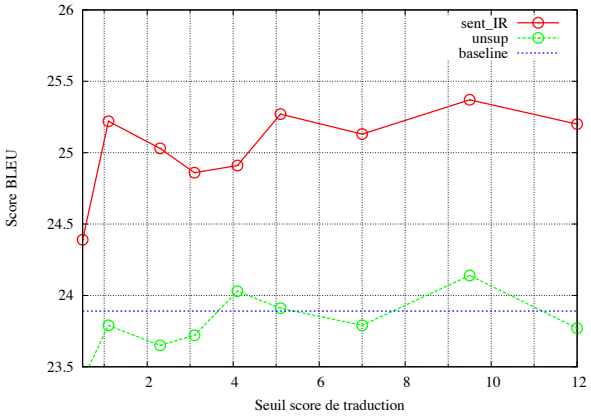

FIGURE 6.3 – Evolution du score BLEU calculé sur les données TestTED après l'adaptation d'un système de base appris sur eparl7nc7.

dans le tableau 6.2, où la différence est de 5 points de BLEU, à peu près, entre la méthode d'ajout et les autres méthodes.

6.3.2 Adaptation non supervisée et recherche d'information

Dans nos systèmes, nous utilisons un module RI pour filtrer les phrases obtenues afin d'exclure les paires qui sont très bruitées. Nous comparons dans cette partie cette méthode avec la technique *unsup* où le choix de la phrase se fait, comme nous avons mentionné précédemment, en utilisant le score de traduction. Nous comparons ces deux méthodes, en mesurant leurs impacts sur les performances des données extraites.

Pour cela, nous avons réalisé des expériences d'adaptation d'un système de base appris avec des données génériques. Nous avons utilisé les données TED, où nous avons choisi un seuil de TER 80 qui a été déterminé empiriquement dans le chapitre

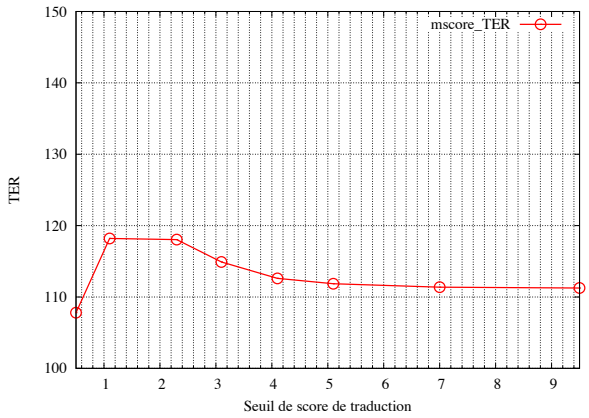

FIGURE 6.4 – Courbe du TER moyen en fonction du seuil de score de traduction.

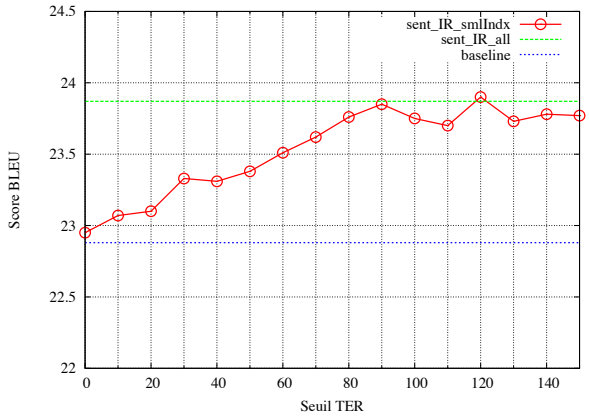

FIGURE 6.5 – Courbe de l'évolution du score BLEU en fonction des seuils TER sur les données DevTED.

4. Les résultats sont présentés dans les figures suivantes. Nous remarquons dans la figure 6.2 que le changement du seuil de traduction n'a pas un effet sur les résultats BLEU avant (en unsup) ou après la RI. Pour tous les seuils, la RI améliore le système de base mieux que l'*unsup* pour ce cadre de travail. Ce résultat est validé par l'évolution du score BLEU sur les données de test (figure 6.3) et l'exemple du tableau 6.4.

Pour comparer le filtrage de la méthode *unsup* basée sur le score de traduction avec le filtrage utilisé dans la méthode de RI, nous avons tracé les courbes de TER moyen en fonction du score de traduction présentées dans la figure 6.4. Nous remarquons que cette courbe est presque linéaire, ce qui montre que ce filtrage ne donne pas une idée exacte sur la qualité des données extraites dans ce cas. Au contraire, les figures 6.5 et 8.8 montrent que le filtrage TER donne une idée sur la qualité des données filtrées.

Ce résultat a été validé aussi dans les expériences précédentes du chapitre 4,

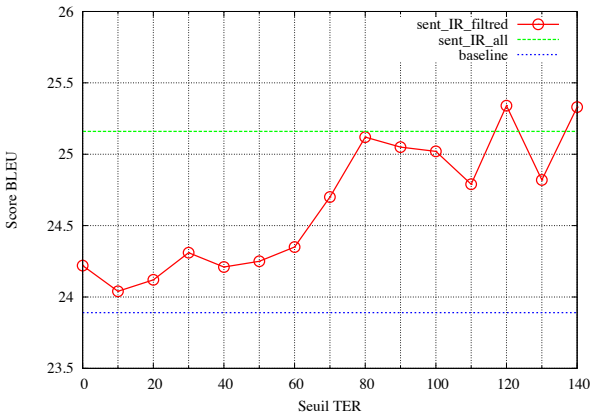

FIGURE 6.6 – Courbe de l'évolution du score BLEU en fonction des seuils TER sur les données TestTED.

Sortie	
Sys. RAP	for me it's a necessity to greece stays in the euro zone and that greece gets the chance to get back on track the problem
Sys. TAS	pour moi une nécessité pour la grèce reste dans la zone euro et que la grèce aura la chance de revenir sur la piste problème
Sys. RI	Je vois la nécessité que la Grèce reste dans la zone euro et que la Grèce aura la chance de se remettre sur pieds .

TABLE 6.4 – Exemple d'amélioration du système de base après l'utilisation du module de la RI.

et dans l'exemple du tableau 6.4. Nous pouvons remarquer dans cet exemple l'importance du module de la recherche d'information dans la correction de la phrase proposée par le module de traduction.

6.4 Conclusion

Dans ce chapitre nous avons présenté nos expériences sur l'adaptation des systèmes de traduction automatique statistiques avec les données extraite à partir des corpus comparables multimodales en utilisant des différentes techniques. Ces expériences ont permis de comparer ces différentes méthodes en fonction des résultats d'amélioration des systèmes après l'utilisation des nouvelles données.

Chapitre 7

Conclusion et perspectives

7.1 Coclusion

Les travaux fondateurs en extraction des données parallèles ont été entrepris pour pallier les problèmes rencontrés en traduction automatique statistique avec les bitextes, notamment en ce qui concerne leur disponibilité. Partant de ce constat, nous avons proposé dans ces travaux d'utiliser des données multimodales multilingues comme ressources pour l'amélioration des systèmes de traduction automatique.

Notre axe de recherche nous a été inspiré à partir des ressources et technologies disponibles au sein du LIUM telles que le système de reconnaissance automatique de la parole, le système de traduction automatique, et les différents corpus parallèles et comparables dans les modalités texte et audio. La tâche retenue tout au long de cette thèse est l'amélioration des systèmes de traduction de l'anglais vers le français. Cette tâche s'inscrit dans le cadre du projet DEPART dont l'un des objectifs est l'exploitation de données multimodales et multilingues pour la traduction automatique.

Les systèmes de traduction automatique développés au cours de cette thèse reposent sur des modèles statistiques, correspondant à l'état de l'art. Nous avons utilisé le système de LIUM de reconnaissance automatique de la parole appris sur les données TED-LIUM. Ce système multi-passes utilise des modèles acoustiques basés sur les modèles de Markov caché. En recherche d'information, nous avons utilisé un système basé sur l'outil *Lemur* qui contient différentes fonctionnalités pour l'indexation et la recherche.

Les méthodes explorées durant ces travaux reposent sur un corpus comparable multimodal, constitué de données audio en langue source et de données textuelles en langue cible. Les données audio sont tout d'abord transcrites par un système de reconnaissance automatique de la Puis, ce système produit une hypothèse de transcription qui est ensuite traduite par le système de TAS. La meilleure hypothèse de traduction est alors utilisée comme requête dans le système de RI, dont le corpus indexé correspond à la partie textuelle en langue cible du corpus comparable multimodal. A noter que la RI s'applique normalement à un ensemble de documents et non à des phrases. Dans notre application, chaque phrase est considérée comme un document, et le paradigme de recherche reste inchangé. Les

résultats de la RI ne sont pas toujours satisfaisants, il est donc nécessaire de filtrer ces résultats afin de ne pas ajouter de phrases non parallèles dans le bitexte final. Nous considérons le Taux d'Edition de la Traduction (Translation Edit Rate - TER) calculé entre les phrases retournées par la RI et la requête comme mesure de filtrage des phrases trouvées. Les phrases ayant un TER supérieur à un certain seuil (déterminé empiriquement) sont exclues. Au final, nous obtenons un bitexte constitué d'une part de la transcription automatique et, d'autre part du résultat de la RI, qui pourra être réinjecté dans le système de base.

Dans une première partie de ces travaux, nous avons mis en évidence la faisabilité de l'approche en appliquant notre chaîne de traitements sur les données TED de la campagne d'évaluation IWSLT'11. Les données bilingues multimodales sont disponibles dans le cadre de cette évaluation. Ces expériences ont montré l'efficacité de notre méthode dans l'extraction des textes parallèles à partir d'un corpus comparable multimodal. Il en ressort que l'enchaînement des modules n'altère que faiblement les résultats, mais le filtrage des résultats de la RI est nécessaire. En effet, ces données ont permis d'adapter un système de traduction de base à un domaine, tel que le domaine TED dans notre cas. Ces résultats préliminaires prometteurs sont à l'origine de l'ouverture de plusieurs perspectives d'amélioration telles que :

- l'application à des domaines réels ;
- le niveau d'extraction (unités plus petites que des phrases) ;
- une meilleure technique de filtrage
- la recherche d'une meilleure méthode d'intégration des données générées.

Nous avons validé notre hypothèse dans une deuxième partie, en effectuant des expériences sur d'autres données d'un domaine différent. Pour ces expériences nous avons construit un corpus comparable multimodale à partir des données du site web Euronews. L'ensemble des documents collectés relève de sept domaines de dépêches. Nous avons développé un outil pour sélectionner automatiquement les vidéos diffusées et leurs textes correspondants (des actualités journalières) pour chaque catégorie dans la période allant de 2010 à 2012. Les vidéos téléchargées sont automatiquement transcrites en utilisant un système de reconnaissance automatique de la parole. Nous avons ainsi automatiquement collecté environ 2,2 millions de mots transcrits et environ 6,2 millions de mots des textes liés aux transcriptions. Ces données seront mises à la disposition de la communauté scientifique. Quant aux données parallèles extraites, elles ont permis l'amélioration du système de traduction de base.

Au cours d'une troisième étape, l'étude a porté sur l'extraction des segments parallèles. Nous avons été motivés par ce choix via l'étude sur les types d'erreurs obtenues avec l'extraction des phrases parallèles, notamment, le nombre de mots à éditer, insérés ou supprimés pour chaque phrase extraite. Nous avons remarqué que les phrases longues ont généralement un TER élevé, et sont donc écartées par le filtrage. Ceci a pour conséquence une réduction de la quantité de données sélectionnées (et donc une perte de couverture) alors que des segments de plus petite taille

sont exploitables à l'intérieur de ces longues phrases.

La différence par rapport à l'extraction des phrases réside dans les segments de données transcrites proposés en groupe de mots avant la traduction. Afin de maintenir une cohérence lors de la recherche d'information, nous avons également segmenté les phrases de la langue cible avec la même méthode. La taille des segments varie de deux à dix mots. Les résultats de ces expériences nous ont permis de souligner l'intérêt de cette approche pour mieux exploiter ce type de données comparables.

Une extension à notre système d'extraction des segments parallèles a été réalisée à travers l'utilisation de la technique de filtrage avec un *rapport de vraisemblance* (LLR Log-Likelihood-Ratio) afin de remplacer le module de filtrage TER. Cette proposition consiste, dans une première étape, en l'utilisation du système d'extraction des phrases quasi parallèles. Puis, dans une seconde étape, de détecter les fragments parallèles dans les phrases générées via l'utilisation du module de filtrage basé sur la *LLR*. Le lexique utilisé dans cette méthode est construit dans notre cas, à partir des bitextes du système de TA de base. Les résultats acquis par cette approche sont comparés avec les meilleurs résultats obtenus par filtrage TER. L'avantage de cette méthode est qu'elle ne repose pas sur un seuil déterminé empiriquement, comme c'est le cas pour la technique utilisant le filtrage TER.

Notre but est l'adaptation des systèmes de TAS. Ainsi, l'approche classique est de réinjecter les données parallèles extraites dans le système de traduction de base, qui est ensuite utilisé pour retraduire les données de test. L'évaluation peut ensuite se faire avec une mesure automatique comme le score BLEU. Nous avons également envisagé deux autres méthodes d'adaptation : la méthode *Fill-up* et l'utilisation de plusieurs tables de traduction dans le décodeur Moses. Nous avons comparé ces méthodes. Les résultats obtenus ont permis de montrer que notre approche actuelle, d'inclusion des données dans les données d'entraînement est celle qui permet de mieux utiliser les données extraites à partir d'un corpus comparable dans le système de TAS.

L'ensemble de nos expériences réalisées a permis de montrer sous quelles conditions un corpus extrait des données multimodales améliore un système de TAS de base. Les approches proposées dans cette thèse peuvent être étendues à d'autres modalités de corpus. Ainsi, toutes les méthodes que nous avons proposées peuvent être généralisées à des données audio parallèles ou autres modalités. Plus généralement, l'application de ces techniques à l'exploitation de tout type de données disponibles en grande quantité comme les vidéos pourrait être une piste intéressante à approfondir.

Cette piste nécessite un pré-filtrage des données avant de commencer le processus d'extraction. Ainsi, un module d'évaluation du degré de similitude des données initiales sera indispensable. Ce module va permettre de mieux choisir les couples de vidéo et textes exploités, afin d'assurer une meilleure comparabilité entre eux.

Plusieurs autres extensions de nos approches sont intéressantes à étudier, comme l'utilisation de l'adaptation incrémentale dans le processus d'extraction pour corriger les erreurs de traduction du système de base, ou l'ajout d'un module d'alignement qui précède le module de traduction.

Pour conclure, nous avons présenté dans cette thèse les premiers travaux, selon notre connaissance, dans l'exploitation des données comparables multimodales et espérons que cet ouvrage ouvrira la voie à ce type de données en vue d'améliorer la traduction automatique.

7.2 Publications

Haithem Afli, Loïc Barrault and Holger Schwenk. Multimodal Comparable Corpora for Machine Translation. 7th Workshop on Building and Using Comparable Corpora, Building Resources for Machine Translation Research, Reykjavik (Iceland) 2014.

Haithem Afli, Loïc Barrault and Holger Schwenk. Multimodal Comparable Corpora as Resources for Extracting Parallel Data : Parallel Phrases Extraction. International Joint Conference on Natural Language Processing, Nogoya(Japan), 14-18 oct 2013.

Haithem Afli, Loïc Barrault and Holger Schwenk . Parallel text extraction from multimodal comparable corpora. 8th International Conference on Natural Language Processing In LNAI 7614, edited by Springer, Heidelberg (2012), Kanazawa(Japan), p.40-51, 22-24 oct 2012.

Haithem Afli, Loïc Barrault et Holger Schwenk. Traduction automatique à partir de corpus comparables : extraction de phrases parallèles à partir de données comparables multimodales. TALN, Grenoble(France), 4-8 juin 2012.

Schwenk Holger, Lambert Patrik, Barrault Loïc, Servan Christophe, Abdul-Rauf Sadaf, Afli Haithem, Shah Kashif. LIUM's SMT Machine Translation Systems for WMT 2011. System description for WMT 2011.

Laurent Besacier, Haithem Afli, Do Thi Ngoc Diep, Hervé Blanchon and Marion Potet . LIG Statistical Machine Translation Systems for IWSLT 2010. International Workshop on Spoken Language Translation , 2-3 December 2010.

Bibliographie

[AbduI-Rauf et Schwenk, 2009] AbduI-Rauf, S. et Schwenk, H. (2009). On the use of comparable corpora to improve smt performance. *In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 16–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Abdul-Rauf et Schwenk, 2011] Abdul-Rauf, S. et Schwenk, H. (2011). Parallel sentence generation from comparable corpora for improved smt. *Machine Translation*.

[Afli *et al.*, 2012] Afli, H., Barrault, L. et Schwenk, H. (2012). Traduction automatique à partir de corpus comparables : extraction de phrases parallèles à partir de données comparables multimodales. *TALN'12*.

[Afli *et al.*, 2013] Afli, H., Barrault, L. et Schwenk, H. (2013). Multimodal comparable corpora as resources for extracting parallel data : Parallel phrases extraction. *International Joint Conference on Natural Language Processing*.

[Afli *et al.*, 2014] Afli, H., Barrault, L. et Schwenk, H. (2014). Multimodal comparable corpora for machine translation. *LREC 2014, 7th Workshop on Building and Using Comparable Corpora, Building Resources for Machine Translation Research*.

[Al-Onaizan *et al.*, 1999] Al-Onaizan, Yaser, Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, F. J. O. D., Purdy, D., Smith, N. et Yarowsky, D. (1999). Statistical machine translation. Rapport technique, Johns Hopkins University.

[Besacier *et al.*, 2010] Besacier, L., Afli, H., Diep, D. T. N., Blanchon, H. et Potet, M. (2010). Lig statistical machine translation systems for iwslt 2010. *International Workshop on Spoken Language Translation*.

[Besacier *et al.*, 2006] Besacier, L., Zhou, B. et Gao, Y. (2006). Towards speech translation of non written languages. *SLT*.

[Bisazza *et al.*, 2011] Bisazza, A., Ruiz, N. et Federico, M. (2011). Fill-up versus interpolation methods for phrase-based smt adaptation. *International Workshop on Spoken Language Translation 2011*.

[Brown *et al.*, 1990] Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L. et Roossin, P. S. (1990). A statistical approach to machine translation. *Comput. Linguist.*, 16:79–85.

[Brown *et al.*, 1991] Brown, P. F., Lai, J. C. et Mercer, R. L. (1991). Aligning sentences in parallel corpora. *In Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, pages 169–176.

[Brown *et al.*, 1993] Brown, P. F., Pietra, V. J. D., Pietra, S. A. D. et Mercer, R. L. (1993). The mathematics of statistical machine translation : parameter estimation. *Comput. Linguist.*, 19:263–311.

[Caroline *et al.*, 2007] Caroline, L., Smaïli, K. et Langlois, D. (2007). Building parallel corpora from movies. *Natural Language Processing and Cognitive Science*.

[Cettolo *et al.*, 2010] Cettolo, M., Federico, M. et Bertoldi, N. (2010). Mining parallel fragments from comparable texts. *Proceedings of the 7th International Workshop on Spoken Language Translation*.

[Chiang, 2007] Chiang, D. (2007). Hierarchical phrase-based translation. *Comput. Linguist.*, 33:201–228.

[Civera et Juan, 2007] Civera, J. et Juan, A. (2007). Domain adaptation in statistical machine translation with mixture modelling. *In Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 177–180.

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas*, 20:27–46.

[Davis et Dunning, 1995] Davis, M. W. et Dunning, T. E. (1995). A trec evaluation of query translation methods for multi-lingual text retrieval. *Fourth Text Retrieval Conference*, pages 483–498.

[Déjean et Gaussier, 2002] Déjean, H. et Gaussier, É. (2002). Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables. *VÉRONIS J., directeur de la publication : Lexicometrica, Alignement lexical dans les corpus multilingues*, pages 1–22.

[Deléglise *et al.*, 2009] Deléglise, P., Estève, Y., Meignier, S. et Merlin, T. (2009). Improvements to the LIUM french ASR system based on CMU Sphinx : what helps to significantly reduce the word error rate ? *In Interspeech 2009*, Brighton (United Kingdom).

[Diab et Resnik, 2002] Diab, M. et Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 255–262.

[Do *et al.*, 2010] Do, T. N. D., Besacier, L. et Castelli, E. (2010). Apprentissage non supervisé pour la traduction automatique : application à un couple de langues peu doté. *TALN 2010, Montréal*.

[Dunning, 1993] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74.

[Estève, 2009] Estève, Y. (2009). *Traitementautomatiquedelaparole :contributions,dansHabilitation à Diriger des Recherches (HDR)*. Thèse de doctorat, LIUM, Université du Maine, France.

[Foster et Kuhn, 2007] Foster, G. et Kuhn, R. (2007). Mixture-model adaptation for smt. *EMNLP*.

[Fung et Cheung, 2004] Fung, P. et Cheung, P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. *In Proceedings of the 20th international conference on Computational Linguistics*, COLING '04.

[Fung et Yee, 1998] Fung, P. et Yee, L. Y. (1998). An ir approach for translating new words from nonparallel, comparable texts. *In Proceedings of the 17th international conference on Computational linguistics - Volume 1*, COLING '98, pages 414–420.

[Gahbiche-Braham *et al.*, 2011] Gahbiche-Braham, S., Bonneau-Maynard, H. et Yvon, F. (2011). Two ways to use a noisy parallel news corpus for improving statistical machine translation. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, BUCC '11, pages 44–51.

[Gale et Church, 1991] Gale, W. A. et Church, K. W. (1991). Identifying word correspondence in parallel texts. *In Proceedings of the workshop on Speech and Natural Language*, HLT '91, pages 152–157, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Gale et Church, 1993] Gale, W. A. et Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19:75–102.

[Gaussier *et al.*, 2004] Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C. et Déjean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04.

[Harris, 1988] Harris, B. (1988). 'bi-text, a new concept in translation theory'. *Language Monthly*, pages 8–10.

[Hewavitharana et Vogel, 2011] Hewavitharana, S. et Vogel, S. (2011). Extracting parallel phrases from comparable data. *In Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, BUCC '11, pages 61–68.

[Hutchins, 2004] Hutchins, J. (2004). Machine translation : from real users to research. *6th conference of the Association for Machine Translation in the Americas, AMTA 2004, Washington, DC, USA*, pages 102–114.

[Koehn, 2004] Koehn, P. (2004). Pharaoh : A beam search decoder for phrase-based statistical machine translation models. *AMTA*.

[Koehn, 2005] Koehn, P. (2005). Europarl : A parallel corpus for statistical machine translation. *Proceedings of MT Summit*.

[Koehn *et al.*, 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. et Herbst, E. (2007). Moses : open source toolkit for statistical machine translation. *In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180.

[Koehn et Knight, 2000] Koehn, P. et Knight, K. (2000). Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. *In Proceedings of the Seventeenth National Conference on Artificial Intelligence*

and Twelfth Conference on Innovative Applications of Artificial Intelligence, pages 711–715. AAAI Press.

[Koehn *et al.*, 2003] Koehn, P., Och, F. J. et 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*, NAACL '03, pages 48–54.

[Koehn et Schroeder, 2007] Koehn, P. et Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. *In Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 224–227, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Lavie et Agarwal, 2007] Lavie, A. et Agarwal, A. (2007). Meteor : an automatic metric for mt evaluation with high levels of correlation with human judgments. *In Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231.

[Lee *et al.*, 1989] Lee, K.-F., Hon, H.-W. et Hwang, M.-Y. (1989). Recent progress in the sphinx speech recognition system. *In Proceedings of the workshop on Speech and Natural Language*, HLT '89, pages 125–130.

[Levenshtein, 1966] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, pages 10 :707–710.

[Li et Gaussier, 2010] Li, B. et Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *In Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 644–652.

[Munteanu et Marcu, 2005] Munteanu, D. S. et Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, 31(4):477–504.

[Munteanu et Marcu, 2006] Munteanu, D. S. et Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. *In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 81–88.

[Och, 1999] Och, F. J. (1999). An efficient method for determining bilingual word classes. *In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, pages 71–76.

[Och, 2003] Och, F. J. (2003). Minimum error rate training in statistical machine translation. *In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Och et Ney, 2003] Och, F. J. et Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29:19–51.

[Papineni *et al.*, 2002] Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. *In Proceedings*

of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pages 311–318.

[Paulik et Waibel, 2009] Paulik, M. et Waibel, A. (2009). Automatic translation from parallel speech : Simultaneous interpretation as mt training data. *ASRU, Merano, Italy.*

[Paulik et Waibel, 2013] Paulik, M. et Waibel, A. (2013). Training speech translation from audio recordings of interpreter-mediated communication. *Comput. Speech Lang.*, 27(2):455–474.

[Prochasson, 2009] Prochasson, E. (2009). *Alignement multilingue en corpus comparables spécialisés Caractérisation terminologique multilingue.* Thèse de doctorat, Université de Nantes.

[Quirk *et al.*, 2007] Quirk, Q., Udupa, R. et Menezes, A. (2007). Generative models of noisy translations with applications to parallel fragment extraction. *In In Proceedings of MT Summit XI, European Association for Machine Translation.*

[Ravishankar *et al.*, 2000] Ravishankar, M., Singh, R., Raj, B. et Stern, R. M. (2000). The 1999 cmu 10x real time broadcast news transcription system. *DARPA Workshop on Automatic Transcription of Broadcast News.*

[Resnik et Smith, 2003] Resnik, P. et Smith, N. A. (2003). The web as a parallel corpus. *Comput. Linguist.*, 29:349–380.

[Riesa et Marcu, 2012] Riesa, J. et Marcu, D. (2012). Automatic parallel fragment extraction from noisy data. *In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, NAACL HLT '12, pages 538–542.

[Rousseau *et al.*, 2011] Rousseau, A., Bougares, F., Deléglise, P., Schwenk, H. et Estève, Y. (2011). LIUM's systems for the IWSLT 2011 speech translation tasks. *International Workshop on Spoken Language Translation 2011.*

[Sarikaya *et al.*, 2009] Sarikaya, R., Maskey, S., Zhang, R., Jan, E., Wang, D., Ramabhadran, B. et Roukos, S. (2009). Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. *Interspeech.*

[Schwenk, 2007] Schwenk, H. (2007). Continuous space language models. *Comput. Speech Lang.*, 21:492–518.

[Schwenk et Senellart, 2009] Schwenk, H. et Senellart, J. (2009). Translation model adaptation for an arabic/french news translation system by lightly-supervised training. *MT Summit.*

[Sennrich, 2012] Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 539–549.

[Shah *et al.*, 2011] Shah, K., Barrault, L. L. et Schwenk, H. (2011). Parametric weighting of parallel data for statistical machine translation. *The 5th International Joint Conference on Natural Language Processing.*

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

[Shannon, 2001] Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5:3–55.

[Simard *et al.*, 1993] Simard, M., Foster, G. F. et Isabelle, P. (1993). Using cognates to align sentences in bilingual corpora. *In Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research : distributed computing - Volume 2*, CASCON '93, pages 1071–1082. IBM Press.

[Sinclair, 1996] Sinclair, J. (1996). Preliminary recommendations on corpus typology. Rapport technique, EAGLES (Expert Advisory Group on Language Engineering Standards).

[Snover *et al.*, 2009] Snover, M. G., Madnani, N., Dorr, B. et Schwartz, R. (2009). Ter-plus : paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation*, 23:117–127.

[Snover *et al.*, 2006] Snover, S., Dorr, B., Schwartz, R., Micciulla, M. et Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

[Su et Babych, 2012] Su, F. et Babych, B. (2012). Measuring comparability of documents in non-parallel corpora for efficient extraction of (semi-)parallel translation equivalents. *In Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, EACL 2012, pages 10–19. Association for Computational Linguistics.

[Ueffing, 2006] Ueffing, N. (2006). Using monolingual source-language data to improve mt performance. *IWSLT*.

[Utiyama et Isahara, 2003] Utiyama, M. et Isahara, H. (2003). Reliable measures for aligning japanese-english news articles and sentences. *In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 72–79.

[Vanden Berghen et Bersini, 2005] Vanden Berghen, F. et Bersini, H. (2005). Condor, a new parallel, constrained extension of powell's uobyqa algorithm : Experimental results and comparison with the dfo algorithm. *J. Comput. Appl. Math.*, 181(1):157–175.

[Vauquois et Boitet, 1985] Vauquois, B. et Boitet, C. (1985). Automated translation at grenoble university. *Comput. Linguist.*, 11:28–36.

[Vogel, 2003] Vogel, S. (2003). Using noisy bilingual data for statistical machine translation. *In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 175–178.

[Walker *et al.*, 2004] Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P. et Woelfel, J. (2004). Sphinx-4 : A flexible open source framework for speech recognition. *SUN MICROSYSTEMS*.

[Yang et Li, 2003] Yang, C. C. et Li, K. W. (2003). Automatic construction of english/chinese parallel corpora. *J. Am. Soc. Inf. Sci. Technol.*, 54:730–742.

[Zhao *et al.*, 2004] Zhao, B., ECK, M. et S., V. (2004). Language model adaptation for statistical machine translation with structured query models. *COLING '04 Proceedings of the 20th international conference on Computational Linguistics*.

[Zhao et Vogel, 2002] Zhao, B. et Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. *In Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, Washington, DC, USA. IEEE Computer Society.

CHAPITRE 8

Annexe

8.1 Résultats des expériences de faisabilité

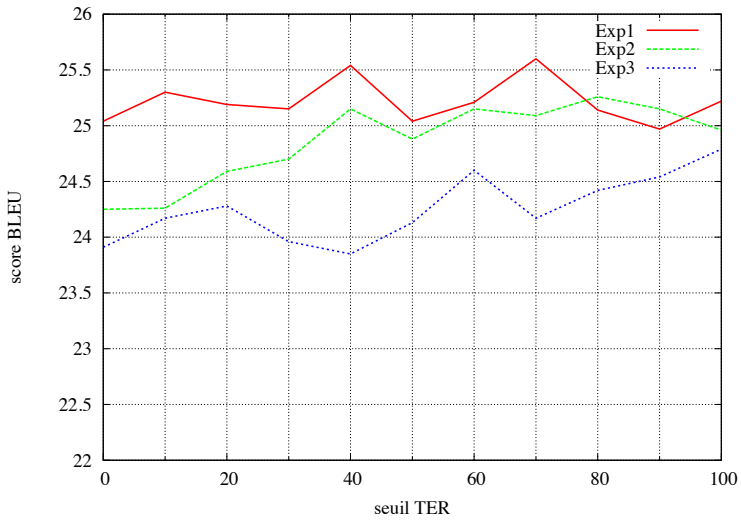

FIGURE 8.1 – Score BLEU obtenu sur TestTED dans les conditions *Exp1*, *Exp1* et *Exp1* pour chaque seuil TER et avec les données génériques + 75% TEDbi.

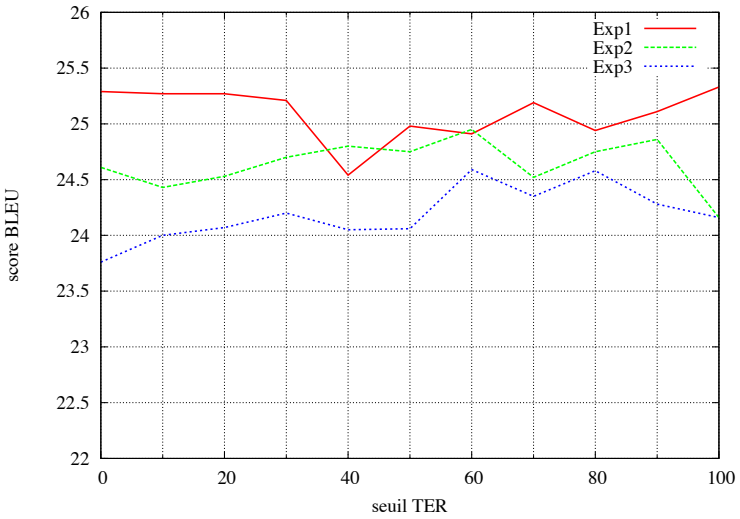

FIGURE 8.2 – Score BLEU obtenu sur TestTED dans les conditions *Exp1*, *Exp1* et *Exp1* pour chaque seuil TER et avec les données génériques + 50% TEDbi.

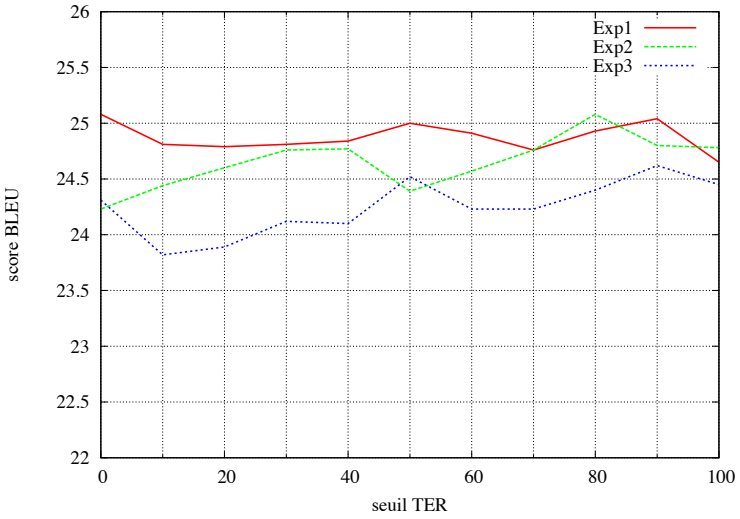

FIGURE 8.3 – Score BLEU obtenu sur TestTED dans les conditions *Exp1*, *Exp1* et *Exp1* pour chaque seuil TER et avec les données génériques + 25% TEDbi.

8.2 Résultats de l'étude des erreurs dans les phrases rejetés par le filtrage TER

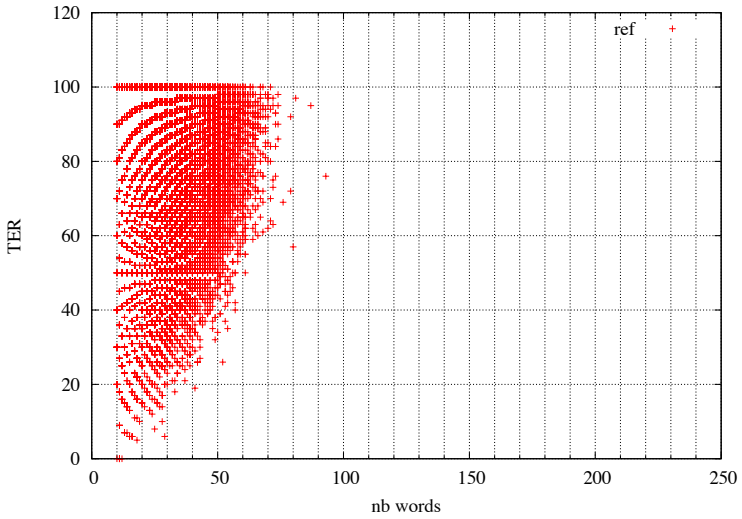

FIGURE 8.4 – Distribution du seuil TER par rapport au nombre de mots de chaque phrase (Corpus TED).

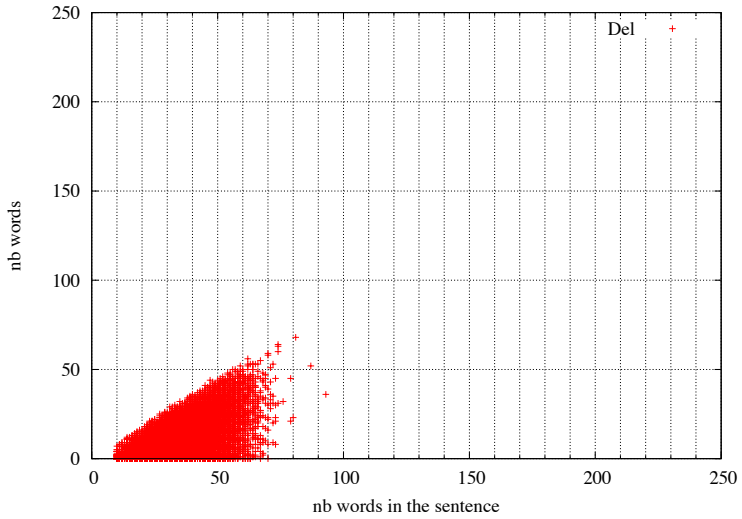

FIGURE 8.5 – Distribution de nombre de mots supprimés pour chaque phrase (Corpus TED).

TER	mots En	mots Fr	nb phrases
40	84	105	3
50	580	712	19
60	3066	3543	85
70	13614	15117	341
80	51997	54524	1183
90	219304	205539	4529
100	798405	665328	15820
All	2148347	4462663	64757

TABLE 8.1 – Nombre de mots et phrases par seuil de TER (données Euronews).

TER	nb phrases	%
0-40	3	0%
40-50	16	0.02%
50-60	66	0.10%
60-70	256	0.39%
70-80	842	1.30%
80-90	3346	5.16%
90-100	11291	17.43%
All	64757	100%

TABLE 8.2 – Nombre de phrases par seuil de TER (données Euronews).

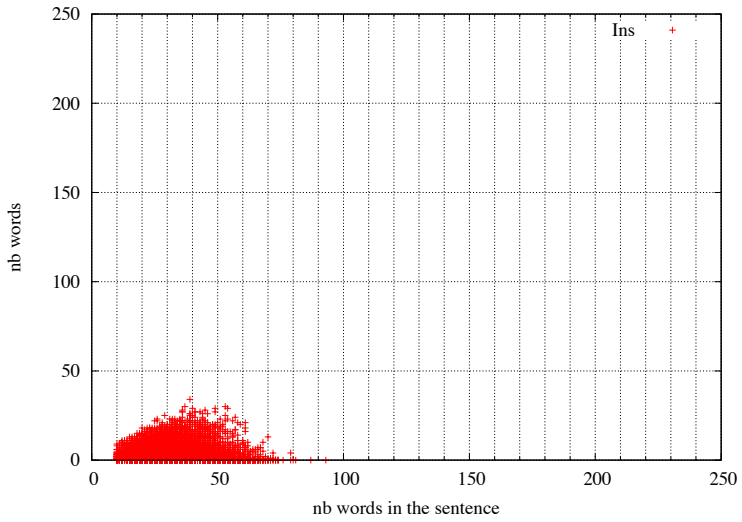

FIGURE 8.6 – Distribution de nombre de mots insérés pour chaque phrase (Corpus TED).

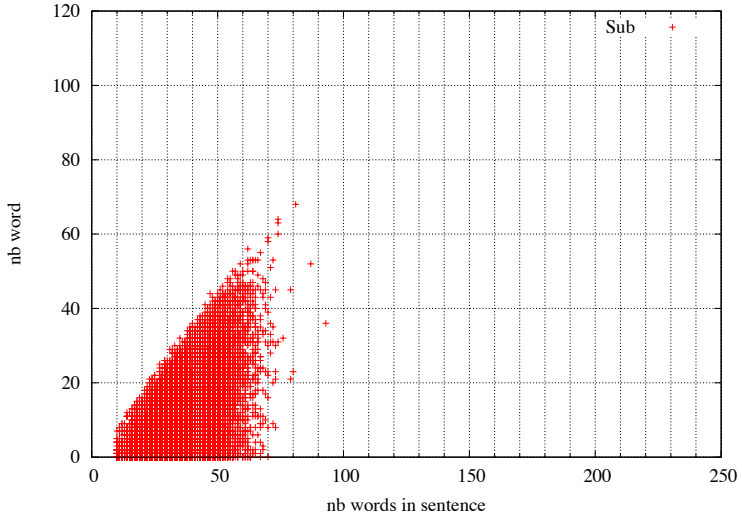

FIGURE 8.7 – Distribution de nombre de mots de substitution pour chaque phrase (Corpus TED)

8.3 Résultats des expériences complémentaires d'adaptation

Bitexts	DevTED	TestTED	
Millnc7	19.50	21.80	1.1 M
TED_ter80	21.00	21.82	461 k
Millnc7 + TED_ter80	22.47	23.39	1.59 M
Millnc7 + TED_unsup	22.48	23.92	3.05 M

TABLE 8.3 – Scores BLEU obtenus avec les systèmes adaptés avec différents bitextes dans les expériences TED (quantité ajouté est comparable à la quantité des données de base).

Bitexts	newstest11	newstest10	# mots
euronews	9.41	9.17	2.14 M
euronews_TER100	9.60	8.33	798 k
euronews_TER90	10.77	9.78	219 k
euronews_unsup	24.48	21.46	2.14 M

TABLE 8.4 – Scores BLEU obtenus sur les données de développement et de test en utilisant que les données extraites avec différentes conditions dans l'apprentissage. (données Euronews)

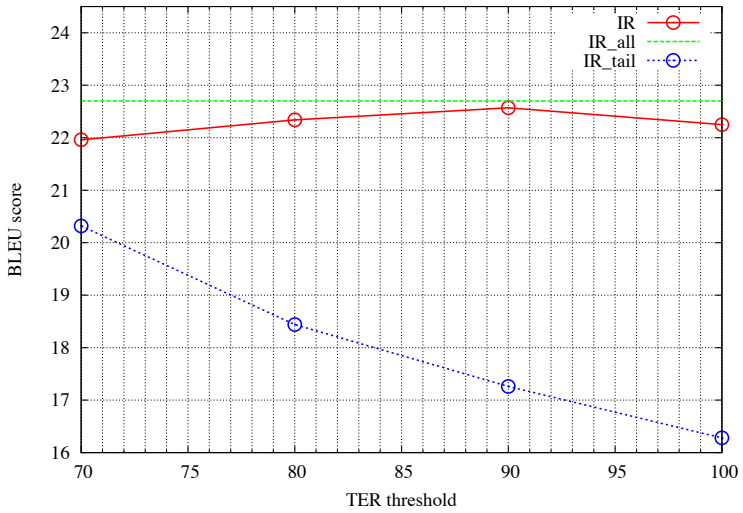

FIGURE 8.8 – Expérience de comparaisons des systèmes construits à partir des bitextes extraies pour chaque seuil et le reste des données non sélectionnées pour le même seuil. Un résultat attendu est que la courbe des données filtrées pour chaque seuil augmente à l'inverse de la courbe du reste des données qui diminuent. (données DevTED)

Résumé :

Les performances des systèmes de traduction automatique statistique dépendent de la disponibilité de textes parallèles bilingues, appelés aussi bitextes. Cependant, les textes parallèles librement disponibles sont aussi des ressources rares : la taille est souvent limitée, la couverture linguistique insuffisante ou le domaine des textes n'est pas approprié. Il y a relativement peu de paires de langues pour lesquelles des corpus parallèles de tailles raisonnables sont disponibles pour certains domaines. L'une des façons pour pallier au manque de données parallèles est d'exploiter les corpus comparables qui sont plus abondants. Les travaux précédents dans ce domaine n'ont été appliqués que pour la modalité texte. La question que nous nous sommes posée durant cette thèse est de savoir si un corpus comparable multimodal permet d'apporter des solutions au manque de données parallèles dans le domaine de la traduction automatique. Dans cette thèse, nous avons étudié comment utiliser des ressources provenant de différentes modalités (texte ou parole) pour le développement d'un système de traduction automatique statistique. Une première partie des contributions consiste à proposer une technique pour l'extraction des données parallèles à partir d'un corpus comparable multimodal (audio et texte). Les enregistrements sont transcrits avec un système de reconnaissance automatique de la parole et traduits avec un système de traduction automatique. Ces traductions sont ensuite utilisées comme requêtes d'un système de recherche d'information pour sélectionner des phrases parallèles sans erreur et générer un bitexte. Dans la deuxième partie des contributions, nous visons l'amélioration de notre méthode en exploitant les entités sous-phrastiques créant ainsi une extension à notre système en vue de générer des segments parallèles. Nous améliorons aussi le module de filtrage. Enfin, nous présentons plusieurs manières d'aborder l'adaptation des systèmes de traduction avec les données extraites. Nos expériences ont été menées sur les données des sites web TED et Euronews qui montrent la faisabilité de nos approches.

Mots clés : traduction automatique statistique, corpus multimodal bilingue, extraction de données parallèles.