Traduire la parole: le cas des TED Talks

Natalia Segal, Hélène Bonneau-Maynard, François Yvon

Traduire la parole: le cas des *TED Talks*

Natalia Segal* — **Hélène Bonneau-Maynard*,**** — **François Yvon***

* *LIMSI-CNRS, 91403 Orsay*
** *Université Paris-Sud, 91400 Orsay*
{natalia.segal,helene.maynard,francois.yvon}@limsi.fr

RÉSUMÉ. *L'amélioration continue des performances des systèmes statistiques de traduction automatique, comme celle des outils de reconnaissance vocale, ouvrent de nouvelles perspectives pour le développement d'applications de traduction automatique de la parole. En nous appuyant sur notre expérience de développement de systèmes pour des traductions de conférences, nous analysons et essayons de quantifier les difficultés principales auxquelles continuent de faire face les développeurs d'outils de traduction de parole, en tout premier lieu le manque de corpus oraux parallèles, et présentons diverses manières de les contourner.*

ABSTRACT. *The continuous improvement of the quality of machine translation and of speech recognition systems opens new perspectives for the development of spoken translation applications. In this study, based on our own experience with the development of Spoken Translation Systems (STS) for conferences, we analyze and quantify the main difficulties raised by STSs, and discuss possible strategies to migitate these issues.*

MOTS-CLÉS : *traduction automatique statistique, traduction automatique de la parole, reconnaissance vocale*

KEYWORDS: *Statistical Machine Translation, Spoken Translation Systems, Automatic Speech Recognition*

1. Introduction

L'arrivée à maturité d'une nouvelle génération de systèmes statistiques de traduction automatique (Koehn, 2010 ; Allauzen et Yvon, 2011), capables de tirer partie des immenses mémoires de traduction accumulées[1] par les institutions internationales (Koehn, 2005), ou bien encore obtenues en exploitant diverses sources multilingues (Skadiņš *et al.*, 2014), a permis de diffuser ces technologies auprès d'un très large public et d'en augmenter l'acceptabilité et l'utilisabilité, en dépit d'une qualité toujours insuffisante. Les technologies vocales (reconnaissance et synthèse) ayant également franchi ce cap, avec la dissémination d'assistants vocaux tels que SIRI[2], le développement de systèmes de traduction automatique de parole (TrAP) pour des domaines ouverts redevient une perspective envisageable.

De fait, alors que les recherches en TrAP s'étaient initialement focalisées sur des domaines extrêmement restreints comme les interactions entre un touriste et un fournisseur de services ou de renseignements (Levin *et al.*, 2000 ; Rayner *et al.*, 2000 ; Wahlster, 2000), ou bien, dans un tout autre domaine, entre des soldats et des civils (Zhou *et al.*, 2013), les travaux plus récents s'attaquent à des énoncés moins contraints : traductions d'émissions radiotélévisées dans le cadre du projet GALE (Olive *et al.*, 2011), traduction automatique de cours et conférences (voir, par exemple, les travaux menés au sein des projets européens EuBridge[3] et TransLectures[4]), de sous-titres (Driesen et Renals, 2013) ou bien encore de conversations libres entre proches (Bangalore *et al.*, 2012).

La TrAP se distingue de la TA classique par bien des aspects et en premier lieu par la modalité de communication, qui non seulement questionne la nature même de ce qui doit être traduit (le seul contenu verbal ? les disfluences ? la prosodie ? les gestes et expressions faciales ?) mais également impose de nouvelles contraintes à la fois en matière de modélisation (intégration de l'ensemble des niveaux de traitement, depuis le niveau phonétique jusqu'au niveau discursif (Seligman, 2000)) et en matière de calcul (optimisation de la vitesse de traitement pour permettre des interactions plus naturelles, amélioration de la robustesse au bruit et aux disfluences, limitation des capacités de calcul sur des dispositifs mobiles, etc.) ; en revanche, elle autorise, dans certaines situations d'interaction, des stratégies dialogiques permettant de lever certaines ambiguïtés (Ayan *et al.*, 2013 ; Hewavitharana *et al.*, 2014).

Du point de vue de l'organisation des traitements automatiques, la TrAP est le plus souvent envisagée comme un *pipe-line* impliquant l'enchaînement successif de modules spécialisés gérant les opérations de reconnaissance automatique de la parole (RAP), de traduction automatique, enfin de synthèse vocale. Un enjeu important est alors celui des interfaces entre ces différentes couches de traitement, qu'il s'agisse

1. Au moins pour les principales langues « à vocation internationale », la situation étant bien moins satisfaisante pour les langues dites « peu dotées » (Besacier *et al.*, 2006).

2. `http://www.apple.com/ios/siri/`

3. `http://www.eu-bridge.eu/`

4. `https://www.translectures.eu/`

de modéliser les interactions entre RAP et traduction (voir, par exemple, (Matusov et Ney, 2011)) ou entre traduction et synthèse (Hashimoto *et al.*, 2012).

Dans cet article, nous nous intéressons à l'analyse de ces différences entre les ressources (souvent textuelles) disponibles pour apprendre les systèmes de TrAP, et les productions vocales qui sont traduites. Nous le faisons dans un contexte simplifié par rapport à des situations d'interaction en face-à-face, celui de la traduction des conférences *TED Talks*, pour lesquelles un certain nombre de difficultés sont affaiblies : il s'agit ici de parole préparée plutôt que spontanée, il n'y a pas d'impératif de traitement en temps réel, pas de restitution sonore des documents traduits, etc.

Notre contribution principale est une tentative de quantifier l'impact de ces différences sur les performances des systèmes et de pointer les voies d'amélioration les plus prometteuses, en nous focalisant sur deux sources de divergences : l'une, linguistique, induite par les différences entre langue écrite et langue orale, telles qu'elle se manifestent, par exemple, dans la structuration des énoncés ; l'autre, induite par l'architecture en cascade des systèmes de TrAP, impliquant que des erreurs du module de reconnaissance, plus ou moins sérieuses, viennent compliquer la tâche du module de traduction. Au rebours, nous discuterons peu des interfaces entre modules, en nous contentant d'exploiter une architecture standard qui enchaîne les deux traitements.

Cet article est organisé comme suit : nous commençons (section 2) par brosser un large panorama des recherches en traduction de la parole, en insistant sur les divergences entre transcriptions automatiques d'une part, données parallèles d'autre part, et sur les différentes manières de les combler. La partie expérimentale de l'article débute par l'introduction des principales sources de données et des outils utilisés (section 3), suivie de l'analyse des sources d'erreurs (section 4). Nous présentons ensuite des expériences qui visent à mettre en évidence et à limiter l'effet des erreurs de ponctuation (section 5), de reconnaissance vocale (section 6), et de segmentation (section 7). Diverses perspectives ouvertes par cette étude sont listées à la section 8.

2. Traduction de la parole : quelques problèmes et solutions

2.1. *Limites de l'étude*

En toute généralité, la conception d'applications de traduction de parole pose de nombreuses questions à la fois de nature linguistique et de nature plus technologique. Comme annoncé *supra*, nous l'abordons ici en posant un certain nombre d'hypothèses et de contraintes, qui permettront de mieux sérier les grandes difficultés et de présenter les manières dont ces difficultés sont abordées dans les travaux de l'état de l'art. Les principales hypothèses qui circonscrivent notre cadre de travail sont les suivantes :

(a) l'application visée se limite à produire une retranscription écrite en langue cible des enregistrements en langue source, ce qui nous autorise à mettre de côté les questions liées au transfert en langue cible du matériau suprasegmental et non verbal ; ainsi que les questions plus techniques relatives à l'enchaînement des opérations de traduc-

tion et de synthèse (Hashimoto *et al.*, 2012) ;

(b) la situation étudiée est celle de l'exposé oral donné par un locuteur unique, bien rompu à l'exercice et présentant un contenu très préparé, capté, qui plus est, dans de bonnes conditions d'enregistrement. Ceci nous dispensera en particulier d'approfondir d'une part les questions liées au traitement d'énoncés spontanés ou encore à la reconnaissance dans le bruit, qui posent toutes deux de grandes difficultés à la RAP, d'autre part les problèmes relatifs à la gestion du dialogue et de l'intrication des tours de parole ; d'un point de vue plus technique, ceci nous permettra de faire l'impasse sur les questions relatives à la traduction en temps réel, et plus généralement à la gestion de la « désynchronisation » temporelle des énoncés en langue source et cible. Cela explique également que nous n'aborderons pas la possibilité de mettre en œuvre des méthodes de compréhension de parole (Lefèvre *et al.*, 2012) sur les énoncés traduits ;

(c) les systèmes de traduction automatique que nous étudions sont des *systèmes statistiques*, dont l'apprentissage présuppose l'existence de données parallèles. Ceci justifie l'angle principal selon lequel nous organiserons notre présentation du domaine, celui de la réconciliation des entrées du système de traduction telles qu'elles sont délivrées par la RAP, avec les données parallèles, en leur immense majorité des textes écrits, qui servent à entraîner les systèmes de traduction statistique.

Conformément à l'hypothèse (c), notre présentation commencera par rapidement représenter les grandes différences entre langue orale et langue écrite, qui correspondent d'un certain point de vue à des écarts qui ne sont pas imputables aux dispositifs technologiques employés pour réaliser la traduction de la parole, et qu'il serait illusoire de chercher à combler en améliorant ces derniers (section 2.2). Nous aborderons ensuite les couches de traitement appliquées pour transformer le signal enregistré en un écrit conforme aux attentes de la traduction, en abordant tout d'abord l'étape de reconnaissance (section 2.3), de segmentation (section 2.5), puis finalement de reponctuation (section 2.6).

2.2. *Traduire la parole*

Le traitement automatique d'énoncés oraux, que ce soit à des fins de compréhension ou de traduction, présente de nombreuses difficultés. C'est tout particulièrement vrai lorsque l'on considère des énoncés spontanés, qui se distinguent de la parole préparée sous de nombreux aspects : par leur structure syntaxique et prosodique, par la présence de disfluences (dysfonctionnements), ainsi que par leur style (de Mareüil *et al.*, 2005 ; Bazillon *et al.*, 2008 ; Blanche-Benveniste et Martin, 2010).

Les systèmes de RAP, développés le plus souvent à partir d'énoncés préparés, ont besoin d'être adaptés à ce type de parole, et le traitement de la parole spontanée reste un des grands défis du domaine. Les études comprennent des tentatives pour détecter ces diverses disfluences de façon automatique (Liu *et al.*, 2005 ; Lease et Charniak, 2006 ; Constant et Dister, 2010), avant de les retirer ou de les réparer (Honal et Schultz, 2005 ; Fitzgerald *et al.*, 2009). Certaines de ces tentatives ont été faites plus

spécifiquement dans un contexte de traduction de parole, dans le but de nettoyer ou d'enrichir le texte source (Wang *et al.*, 2010 ; Cho *et al.*, 2014). Une autre approche pour améliorer la qualité de la reconnaissance de la parole consiste à adapter différents composants du système de reconnaissance à la parole spontanée. Cette adaptation se fonde notamment sur la construction des grands corpus de parole spontanée pour entraîner ou pour adapter incrémentalement des modèles acoustiques et des modèles de langue (Furui *et al.*, 2005).

Les mêmes difficultés se reposent pour la traduction de la parole spontanée, domaine pour lequel le manque de données parallèles appropriées pour apprendre ou adapter des modèles de traduction est encore plus important. La construction de corpus de parole comportant une traduction en langue étrangère est en effet une tâche coûteuse et difficile, car même pour les traducteurs professionnels traduire la parole spontanée en respectant le registre de langue et la structure syntaxique constitue un défi. Les données de sous-titrage, qui commencent à être massivement disponibles (Tiedemann, 2007 ; Lavecchia *et al.*, 2007), constituent une approximation utile de telles données, bien que les contraintes propres à cette activité diffèrent assez fortement des contraintes de la traduction, sans parler du caractère extrêmement bruité des ressources publiquement disponibles. De nombreuses méthodes ont été proposées pour essayer de combler ce manque relatif de corpus parallèles, soit par l'utilisation de corpus comparables (Paulik et Waibel, 2009 ; Afli *et al.*, 2012), soit par l'application de techniques d'adaptation nécessitant moins de données, comme l'apprentissage semi-supervisé (Schwenk, 2008).

Les données utilisées dans cette étude sont des captations de *TED Talks*[5]. Ces conférences grand public, portant sur les sujets technologiques les plus variés, sont délivrées par des experts reconnus et sont le plus souvent extrêmement bien préparées et présentées. Comme discuté *supra*, nous sommes donc assez loin ici de la parole spontanée et les énoncés présentent relativement peu des dysfonctionnements décrits ci-dessus. Ils sont également le plus souvent assez bien structurés du point de vue syntaxique et contiennent très peu d'hésitations et d'autres disfluences.

La suite de cette étude se concentre donc sur l'adaptation des systèmes de traduction pour des données issues d'une retranscription automatique sans qu'il nous ait semblé nécessaire de mettre en œuvre de traitement spécifique pour la parole spontanée, tel que la réparation des disfluences, qui pourrait dans d'autres circonstances faciliter la traduction.

2.3. *Les erreurs de la reconnaissance de la parole*

Les erreurs de reconnaissance vocale constituent naturellement une source majeure de dégradation de qualité pour la traduction automatique, car un texte erroné ne peut pas, dans la plupart des cas, être traduit correctement. Nous commençons donc

5. http://www.ted.com/

cette section par une présentation des erreurs typiques de la RAP ; nous la poursuivons par une étude des différentes manières de limiter l'impact de ces erreurs, en effectuant un couplage plus serré entre ces deux modules de traitement. L'analyse d'erreur sera complétée à la section 4 par un examen des erreurs de notre propre système de reconnaissance et de leur impact sur un système de traduction non adapté aux transcriptions automatiques.

Les travaux consacrés à l'analyse et au typage des erreurs de reconnaissance automatique les plus fréquentes et les plus difficiles à éviter, ont permis de discerner un certain nombre de problèmes communs aux différents systèmes de reconnaissance.

En premier lieu, les mots inconnus (hors vocabulaire) ou peu fréquents, tels que les entités nommées peu courantes, présentent des difficultés pour la reconnaissance (Shinozaki et Furui, 2001). Les diverses propriétés acoustiques du signal de la parole, telles que des variations importantes de la fréquence fondamentale, de la vitesse, de l'énergie, ainsi que la présence de bruit, contribuent également à l'augmentation des taux d'erreurs (Hirschberg *et al.*, 2004 ; Goldwater *et al.*, 2010). Une troisième source de difficulté pour les systèmes de RAP, mise en évidence notamment par Goldwater *et al.* (2010) et Vasilescu *et al.* (2011) est la confusion entre les homonymes ou les mots phonétiquement très similaires. D'après les études psycholinguistiques en reconnaissance de stimuli, plus le mot possède de « voisins » phonétiques et plus ces voisins sont fréquents, plus il est difficile à reconnaître pour les auditeurs humains (Vitevitch et Luce, 1998). En reconnaissance automatique, les résultats de Goldwater *et al.* (2010) montrent que la probabilité qu'un mot soit mal reconnu ne dépend pas seulement du nombre d'homophones ou de mots voisins, mais également du fait que dans certains cas le contexte n'est pas suffisant pour désambiguïser les mots ayant une prononciation similaire. Ces couples de mots, qu'on appelle mots « doublement portant à confusion », possèdent à la fois une prononciation similaire et des contextes d'occurrences semblables. C'est le cas, par exemple, en anglais des couples composés de formes verbales au présent et au passé (*ask/asked*, *says/said*, *watch/watched* et *want/wanted*), ou bien encore de paires dont la forme acoustique réduite est similaire, comme *than/and* et *him/them*.

L'intuition linguistique suggère, bien entendu, que tous les types d'erreurs rapportés ci-dessus ne doivent pas influencer la traduction de la même façon (He *et al.*, 2011). Les substitutions telles que *going/gonna*, ou encore *a/the* sont probablement moins pénalisantes que des remplacements tels que *is/as* ou *I/they* qui peuvent être une source plus importante de dégradation (voir la section 4).

Plusieurs stratégies ont été proposées pour réduire l'effet négatif des erreurs de RAP sur la traduction. Dans un cadre de traduction statistique, nous en distinguons trois : la première consiste à agir sur les données ; la deuxième à agir sur les modèles et leur entraînement ; la troisième, probablement la plus ambitieuse, consiste enfin à enrichir l'interface entre les deux étapes de traitement, en propageant vers le module de traduction tout ou partie des indécisions de la RAP. Ces stratégies sont discutées ci-dessous.

2.4. *Limiter l'impact d'une reconnaissance imparfaite*

2.4.1. *En agissant sur les données*

Rapprocher les transcriptions automatiques imparfaites de textes sources corrects peut être envisagé de deux manières : en corrigeant les transcriptions, ou bien en utilisant des textes sources bruités.

Des tentatives de corriger la sortie de la RAP utilisent des connaissances et des techniques variées. Certains travaux s'inspirent ainsi de méthodes de traduction automatique, où un système de traduction monolingue est entraîné pour « traduire » le texte erroné vers le texte corrigé (Cucu *et al.*, 2013). D'autres s'appuient sur des techniques d'apprentissage automatique qui permettent de capturer diverses caractéristiques linguistiques, syntaxiques ou sémantiques de haut niveau, à l'aide d'un modèle probabiliste conditionnel : les premiers travaux sur ce thème utilisent des modèles « maximum d'entropie » pour lesquels les erreurs sont corrigées localement (Jeong *et al.*, 2005) ; un modèle de correction global utilisant les champs conditionnels aléatoires (CRF, *Conditional Random Fields*) est proposé par Béchet et Favre (2013).

La démarche inverse, consistant à apprendre le système de traduction automatique statistique avec des textes bruités est étudiée par Peitz *et al.* (2012) ; elle demande toutefois des données de parole assorties d'une traduction. En l'absence de données idoines, il est possible de produire des données bruitées artificielles en mettant en œuvre des techniques de traduction automatique pour convertir un texte correct en texte bruité : dans ce nouveau type de traduction monolingue, la source est le texte correct et la cible le texte bruité, le modèle de bruitage étant entraîné en exploitant des corpus de transcriptions automatiques imparfaites. Inversement, Bonneau-Maynard *et al.* (2014) s'inspirent de techniques d'apprentissage non supervisé et proposent d'entraîner un système de traduction sur des textes écrits, puis d'enrichir le corpus d'apprentissage avec des traductions automatiques de transcriptions de corpus oraux.

2.4.2. *En agissant sur les modèles*

Dans ces approches, les système de reconnaissance et de traduction continuent de prendre leurs décisions indépendamment, mais les données et les techniques d'entraînement des paramètres sont adaptées pour rapprocher les deux systèmes (He et Deng, 2011). L'adaptation peut être appliquée aux deux systèmes : le système de reconnaissance peut être modifié pour produire en sortie les données mieux adaptées à la tâche de traduction, et le système de traduction peut être adapté aux données en sortie de reconnaissance.

He *et al.* (2011) montrent ainsi que l'optimisation globale des paramètres du système de RAP en utilisant comme métrique de performances une mesure de qualité de la traduction (BLEU) plutôt que la mesure traditionnelle de qualité de transcription (WER, pour *Word Error Rate*), permet d'obtenir des gains significatifs pour la tâche de traduction de la parole. Une adaptation plus dynamique du système de RAP est présentée par Ng *et al.* (2013), qui envisagent un système en deux passes dans lequel les

meilleures hypothèses de traduction de la première passe sont rétrotraduites en langue source et utilisées pour apprendre un modèle de langue (source) spécifique à chaque phrase, qui sert alors pour réévaluer les treillis en sortie de la reconnaissance vocale.

2.4.3. *En agissant sur les systèmes*

Les méthodes précédentes s'appuient sur une architecture simple qui consiste à enchaîner les étapes de reconnaissance de la parole et de traduction automatique, considérées comme deux traitements indépendants, qui interviennent l'un après l'autre : aucune adaptation de ces composants n'est effectuée et la séquence de mots la plus probable produite par le système de RAP est directement passée au système de traduction. Cette approche, qui découple les deux traitements, est sous-optimale, car elle ne permet pas aux modèles de s'adapter à la tâche et aux données spécifiques de traduction de la parole, ni de bénéficier des connaissances supplémentaires qu'ils pourraient se transmettre mutuellement. Pour améliorer cette approche de base, plusieurs architectures ont été explorées dans la littérature.

– La *propagation d'incertitude* vise à enrichir l'information transmise entre le système RAP et le système de traduction. Le module de RAP peut, par exemple, produire plusieurs hypothèses de reconnaissance sous forme d'une liste de n-meilleures solutions, de treillis de mots ou de réseaux de confusion. Il est également possible de transmettre au système de traduction des informations sur la qualité de reconnaissance sous forme de mesures de confiance. Le système de traduction peut alors utiliser ces informations complémentaires pour choisir la meilleure hypothèse de transcription pour calculer la séquence de traduction la plus probable (Matusov et Ney, 2011).

– L'utilisation d'une *interface phonétique* permet d'impliquer encore davantage le système de traduction (et le modèle de la langue cible qu'il intègre) dans la résolution des incertitudes acoustiques. C'est ce que proposent Jiang *et al.* (2011) et Raybaud (2012) en s'appuyant sur un modèle de traduction qui apparie des séquences de phonèmes en langue source avec des mots ou des groupes de mots en langue cible.

– L'*intégration des modèles* pousse cette logique à son terme et fusionne l'étape de reconnaissance acoustique et l'étape de traduction qui sont réalisées simultanément lors d'une passe de décodage unique (Casacuberta *et al.*, 2002 ; Bangalore et Riccardi, 2002 ; Perez *et al.*, 2012). Le transfert s'effectue alors entre des modèles acoustiques (en source) et des mots (en cible), ce qui rend envisageable d'apprendre le modèle acoustique avec un critère de performances en traduction. Cette approche implique toutefois d'utiliser (en traduction) des modèles de distortion simplistes[6], et nécessite une modification globale des techniques de modélisation et de décodage impliquées, empêchant la réutilisation de composants existants.

6. Pour que la recherche de la meilleure traduction reste computationnellement faisable, il importe de limiter l'espace des réordonnancements possibles de la phrase source.

2.5. *Questions de segmentation*

Les corpus utilisés pour entraîner les systèmes de traduction automatique sont segmentés et alignés au niveau des phrases. Cette segmentation est naturelle pour des textes écrits et relativement facile à déterminer de manière automatique à partir d'indices exploitant la ponctuation et la structure syntaxique.

En revanche, la parole, surtout lorsqu'elle est spontanée, s'organise selon des principes relativement différents et les unités identifiées par les systèmes de reconnaissance vocale sont de nature très différente de celles qui organisent le discours écrit. Un traitement supplémentaire des transcriptions automatiques est donc nécessaire pour reconstruire une segmentation plus semblable à celle qui existe dans les documents écrits. Il est, en particulier, essentiel de considérer des segments qui pourront être traduits de manière indépendante des segments qui les entourent, et qui doivent donc inclure tous les fragments nécessaires à la construction de la structure cible. Parallèlement, il est important de ne pas découper les entrées en unités trop longues, qui sont plus difficiles à traiter pour le système de traduction et engendrent davantage d'erreurs de recherche que les traductions courtes.

Dans le cas particulier de la traduction simultanée de la parole, un enjeu supplémentaire porte sur la latence de la traduction, c'est-à-dire le temps d'attente de l'auditeur entre l'énonciation en langue source et sa traduction en langue cible : l'impératif de devoir commencer à traduire vite demande, en sus, que les segments reconnus soient aussi brefs que possible. Cettolo et Federico (2006) étudient la segmentation optimale de documents écrits, s'appuyant sur des critères de longueur et de ponctuation ; alors que Fügen *et al.* (2007), Rangarajan Sridhar *et al.* (2013), et Finch *et al.* (2014) s'intéressent à des données orales et intègrent également des indices prosodiques (silences, etc) dans leur décision de segmentation.

La question de la segmentation pose également des difficultés au moment de l'évaluation automatique : les traductions de référence s'appuient sur une segmentation de référence, qui n'est pas nécessairement celle que considère le système de reconnaissance vocale. Les deux doivent pourtant être réconciliées, surtout lorsque l'on s'appuie sur des métriques qui comparent une traduction automatique et une traduction humaine.

Deux stratégies prédominent dans la littérature pour contourner cette difficulté : soit le réalignement de toutes les sorties sur une même segmentation de référence, en utilisant par exemple la méthode de resegmentation proposée par Matusov *et al.* (2005), soit l'utilisation d'un score BLEU calculé en aggrégeant toutes les hypothèses de traduction au sein d'un document (Rangarajan Sridhar *et al.*, 2013). Cette dernière approche est celle que nous utiliserons dans les expériences de la section 7, en conservant à l'esprit qu'elle fournit des scores absolus qui surestiment les performances réelles des systèmes.

2.6. *La place de la ponctuation*

Une source supplémentaire d'écart entre les données produites par les systèmes de reconnaissance vocale et les textes parallèles utilisés en traduction automatique statistique est due à l'absence de ponctuation dans les sorties de la RAP. Dans toutes les évaluations en traduction, les références en cible sont ponctuées, pour en faciliter la lecture. Diverses stratégies pour atteindre cet objectif ont été envisagées dans la littérature.

La première consiste à ajouter une ponctuation à la transcription automatique en langue source, en utilisant un modèle de ponctuation dans un traitement intermédiaire entre l'étape de reconnaissance et l'étape de traduction. Les techniques utilisées pour ce faire sont très variées, et peuvent prendre en compte des indices, en particulier prosodiques, extraits du signal de parole, ou bien prendre des décisions uniquement sur la base d'informations textuelles, ou bien encore en fusionnant ces deux sources d'information (Christensen *et al.*, 2001 ; Kim et Woodland, 2001 ; Liu *et al.*, 2006 ; Favre *et al.*, 2009). Dans la mesure où l'ajout de ponctuation peut être vu comme une tâche d'annotation de séquences, il est également possible de l'envisager comme une traduction d'une langue source (non ponctuée) et une langue cible (ponctuée), et mettre en œuvre des techniques de TAS (Cho *et al.*, 2012). D'autres techniques de classification statistique ont également été exploitées pour la même tâche, comme l'utilisation de modèles de maximum d'entropie (Huang et Zweig, 2002). Notons que la ponctuation ainsi produite reste assez limitée et ne correspond qu'approximativement à la ponctuation que l'on trouve dans la partie source des textes écrits utilisés pour l'entraînement des systèmes de traduction. Dans ce schéma, un raffinement supplémentaire consistera donc à déponctuer les données d'apprentissage, puis à les reponctuer automatiquement pour diminuer encore l'écart entre les deux types de textes.

La deuxième manière de combler cet écart consiste, inversement, à supprimer la ponctuation des textes sources utilisés pour l'entraînement, tout en la conservant dans la partie cible. C'est alors le système de traduction qui aura la charge d'insérer des ponctuations de manière conjointe au reste du processus de traduction. Cette stratégie de modélisation implicite constitue une sorte d'adaptation du modèle de traduction (Peitz *et al.*, 2011 ; Ha *et al.*, 2013). Elle présente l'avantage de supprimer une étape de traitement, potentiellement source d'erreurs ; elle rend, au rebours, la tâche de traduction plus complexe, en supprimant de la source des indices potentiellement utiles.

Il est enfin possible d'introduire la ponctuation comme un post-traitement de l'étape de traduction. Le modèle de traduction est ainsi entraîné sur des corpus qui ne sont ponctués ni en source ni en cible, et produit en sortie une traduction sans ponctuation. Un autre modèle monolingue se chargera ensuite de remettre les ponctuations dans le texte traduit, de la même façon que pour la modélisation en source. La différence principale est toutefois que dans cette approche, le modèle de ponctuation ne dispose plus du signal de parole et doit s'appuyer uniquement sur des indices textuels pour effectuer la ponctuation. Une deuxième différence avec la ponctuation en source est que le modèle de ponctuation en cible doit insérer la ponctuation correcte dans un

texte bruité par des erreurs de traduction, en plus des erreurs de reconnaissance. Cette dernière solution s'étant avérée beaucoup moins efficace que les deux autres (Peitz *et al.*, 2011), nous nous sommes contentés, pour cette étude, de comparer l'approche de modélisation explicite en source avec l'approche de modélisation implicite pendant la traduction. Les expériences correspondantes sont décrites à la section 5.

3. Conditions expérimentales

3.1. *Données*

Pour réaliser les expériences présentées dans cette étude, nous avons utilisé les données provenant du site *TED Talks*[7], qui diffuse des conférences de vulgarisation, délivrées par des experts renommés, portant sur des sujets technologiques et sociétaux très variés. Les enregistrements sont préparés et contiennent peu des phénomènes spécifiques à la parole spontanée. Les orateurs sont majoritairement des locuteurs natifs de l'anglais américain, mais une proportion significative des locuteurs est originaire d'autres pays anglophones, voire d'autres zones linguistiques, ce qui fait que le corpus contient une large palette d'accents étrangers. La durée moyenne des présentations est d'environ 20 minutes et chaque exposé est en général présenté par un seul locuteur (quelques conférences très rares impliquent plusieurs intervenants et des éléments de dialogues). Les discours sont retranscrits, puis traduits dans de nombreuses langues par des traducteurs amateurs bénévoles, dont la traduction est toutefois révisée et contrôlée[8]. Nous avons utilisé pour nos expériences les corpus bilingues anglais-français issus des *TED Talks* et préparés dans le cadre de la campagne d'évaluation IWSLT'2013 (Cettolo *et al.*, 2013) pour l'entraînement, le développement et le test des systèmes de traduction.

Les transcriptions manuelles et les traductions de ces corpus ont été resegmentées par les organisateurs des évaluations IWSLT de façon à produire des corpus parallèles. La segmentation ainsi produite est différente de celle des transcriptions et des traductions initiales disponibles sur le site Internet de TED.

Corpus	*Talks*	Lignes	Tokens EN	Tokens FR
Train IWSLT 2013	1 416	179 409	3 534 226	3 883 550
Dev IWSLT 2010	8	887	19 694	20 258
Test IWSLT 2010	11	1 164	31 077	33 887

Tableau 1. *Statistiques des corpus TED Talks*

7. `http://www.ted.com/`
8. Voir `https://www.ted.com/participate/translate/`

3.2. *Boîtes à outils*

3.2.1. *Transcription automatique*

Pour les expériences décrites ci-dessous nous avons utilisé deux systèmes de transcription de la parole développés au LIMSI.

Le système de transcription de base met en œuvre des modèles et stratégies de décodage semblables à ceux du système *Broadcast News* pour l'anglais (Gauvain *et al.*, 2002 ; Lamel, 2012). Il repose sur deux composants principaux : un partitionneur parole et non parole et un décodeur de parole. Les modèles acoustiques, entraînés sur environ 500 heures de données audio, sont dépendants du genre. Ils comprennent environ 30 000 phones en contexte avec 11 600 états liés. Les modèles de langue n-gramme ont été entraînés sur plus de 1,2 milliard de mots à partir de différents corpus produits par le LDC[9] (*English Gigaword*, transcriptions *broadcast news*, transcriptions commerciales), d'articles de journaux téléchargés sur le Web, et de différentes transcriptions audio. Le vocabulaire comprend 78 000 mots, sélectionnés par interpolation de modèles de langue unigrammes entraînés sur différents sous-ensembles de textes de façon à minimiser le taux de mots hors vocabulaire sur un corpus de développement indépendant. Le décodeur de mots utilise des modèles acoustiques à base de HMMs et un modèle de langue bigramme pour construire un treillis de mots, qui est ensuite réévalué par un modèle de langue quadrigramme. Ce système fonctionne en temps réel et n'a pas été adapté à la tâche. Ses performances sur des données *Broadcast News* sont de l'ordre de 12 % d'erreurs et de 19 % sur les données de TED.

Le deuxième système utilisé a été spécifiquement adapté aux données de *TED Talks*. Cette adaptation a concerné à la fois les modèles acoustiques, les modèles de langue et le dictionnaire de prononciation. Le ML n-gramme est obtenu par interpolation du modèle appris sur les données de TED et du modèle du système *Broadcast News*. Le vocabulaire est étendu et atteint approximativement 95 000 mots. Le décodage est toujours effectué en deux passes : pour la première passe, uniquement le ML et le dictionnaire de prononciation adaptés sont utilisés ; pour la seconde, le même ML est utilisé, avec cette fois-ci des modèles acoustiques entraînés exclusivement sur les données de *TED Talks* (180 heures de signal). Le taux d'erreur de ce système adapté est de l'ordre de 12 % sur les données *TED Talks*. Une description plus complète de ce système est donnée dans (Segal *et al.*, 2014).

3.2.2. *Enrichissement des transcriptions : ponctuation et normalisation des nombres*

Les systèmes de transcription automatique de la parole tels qu'ils sont développés aujourd'hui produisent une transcription dépourvue de ponctuation, et dans laquelle tous les nombres sont écrits en toutes lettres.

La conversion des nombres en chiffres arabes et l'insertion des ponctuations sont développées en post-traitement de la RAP dans l'objectif d'enrichir les transcriptions

9. *Linguistic Data Consortium*, `https://www.ldc.upenn.edu`.

automatiques et d'améliorer le confort de lecture de ces transcriptions. Les mêmes procédures de post-traitement s'avèrent également utiles dans le contexte de la traduction de la parole. En effet, les systèmes de traduction automatique standard sont entraînés sur des textes écrits, ponctués et dans lesquels les nombres sont souvent écrits en chiffres. Ces représentations « enrichies » sont particulièrement adaptées pour la traduction automatique, car elles permettent d'optimiser la traduction vers les mêmes représentations du côté cible. Ainsi, afin d'améliorer la qualité de la traduction de la parole, il apparaît nécessaire d'effectuer de tels traitements.

Pour transformer les nombres, nous avons utilisé un algorithme à base de règles. Pour insérer la ponctuation, nous avons développé un système de traduction monolingue, qui transforme les textes non ponctués en textes ponctués. Les meilleurs résultats sont obtenus par un système produisant des ponctuations simples non appariées (point, virgule, point-virgule, deux-points, points d'exclamation et d'interrogation) avec un taux d'erreur de 2,9 %. La transformation des nombres et l'insertion des ponctuations peuvent être appliquées uniquement aux transcriptions automatiques. Elles ne requièrent pas de modification des systèmes de traduction existants. Néanmoins, ces traitements ne produisent pas les résultats exactement équivalents à ce qui apparaît dans des textes écrits : d'une part, les traitements automatiques introduisent un certain nombre d'erreurs, d'autre part, les transcriptions humaines ne suivent pas toujours les mêmes conventions. Pour harmoniser et rapprocher les formats des données en sortie de RAP et en entrée des systèmes de traduction, il peut être préférable d'appliquer les mêmes transformations automatiques à la partie source des corpus d'entraînement et de développement des systèmes de traduction. Il faudra toutefois, dans ce cas, modifier et réentraîner des nouveaux systèmes de traduction spécifiquement pour la tâche de traduction de la parole.

3.2.3. *Traduction automatique*

Le système de traduction statistique utilisé est Ncode. Ce système développé au LIMSI[10] est fondé sur l'approche n-gramme (Casacuberta et Vidal, 2004 ; Mariño *et al.*, 2006 ; Crego *et al.*, 2011). La particularité de cette approche est que le modèle de traduction décompose chaque couple de phrases parallèles comme une séquence d'unités bilingues, dont la probabilité est calculée par un modèle n-gramme. Dans cette approche, il est nécessaire que les phrases sources et cibles aient été préalablement « synchronisées », de façons que les segments sources et leur traduction apparaissent dans le même ordre. Cette synchronisation est effectuée en réordonnant les mots en langue source à l'aide d'un ensemble de règles automatiquement extraites des données d'alignement. Cette approche a régulièrement montré qu'elle permettait d'obtenir des résultats à l'état de l'art lors de compétitions internationales (voir par exemple (Allauzen *et al.*, 2013)). On notera que Ncode exploite, pour apprendre les règles de réordonnancement, un étiquetage en parties du discours (Schmid, 1994) ; cet étiquetage étant par construction moins fiable pour des transcriptions automatiques, il est attendu que Ncode soit doublement pénalisé lorsqu'il traite des données bruitées.

10. `http://ncode.limsi.fr/`

Lors de l'apprentissage, les données parallèles sont alignées au niveau des mots grâce à l'outil MGIZA++ (Gao et Vogel, 2008), qui implémente efficacement les modèles statistiques d'alignement de mots de Brown *et al.* (1993) et Vogel *et al.* (1996). Le réglage des poids des différents modèles est effectué sur le corpus de développement, en utilisant l'approche de Och (2003).

3.3. *Évaluations*

Les performances des systèmes de transcription vocale sont évaluées, de manière usuelle, par la métrique WER, qui calcule un taux d'erreur par mot : les erreurs de reconnaissance sont identifiées en calculant l'alignement optimal entre la transcription et la référence ; toute déviation par rapport à la référence (substitution, insertion, ou suppression) compte pour une erreur.

La qualité de la traduction est évaluée automatiquement par la métrique BLEU (Papineni *et al.*, 2002), qui repose sur une comparaison de surface entre l'hypothèse de traduction et une ou plusieurs traductions de référence. Formellement, soit $e_1 \dots e_J$ une hypothèse de traduction et $r_1 \dots r_L$ une traduction de référence d'une phrase source $f_1 \dots f_J$, on note c_n le nombre de n-grammes apparaissant dans $e_1 \dots e_J$ et m_n le nombre de n-grammes de $e_1 \dots e_J$ qui apparaissent également dans la référence. La *précision n-gramme* p_n est simplement le rapport $\frac{m_n}{c_n}$[11]. Le score BLEU est alors défini comme la moyenne arithmétique des précisions n-grammes, pour n variant entre 1 et 4, à deux détails près :

– si, pour une valeur de n, c_n est nul, alors la moyenne arithmétique $\sqrt[k]{\prod_{n=1}^{4} p_n}$ sera nulle, et l'hypothèse aura un score BLEU égal à zéro, quelles que soient les valeurs des autres précisions n-grammes. Pour éviter de telles situations, la précision n-gramme est calculée sur des ensembles contenant quelques centaines, voire quelques milliers de couples (référence, hypothèse) ;

– il existe un moyen simple d'obtenir de bonnes précisions n-grammes, consistant à produire des hypothèses très courtes : en produisant moins de mots, on a moins de chance de se tromper. Il est donc nécessaire de corriger les valeurs de la précision moyenne en intégrant un facteur (BP pour *brevity penalty* dans la formule [1]) qui pénalise les hypothèses qui seraient trop courtes par rapport aux références.

Au final, la mesure BLEU de l'ensemble d'hypothèses $\mathbf{E}$ est donc donnée par :

$$BLEU(\mathbf{E}) = BP \times \exp(\sum_{n=1}^{4} \frac{1}{k} \log(p_n)) \qquad [1]$$

Par construction, le score BLEU est donc compris entre 0 et 1, cette dernière valeur étant atteinte lorsque l'hypothèse est identique à une des traductions de référence.

11. p_n est en fait la précision n-gramme modifiée, qui se calcule comme $\max(1, \frac{c_n}{m_n})$, afin de ne pas récompenser (à tort) un système qui produirait plusieurs fois un n-gramme correct.

La métrique METEOR[12] de Banerjee et Lavie (2005) sera utilisée secondairement : cette métrique, qui repose également sur un appariement mot à mot optimal d'une hypothèse de traduction et d'une référence, est en effet plus appropriée pour évaluer la qualité de traduction au niveau de phrases isolées. Comme pour BLEU, ces valeurs sont nominalement comprises entre 0 et 1, mais nous reportons ci-dessous des valeurs entre 0 et 100 ; plus elles sont élevées, meilleure est la traduction.

4. Système de base et analyse de l'impact du couplage simpliste transcription/traduction

4.1. *Système de base*

Le système de base (`baseline`) est construit en utilisant les données d'entraînement *TED Talks* pour apprendre le modèle de traduction (MT). Ces données correspondent aux corpus d'entraînement fournis et segmentés par les organisateurs de la campagne d'évaluation IWSLT'2014 (Cettolo *et al.*, 2014). Le modèle de langue (ML) cible est obtenu par interpolation log-linéaire entre un modèle appris avec les données monolingues de *TED Talks* et un grand modèle entraîné sur des corpus monolingues de la campagne WMT 2011 (1,4 milliard de tokens), Lavergne *et al.* (2011). Les deux modèles ont été appris avec l'outil SRILM (Stolcke, 2002).

Les corpus d'entraînement sont ponctués manuellement en source et en cible, de même que les transcriptions manuelles et les traductions manuelles de référence des corpus de développement et de test. En revanche, dans l'idée de fixer un système de base dans lequel le couplage entre RAP et TA sera le plus simpliste possible, les transcriptions automatiques des mêmes données de test sont produites et présentées à la traduction sans aucune ponctuation (hormis un '.' qui est ajouté systématiquement à la fin des phrases). La segmentation en phrases des données d'entraînement, de développement et de test est celle qui est imposée par la campagne d'évaluation IWSLT'2014.

Le tableau 2 permet de comparer les performances du système de traduction de base évaluées en BLEU sur des transcriptions manuelles et sur des transcriptions automatiques produites par les deux systèmes de RAP décrits à la section 3.2.1.

Système	Trans. manuelle Trans. manuelle	Trans. auto de base (WER≈17 %)	Trans. auto adaptée (WER≈12 %)
`baseline`	33,2	20,5	21,9

Tableau 2. *Performances du système de traduction de base sur les données de test transcrites manuellement et automatiquement*

Le système de reconnaissance vocale de base a un taux d'erreur mots de 17 %, ce qui induit une dégradation de 12,7 points BLEU entre les traductions des transcriptions

12. `http://www.cs.cmu.edu/~alavie/METEOR/`

idéales et des transcriptions bruitées. Cette dégradation est due à la fois aux erreurs produites par le système de RAP et à l'absence de ponctuation dans les transcriptions automatiques et sera analysée en détail *supra*. On note que le travail sur l'amélioration de la RAP a permis de réduire de près de 30 % le taux d'erreur ; l'impact en BLEU est bien plus modeste, même s'il permet d'améliorer très sensiblement les performances (près de 1,5 point BLEU).

4.2. *Impact des erreurs de RAP sur les performances en traduction*

Pour analyser l'impact des erreurs de reconnaissance vocale sur la traduction, nous avons étudié la relation entre la performance de RAP au niveau des phrases, mesurée en taux d'erreur sur les mots (WER), et la performance de traduction, mesurée par la différence de score sBLEU[13] entre la traduction de la transcription manuelle et la transcription automatique.

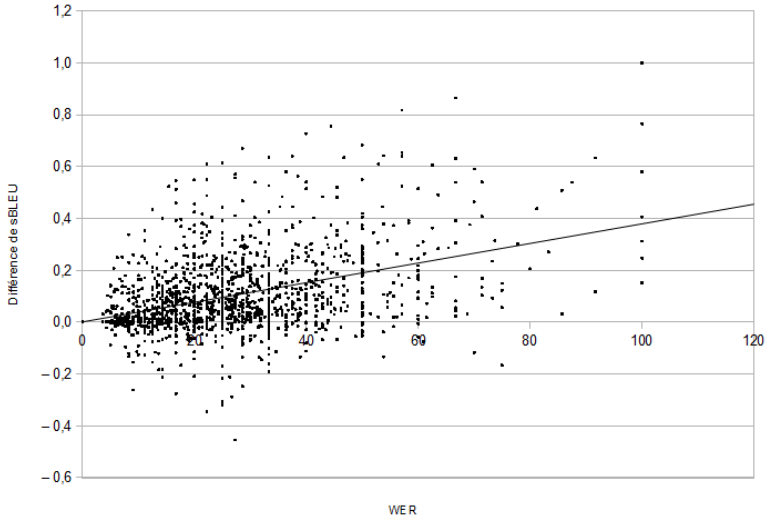

Figure 1. *Pour chaque phrase du corpus de test, on reporte en abscisse le WER et en ordonnée la différence de score sBLEU entre la traduction d'une transcription manuelle et celle de la transcription automatique (avec une ligne de régression). Une valeur positive indique que la première est meilleure que la seconde*

13. La métrique BLEU n'étant pas toujours bien définie au niveau des phrases, nous utilisons une approximation proposée par Lin et Och (2004), qui permet de contourner cette difficulté. Dans la suite cette métrique sera dénotée sBLEU pour *sentence-level BLEU*.

Le graphique 1 permet de visualiser phrase par phrase l'impact des erreurs de transcription sur la qualité de la traduction pour le corpus de test. Globalement, la différence entre les scores sBLEU des traductions des phrases sources propres et bruitées augmente lorsque la qualité de reconnaissance diminue : à partir de WER > 40 %, la traduction de la transcription automatique est presque toujours moins bonne que pour la transcription de référence. Nous pouvons également observer que les valeurs sont assez dispersées, ce qui peut être expliqué en partie par la métrique utilisée (sBLEU), qui n'est pas une mesure très précise de la qualité de traduction.

Notre hypothèse principale est toutefois que toutes les erreurs de reconnaissance n'ont pas le même impact sur la qualité de traduction : dans certains cas, même un faible écart entre les deux sources peut provoquer une différence importante entre les traductions, tandis que dans d'autres cas une différence entre les textes sources peut avoir peu d'impact voire aucun sur la traduction.

Pour pouvoir analyser de façon plus précise l'impact des différentes erreurs sur la qualité de traduction, nous avons considéré séparément les erreurs d'`Ins`(ertion), de `Sub`(stitution), et de `Sup`(pression). Nous avons à cet effet isolé trois sous-ensembles du corpus de test, chacun regroupant toutes les phrases contenant uniquement des erreurs d'un type donné. Nous avons ensuite calculé, pour chaque sous-ensemble, la moyenne de la différence de sBLEU entre la traduction de la source correcte et la traduction de la source erronnée. Nous obtenons ainsi une estimation de la dégradation de la qualité de traduction pour chacun des types d'erreurs. Pour conforter notre analyse, nous avons également inclus ici les scores METEOR. Les résultats sont dans le tableau 3, pour les deux systèmes de RAP décrits à la section 3.2.1.

Type d'erreur	RAP de base			RAP adaptée		
	Nb. phrases	Diff. sBLEU moyenne	Diff. METEOR moyenne	Nb. phrases	Diff. sBLEU moyenne	Diff. METEOR moyenne
`Sub`	347	6,1	9,0	344	4,4	7,7
`Sup`	100	6,2	8,8	120	7,0	12,9
`Ins`	57	1,4	1,2	74	2,8	2,6

Tableau 3. *Impact des erreurs de la RAP sur la traduction automatique, ventilé par type d'erreur*

Les erreurs de RAP les moins pénalisantes pour la traduction sont les insertions (qui correspondent souvent à des mots répétés). Les erreurs les plus gênantes sont les suppressions des mots, qui sont souvent dues au fait que le système de reconnaissance ne conserve pas les mots en dessous d'un seuil de confiance prédéterminé. Ceci permet d'optimiser les performances en termes de WER, au détriment de la traduction : il est en effet difficile de compenser l'absence d'un mot, surtout s'il s'agit d'un mot lexical.

En ce qui concerne la comparaison entre la version de RAP adaptée et non adaptée, notons en préambule que le nombre de phrases par type d'erreur (rappelons qu'il s'agit des phrases contenant *uniquement* des erreurs du type en question) est plus élevé

trans. manuelle	*his name is Paul* **Offit** .
trans. auto	*his name is Paul* **said** .
trad. trans. manuelle	*son nom est Paul Offit* .
trad/ trans. auto	*son nom est Paul a dit* .
trans. manuelle	*which is the history of who invented games and* **why** .
trans. auto	*which is a history of who invented games and* **one** .
trad. trans. manuelle	*c' est l' histoire de qui a inventé les jeux et pourquoi* .
trad. trans. auto	*c' est une histoire qui a inventé les jeux* .
trans. manuelle	*you have not* **changed** *the story* .
trans. auto	*you have not the story* .
trad trans. manuelle	*vous n' avez pas changé l' histoire* .
trad trans. auto	*vous n' avez pas l' histoire* .

Tableau 4. *Des exemples d'erreurs pénalisantes pour la traduction*

pour la version adaptée, bien que le nombre total d'erreurs soit naturellement réduit. L'adaptation de la RAP semble avoir un impact positif sur sBLEU et sur METEOR notamment pour ce qui concerne les substitutions, qui sont à la fois moins nombreuses (le système adapté à taux de substitution de 6,9 %, contre 9,5 % pour le système de base), et moins pénalisantes ; en revanche, on constate que les insertions, quoique moins nombreuses dans la version adaptée (3,1 % contre 4,3 %) restent très gênantes pour la traduction.

Une analyse manuelle des phrases qui ont, pour le système de base, une bonne qualité de reconnaissance (WER < 20 %) et une différence importante entre les scores sBLEU des deux traductions permet d'observer des exemples d'erreurs gênantes pour la traduction (voir le tableau 4) :

– les substitutions, notamment celles qui génèrent des mots hors vocabulaire (comme les entités nommées) ou entre les mots fréquents phonétiquement voisins (*this/the*, *one/why*, *the/a*, *going/gonna*, etc.), mal transcrits par la RAP ;

– les suppressions, notamment les suppressions des mots lexicaux.

Inversement, nous avons également examiné les phrases plutôt bien reconnues (0 < WER < 20 %) et pour lesquelles il n'y avait aucune différence entre les scores sBLEU des deux traductions. Il s'agit notamment de cas d'insertion ou de suppression de mots-outils (voir l'exemple du tableau 5). Ceci illustre le fait que le modèle de traduction et le modèle de langue intègrent implicitement une capacité à compenser certaines erreurs de la reconnaissance.

5. Impact quantitatif de la ponctuation en source

Dans cette section, nous nous intéressons maintenant à quantifier plus précisément d'une part l'impact des incohérences entre les transcriptions automatiques et les transcriptions manuelles (absence de ponctuation, incohérence de format et erreurs de

trans. manuelle	*and* **there** *have been other efforts along those lines .*
trans. auto	*and* **there there** *have been other efforts along those lines .*
trad. trans. manuelle	*et il y a eu d' autres efforts sur ces lignes .*
trad. trans. auto	*et il y a eu d' autres efforts sur ces lignes .*

Tableau 5. *Une erreur sans conséquence*

transcription) (voir § 5.1), d'autre part les effets de l'insertion automatique de ponctuations et de la normalisation des nombres sur la qualité de la traduction (§ 5.2).

5.1. *Mesurer l'effet des incohérences*

Quatre systèmes de traduction, correspondant à quatre manières différentes de traiter la partie source du corpus d'entraînement, sont construits pour mesurer l'impact du format de la source :

– avec ponctuations et nombres d'origine (aucun prétraitement) ;

– avec normalisation des nombres (norm.) et ponctuations d'origine

– avec normalisation et sans aucune ponctuation, hormis un point à la fin des phrases (norm, sans ponct) ;

– avec la normalisation et la ponctuation automatique produite par un système de traduction automatique monolangue (voir § 3.2.2) (norm., ponct. auto).

Nous nous limitons ici à étudier le comportement du système de RAP de base.

Le même prétraitement a été appliqué à la partie source du corpus de développement de chaque système. Des évaluations extensives ont été conduites avec ces quatre systèmes, en prenant comme entrées, respectivement, pour le corpus de test :

– les transcriptions manuelles avec chacun des quatre types de prétraitements ;

– les transcriptions automatiques avec trois types de prétraitements (la ponctuation manuelle n'est pas disponible dans ces conditions).

Le tableau 6 résume l'ensemble des scores BLEU ainsi obtenus. Rappelons que les références de traduction en français sont ponctuées, ce qui signifie que les performances en BLEU prennent en compte la qualité de la ponctuation produite dans la traduction automatique.

On retrouve dans ce tableau les performances du système de base (aucun prétraitement, première ligne) sur les corpus de test sans prétraitement : les transcriptions manuelles ponctuées (BLEU = 33,2) et les transcriptions automatiques non ponctuées (BLEU = 20,5). La part inhérente aux erreurs de RAP peut être quantifiée en observant la différence de BLEU entre la traduction de transcriptions manuelles normalisées et non ponctuées (BLEU = 25,0) et celle de transcriptions automatiques normalisées et non ponctuées (BLEU = 20,6), soit une perte de 4,4 points BLEU.

	Test, trans. manuelle				Test, trans. auto.		
Prétraitement	aucun	norm.	norm. sans ponct.	norm. ponct. auto.	aucun	norm. sans ponct.	norm. ponct. auto.
aucun (`base`)	**33,2**	32,3	25,0	29,3	20,5	20,6	23,9
norm.	32,9	33,1	25,5	29,8	20,3	21,0	24,4
norm. sans ponct	32,6	32,8	29,2	29,3	23,7	24,4	24,1
norm. ponct auto	32,9	33,0	25,1	30,5	20,3	20,8	**25,2**

Tableau 6. *Performances de quatre systèmes de traduction (avec différents prétraitements de la partie source du corpus d'entraînement) sur les transcriptions manuelles et automatiques du corpus de test*

La deuxième ligne du tableau permet de mesurer l'impact de la normalisation en source des nombres. Les nombres dans les corpus d'entraînement sont d'abord transformés vers la représentation en lettres, pour ensuite être retransformés vers la représentation en chiffres plus proche des transcriptions automatiques enrichies. Ceci améliore notamment la qualité de traduction pour les transcriptions automatiques (21,0 contre 20,5).

L'observation des performances du système avec une normalisation sur le test manuel normalisé et ponctué (deuxième colonne, BLEU = 33,1) et non ponctué (troisième colonne, BLEU = 25,5), permet, par comparaison, d'isoler l'impact de l'absence de ponctuation dans les données à traduire (perte de 7,6 points BLEU).

La troisième ligne du tableau correspond au système normalisé et déponctué en source : les ponctuations en cible sont ici produites implicitement par le système de traduction. La comparaison des performances de ce système sur les transcriptions sans ponctuation en test, manuelles et automatiques, permet à nouveau d'isoler la part des erreurs de RAP dans la dégradation de performances, soit une perte de 4,8 points BLEU (29,2 − 24,4). La dégradation liée à l'absence de ponctuation en source est, dans ce cas, considérablement réduite (3,9 points BLEU : 33,1−29,2).

Enfin, la comparaison des performances des deux systèmes normalisés avec et sans ponctuation (deuxième et troisième ligne) confirme que, sans enrichissement des transcriptions automatiques du test par des ponctuations, le modèle de traduction entraîné sur des données non ponctuées en source est meilleur que celui qui bénéficie des ponctuations en source (gain de 3,4 points BLEU : 24,4−21,0).

5.2. *Utilisation de transcriptions « riches »*

Dans cette dernière série d'observations, le module de normalisation des nombres et le système de ponctuation par TA sont utilisés tous les deux pour normaliser et ponctuer automatiquement les textes sources[14].

Le système enrichi (quatrième ligne du tableau 6) est entraîné avec les corpus des données parallèles dont la partie source a été modifiée : les ponctuations manuelles d'origine sont remplacées par les ponctuations automatiques et les nombres sont normalisés. Cette transformation de la partie source des corpus d'entraînement répond à un double objectif : rendre ces corpus plus uniformes et plus proches de ce que l'on obtient en sortie de la RAP (avec post-traitement) tout en conservant, pour la source, un format qui permette une traduction optimale vers le format de la partie cible.

Sur la traduction des transcriptions manuelles ponctuées automatiquement, ce modèle est sans surprise moins performant (BLEU = 30,5) que le système de base avec les transcriptions d'origine pour le test (BLEU = 33,2). Cette configuration permet de quantifier la part de la dégradation engendrée par la ponctuation et la normalisation automatiques (perte de 2,7 points BLEU). En revanche, ce modèle est plus performant (gain de 1,3 point BLEU) pour la traduction des transcriptions automatiques ponctuées automatiquement (BLEU = 25,2) que le système de base (BLEU = 23,9).

Cette série d'expériences a permis de mettre en évidence l'importance de la ponctuation en source pour la traduction automatique (perte de 7,6 points BLEU sur les transcriptions manuelles). Cet impact s'avère en pratique plus important que celui des erreurs de RAP (perte de 4,4 points BLEU, pour un WER de 17 %). Enfin, l'enrichissement des transcriptions à la fois dans la partie source des corpus d'entraînement du système de traduction et dans la partie source du corpus de test permet d'améliorer considérablement les performances en traduction des transcriptions automatiques par rapport au modèle de base appliqué sur les transcriptions non enrichies (BLEU = 25,2 contre BLEU = 20,5, soit un gain de 4,7 points).

6. Adaptation des systèmes de traduction à une reconnaissance imparfaite

Dans cette nouvelle série d'expériences, nous exploitons en tant que corpus d'entraînement pour la TA les transcriptions automatiques de la source sonore des *TED Talks*, afin de rapprocher davantage les textes sources utilisés pour apprendre les systèmes de traduction et les conditions réelles de traduction de la parole. L'ensemble des conférences disponibles pour l'entraînement, le développement et le test a ainsi été transcrit automatiquement par notre système de RAP et aligné avec les références

14. Rappelons que le système de ponctuation ne prédit que les ponctuations de base « impaires ». Les ponctuations produites sont donc à la fois moins riches que les ponctuations manuelles d'origine, et possiblement erronées. Il en va de même pour la normalisation des nombres : ce traitement automatique peut produire des erreurs et ne correspond pas toujours à la représentation des nombres dans les transcriptions manuelles.

transcrites, segmentées et traduites manuellement, ce qui permet de disposer pour l'ensemble des corpus :

– de la transcription manuelle ;

– de la transcription automatique par le système de RAP ;

– de la traduction de référence.

Afin de pourvoir isoler l'impact des erreurs de RAP de l'impact de la ponctuation, nous avons utilisé dans les expériences sur la reconnaissance imparfaite les versions non ponctuées des systèmes. Le tableau 7 résume les résultats de ces expériences d'adaptation.

6.1. *Agir sur les modèles*

Nous nous plaçons dans le cas présenté à la section 2.4.2. Le modèle de traduction est initialement entraîné sur des textes écrits. L'adaptation consiste à utiliser des transcriptions automatiques lors de la phase d'optimisation des paramètres des modèles de traduction. Ainsi, sans changer la métrique d'optimisation des modèles de traduction (BLEU), on agit sur ce qu'elle mesure : la qualité de la traduction des transcriptions automatiques.

Comme l'indique la première ligne du tableau 7, l'utilisation des transcriptions automatiques pour le corpus de développement conduit à une sensible amélioration des performances du système de traduction (BLEU = 24,9), par rapport au système réglé sur des transcriptions manuelles (BLEU = 24,4, voir le tableau 6).

6.2. *Agir sur les données*

Nous nous plaçons maintenant dans la configuration décrite à la section 2.4.1, dans laquelle nous cherchons à agir sur les données pour limiter l'impact d'une reconnaissance imparfaite. Ici, la solution envisagée consiste à rapprocher les données d'entraînement des modèles de traduction des entrées à traduire (transcriptions automatiques bruitées), de façon à introduire dans les modèles de traduction la variabilité engendrée par les erreurs de reconnaissance.

6.2.1. *Entraînement sur les transcriptions automatiques*

L'utilisation des transcriptions automatiques pour l'entraînement des modèles de traduction est envisagée selon trois configurations (lignes 2, 3 et 4 du tableau 7). Le premier système (`auto`) est entraîné spécifiquement sur les transcriptions automatiques. Les deux autres systèmes intègrent à la fois la transcription manuelle et la transcription automatique, chacune étant alignée avec la traduction de référence. Dans le système (`manuel + auto., 1 modèle`), un seul modèle bilingue est construit à partir de l'ensemble des données (transcriptions automatiques et transcriptions manuelles). Le système (`manuel + auto., 2 modèles`) conserve deux modèles bi-

Système	Trans. manuelle	Trans. auto.
`manuel`	29,9	24,9
`auto`	28,8	24,2
`manuel + auto., 1 modèle`	29,5	24,8
`manuel + auto., 2 modèles`	29,3	24,6
`manuel + pseudoauto., 1 modèle`	30,1	24,8

Tableau 7. *Performances des modèles de traduction sans ponctuation adaptés sur les données de test transcrites manuellement et automatiquement. Réglage des modèles sur les transcriptions automatiques du corpus de réglage*

lingues distincts en utilisant la possibilité offerte par Ncode d'effectuer le réglage du système de traduction à partir de plusieurs modèles bilingues.

Dans l'ensemble, les résultats ne montrent pas d'amélioration importante par rapport au système entraîné sur les seules transcriptions manuelles et réglé avec des transcriptions automatiques. Ce résultat peut s'expliquer partiellement par le fait que l'étiquetage en parties du discours, nécessaire à Ncode pour apprendre les règles de réordonnancement, est moins fiable pour des transcriptions automatiques. Par ailleurs, certaines erreurs produites par la RAP sont en fait déjà présentes dans les transcriptions manuelles[15], et donc connues dans le modèle manuel comme des variantes en source.

Comme précédemment, nous avons effectué une analyse de l'impact de l'adaptation du système de traduction sur la traduction des phrases contenant différents types d'erreurs de RAP. Le tableau 8 montre que l'adaptation de la traduction améliore notamment la traduction pour les phrases qui ne contiennent que des substitutions : la moyenne de la différence entre le score sBLEU de la traduction de la source propre et de la traduction de la source bruitée diminue pour ce type d'erreur. En revanche, notre adaptation ne permet pas de compenser les erreurs de traduction dues aux insertions et aux suppressions des mots.

Type d'erreur	Nb. phrases	Diff. sBLEU moyenne TA de base	Diff. sBLEU moyenne TA adaptée
`Sub`	347	6,1	5,5
`Sup`	100	6,2	8,0
`Ins`	57	1,4	2,4

Tableau 8. *Impact de l'adaptation du système de TA sur la traduction des erreurs de la RAP par type d'erreur*

Une analyse plus détaillée des erreurs fréquentes de chaque type permet de mieux comprendre les effets principaux de l'adaptation de la traduction (voir le tableau 9).

15. Ceci s'expliquant par variabilité inhérente de la qualité des transcriptions de TED, réalisées par des transcripteurs bénévoles.

Cette analyse porte sur toutes les phrases contenant l'erreur en question, ce qui permet d'avoir suffisamment d'exemples pour calculer des statistiques interprétables. On peut observer l'impact positif de l'adaptation sur les substitutions de certains mots lexicaux, comme *could* → *can* ou *going to* → *gonna*, tandis que sur les confusions entre les mots-outils cet impact est moins visible.

Pour les suppressions, l'adaptation apporte une légère amélioration pour les mots-outils absents, alors que sur les verbes comme *is*, *are* la tendance est inverse.

En ce qui concerne les insertions, la traduction des phrases correspondantes est très peu corrigée par l'adaptation.

Erreur	Diff. sBLEU moyenne TA de base	Diff. sBLEU moyenne TA adaptée
Substitutions		
and/in	3,7	4,5
a/the	5,4	6,2
going/gonna	1,2	0,6
that/the	1,3	2,3
could/can	6,9	3,0
Suppressions		
to	7,1	6,7
it	6,3	5,1
and	7,0	6,0
of	7,4	6,7
is	7,1	9,1
are	6,0	7,1
Insertions		
the	10,3	10,5
and	11,3	11,8
in	10,4	9,0
you	6,4	6,7

Tableau 9. *Analyse détaillée de l'impact des erreurs de RAP avant et après l'adaptation du système de traduction*

Le tableau 2 illustre différents effets d'une erreur fréquente de transcription (substitution *going to* → *gonna*) sur les traductions produites par les systèmes comparés dans le tableau 7. Pour le premier extrait de phrase, la substitution en transcription n'engendre pas d'erreur en traduction, même sur le modèle de base (car la traduction *gonna* → *allez* est déjà présente dans le modèle bilingue entraîné sur des données transcrites manuellement). Pour le second extrait, la traduction de cette séquence bruitée produite par le modèle adapté (`adapt`) est correcte, ce qui n'était pas le cas pour celle produite par le modèle non adapté (`base`). Sur les 26 occurrences de cette substitution observées dans le test, le modèle `adapt` améliore une fois sur quatre la traduction obtenue avec le modèle `manuel`.

trans. manuelle	*[...] whether or not you're* **going to** *meet adversity [...]*
trans. auto.	*[...] whether or not you're* **gonna** *meet adversity [...]*
modèle base	*[...] si oui ou non vous allez rencontrer l' adversité [...]*
modèle adapt	*[...] si oui ou non vous allez rencontrer l' adversité [...]*
référence	*[...] si vous allez ou non rencontrer l' adversité [...]*
trans. manuelle	**he's going to** *land in a couple of hours , he's going to rent a car [...]*
trans. auto.	**it is gonna** *land in a couple hours rent a car [...]*
modèle base	*c' est la terre allait dans quelques heures une voiture [...]*
modèle adapt	*il va atterrir dans quelques heures louer une voiture [...]*
référence	*il va atterrir dans quelques heures , louer une voiture [...]*

Figure 2. *Exemples de traduction dans le cas d'erreurs de transcription automatique avec 2 modèles : modèle de base et modèle adapté. Phrase 1 : traduction déjà correcte avec modèle de base. Phrase 2 : traduction améliorée avec le modèle adapté.*

6.2.2. *Entraînement sur les pseudotranscriptions automatiques*

Dans la section précédente, l'adaptation est réalisée en utilisant à la fois la source sonore et sa retranscription automatique, la transcription manuelle en langue source, ainsi que la traduction en langue cible ; ceci pour l'ensemble du corpus d'entraînement des modèles de traduction. Des ressources aussi complètes sont cependant rares, même pour la paire de langues anglais/français. Par opposition, la communauté dispose maintenant de centaines d'heures d'audio transcrites (mais non traduites) pour l'entraînement des modèles acoustiques de la RAP. Pour la traduction de textes, les corpus bilingues anglais/français comprennent des millions de paires de phrases.

Dans cette dernière section sur l'adaptation, nous nous intéressons à l'exploitation de corpus bilingues textuels pour lesquels aucune source sonore n'est disponible. Notre proposition consiste à bruiter artificiellement la partie source des bitextes, afin de la rapprocher des transcriptions automatiques produites par la RAP. Le bruitage est réalisé par un système de traduction monolingue, dans lequel la source est le texte correct et la cible le texte bruité obtenu par transcription automatique.

Ce modèle de bruitage est ensuite appliqué sur la partie source du corpus d'entraînement bilingue. Ces sources bruitées, ainsi que les sources propres du même corpus, sont utilisées pour l'entraînement d'un nouveau système de traduction (*manuel + pseudoauto.*), comparable au système *manuel + auto* décrit dans la section précédente. Les performances du système *manuel + pseudoauto.*, reportées à la dernière ligne du tableau 7, sont comparables à celles produites par le système *manuel + auto*.

Cette expérience préliminaire, qui valide indirectement le modèle de bruitage, nous permet d'envisager à terme d'appliquer cette technique pour construire en quantité des données d'entraînement bruitées à partir de grands corpus bilingues.

Segmentation	Nb. segments	BLEU test man.
IWSLT	1664	38,1
RAP	524	37,7
site TED	3577	34,9

Tableau 10. *Impact de la segmentation*

7. Les incohérences de segmentation

Dans cette dernière section, nous nous intéressons à une troisième source majeure d'incohérences entre les transcriptions automatiques et les données d'apprentissage des systèmes de traduction, à savoir l'incohérence des segmentations. Les expériences ci-dessus ont essentiellement utilisé les corpus parallèles préparés par les organisateurs des campagnes IWSLT, pour lesquels un grand nombre d'ajustements ont été opérés pour resegmenter les transcriptions manuelles de manière à produire des unités plus cohérentes pour la traduction.

Dans une utilisation réelle des systèmes de traduction de la parole, de tels ajustements ne sont pas possibles, et il faut s'accommoder de la segmentation automatique produite par le système de RAP. Dans cette étude préliminaire, nous nous limitons à mesurer l'impact de cette resegmentation. Le tableau 10 donne ainsi le score BLEU[16] pour différentes stratégies de segmentation, ainsi que le nombre de segments (pour mémoire, le test contient environ 31 000 tokens, soit une moyenne proche de 18 mots par segment). Le système de traduction utilisé pour ces tests est celui qui est entraîné avec les transcriptions manuelles enrichies.

Les résultats reportés dans le tableau 10 mettent en évidence l'impact considérable (bien que sous-estimé par la métrique) des différences de segmentation : entre la segmentation produite par les transcripteurs et la segmentation révisée pour la traduction, la différence est de plus de 3 points BLEU. La segmentation à très gros grains fournie par notre système de RAP obtient des résultats intermédiaires. Le risque principal est celui de sursegmenter les transcriptions automatiques, créant ainsi des segments auxquels il manque du contexte pour pouvoir être traduits correctement. Inversement, si l'on s'abstrait des contraintes de production en temps réel des traductions, alors le traitement différent de longs segments ne semble pas causer trop d'erreurs de recherche.

Les expériences ci-dessus suggèrent que la production de segmentations conformes aux segmentations manuelles est une troisième voie d'amélioration des systèmes existants. Nous envisageons, parmi les suites possibles à ce travail, d'étudier la possibilité d'optimiser la segmentation de la RAP pour minimiser les erreurs de traduction, à la manière des propositions de Oda *et al.* (2014).

16. Calculé ici en considérant chaque document comme une seule phrase, ce qui est le moyen le plus simple de comparer des segmentations différentes, cf. la discussion de la section 2.5.

8. Conclusion

Dans cet article, nous avons présenté une analyse des principales difficultés qui se posent aux concepteurs de systèmes de TrAP, en les illustrant sur une application particulière, la traduction automatique des *TED Talks*. Pour l'essentiel, ces difficultés trouvent leur origine dans la grande rareté des données de parole traduites, qui implique d'utiliser, pour l'apprentissage des modèles de traduction, des données parallèles textuelles qui ne ressemblent que de loin aux réelles sorties des systèmes de RAP. Comme nous l'avons montré, les divergences entre ces données et transcriptions automatiques sont importantes et se matérialisent de multiples manières : dans la structuration des énoncés et dans les marques (orales et écrites) qui reflètent cette structuration d'une part, dans les différences du format entre les transcriptions manuelles et automatiques, et dans l'existence « d'erreurs » de production et de reconnaissance vocale, d'autre part.

Pour la tâche considérée et les systèmes de base utilisés, il apparaît que cette divergence se traduit par un écart de plus de 10 points BLEU sur la qualité de la traduction automatique, dont environ 5 points sont directement imputables aux erreurs de la RAP (en utilisant une segmentation de référence).

Parmi les principales perspectives ouvertes par ce travail, hormis la reproduction de ces tendances pour d'autres versions de systèmes de RAP et de traduction automatique, nous envisageons de faire porter notre effort principalement dans trois directions : en premier lieu, sur l'optimisation du système de RAP pour des tâches de traduction, afin d'obtenir des points de fonctionnement (compromis entre substitutions, suppressions et insertions) qui soient le moins pénalisants possible pour la traduction automatique ; ensuite, sur l'utilisation de sources complémentaires de données (des enregistrements non traduits, des traductions de textes, cf. la discussion de la section 6.2) ; enfin, sur la génération automatique de segmentations des énoncés reconnus qui soient plus proches de celles qui sont utilisées pour les traductions de référence.

Remerciements

Les auteurs remercient Jean-Luc Gauvain et Lori Lamel pour le développement des systèmes de transcription enrichie pour les données *TED Talks* ; Alexandre Allauzen pour le développement des modèles de langue ; ainsi que Jan Niehues (Karlsruhe Institute of Technology) pour son aide dans la préparation des données de la tâche.

9. Bibliographie

Afli H., Barrault L., Schwenk H., « Traduction automatique à partir de corpus comparables : extraction de phrases parallèles à partir de données comparables multimodales », *Actes de la* 19me *conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Association pour le Traitement Automatique des Langues, Grenoble, France, p. 447-454, 2012.

Allauzen A., Pécheux N., Do Q. K., Dinarelli M., Lavergne T., Max A., Le H.-S., Yvon F., « LIMSI @ WMT13 », *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, p. 62-69, 2013.

Allauzen A., Yvon F., « Méthodes statistiques pour la traduction automatique », *in* E. Gaussier, F. Yvon (eds), *Modèles Probabilistes pour l'accès à l'information*, Hermès, Paris, chapter 7, p. 271-356, 2011.

Ayan N. F., Mandal A., Frandsen M. W., Zheng J., Blasco P., Kathol A., Béchet F., Favre B., Marin A., Kwiatkowski T., Ostendorf M., Zettlemoyer L. S., Salletmayr P., Hirschberg J., Stoyanchev S., « "Can you give me another word for hyperbaric ?" : Improving speech translation using targeted clarification questions », *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, p. 8391-8395, 2013.

Banerjee S., Lavie A., « METEOR : An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments », *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, Ann Arbor, Michigan, p. 65-72, 2005.

Bangalore S., Rangarajan Sridhar V. K., Kolan P., Golipour L., Jimenez A., « Real-time Incremental Speech-to-Speech Translation of Dialogs », *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Montréal, Canada, p. 437-445, 2012.

Bangalore S., Riccardi G., « Stochastic Finite-State Models for Spoken Language Machine Translation », *Machine Translation*, vol. 17, p. 165-184, 2002.

Bazillon T., Jousse V., Béchet F., Estève Y., Linarès G., Luzzati D., « La parole spontanée : transcription et traitement », *Traitement Automatique des Langues (TAL)*, vol. 49, p. 47-67, 2008.

Besacier L., Le V.-B., Boitet C., Berment V., « ASR and translation for under-resourced languages », *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. V, Toulouse, France, 2006.

Blanche-Benveniste C., Martin P., *Le Français : Usages de la langue parlée*, Peeters, 2010.

Bonneau-Maynard H., Segal N., Bilinski E., Gauvain J.-L., Gong L., Lamel L., Laurent A., Yvon F., Despres J., Josse Y., Le V. B., « Traduction de la parole dans le projet RAPMAT », *Journées d'Études sur la Parole (JEP)*, Le Mans, France, 2014.

Brown P. F., Pietra S. A. D., Pietra V. J. D., Mercer R. L., « The Mathematics of Statistical Machine Translation : Parameter Estimation », *Computational Linguistics*, vol. 19, n° 2, p. 263-311, 1993.

Béchet F., Favre B., « ASR error segment localization for spoken recovery strategy », *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, p. 6837-6841, 2013.

Casacuberta F., Vidal E., « Machine Translation with Inferred Stochastic Finite-State transducers », *Computational Linguistics*, vol. 30, n° 3, p. 205-225, 2004.

Casacuberta F., Vidal E., Vilar J. M., « Architectures for Speech-to-speech Translation Using Finite-state Models », *Proceedings of the ACL-02 Workshop on Speech-to-speech Translation : Algorithms and Systems - Volume 7*, S2S '02, p. 39-44, 2002.

Cettolo M., Federico M., « Text segmentation criteria for statistical machine translation », *Advances in Natural Language Processing*, Springer, p. 664-673, 2006.

Cettolo M., Niehues J., Stüker S., Bentivogli L., Federico M., « Report on the 10th IWSLT Evaluation Campaign », *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

Cettolo M., Niehues J., Stüker S., Bentivogli L., Federico M., « Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014 », *in* M. Federico, S. Stüker, F. Yvon (eds), *Proceedings of the eleventh International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, p. 2-17, 2014.

Cho E., Niehues J., Waibel A., « Segmentation and Punctuation Prediction in Speech Language Translation Using a Monolingual Translation System », *Proceeding of the 9th International Workshop on Spoken Language Translation (IWSLT)*, 2012.

Cho E., Niehues J., Waibel A., « Tight Integration of Speech Disfluency Removal into SMT », *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Gothenburg, Sweden, 2014.

Christensen H., Gotoh Y., Renals S., « Punctuation annotation using statistical prosody models », *in Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, p. 35-40, 2001.

Constant M., Dister A., « Automatic detection of disfluencies in speech transcriptions », *in* M. Pettorino, A. Giannini, I. Chiari, F. M. Dovetto (eds), *Spoken Communication*, Cambridge Scholars Publishing, p. 259—272, 2010.

Crego J. M., Yvon F., Mariño J. B., « N-code : an open-source Bilingual N-gram SMT Toolkit », *Prague Bulletin of Mathematical Linguistics*, vol. 96, p. 49-58, 2011.

Cucu H., Buzo A., Besacier L., Burileanu C., « Statistical error correction methods for domain-specific ASR systems », *Proceedings of the First international conference on Statistical Language and Speech Processing*, Tarragona, Spain, p. 83-92, 2013.

de Mareüil P. B., Habert B., Bénard F., Adda-Decker M., Barras C., Adda G., Paroubek P., « A quantitative study of disfluencies in French broadcast interviews », *Workshop Disfluency In Spontaneous Speech (DISS)*, Aix-en-Provence, France, 2005.

Driesen J., Renals S., « Lightly supervised automatic subtitling of weather forecasts », *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, p. 452-457, 2013.

Favre B., Hakkani-Tur D., Shriberg E., « Syntactically-informed models for comma prediction », *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, 2009.

Finch A., Wang X., Sumita E., « An Exploration of Segmentation Strategies in Stream Decoding », *in* M. Federico, S. Stüker, F. Yvon (eds), *Proceedings of the eleventh International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, p. 206-213, 2014.

Fitzgerald E., Hall K., Jelinek F., « Reconstructing false start errors in spontaneous speech text », *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, p. 255-263, 2009.

Fügen C., Waibel A., Kolss M., « Simultaneous translation of lectures and speeches », *Machine Translation*, vol. 21, n° 4, p. 209-252, 2007.

Furui S., Nakamura M., Ichiba T., Iwano K., « Why is the recognition of spontaneous speech so hard ? », *Text, Speech and Dialogue*, vol. 3658, p. 9-–22, 2005.

Gao Q., Vogel S., « Parallel Implementations of Word Alignment Tool », *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, Columbus, Ohio, p. 49-57, 2008.

Gauvain J.-L., Lamel L., Adda G., « The LIMSI broadcast news transcription system », *Speech Communication*, vol. 37, p. 89-108, 2002.

Goldwater S., Jurafsky D., Manning C. D., « Which words are hard to recognize ? Prosodic, lexical, and disfluency factors that increase speech recognition error rates », *Speech Communication*, vol. 52, n° 3, p. 181-200, 2010.

Ha T.-L., Herrmann T., Niehues J., Mediani M., Cho E., Zhang Y., Slawik I., Waibel A., « The KIT Translation Systems for IWSLT 2013 », *Proceeding of the 11th International Workshop on Spoken Language Translation (IWSLT)*, 2013.

Hashimoto K., Yamagishi J., Byrne W., King S., Tokuda K., « Impacts of machine translation and speech synthesis on speech-to-speech translation », *Speech Communication*, vol. 54, n° 7, p. 857-866, 2012.

He X., Den L., Acero A., « Why Word Error Rate is not a Good Metric for Speech Recognizer Training for the Speech Translation Task ? », *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, p. 5632-5635, 2011.

He X., Deng L., « Robust Speech Translation by Domain Adaptation », *Conference of the International Speech Communication Association (INTERSPEECH)*, Florence, Italy, p. 2105-2108, 2011.

Hewavitharana S., Mehay D., Ananthakrishnan S., Kumar R., Makhoul J., « Anticipatory Translation Model Adaptation for Bilingual Conversations », *in* M. Federico, S. Stüker, F. Yvon (eds), *Proceedings of the eleventh International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, p. 230-235, 2014.

Hirschberg J., Litman D., Swerts M., « Prosodic and other cues to speech recognition failures », *Speech Communication*, vol. 43, p. 155–175, 2004.

Honal M., Schultz T., « Automatic disfluency removal on recognized spontaneous speech – rapid adaptation to speaker dependent dislfuencies », *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Philadelphia, USA, p. 969–972, 2005.

Huang J., Zweig G., « Maximum Entropy Model for Punctuation Annotation from Speech », *Proc. International Conference on Spoken Language Processing (ICSLP)*, 2002.

Jeong M., Eun J., Jung S., Lee G. G., « An Error-Corrective Language-Model Adaptation For Automatic Speech Recognition », *Conference of the International Speech Communication Association (InterSpeech)*, Lisbon, Portugal, 2005.

Jiang J., Ahmed Z., Carson-Berndsen J., Cahill P., Way A., « Phonetic representation-based speech translation », *Proceedings of the 13th Machine Translation Summit*, Xiamen, China, 2011.

Kim J.-H., Woodland P. C., « The use of prosody in a combined system for punctuation generation and speech recognition », *Proceedings of the Seventh European Conference on Speech Communication and Technology (Eurospeech)*, p. 2757-2760, 2001.

Koehn P., « Europarl : A Parallel Corpus for Statistical Machine Translation », *Proceedings of the 10th Machine Translation Summit*, AAMT, AAMT, Phuket, Thailand, p. 79-86, 2005.

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

Lamel L., « Multilingual Speech Processing Activities in Quaero : Application to Multimedia Search in Unstructured Data », *Proceedings of the fifth International Conference Human Language Technologies*, Tartu, Estonia, 2012.

Lavecchia C., Smaïli K., Langlois D., « Building parallel corpora from movies », *The 4th International Workshop on Natural Language Processing and Cognitive Science-NLPCS 2007*, Funchal, Madeira, Portugal, 2007.

Lavergne T., Allauzen A., Le H.-S., Yvon F., « LIMSI's experiments in domain adaptation for IWSLT11 », *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Franciso, USA, 2011.

Lease M., Charniak E., « Recognizing Disfluencies in Conversational Speech », *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, p. 1566-–1573, 2006.

Lefèvre F., Mostefa D., Besacier L., Esteve Y., Camelin N., Favre B., Jabaian B., « Robustness and adaptation of spoken language understanding systems among languages and domains : the PORTMEDIA project », *Language Resources and Evaluation Conference (LREC)*, 2012.

Levin L., Lavie A., Woszczyna M., Gates D., Gavaldá M., Koll D., Waibel A., « The Janus-III Translation System : Speech-to-Speech Translation in Multiple Domains », *Machine Translation*, vol. 15, n° 1-2, p. 3-25, 2000.

Lin C.-Y., Och F. J., « ORANGE : a method for evaluating automatic evaluation metrics for machine translation », *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Geneva, Switzerland, 2004.

Liu Y., Shriberg E., Stolcke A., Harper M., « Comparing HMM, Maximum Entropy, and Conditional Random Fields for Disfluency Detection », *Conference of the International Speech Communication Association (INTERSPEECH)*, Lisbonne, Portugal, p. 3033—3036, 2005.

Liu Y., Shriberg E., Stolcke A., Hillard D., Ostendorf M., Harper M., « Enriching speech recognition with automatic detection of sentence boundaries and disfluencies », *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, n° 5, p. 1526-1540, 2006.

Mariño J. B., Banchs R. E., Crego J. M., de Gispert A., Lambert P., Fonollosa J. A., Costa-Jussà M. R., « N-gram-based machine translation », *Computational Linguistics*, vol. 32, n° 4, p. 527-549, 2006.

Matusov E., Leusch G., Bender O., Ney H., « Evaluating Machine Translation Output with Automatic Sentence Segmentation », *International Workshop on Spoken Language Translation*, Pittsburgh, PA, USA, p. 148-154, 2005.

Matusov E., Ney H., « Lattice-Based ASR-MT Interface for Speech Translation », *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, p. 721-732, 2011.

Ng R. W. M., Hain T., Cohn T., « Adaptation of lecture speech recognition system with machine translation output », *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, p. 8401-8405, 2013.

Och F. J., « Minimum error rate training in statistical machine translation », *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, p. 160-167, 2003.

Oda Y., Neubig G., Sakti S., Toda T., Nakamura S., « Optimizing Segmentation Strategies for Simultaneous Speech Translation », *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Baltimore, Maryland, p. 551-556, 2014.

Olive J., Christianson C., McCary J. (eds), *Handbook of Natural Language Processing and Machine Translation : DARPA Global Autonomous Language Exploitation*, Springer, New-York, Dordrecht, Heidelberg, London, 2011.

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

Paulik M., Waibel A., « Automatic translation from parallel speech : Simultaneous interpretation as mt training data », *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, p. 496-501, 2009.

Peitz S., Freitag M., Mauser A., Ney H., « Modeling Punctuation Prediction as Machine Translation », *Proceedings of the 8th International Workshop on Spoken Language Translation (IWSLT)*, 2011.

Peitz S., Wiesler S., Nussbaum-Thom M., Ney H., « Spoken Language Translation Using Automatically Transcribed Text in Training », *Proceeding of the 9th International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, China, 2012.

Perez A., Torres I., Casacuberta F., « Finite-state acoustic and translation model composition in statistical speech translation : empirical assessment », *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, Donostia-San Sebastian, Spain, 2012.

Rangarajan Sridhar V. K., Chen J., Bangalore S., Ljolje A., Chengalvarayan R., « Segmentation Strategies for Streaming Speech Translation », *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Atlanta, Georgia, p. 230-238, 2013.

Raybaud S., De l'utilisation de mesures de confiance en traduction automatique : évaluation, post-édition et application à la traduction de la parole, PhD thesis, Université de Lorraine, Nancy, 2012.

Rayner M., Carter D., Bouillon P., Digalakis V., Wirén M. (eds), *The spoken language translator*, Cambridge University Press, Cambridge, UK, 2000.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, p. 44-49, 1994.

Schwenk H., « Investigations on large scale lightly-supervised training for statistical machine translation », *Proceedings of the International Workshop on Spoken Language Translation*, Hawaii, USA, 2008.

Segal N., Bonneau-Maynard H., Do Q., Allauzen A., Gauvain J.-L., Lamel L., Yvon F., « LIMSI English-French Speech Translation System », *in* M. Federico, S. Stüker, F. Yvon (eds), *Proceedings of the eleventh International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, p. 106-112, 2014.

Seligman M., « Nine issues in speech translation », *Machine Translation*, vol. 15, n° 1-2, p. 149-186, 2000.

Shinozaki T., Furui S., « Error analysis using decision trees in spontaneous presentation speech recognition », *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Madonna di Campiglio, Italy, 2001.

Skadiņš R., Tiedemann J., Rozis R., Deksne D., « Billions of parallel words for free : building and using the EU bookshop corpus », *Proceedings of the 9th International Conference on*

Language Resources and Evaluation (LREC-2014), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014.

Stolcke A., « SRILM - An extensible language modeling toolkit », *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, p. 901-904, 2002.

Tiedemann J., « Building a Multilingual Parallel Subtitle Corpus », *Proceedings of the conference on Computational Linguistics in the Netherlands (CLIN'17)*, Leuven, Belgium, 2007.

Vasilescu I., Yahia D., Snoeren N. D., Adda-Decker M., Lamel L., « Cross-lingual study of ASR errors : on the role of the context in human perception of near-homophones », *Conference of the International Speech Communication Association (InterSpeech)*, Florence, Italy, 2011.

Vitevitch M. S., Luce P. A., « Probabilistic phonotactics and neighborhood activation in spoken word recognition », *Journal of Memory and Language*, vol. 40, p. 374—408, 1998.

Vogel S., Ney H., Tillmann C., « HMM-based word alignment in statistical translation », *Proceedings of the 16th conference on Computational linguistics*, COLING, Copenhagen, Denmark, p. 836-841, 1996.

Wahlster W. (ed.), *Verbmobil : Foundations of speech-to-speech translation*, Artificial Intelligence series, Springer Verlag, Berlin Heidelberg New-York, 2000.

Wang W., Tur G., Zheng J., Ayan N. F., « Automatic Disfluency Removal for Improving Spoken Language Translation », *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Dallas, USA, p. 5214-5217, 2010.

Zhou B., Cui X., Huang S., Cmejrek M., Zhang W., Xue J., Cui J., Xiang B., Daggett G., Chaudhari U., Maskey S., Marcheret E., « The IBM speech-to-speech translation system for smartphone : Improvements for resource-constrained tasks », *Computer Speech & Language*, vol. 27, n° 2, p. 592-618, 2013. Special Issue on Speech-speech translation.