Micro-adaptation lexicale en traduction automatique statistique [1]

Josep Maria Crego[*] **— Gregor Leusch**[***] **— Aurélien Max**[*,**] **— Hermann Ney**[***] **— François Yvon**[*,**]

(*) LIMSI-CNRS, B.P. 133, 91403 Orsay cedex, France
(**) Université Paris Sud 11, 91405 Orsay cedex, France
(***) Lehrstuhl für Informatik 6 - RWTH Aachen University, Ahornstr. 55, D-52056 Aachen, Allemagne

RÉSUMÉ. *Nous présentons un cadre générique en traduction automatique statistique (TAS) dans lequel des prédictions lexicales, sous forme d'un modèle de langue local à la phrase à traduire, sont exploitées pour guider la recherche de la meilleure hypothèse de traduction, ce qui a pour effet d'opérer une micro-adaptation lexicale. Nous proposons une instanciation de ce cadre qui est évaluée sur trois paires de langues : les prédictions auxiliaires proviennent d'autres systèmes de TAS qui réalisent une triangulation* via *une langue auxiliaire. Une première configuration met en jeu neuf langues auxiliaires, ce qui permet de mesurer la contribution relative de chaque langue. Nous proposons ensuite d'utiliser simultanément ces neuf systèmes, en les combinant par consensus. Nos résultats montrent qu'il est possible d'augmenter les performances d'un système de TAS de manière entièrement automatique en exploitant des sources auxiliaires.*

ABSTRACT. *We introduce a generic framework in Statistical Machine Translation (SMT) in which lexical hypotheses, in the form of a target language model local to the input sentence, are used to guide the search for the best translation, thus performing a* lexical microadaptation. *An instantiation of this framework is presented and evaluated on three language pairs, where these auxiliary hypotheses are derived through triangulation* via *an auxiliairy language. Our first experiments consider nine auxiliary languages, allowing us to measure their individual contribution. We then combine all their hypotheses through a decoding by consensus. Our experiments show that SMT systems can be improved by automatically produced auxiliary hypotheses.*

MOTS-CLÉS : *traduction automatique statistique, traduction par pivot.*

KEYWORDS: *statistical machine translation, pivoting in translation.*

1. Research conducted in the scope of the European Associated Laboratories IMMI-Labs/ Recherche conduite dans le cadre des Laboratoires Européens Associés IMMI-Labs.

1. Introduction

Les systèmes de traduction automatique, en particulier les systèmes utilisant des méthodes statistiques, ont accompli, ces dernières années, des progrès suffisants pour pouvoir s'exposer au grand jour, et rendre des services de traduction (approximative) au plus grand nombre, notamment au travers d'interfaces de traduction en accès libre sur Internet. Le développement de systèmes performants exige toutefois de disposer de corpus bilingues parallèles, une ressource relativement rare, du moins pour certaines paires de langues. Pour pallier ce manque, de nombreux travaux récents s'intéressent à l'intégration de ressources qui viendraient complémenter ces corpus, que ces ressources correspondent à des logiciels d'analyse linguistique, à des dictionnaires ou à des terminologies bilingues, ou encore à d'autres collections documentaires (parallèles, comparables, voire monolingues).

Dans cet article, nous nous intéressons à l'amélioration de la performance des systèmes de traduction automatique statistique par l'exploitation de ressources complémentaires qui prennent ici la forme d'hypothèses de traductions auxiliaires. Nous présentons un cadre générique dans lequel des prédictions lexicales sur le texte cible sont exploitées pour guider la recherche de la meilleure hypothèse de traduction. Ces prédictions sont faites pour chaque phrase à traduire et sont intégrées sous la forme d'un modèle de langue additionnel utilisé par le décodeur statistique, ce qui a pour effet d'opérer une *micro-adaptation lexicale* en renforçant la probabilité de certains mots.

Nous proposons ensuite une instanciation de ce cadre qui est décrite et évaluée sur trois paires de langues (français $\rightarrow$ anglais, français $\rightarrow$ allemand, et allemand $\rightarrow$ anglais). Dans cette instanciation, les prédictions auxiliaires proviennent d'autres systèmes de traduction automatique qui réalisent une triangulation *via* une langue auxiliaire. Cette situation est particulièrement digne d'intérêt, puisqu'elle permet d'envisager d'améliorer indirectement un système traduisant depuis A vers B lorsque l'on améliore les systèmes impliqués dans la triangulation.

Diverses configurations expérimentales sont analysées. Une première configuration met en jeu des systèmes pour neuf langues auxiliaires appris sur les mêmes données, ce qui permet de mesurer la contribution relative de chaque langue. Nous nous intéressons ensuite à l'exploitation simultanée de ces neuf traductions auxiliaires, en proposant de les recombiner par consensus. Dans les deux cas, les résultats quantitatifs obtenus sont discutés, et complétés par une analyse plus fine des améliorations observées, qui s'appuie sur une méthode originale destinée à mesurer la qualité des prédictions lexicales d'un système.

La contribution de ce travail est donc triple : d'une part, nous définissons un cadre générique permettant d'intégrer des connaissances auxiliaires dans un système de traduction statistique, d'autre part, nous étudions différentes manières d'améliorer un système de base en intégrant ces prédictions auxiliaires (ici obtenues par triangulation), et montrons qu'elles conduisent à des améliorations significatives. Nous montrons également, à travers une série d'analyses complémentaires, que ces améliora-

tions sont, pour une large part, attribuables à une réelle amélioration des prédictions lexicales du système.

Le reste de l'article est organisé comme suit : la section 2 présente un survol de l'état de l'art en traduction statistique, en s'attardant plus longuement sur les travaux relatifs à la combinaison de systèmes (section 2.2). Nous introduisons ensuite à la section 3 un cadre générique pour l'adaptation lexicale et discutons une instanciation particulière, qui utilise des techniques de pivot. La section 4 présente un autre cadre générique de combinaison de systèmes, qui utilise des réseaux de consensus. Ce cadre servira de point de comparaison avec notre méthode, mais également à évaluer l'intérêt de combiner simultanément plusieurs sources auxiliaires. Les protocoles expérimentaux et les résultats obtenus sont présentés et analysés en section 5. Un bilan de ces expériences est tiré à la section 6, dans laquelle nous évoquons également diverses pistes pour prolonger ce travail.

2. Traduction automatique statistique

2.1. *Un survol de l'état de l'art*

Les recherches et développements en traduction automatique connaissent un renouveau, du fait de l'arrivée à maturité de technologies qui, s'appuyant sur des modèles probabilistes simples, permettent de tirer efficacement partie des grandes masses de données, en particulier de données parallèles bilingues, qui sont disponibles dans des quantités croissantes sur Internet (voir (Lopez, 2008 ; Koehn, 2010) pour des états de l'art récents sur ces modèles).

On peut dater ce renouveau aux travaux pionniers (Brown *et al.*, 1990 ; Brown *et al.*, 1993) conduits au sein des équipes d'IBM au début des années 90, qui ont été les premiers à proposer des modèles probabilisant le processus de traduction. Ces modèles, ainsi que les techniques d'estimation afférentes, ont été initialement conçus à l'imitation des techniques utilisées en reconnaissance vocale (*modèle du canal bruité*) et envisagent la traduction comme une suite de décisions probabilistes s'appliquant sur les « mots-formes » d'une phrase. Pour chacun des mots, on modélise le choix du ou des équivalents de traduction, l'ordre dans lequel ces équivalents doivent être positionnés, relativement les uns aux autres ainsi que relativement aux équivalents des mots voisins. L'ensemble de ces mécanismes, dûment probabilisé, constitue le *modèle de traduction* (de mots) et permet d'évaluer des probabilités conditionnelles de la forme $P(\mathbf{f}|\mathbf{e})$, où $\mathbf{f}$ est une phrase en langue source résultant de l'application du modèle à une phrase $\mathbf{e}$ en langue cible [2]. Ces modèles sont appris sur des corpus bilingues parallèles. Un modèle de langue n-grammes, estimé sur un corpus de textes monolingues, permet de probabiliser l'espace des phrases cibles. En combinant les deux modèles, on obtient une règle de décision pour la traduction statistique d'une

2. Le paradoxe n'est qu'apparent : dans le modèle du canal bruité, on probabilise la traduction de $\mathbf{e}$ vers $\mathbf{f}$ pour traduire de $\mathbf{f}$ vers $\mathbf{e}$.

phrase source $\mathbf{f}$. Cette règle consiste à choisir la phrase cible $\mathbf{e}^*$ qui maximise la probabilité jointe $P(\mathbf{e}, \mathbf{f})$, soit encore :

$$\mathbf{e}^* = \underset{\mathbf{e}}{\operatorname{argmax}}\, P(\mathbf{e}|\mathbf{f}) = \underset{\mathbf{e}}{\operatorname{argmax}}\, P(\mathbf{e}, \mathbf{f}) = \underset{\mathbf{e}}{\operatorname{argmax}}\, P(\mathbf{f}|\mathbf{e})P(\mathbf{e})$$

Ces modèles ont été progressivement sophistiqués, pour donner lieu à une famille de modèles qui aujourd'hui représente l'état de l'art pour de nombreuses paires de langues, les modèles de *segments* (*phrase-based models*[3]) (Zens *et al.*, 2002 ; Koehn *et al.*, 2003). Dit simplement, ces modèles probabilisent non plus des appariements entre mots, mais des appariements entre des *séquences de taille variable en langues source et cible*. Passer des modèles de mots à des modèles de séquences permet de capturer passivement un certain nombre de phénomènes linguistiques de portée locale, comme certains accords ou encore des différences dans l'ordre des mots entre langues source et cible.

Selon cette approche, pour traduire une phrase source il faudra alors (i) envisager toutes les manières de la découper en segments, (b) envisager tous les équivalents de traduction possibles de ces segments et (c) envisager tous les arrangements possibles des segments en langue cible. Chaque hypothèse de traduction $\mathbf{e}$ d'une phrase $\mathbf{f}$ correspond donc à une segmentation conjointe de la source et de la cible et à un alignement entre segments cible et source ; nous notons dans la suite $\mathbf{a}$ cette segmentation et l'alignement qui lui est associé. La meilleure hypothèse possible est alors celle qui maximise une combinaison de scores qui évaluent chacun différents aspects de l'hypothèse.

Dans la mesure où l'extraction automatique de ces appariements, l'estimation des modèles correspondants et la recherche de la traduction optimale (au sens des opérations (a-c)) posent des problèmes combinatoires difficiles (Knight, 1999), l'apprentissage comme le décodage reposent sur des heuristiques. Typiquement, l'apprentissage consiste à produire des alignements mot à mot symétrisés entre phrases parallèles, alignements desquels sont extraits des segments bilingues ainsi que diverses statistiques afférentes à ces unités. Ces statistiques mesurent notamment (i) la fréquence d'apparition de ces unités et (ii) la fréquence des arrangements relatifs de ces unités dans les phrases source et cible. Les premières permettront d'estimer le *modèle de traduction*, les secondes le *modèle de réordonnancement*. La figure 1 montre un exemple d'alignement symétrisé représenté sous la forme d'une matrice, ainsi que les segments bilingues qui peuvent en être extraits. L'analyse de cette matrice permet d'extraire et d'incrémenter les statistiques relatives à des segments bilingues simples tels que $\begin{bmatrix} \text{alexander} \\ \text{alexandre} \end{bmatrix}$ et $\begin{bmatrix} \text{nikitin} \\ \text{nikitin} \end{bmatrix}$, mais également à des segments plus longs tels que $\begin{bmatrix} \text{cas d' alexander nikitin} \\ \text{case of alexandre nikitin} \end{bmatrix}$ ou $\begin{bmatrix} \text{is the case} \\ \text{s'agit du cas} \end{bmatrix}$.

3. La terminologie anglaise est un peu trompeuse, dans la mesure où les segments appariés ne correspondent pas nécessairement à des constituants syntaxiques.

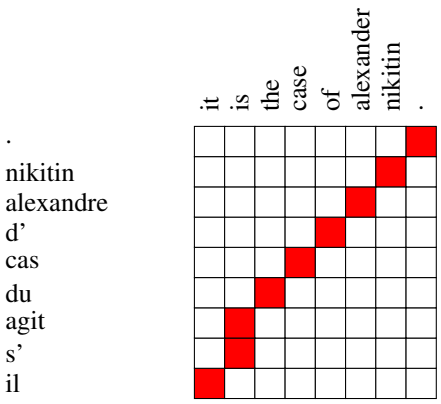

Figure 1. *Une matrice d'alignement de mots : chaque carré plein représente un lien d'alignement*

Pendant la traduction de $\mathbf{f}$, chaque hypothèse $(\mathbf{e}, \mathbf{a})$ reçoit un score $s(\mathbf{e}, \mathbf{a}, \mathbf{f})$ qui combine linéairement les évaluations fournies par les différents modèles disponibles (y compris le modèle de langue cible), selon $s(\mathbf{e}, \mathbf{a}, \mathbf{f}) = \sum_k \lambda_k s_k(\mathbf{e}, \mathbf{f}, \mathbf{a})$. Dans cette formule, chaque score s_k est fourni par un modèle, les coefficients λ_k permettant de pondérer les contributions des différents modèles. La meilleure traduction est celle qui maximise ce critère. Pour des raisons d'efficacité, la recherche du maximum ne considère qu'un petit sous-ensemble, défini de manière heuristique, des alignements et des traductions possibles.

Ces progrès récents en traduction statistique n'auraient pu avoir lieu sans l'émergence d'un consensus fragile autour d'une ou plusieurs métriques automatiques permettant de mesurer la qualité des systèmes. Dans les expériences reportées ci-dessous, nous avons choisi de faire apparaître les performances au sens de la métrique BLEU (Papineni *et al.*, 2002), qui mesure la ressemblance de surface entre les hypothèses de traduction produites automatiquement et des références de traduction produites par des humains.

2.2. *Combiner des langues et des systèmes*

La mise en œuvre des techniques statistiques présentées ci-dessus repose crucialement sur la disponibilité, en quantité suffisante, de ressources bilingues parallèles représentatives du domaine, condition qui n'est aujourd'hui remplie que pour quelques types de textes et paires de langues. Par ailleurs, l'utilisation de multiples

heuristiques lors des diverses étapes d'estimation des modèles conduit à des modèles très bruités et à des systèmes relativement peu robustes à de nouvelles données ou conditions expérimentales, ce qui justifie le recours à des ressources auxiliaires lors de l'apprentissage ou de la traduction. Dans ce contexte, de nombreux travaux récents ont recours à des méthodes *d'ensemble*, qui sont à la fois fondées d'un point de vue statistique, et ont fait leurs preuves dans le domaine de la reconnaissance vocale (Fiscus, 1997). Ces méthodes visent à faire collaborer plusieurs systèmes de traduction automatique, de façon à construire un système plus performant que chacun des systèmes pris isolément. Une manière d'implanter une telle collaboration consiste à recombiner les sorties de plusieurs systèmes en utilisant des mécanismes simulant un vote entre systèmes. L'intuition est que si plusieurs systèmes s'accordent sur un fragment de traduction, alors la plausibilité de ce fragment doit s'en trouver renforcée. D'un point de vue technique, la mise en œuvre de cette intuition est rendue difficile par les différences dans l'ordre des mots qui peuvent exister entre les hypothèses à recombiner. Diverses manières de contourner cette difficulté sont présentées dans (Kumar *et al.*, 2007 ; Rosti *et al.*, 2007 ; Matusov *et al.*, 2008 ; Hildebrand et Vogel, 2008), qui ont permis d'obtenir des améliorations quantitatives très notables des performances en traduction (Callison-Burch *et al.*, 2008).

Les mêmes techniques s'appliquent, avec le même succès (Crego *et al.*, 2009 ; Schroeder *et al.*, 2009), lorsque l'on dispose de systèmes traduisant depuis plusieurs langues sources vers la même langue cible, un cadre souvent désigné sous le nom de *traduction multisource*. Plusieurs auteurs ont, de longue date, noté que les différentes paires de traductions présentent des degrés de difficulté inégaux et que l'utilisation simultanée de plusieurs sources pouvait contribuer à lever certaines ambiguïtés (Kay, 2000). Le travail décrit dans (Och et Ney, 2001) est le premier à proposer une implémentation très simple de cette idée, fondée sur la sélection automatique, pour chaque phrase, de la meilleure source possible. Cette idée a été revisitée et améliorée à plusieurs reprises, en particulier dans (Nomoto, 2004), qui propose de réévaluer les différentes hypothèses à l'aune d'un modèle de langue cible. Les expériences plus récentes de Schwartz (2008), si elles mettent clairement en évidence les potentialités de l'approche multisource, soulignent les limites de la démarche de Och et Ney, dont les gains (pour la métrique BLEU) s'avèrent modestes et permettent de mieux apprécier les approches alternatives.

Un autre cadre s'est avéré fructueux pour faire collaborer plusieurs systèmes traduisant depuis et vers plusieurs langues [4], qui consiste à les utiliser *en série* plutôt qu'en parallèle. Nous désignerons cette démarche générique sous le nom de *traduction par pivot*. Ainsi, pour traduire de A vers C, on pourra mettre à profit la disponibilité de systèmes dits *auxiliaires* pour les paires $A \rightarrow B$ et $B \rightarrow C$ pour traduire en deux étapes.

4. Dans une large mesure, les techniques présentées dans cette section peuvent également servir dans un cadre d'adaptation de systèmes à de nouveaux domaines. Disposant d'un système appris sur un corpus parallèle représentatif de la tâche A, on pourra ainsi synthétiser un nouveau corpus parallèle pour la tâche B, pour autant qu'on dispose d'un corpus représentatif de cette tâche en langue source. Voir, par exemple, (Schwenk, 2008 ; Bertoldi et Federico, 2009).

Cette méthode s'avère d'autant plus efficace que l'on dispose comparativement de peu de données (voire pas de données du tout) pour construire le système *direct*, alors que les ressources existent en des quantités suffisantes pour les systèmes auxiliaires. Trois manières de procéder ont été principalement considérées dans la littérature (Utiyama et Isahara, 2007 ; Wu et Wang, 2007 ; Wu et Wang, 2009) :

– la *méthode par transfert*, qui consiste à passer de A à C en construisant une ou plusieurs traductions dans la langue B, puis en traduisant ces traductions pour obtenir le résultat souhaité ;

– la *méthode synthétique* consiste à utiliser les systèmes auxiliaires pour synthétiser les corpus nécessaires à l'entraînement du système direct. Par exemple, on utilisera le système $B \rightarrow C$ pour traduire la partie cible du corpus parallèle $A \rightarrow B$ et ainsi obtenir un corpus parallèle synthétique pour la paire $A \rightarrow C$;

– une démarche intermédiaire, dite par *triangulation*, consiste à travailler au niveau des modèles, en combinant les modèles de traduction estimés pour les paires $A \rightarrow B$ et $B \rightarrow C$ de manière à construire les modèles nécessaires à la traduction $A \rightarrow C$.

Si le manque de données est souvent invoqué pour justifier le recours à ces techniques, un autre cadre applicatif qui en bénéficie de manière évidente est celui dans lequel il est nécessaire de savoir traduire un grand nombre de paires de langues (par exemple depuis et vers toutes les langues officielles de l'Union européenne). Dans ce contexte, l'utilisation d'une langue pivot, qui joue le rôle d'une *interlingua*, permet de réduire considérablement le nombre de systèmes à développer (Ignat, 2009 ; Koehn *et al.*, 2009). La méthode présentée section 4 et les expériences décrites section 5.4 illustrent le potentiel de cette approche.

Enfin, de façon assez comparable aux méthodes que nous décrivons dans la section suivante, Simard et Isabelle (2009) exploitent la traduction de la meilleure correspondance issue d'une mémoire de traduction sous forme d'un modèle de langue ajouté à la combinaison linéaire de scores calculée par le décodeur d'un système statistique. Cependant, leur approche peut tirer profit de prédictions relativement étendues, alors que nous proposerons notamment d'exploiter des sources de connaissances automatiques, pour lesquelles les prédictions ne peuvent être introduites que sous forme de prédictions locales (*lexicales*).

3. Un cadre pour la micro-adaptation lexicale

3.1. *Présentation du cadre général*

Comme nous l'avons présenté dans la section 2.2, de nombreux travaux en traduction automatique statistique ont porté sur l'exploitation conjointe de plusieurs systèmes impliquant au moins deux langues pour améliorer collectivement la qualité d'un ensemble de systèmes. Le présent travail se démarque de ces travaux par le fait qu'il vise à augmenter les performances d'un système particulier, que nous appellerons le *système direct*. Pour des raisons qui seront exposées par la suite, le système direct

implémente une approche statistique de la traduction automatique. Dans cette section, nous décrivons tout d'abord le cadre général dans lequel nous nous situons, puis nous introduisons notre approche, qui se fonde sur l'exploitation de traductions obtenues par triangulation *via* des langues auxiliaires. Ce cadre est présenté figure 2 : sur la figure, le système direct correspond à la configuration **1**, la paire de langues allemand → anglais (de → en) ayant été choisie à titre d'illustration.

Un système de traduction statistique effectue une recherche heuristique visant à construire la meilleure hypothèse atteignable. Cette recherche est guidée par les informations fournies par différents modèles probabilistes, dont les paramètres sont estimés lors de l'apprentissage. Notre objectif est de permettre à des sources externes, telles que d'autres systèmes de traduction complets, des modules de TAL spécialisés ou des ressources telles que des dictionnaires ou des terminologies bilingues, voire des traducteurs humains, d'orienter ce décodage en indiquant des éléments lexicaux (mots ou groupes de mots) qu'il est souhaitable de faire figurer dans la meilleure solution. Néanmoins, ces préférences ne peuvent être satisfaites que si elles appartiennent à l'espace de recherche du système principal [5].

Une manière naturelle de procéder consiste à demander que les informations fournies par les sources externes prennent la forme d'hypothèses auxiliaires de traduction, qui sont utilisées pour renforcer la vraisemblance de certains mots. Ce renforcement est implémenté en dérivant de ces hypothèses auxiliaires un modèle de langue spécifique à la phrase à traduire qui sera combiné avec les autres modèles de langue disponibles : en ce sens, notre démarche utilise des outils classiques en adaptation de modèles de langue (Bellagarda, 2001), en se situant toutefois à un niveau microscopique, celui de la phrase (plutôt que d'adapter sur tout un document), de façon à renforcer les choix lexicaux. L'ordre du modèle de langue utilisé dépendra de la capacité de la source de prédictions à générer des segments bien formés. Nous désignons cette approche sous le terme de *micro-adaptation lexicale*.

Différentes sources pouvant engendrer les informations utilisées pour réaliser cette micro-adaptation lexicale sont illustrées par les configurations **2** à **6** sur la figure 2. Notre cadre autorise naturellement l'utilisation simultanée de plusieurs modèles auxiliaires, dont la contribution peut être pondérée à l'aide des techniques habituellement utilisées pour optimiser les systèmes de traduction statistique. Concrètement, un score additionnel correspondant au modèle de langue pour les meilleures hypothèses de traduction produites par une source auxiliaire est ajouté à la combinaison linéaire de scores maximisée par le décodeur statistique utilisé. Ceci permet, à l'extrême, de ne pas utiliser le modèle provenant d'une source auxiliaire si celle-ci ne permet pas d'améliorer la qualité des traductions pour des textes de *développement*.

La description des configurations de la figure 2 va nous permettre d'établir des liens de façon explicite avec d'autres approches évoquées dans la section 2.2.

5. Une extension de ce travail pourrait porter sur l'acquition dynamique de connaissances, en particulier dans un cadre où interviendraient des traducteurs humains.

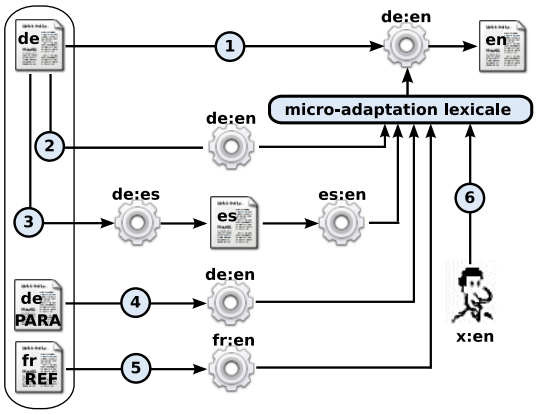

Figure 2. *Cadre général pour la micro-adaptation lexicale impliquant plusieurs configurations pour la génération de prédictions lexicales auxiliaires*

Dans la configuration **2**, le fichier source traduit par le système direct est également traduit par un autre système pour la même paire de langues. L'exploitation conjointe des hypothèses proposées par plusieurs systèmes traduisant le même fichier source correspond exactement au contexte de la combinaison de systèmes, qui prend une importance croissante dans les campagnes d'évaluation internationales [6]. Les méthodes de combinaison actuelles se divisent en deux catégories. D'une part, celles qui sont capables de produire, par combinaison, des hypothèses originales (c'est-à-dire qui n'apparaissent pas parmi les hypothèses proposées), au risque de produire des hypothèses très erronées ; la méthode décrite section 4 entre dans cette catégorie. D'autre part, celles qui ne font que *reclasser* les hypothèses existantes (par exemple (Rosti *et al.*, 2007)), qui sont plus conservatives, mais ont des potentiels de gains plus faibles. La méthode décrite ci-dessus constitue un troisième mode de recombinaison, qui va au-delà du reclassement simple d'hypothèses, tout en offrant les mêmes « garanties » de cohérence des traductions produites que le système direct, puisque les hypothèses produites sont nécessairement atteignables par le système direct. Bien que ceci puisse être présenté comme une limitation, le fait d'avoir la construction de la nouvelle hypothèse sous le contrôle direct d'un seul système permet, outre un diagnostic facilité du décodage effectué, d'éviter les effets indésirables induits par la fusion d'hypothèses partielles [7]. Pour conclure sur cette configuration, notons qu'il est tout à fait possible que le système auxiliaire utilise une autre technologie de traduction, auquel cas notre cadre fournit une instance nouvelle d'hybridation entre systèmes de nature différente.

6. Voir par exemple `http://www.statmt.org/wmt10/system-combination-task.html`
7. Nos travaux futurs porteront notamment sur l'étude de cette configuration, qui semble particulièrement intéressante lorsque tous les systèmes utilisés ont accès aux mêmes données d'apprentissage, ce qui est typiquement le cas dans les campagnes d'évaluation mentionnées.

La configuration **3**, qui sera décrite en détail section 3.2, utilise des traductions obtenues par triangulation *via* une langue auxiliaire (l'espagnol dans notre exemple). L'utilisation d'une double traduction entre une langue source et une langue auxiliaire, puis de cette langue auxiliaire vers une langue cible, est, comme expliqué ci-dessus, une manière d'implanter une *traduction par pivot*. À la différence des approches de l'état de l'art, les hypothèses obtenues par pivot, qui sont souvent de qualité moindre que celles d'un système direct quand des quantités comparables de données d'apprentissage sont disponibles (voir des résultats détaillés en section 5.3, et en particulier dans le tableau 3), ne sont ici utilisées que pour guider la recherche de la meilleure hypothèse relativement à l'espace de recherche du système direct.

Les configurations **4** et **5** sont des instances de *traduction multisource*, dans lesquelles une réécriture (ou *paraphrase*) du texte source (configuration **4**) ou encore une traduction dans une troisième langue (le français dans la configuration **5**) de haute qualité est déjà disponible. Finalement, la configuration **6** représente un scénario dans lequel un traducteur humain ayant des connaissances dans l'une des langues sources disponibles spécifie des préférences lexicales [8]. Une telle approche pourrait être utilisée par un traducteur qui, étant donné une phrase à traduire, anticiperait les difficultés de traduction du système en lui fournissant en amont des éléments lexicaux utiles. La possibilité alors offerte de spécifier de façon itérative de tels éléments en temps réel lorsque le système propose de nouvelles hypothèses rentre dans le cadre décrit par (Dymetman *et al.*, 2003), dans lequel un expert humain supervise un processus d'analyse automatique par interaction avec un *texte de contrôle* (*feedback text*).

La description de ce nouveau type de systèmes de traduction automatique soulève plusieurs questions. La première porte sur la validité d'une telle approche qui, dans certaines configurations (**2-5**), vise à améliorer les performances d'un système de traduction statistique en utilisant d'autres systèmes automatiques. Si cette hypothèse est confirmée, une seconde question importante consiste à évaluer l'ampleur des gains que l'on pourrait espérer obtenir par cette voie indirecte, qui consiste à améliorer la traduction de A vers B en améliorant la qualité des paires $A \rightarrow C$ et $C \rightarrow B$.

3.2. *Utilisation de traductions obtenues par triangulation* via *une langue auxiliaire*

Dans ce travail, nous souhaitons étudier la possibilité d'utiliser comme sources auxiliaires pour guider la micro-adaptation lexicale des traductions obtenues par triangulation *via* des langues auxiliaires. Cette configuration apparaît difficile de prime

8. On ne peut pas parler ici de *contraintes* lexicales, car les mots que spécifierait un traducteur ne pourraient être effectivement employés que s'ils appartiennent à des solutions atteignables par le décodeur utilisé. Cependant, des adaptations pourraient être envisagées pour rendre ce cadre plus flexible, comme celle qui renforcerait la sélection de mots de même lemme ou de même sens. Ainsi, notre proposition pourrait intégrer des caractéristiques de *désambiguïsation interactive*.

abord, car elle repose sur l'utilisation d'au moins deux autres systèmes automatiques. Cependant, l'intuition principale qui motive notre choix, également mise en avant dans les travaux en traduction multisource, est que le passage par d'autres langues auxiliaires permettra, dans certains cas, de désambiguïser le texte à traduire. Contrairement au cas de la traduction multisource, où des traductions de haute qualité dans d'autres langues sont disponibles, nous voyons le fait que les traductions soient produites automatiquement comme un argument fort en faveur de notre approche si une implémentation de celle-ci parvient à améliorer la qualité des traductions délivrées par le système direct. En outre, ce choix nous permettra de répondre à la question de la transposition des gains que nous avons évoquée plus haut : il est en effet naturel de simuler l'amélioration de la qualité d'un système auxiliaire en augmentant artificiellement la quantité de données d'apprentissage utilisée pour le construire.

Algorithme 1 : traduction d'une phrase source par micro-adaptation avec des traductions obtenues par triangulation *via* une langue auxiliaire

pour chaque *phrase à traduire dans la langue SRC* **faire**
  Traduire la phrase source avec le système SRC→AUX et extraire les n meilleures hypothèses
  pour chaque n *meilleure hypothèse dans la langue AUX* **faire**
    Traduire l'hypothèse avec le système AUX→CIB et extraire les m meilleures hypothèses
  fin
  Construire un corpus à partir des $n * m$ hypothèses en langue cible CIB
  Apprendre un modèle de langue bigramme L_{aux} sur ce corpus
  Utiliser la probabilité fournie par L_{aux} comme score additionnel pour traduire la phrase source avec le système SRC→CIB
fin

Nous utilisons donc ici les traductions auxiliaires obtenues par triangulation comme données d'adaptation pour construire un modèle de langue cible additionnel, comme décrit dans l'algorithme 1. La présence de mots, ou plus généralement de *n*-grammes, dans ces prédictions peut servir à renforcer le score des hypothèses les incluant pour le décodeur du système direct [9]. Cette idée a déjà été mise en œuvre avec succès en reconnaissance automatique de la parole, *via* l'utilisation de sous-titres (Placeway et Lafferty, 1996) ou de traductions automatiques de documents comparables (Paulik *et al.*, 2005) en tant que données d'adaptation représentatives du domaine pour le modèle de langage du système de reconnaissance.

Une implémentation simple de cette stratégie consiste à utiliser les n meilleures hypothèses (communément appelées *listes des* n-*meilleures traductions*) résultant de

9. Nous nous limitons ici au renforcement au niveau de chaque phrase à traduire. La prise en compte de prédictions provenant du même discours (par exemple, un document *cohérent*) pourra être étudiée par la suite, en particulier si l'on prend en considération le fait que l'ambiguïté lexicale peut être résolue en combinant des informations au niveau du document (Carpuat, 2009).

la triangulation *via* une langue auxiliaire pour apprendre un modèle de langue (Crego *et al.*, 2010). Toutefois, s'agissant de données traduites automatiquement, certains groupes de mots produits ne seront pas fiables, en particulier ceux qui apparaissent à la frontière des segments. Il est cependant facile de les filtrer en utilisant des informations d'alignement de segments que fournissent les décodeurs statistiques. Plus généralement, la qualité attendue des prédictions obtenues par triangulation suggère de limiter l'empan des modèles de langue appris à de petites valeurs. Nous nous sommes ainsi limités à des modèles bigrammes pour nos expériences.

La configuration que nous avons décrite est illustrée figure 3 : le système direct (configuration **1**) exploite les prédictions issues de la triangulation *via* une langue auxiliaire (configuration **2**) sous forme de listes des *n*-meilleures hypothèses. Afin de connaître la performance maximale que notre approche permettrait d'atteindre si l'on disposait de prédictions lexicales optimales relativement à la mesure de performance fondée sur la comparaison avec une traduction de référence, nous pouvons faire l'expérience artificielle dans laquelle une telle traduction de référence est utilisée comme source d'information « parfaite » (configuration **3**). Par ailleurs, la performance des systèmes triangulés, ou pivots (configuration **4**), donnera une indication de la qualité de la source d'information effectivement utilisée pour l'adaptation par modèle de langue.

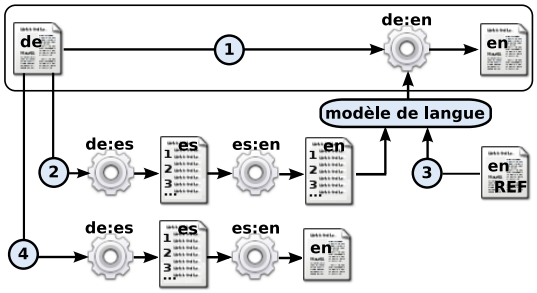

Figure 3. *Architecture d'un système traduisant de l'allemand vers l'anglais et réalisant une micro-adaptation lexicale par exploitation d'une traduction automatique obtenue par triangulation* via *l'espagnol*

4. Amélioration de la source auxiliaire par combinaison de systèmes multipivots

La qualité de la source d'information utilisée apparaît bien évidemment comme cruciale pour le succès de l'approche que nous avons décrite dans les sections précédentes. En particulier, les effets attendus de la triangulation *via* une langue auxiliaire devraient pouvoir se composer, dans une certaine mesure, si plusieurs langues auxiliaires sont utilisées au lieu d'une seule comme illustré sur la figure 3. Néanmoins, l'exploitation effective de plusieurs langues auxiliaires reste à considérer.

Une première solution, qui s'intégrerait directement dans l'architecture décrite, consisterait à exploiter l'union des n meilleures prédictions de plusieurs systèmes pivots pour construire le modèle de langue auxiliaire. Quoique simple à implémenter, cette approche ne permettrait pas d'apprendre de façon directe quelles langues contribuent aux prédictions de meilleure qualité. Une seconde solution consisterait donc à utiliser simultanément autant de sources auxiliaires, et donc de modèles de langue, que de systèmes pivots disponibles. De la sorte, l'optimisation du système direct exploitant ces nouveaux modèles permettrait de donner plus d'importance aux systèmes pivots réalisant les prédictions les plus utiles pour améliorer la qualité des traductions.

Il est toutefois possible qu'une langue pivot qui serait généralement utile réalise localement de mauvaises prédictions. Une façon de prévenir ce type de problème consiste à examiner l'ensemble des hypothèses produites par tous les systèmes pivots disponibles et à réaliser une combinaison fondée sur un consensus entre systèmes. Cette approche a en effet montré d'importants gains relativement au meilleur système utilisé dans la combinaison sur de nombreuses tâches (Matusov *et al.*, 2008 ; Schroeder *et al.*, 2009 ; Leusch *et al.*, 2009). La nouveauté consiste ici dans le fait de combiner plusieurs systèmes pivots, ce qui n'a, à notre connaissance, jamais été fait.

La figure 4 présente l'architecture d'un système obtenu par combinaison de plusieurs systèmes pivots, reprenant l'approche de combinaison fondée sur le consensus décrite dans (Leusch *et al.*, 2009). Le même texte source, en allemand sur la figure, est tout d'abord traduit en parallèle par triangulation *via* plusieurs langues (par exemple danois, espagnol et français). Les meilleures hypothèses de chaque système [10] sont ensuite alignées automatiquement au niveau des mots puis un réseau de confusion (*confusion network*) est construit en prenant successivement chaque hypothèse comme ossature. L'ensemble des réseaux de confusion construit est ensuite fusionné dans un unique treillis de mots, qui est alors réévalué à l'aide d'un modèle de langue cible [11] et de poids associés à chacun des systèmes. La meilleure hypothèse est finalement extraite et est utilisée comme hypothèse du système de combinaison. Par construction, cette hypothèse ne correspond pas nécessairement à l'hypothèse d'un des systèmes et donc peut être globalement originale.

S'il semble raisonnable d'espérer une amélioration de la source auxiliaire utilisée pour la micro-adaptation lexicale par l'approche décrite, des réserves peuvent être formulées sur le coût de construction des systèmes pivots nécessaires. Notre principale réponse consiste à mettre en avant la réutilisation de systèmes lorsque ceux-ci ont déjà été construits pour d'autres raisons. Cependant, la construction de systèmes supplémentaires pourrait trouver sa justification si les gains obtenus s'avéraient importants.

10. Les expériences conduites jusqu'à présent n'ont pas mené à une amélioration des performances par l'utilisation des n meilleures hypothèses de chaque système.

11. Ce modèle de langue cible est appris sur les hypothèses en entrée de la combinaison, voir (Matusov *et al.*, 2008).

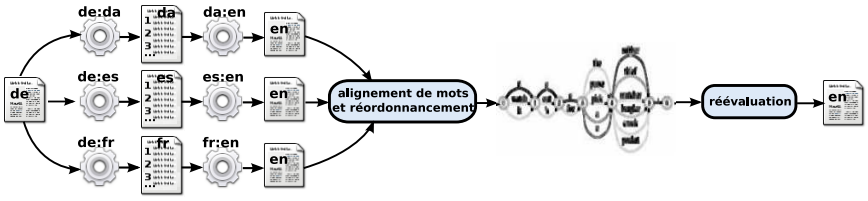

Figure 4. *Système de traduction multipivots traduisant de l'allemand vers l'anglais obtenu par combinaison de systèmes pivots* via *le danois, l'espagnol et le français*

5. Expériences

5.1. *Description des données*

Comme source de textes parallèles multilingues, nous avons utilisé le corpus de débats parlementaires européens Europarl (Koehn, 2005) pour les onze langues suivantes [12] : allemand (de), anglais (en), danois (da), espagnol (es), finnois (fi), français (fr), grec (el), italien (it), néerlandais (nl), portugais (pt) et suédois (sv). Afin d'utiliser des systèmes aux performances comparables, nous avons retenu la partie commune à toutes les langues du corpus, pour un total de 318 804 lignes (soit environ 10,3 millions d'occurrences pour la partie française du corpus). Pour ces expériences, désignées ci-dessous comme correspondant à la condition *intersection*, nous avons construit des corpus de développement et de test en sélectionnant respectivement 500 et 1 000 phrases, qui sont donc exclues du corpus d'apprentissage. Diverses statistiques concernant ce corpus et qui détaillent en particulier la taille des données d'apprentissage, de développement et de test, sont reportées dans le tableau 1.

Nos expériences portent sur trois paires de langues : en étudiant la traduction depuis l'allemand vers l'anglais (de $\rightarrow$ en) ou depuis le français vers l'allemand (fr $\rightarrow$ de), nous nous plaçons dans une situation où la traduction est difficile et peut bénéficier de systèmes auxiliaires bien meilleurs. Nous avons également considéré la paire fr $\rightarrow$ en, pour laquelle nous disposons déjà de systèmes très performants, afin de voir si les améliorations obtenues persistent quand le système direct obtient des performances de base élevées.

Pour valider expérimentalement notre méthode dans des conditions plus réalistes, nous avons également reproduit ces expériences pour la direction fr $\rightarrow$ de (en nous limitant à la meilleure langue pivot, à savoir l'espagnol) avec des corpus plus volumineux, correspondant aux données de la campagne d'évaluation du *Workshop on Statistical Machine Translation* [13]. Pour ces expériences, identifiées ci-après sous le nom de *condition complète*, le corpus d'apprentissage utilise l'ensemble des données

12. Nous rappelons entre parenthèses les codes de langue ISO qui seront utilisés par la suite.
13. http ://www.statmt.org

	Apprentissage		*Développement*			*Test*		
	occurrences	*vocables*	*occurrences*	*vocables*	*HV*	*occurrences*	*vocables*	*HV*
da	8,5 M	133,5 k	13,4 k	3,2 k	104	25,9 k	5,1 k	226
de	8,5 M	145,3 k	13,5 k	3,5 k	120	26,0 k	5,5 k	245
en	8,9 M	53,7 k	14,0 k	2,8 k	39	27,2 k	4,0 k	63
es	9,3 M	85,3 k	14,6 k	3,3 k	56	28,6 k	5,0 k	88
fi	6,4 M	274,9 k	10,1 k	4,3 k	244	19,6 k	7,1 k	407
fr	10,3 M	67,8 k	16,1 k	3,2 k	47	31,5 k	4,8 k	87
el	8,9 M	128,3 k	14,1 k	3,9 k	72	27,2 k	6,2 k	159
it	9,0 M	78,9 k	14,3 k	3,4 k	61	28,1 k	5,1 k	99
nl	8,9 M	105,0 k	14,2 k	3,1 k	76	27,5 k	4,8 k	162
pt	9,2 M	87,3 k	14,5 k	3,4 k	49	28,3 k	5,2 k	118
sv	8,0 M	140,8 k	12,7 k	3,3 k	116	24,5 k	5,2 k	226

Tableau 1. *Taille des corpus d'apprentissage, de développement et de test utilisés pour la condition « intersection ». Les nombres d'occurrences de formes hors vocabulaire apparaissent dans la colonne « HV »*

de la version 5 du corpus Europarl ; les données d'entraînement et de test sont également celles distribuées par les organisateurs pour la campagne d'évaluation de 2008 (voir le tableau 2).

Corpus	**Nb. phrases**	**Nb. mots**	**Taille vocab.**	**Taille hors voc.**	**Taille max.**	**Taille moy.**
train.de	1,2 M	33,1 M	297 k	-	99	26,1
train.fr		39 M	116 k	-	99	30,7
dev.de	500	13,6 k	3,4 k	49	107	27,2
dev.fr		15,6 k	3,1 k	25	113	31,2
tst.de	2 k	56,6 k	8,8 k	268	143	28,3
tst.fr		65,6 k	7,3 k	97	156	32,8
train.es	1,2 M	36,3 M	144,7 k	-	99	28,6
train.fr		39 M	116,9 k	-	99	30,7
dev.fr	500	15,6 k	3,1 k	26	113	31,2
dev.es		15 k	3,2 k	27	155	30,1
tst.fr	2 k	65,6 k	7,3 k	101	156	32,8
tst.es		61,8 k	7,8 k	145	138	30,9
train.de	1,2 M	32,8 M	296,6 k	-	99	26,1
train.es		35,9 M	144,6 k	-	99	28,6
dev.de	500	13,6 k	3,4 k	51	107	27,2
dev.es		15 k	3,2 k	27	155	30,1
tst.de	2 k	56,6 k	8,8 k	273	143	28,3
tst.es		61,8 k	7,8 k	146	138	30,9

Tableau 2. *Taille des corpus d'apprentissage, d'entraînement et de test pour la condition complète*

5.2. *Description des systèmes*

Tous les systèmes de traduction utilisés pour ces expériences utilisent des systèmes statistiques qui implémentent une variante de l'approche à base de segments décrite à la section 2. La principale différence avec les modèles à base de segments conventionnels est le découplage entre réordonnancement et traduction : dans l'approche par n-grammes (Casacuberta et Vidal, 2004 ; Mariño *et al.*, 2006), l'espace de recherche contenant les réordonnancements possibles de la source est engendré avant la recherche de la meilleure traduction possible, qui est elle effectuée de façon monotone et prend appui sur des modèles n-grammes d'unités bilingues, appelés *tuples*. Dans cette approche, les modèles de traduction sont donc des modèles de langue conventionnels (n-grammes), à ceci près que les unités manipulées sont des paires (bilingues) de séquences de taille variable. Lors du décodage, seuls un petit nombre de réordonnancements possibles, précalculés et représentés sous la forme d'un treillis de mots en langue source, sont considérés.

Un modèle de réordonnancement en source, qui permet d'engendrer ces treillis à partir d'une phrase source, est appris automatiquement depuis un corpus parallèle bilingue aligné au niveau des mots et dans lequel les mots en langue source ont été réordonnés pour suivre l'ordre des mots en langue cible (Crego et Marino, 2007). Ce modèle prend la forme d'un ensemble de règles qui expriment les réarrangements, en langue source, qui permettent de restituer l'ordre des mots en langue cible. Pour augmenter la capacité de généralisation du modèle, ces règles portent, lorsque cette information est disponible, sur les catégories morphosyntaxiques plutôt que sur des formes brutes. Ainsi, par exemple, une règle telle que $\mathrm{NN\ JJ} \rightsquigarrow \mathrm{JJ\ NN}$ permet d'exprimer l'inversion des positions relatives des adjectifs et noms lorsque l'on passe du français à l'anglais. Les phrases du corpus d'apprentissage en français, en anglais, en espagnol et en allemand ont été analysées avec le TreeTagger [14] pour obtenir les catégories morphosyntaxiques ; pour les autres langues, les règles de réordonnancement sont apprises directement sur des séquences de formes, ce qui conduit, en pratique, à des systèmes un peu moins performants.

En plus du modèle n-grammes bilingue, nos systèmes de traduction utilisent huit scores qui sont combinés linéairement ; les poids des différents modèles sont optimisés sur les corpus de développement par maximisation du score BLEU sur un corpus de développement, selon la méthodologie initialement proposée dans (Och, 2003) [15]. Ces huit scores sont :

– deux modèles de traduction lexicaux, qui complètent le modèle de *tuples* ;

– deux modèles de réordonancement lexical (Tillmann, 2004), qui visent à évaluer la vraisemblance du réordonnancement de la phrase source proposée. Ces modèles se fondent pour leur évaluation sur l'ordre relatif (en source) de deux unités voisines (en

14. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`.

15. Nous utilisons ici l'implémentation disponible dans le toolkit Moses (`http://www.statmt.org/moses`), légèrement adaptée à nos besoins.

cible) ;

– un modèle de distorsion fondé sur la distance (en source) entre deux mots voisins (en cible) ;

– un modèle de langage n-grammes de la langue cible, qui permet d'évaluer la qualité syntaxique des hypothèses de traduction ;

– deux scores permettant de réguler la longueur des hypothèses produites, et qui visent à modérer la propension du modèle à produire des traductions brèves ; en dernière analyse, cette préférence découle de l'utilisation de modèles n-grammes.

Dans cette étude, les deux modèles de langue (le modèle de traduction et le modèle de la langue cible) sont des modèles d'ordre 2 (3-grammes), estimés avec la boîte à outils SRILM (Stolcke, 2002), en choisissant le schéma de lissage connu sous le nom de *modified Kneser-Ney* (Chen et Goodman, 1996). Les seules ressources utilisées sont les corpus parallèles, dont la partie cible sert à apprendre les modèles de langue. Pour estimer les autres modèles, nous mettons en œuvre une chaîne de traitements standard, consistant à segmenter en mots, à réaliser un alignement mot à mot avec le logiciel *GIZA++* [16], puis à construire des alignements de blocs par symétrisation des alignements mot à mot obtenus dans les deux directions. Comme expliqué *supra*, les unités de traduction bilingues et les modèles de réordonnancement prennent appui sur ces alignements symétrisés.

5.3. *Adaptation lexicale par modèles de langue*

Dans cette section, nous présentons les résultats d'une série d'expériences qui montrent l'intérêt de l'introduction d'un modèle de langue auxiliaire pour renforcer les choix lexicaux opérés par les systèmes auxiliaires traduisant par pivot. Un score de modèle de langue, calculé pour chaque énoncé à traduire sur les hypothèses obtenues automatiquement par triangulation *via* une langue auxiliaire donnée, est ajouté à la combinaison linéaire maximisée par le décodeur du système direct. La contribution de ce modèle peut ainsi être optimisée avec celles des autres modèles utilisés par le système direct sur le corpus de développement.

Les résultats sont rassemblés dans le tableau 3, qui donne les performances quantitatives des systèmes pour les trois paires de langues principales considérées dans cette étude, ainsi que pour les différents systèmes auxiliaires utilisés. Pour chaque triplet de langues (*src*, *aux*, *cib*), les colonnes 4 à 6 contiennent les scores BLEU respectivement des systèmes *src* $\rightarrow$ *aux*, *aux* $\rightarrow$ *cib* et *src* $\rightarrow$ *aux* $\rightarrow$ *cib* (pivot) ; ces scores permettent d'apprécier la qualité des systèmes auxiliaires.

Les deux dernières colonnes donnent les scores BLEU pour les systèmes de traduction principaux obtenus respectivement sans (système de base pour la traduction directe) et avec utilisation d'un système auxiliaire. Cette dernière valeur est présentée

16. http ://www.fjoch.com/GIZA++.html.

sous la forme de l'incrément par rapport à l'utilisation du système de base [17]. Nous présentons enfin les résultats obtenus lorsque les modèles de langue auxiliaires sont construits à partir des traductions de référence (ligne « référence »), ce qui permet d'apprécier le potentiel théorique de cette méthode.

Comme expliqué ci-dessus, les modèles de langue auxiliaires sont appris sur des pseudo-références produites automatiquement en passant par une langue auxiliaire. Il est important de souligner que ces modèles de langue ne sont pas nécessairement appris sur une unique pseudo-référence, puisque nous avons considéré la possibilité de conserver les n-meilleures traductions pour la paire *src* $\rightarrow$ *aux* et les m-meilleures pour la paire *aux* $\rightarrow$ *cib*, soit au maximum $n * m$ pseudo-références, qui sont toutes considérées comme équiprobables pour estimer le modèle de langue. Compte tenu de la très petite taille des corpus sur lesquels ce modèle est appris, le lissage de *Kneser-Ney* n'est pas possible. Nous utilisons une technique de lissage plus simple, connue sous le nom de lissage de *Witten-Bell* [18].

Première observation : tous les systèmes pivots s'avèrent moins bons que les systèmes directs correspondants, ce qui était attendu dans la mesure où tous les systèmes sont appris sur des ensembles de données de taille comparable. Néanmoins, l'ajout d'un modèle auxiliaire prenant appui sur les hypothèses produites par pivot permet, dans de nombreux cas, d'améliorer, parfois de manière sensible, les performances du système direct. C'est particulièrement vrai pour les paires de $\rightarrow$ en et fr $\rightarrow$ de, qui bénéficient de la combinaison avec des systèmes obtenus en pivotant *via* des langues pour lesquelles les systèmes *src* $\rightarrow$ *aux* ou *aux* $\rightarrow$ *cib* sont meilleurs que le système direct. Mais une amélioration est également observée pour notre système fr $\rightarrow$ en, dont les performances de base sont déjà très élevées. On notera également que la corrélation entre l'amplitude des améliorations obtenues et les performances des systèmes auxiliaires est très lâche : des systèmes auxiliaires en apparence très proches peuvent ainsi conduire à des incréments de performances très divers. Inversement, un système qui semble médiocre peut s'avérer utile : utiliser le finnois comme langue auxiliaire apporte des améliorations, alors même que les scores des systèmes auxiliaires depuis et vers le finnois sont très mauvais [19]. On notera enfin que l'apport des langues auxiliaires dépend des langues considérées : ainsi, par exemple, le grec s'avère utile pour les systèmes qui traduisent depuis et vers l'allemand ; le suédois pour traduire vers l'anglais, etc. Plus que la qualité intrinsèque des systèmes auxiliaires, il semble [20]

17. Pour tous les résultats présentés sur la condition *intersection*, une différence de 1 point est considérée comme significative à 95 % de précision.

18. Nous ne pensons pas que ce choix particulier influence de manière significative les résultats.

19. Ces scores sont toutefois à interpréter avec prudence : le finnois figure typologiquement parmi les langues agglutinatives, ce qui transparaît dans le comparativement petit nombre d'occurrences et grand nombre de vocables dans les corpus finnois (voir le tableau 1). En conséquence, prédire correctement un mot en finnois s'avère aussi difficile que prédire correctement plusieurs mots dans d'autres langues, d'où de très mauvais scores BLEU.

20. Il est difficile de tirer de ces quelques chiffres des leçons très claires sur les raisons linguistiques de l'apport de telle ou telle langue, dans la mesure où de nombreux artefacts, qui tiennent à la manière dont le corpus Europarl est produit, rendent les analyses très fragiles.

src aux cib	*src-aux*	*aux-cib*	*pivot*	*src-cib*	*+auxLM*
fr - de	-	-	-	18,02	
es	**34,94**	18,31	**16,76**		**+ 0,96**
el	24,54	18,51	15,86		+ 0,76
fi	10,71	14,15	11,39		+ 0,65
nl	22,71	**21,44**	**16,76**		+ 0,55
en	29,53	17,31	15,69		+ 0,50
da	22,78	20,02	16,27		+ 0,44
it	31,60	16,86	16,54		− 0,05
pt	33,61	17,47	16,34		− 0,12
sv	20,73	19,59	13,73		− 0,14
référence	-	-	-		+ 6,46
fr - en	-	-	-	29,53	
es	**34,94**	**31,05**	**27,76**		**+ 0,61**
sv	20,73	30,98	23,74		+ 0,50
fi	10,71	20,56	19,15		+ 0,44
it	31,60	25,75	25,79		+ 0,32
el	24,54	29,37	25,31		+ 0,07
de	18,02	24,66	23,50		+ 0,05
da	22,78	29,54	25,48		+ 0,02
nl	22,71	24,49	25,15		+ 0,01
pt	33,61	29,44	27,27		+ 0,01
référence	-	-	-		+ 11.30
de - en	-	-	-	24,66	
el	19,72	29,37	20,88		**+ 1,02**
da	24,59	29,54	**2273**		**+ 0,96**
fi	12,42	20,56	18,02		**+ 0,94**
pt	23,15	29,44	21,93		+ 0,87
es	**25,48**	**31,05**	21,23		+ 0,77
sv	19,80	**30,98**	21,35		+ 0,69
nl	24,97	24,49	22,62		+ 0,64
fr	25,93	29,53	21,55		+ 0,19
it	18,82	25,75	18,05		+ 0,19
référence	-	-	-		+ 9,53

Tableau 3. *Résultats (scores BLEU) pour nos systèmes implémentant la micro-adaptation lexicale pour les trois paires de langues français → allemand, français → anglais et allemand → anglais sur le corpus* intersect

bien que ce soit leur complémentarité avec les autres systèmes qui soit déterminante pour obtenir des améliorations.

src aux trg	*src-aux*	*aux-trg*	*pivot*	*src-trg*	*+auxLM*
fr - de	-	-	-	19,94	
es	38,76	20,18	19,36		+ 0,61

Tableau 4. *Résultats (scores BLEU) pour le système implémentant la micro-adaptation lexicale par triangulation* via *l'espagnol pour la paire de langues français → allemand sur le corpus* complet

Le tableau 4, qui fait référence à la condition *complète*, permet de constater que ces améliorations subsistent *y compris lorsque l'on injecte l'ensemble des données disponibles à l'apprentissage*. Ce résultat est particulièrement intéressant, puisque, partant d'un système de base de deux points meilleur, nous observons une amélioration nette du système utilisant des ressources auxiliaires.

5.4. *Combinaison de systèmes multipivots*

Comme nous l'avons décrit dans la section 4, il est possible d'espérer améliorer la qualité d'une source auxiliaire en utilisant simultanément les prédictions réalisées par plusieurs triangulations *via* différentes langues dont les complémentarités pourraient se conjuguer. Nous avons donc essayé d'obtenir de telles améliorations en implémentant l'architecture décrite figure 4, et en distinguant le cas où le système direct ne participe pas à la combinaison (**multipiv-direct**) et le cas où il y participe (**multipiv+direct**) [21]. Les langues auxiliaires ont été ajoutées à la combinaison par caractère prometteur décroissant (fondé sur le score BLEU de la meilleure sortie pivot [22]), et l'optimisation des paramètres de la combinaison repose sur la recherche d'une valeur maximisant un score (TER-BLEU) permettant d'obtenir des résultats stables pour ce type de combinaison.

Le tableau 5 montre, pour chaque paire de langues considérée, le score BLEU obtenu pour la meilleure hypothèse de la combinaison en considérant l'apport d'un ensemble de langues auxiliaires, ainsi que les valeurs obtenues pour les systèmes directs et les différents systèmes pivots. La comparaison avec la performance du système direct permet de constater que la sortie obtenue par combinaison peut être améliorée de façon significative (assez nettement pour la paire fr → de où une amélioration de 1,4 point BLEU pour **multipiv-direct** et 1,6 pour **multipiv+direct** est obtenue en utilisant toutes les langues auxiliaires), bien que les systèmes auxiliaires utilisés aient été appris sur les mêmes données d'apprentissage que les systèmes directs. On notera que

21. Les hypothèses du système direct utilisées sont ici, bien entendu, celles proposées avant toute combinaison ou micro-adaptation.

22. Les petites variations d'ordre sont dues à des différences très sensibles entre deux scripts différents de calcul de score BLEU.

l'utilisation du système direct dans la combinaison joue un rôle plus important pour la paire de langues fr → en, qui est *a priori* plus difficile à améliorer.

src aux trg	*src-aux*	*aux-trg*	*pivot*	*direct*	*multipiv-direct*	*multipiv+direct*
fr - de	-	-	-	18,02		
es	**34,94**	18,31	**16,76**		|	|
nl	22,71	**21,44**	**16,76**		|	+ 0,83
pt	33,61	17,47	16,34		− 0,01	+ 0,84
it	31,60	16,86	16,54		+ 0,38	+ 0,80
en	29,53	17,31	15,69		+ 0,46	+ 1,13
da	22,78	20,02	16,27		+ 0,69	+ 1,28
el	24,54	18,51	15,86		+ 0,95	+ 1,57
sv	20,73	19,59	13,73		+ 1,27	**+ 1,61**
fi	10,71	14,15	11,39		**+ 1,39**	**+ 1,61**
fr - en	-	-	-	29,53		
es	**34,94**	**31,05**	**27,76**		|	|
pt	33,61	29,44	27,27		|	+ 0,01
el	24,54	29,37	25,31		+ 0,19	+ 0,43
it	31,60	25,75	25,79		+ 0,48	+ 0,52
da	22,78	29,54	25,48		− 0,07	**+ 1,34**
nl	22,71	24,49	25,15		**+ 0,39**	+ 0,95
de	18,02	24,66	23,50		+ 0,05	+ 0,88
sv	20,73	30,98	23,74		+ 0,22	+ 1,04
fi	10,71	20,56	19,15		+ 0,25	+ 1,25
de - en	-	-	-	24,66		
nl	24,97	24,49	22,62		|	|
da	24,59	29,54	**22,73**		|	− 0,03
pt	23,15	29,44	21,93		− 1,06	+ 0,70
fr	25,93	29,53	21,55		− 0,19	**+ 0,85**
es	**25,48**	**31,05**	21,23		− 0,23	+ 0,71
el	19,72	29,37	20,88		+ 0,00	+ 0,64
sv	19,80	**30,98**	21,35		+ 0,39	+ 0,88
fi	12,42	20,56	18,02		+ 0,20	+ 0,67
it	18,82	25,75	18,05		**+ 0,54**	+ 0,66

Tableau 5. *Résultats (scores BLEU) pour les configurations multi-pivots pour les trois paires de langues français → allemand, français → anglais et allemand → anglais*

Ces résultats sont tout à fait satisfaisants en tant que tels : ils font clairement apparaître que l'utilisation d'hypothèses obtenues par triangulation *via* des langues auxiliaires dans un cadre de combinaison de systèmes permet des gains significatifs [23].

23. Il est toutefois important de noter que ces résultats ne peuvent pas être comparés ligne à ligne avec ceux du tableau 3, car dans ce dernier les résultats sont obtenus en utilisant une

Système	direct	multipiv-direct	multipiv+direct
fr → de	18,02	+ 1,22	+ 0,96
fr → en	29,53	+ 0,27	+ 0,00
de → en	24,66	+ 0,38	+ 0,85

Tableau 6. *Résultats (scores BLEU) pour les systèmes directs (***direct***) ainsi que pour les systèmes implémentant la micro-adaptation lexicale en exploitant une sortie de combinaison de systèmes n'incluant pas le système direct (***multipiv-direct***) et incluant le système direct (***multipiv+direct***)*

Il est par exemple remarquable qu'en combinant les sorties des systèmes pivots entre le français et l'allemand pour trois langues (espagnol, néerlandais et portugais) on obtienne des résultats comparables à ceux du système direct. En d'autres termes, si l'on dispose des systèmes pivots adéquats, il n'est plus nécessaire de construire certains systèmes directs.

5.5. *Micro-adaptation utilisant des résultats de combinaisons de systèmes multipivots*

Le caractère prometteur de ces résultats en termes de qualité de telles sources auxiliaires suggère donc de mesurer l'impact que leur utilisation en micro-adaptation lexicale pourrait avoir sur la performance des systèmes. Ces résultats sont présentés tableau 6, en distinguant à nouveau les conditions où le système direct a été utilisé ou non pour la combinaison multi-pivots. Nous avons à chaque fois choisi la combinaison de langues menant aux meilleurs gains sur le corpus de développement et avons comme précédemment retenu pour nos systèmes la meilleure configuration (impliquant notamment un certain nombre d'hypothèses pour les systèmes auxiliaires) telle que trouvée sur ce même corpus de développement.

Les résultats obtenus sont plutôt satisfaisants, sans permettre toutefois un gain significatif par rapport aux contributions individuelles des meilleures langues auxiliaires. Il semble donc qu'une amélioration significative de la source auxiliaire utilisée n'implique pas toujours des gains pour les systèmes utilisant la micro-adaptation lexicale. Un travail supplémentaire d'analyse semble ici nécessaire pour mieux comprendre les phénomènes qui sont en jeu et mieux cerner les conditions dans lesquelles les traductions peuvent être améliorées.

seule langue auxiliaire, là où la combinaison de systèmes utilise plusieurs systèmes de façon incrémentale.

5.6. *Évaluation lexicale contrastive*

Les métriques automatiques telles que BLEU sont souvent critiquées, notamment pour le fait qu'elles ne permettent pas de faire ressortir des améliorations fines qui seraient par ailleurs révélées par une étude manuelle poussée, telle que celle décrite dans (Vilar *et al.*, 2006). En outre, le calcul d'une similarité entre une hypothèse et une traduction de référence porte toute l'attention de telles métriques sur les traductions elles-mêmes, ce qui ne permet pas d'observer la performance des traductions relativement aux mots sources qui sont effectivement traduits. Nous suivons donc ici la méthodologie d'évaluation que nous avons établie (Max *et al.*, 2010), et qui se concentre sur la comparaison de la capacité de deux systèmes à correctement traduire les mots sources en regroupant les mots en fonction de catégories d'intérêt.

Pour cela, les mots sources du corpus de test sont tout d'abord alignés avec les mots cibles de la traduction de référence, en alignant automatiquement à l'aide de GIZA++ le corpus constitué de l'union du corpus d'apprentissage et du corpus de test [24]. Le corpus de test est ensuite analysé à l'aide du TREETAGGER pour identifier les mots pleins, qui sont réputés avoir un impact plus fort sur la conservation du sens en traduction. Lorsque des mots sources sont alignés avec plusieurs mots cibles d'après les alignements de référence construits, chaque paire ⟨mot source, mot cible⟩ doit être cherchée individuellement dans la traduction de référence. De plus, un mot de la référence ne peut être utilisé qu'une seule fois pour les mesures de précision.

Le tableau 7 montre les résultats de l'analyse contrastive au niveau des catégories morphosyntaxiques des mots entre le système direct fr $\rightarrow$ en et nos systèmes utilisant diverses langues auxiliaires. Les valeurs sur les lignes étiquetées « − » indiquent le nombre de mots sources que seul le système de référence (ici le système direct) est parvenu à traduire correctement (selon la référence utilisée), et les lignes étiquetées « + » le nombre de mots sources que seul notre système est parvenu à traduire correctement.

Le résultat qui est peut-être le plus marquant concerne la contribution du grec comme langue auxiliaire : alors qu'aucun gain n'était constaté en termes de score BLEU, cette langue permet de traduire correctement 82 mots pleins de plus que le système direct [25]. Cela peut avoir plusieurs causes : outre les moins bons scores de précision 3-grammes et 4-grammes de notre système utilisant le grec comme langue auxiliaire, on peut supposer que la qualité de la traduction des mots n'apparaissant pas dans les catégories retenues, donc les mots grammaticaux, a diminué. Au contraire, l'italien ne semble pas apporter de gains hormis dans le cas des noms. Par ailleurs, les pertes observées au niveau de la traduction des pronoms ne sont pas très surprenantes : en effet, la traduction correcte d'un pronom nécessite l'identification du référent ap-

24. Les alignements ainsi obtenus sont donc très fortement influencés par le corpus d'apprentissage. On peut par ailleurs remarquer que ces alignements pourraient être corrigés manuellement afin d'améliorer la précision des mesures effectuées.

25. Cette valeur peut sembler faible, mais les mesures qui sont traditionnellement effectuées sur les corpus de test ne portent que sur des quantités très limitées de données.

		Catégories des mots dans la source						Précisions *n*-grams dans la cible				
aux		**ADJ**	**ADV**	**NOM**	**PRO**	**VER**	all	**+Bleu**	**1 g**	**2 g**	**3 g**	**4 g**
el	-	27	21	114	25	99	286	+ 0,07	63,3	36,0	22,7	14,8
	+	62	29	136	27	114	368					
es	-	33	25	106	26	110	300	+ 0,61	63,8	36,6	23,2	15,2
	+	64	38	136	22	117	377					
fi	-	44	40	106	20	92	302	+ 0,44	63,6	36,5	23,1	15,1
	+	49	31	120	23	106	329					
it	-	55	39	128	35	119	376	+ 0,32	63,2	36,3	23,0	15,1
	+	55	39	145	36	121	396					
sv	-	40	30	138	29	109	346	+ 0,50	64,0	36,7	23,2	15,1
	+	69	46	144	23	134	416					

Tableau 7. *Résultats de l'évaluation constrastive sur la qualité de traduction des mots sources entre le système direct fr → en et nos systèmes exploitant la micro-adaptation lexicale* via *plusieurs langues et combinaisons. Les valeurs « - » (resp. « + ») indiquent le nombre de fois où le système direct (resp. notre système) a été le seul à traduire correctement un mot source*

proprié dans la langue cible, ce qui ne fait l'objet d'aucune modélisation particulière dans aucun système, et aucune amélioration ne peut *a priori* être attendue de l'utilisation de langues auxiliaires. Globalement, il apparaît que la traduction des noms et des adjectifs connaît les améliorations les plus notables.

Deux exemples de notre corpus de test sont fournis pour illustrer cette évaluation dans la figure 5. Un type d'amélioration assez notable concerne l'intégration de mots qui peuvent ne pas apparaître dans les traductions du système direct, ce qui peut être interprété comme une conséquence directe de la micro-adaptation réalisée lorsqu'un mot est présent dans une prédiction auxiliaire.

6. Conclusion et perspectives

Dans cet article, nous avons étudié plusieurs façons d'améliorer les performances d'un système de traduction statistique en lui apportant des connaissances auxiliaires pendant l'étape de construction de sa meilleure hypothèse.

Nous avons proposé un cadre général dans lequel les connaissances auxiliaires peuvent prendre de très nombreuses formes. Nous avons considéré qu'une forme simple pour les prédictions, qui s'avère toutefois efficace, consiste à utiliser des prédictions lexicales, qui correspondent simplement à des mots qu'il est souhaitable de trouver dans les hypothèses du système principal. Les expériences conduites dans ce cadre ont porté vers une condition d'apparence difficile, qui consiste à utiliser des systèmes automatiques en pivot pour améliorer un système principal. Il semble *a priori* peu probable d'améliorer la performance d'un système de traduction en lui fournissant des

ref #357	this concession to the unions ignores the reality that all airlines have different safety procedures which even differ between aircrafts within each airline .
direct	this **concession unions ignores** the *fact* that all **airlines have different safety procedures** which are **even within each** of the *companies* in accordance with the types of equipment .
rel. src	cette **concession** aux **syndicats ignore** la *réalité* selon laquelle toutes les **compagnies aériennes ont** des **procédures de sécurité différentes** qui diffèrent **même** au **sein** de chacune des *compagnies* en fonction des types d ' *appareils* .
+aux	this **concession** to the trade **unions ignores** the **reality** according to which all the **airlines have different safety procedures** which **differ even within each** of the *companies* in accordance with the types of equipment .
rel. src	cette **concession** aux **syndicats ignore** la **réalité** selon laquelle toutes les **compagnies aériennes ont** des **procédures de sécurité différentes** qui **diffèrent même** au **sein** de chacune des *compagnies* en fonction des types d ' *appareils* .
ref #500	it is the treaty that must include clear commitments and operational tools so that the coordination of economic policies really does take on board the requirement of cooperation over jobs .
direct	this is the **treaty must** *lay down* **clear commitments** and **operational** *instruments* for *coordinating* **economic policies really** integrated into the need for **cooperation** on *employment* .
rel. src	c ' **est** le **traité** qui **doit** *prévoir* des **engagements clairs** et des *outils* **opérationnels** pour que la *coordination* des **politiques économiques** intègre **vraiment** l ' exigence de **coopération** pour l ' *emploi* .
+aux	it **is** the **treaty** which must *provide* **clear commitments** and **operational** *instruments* for the **coordination** of **economic policies really** *incorporates* the **requirement** of **cooperation** for *employment* .
rel. src	c ' **est** le **traité** qui **doit** *prévoir* des **engagements clairs** et des *outils* **opérationnels** pour que la **coordination** des **politiques économiques** *intègre* **vraiment** l ' **exigence** de **coopération** pour l ' *emploi* .

Figure 5. *Exemples de traductions automatiques pour la paire de langues fr → en avec le système direct (***direct***) et avec notre système avec micro-adaptation utilisant l'espagnol comme langue auxiliaire (***+aux***). Le gras indique les mots sources et cibles qui ont été traduits correctement relativement à la référence utilisée (***ref***), et l'italique indique les mots source et cible dont la traduction pourrait être considérée comme acceptable si d'autres références étaient disponibles.*

hypothèses obtenues par pivot, en particulier lorsque les paires de langues impliquées sont réputées difficiles. Or, et nous l'avons montré aussi bien *via* une métrique automatique traditionnelle que *via* une évaluation lexicale contrastive, et ce sur trois paires de langues de difficulté variable, des gains très intéressants peuvent être obtenus.

La question qui s'est ensuite naturellement posée, tout en restant dans la même condition, est celle de l'impact de l'amélioration de la qualité des prédictions auxiliaires utilisées pour la micro-adaptation lexicale. Une approche naturelle consiste à exploiter simultanément plusieurs langues, en espérant ainsi bénéficier des complé-

mentarités de chacune. Les résultats que nous avons obtenus en construisant de nouvelles hypothèses par consensus à partir des hypothèses de plusieurs systèmes pivots montrent également un net potentiel d'amélioration des systèmes.

Ces résultats posent au moins autant de questions nouvelles qu'ils ont apporté de réponses, comme nous l'avons discuté en essayant d'analyser plus finement ces améliorations de nos systèmes de traduction. Nous terminons cet article par une brève présentation des questions sur lesquelles nous comptons travailler dans le futur proche.

Tout d'abord, nous avons pu constater que les sources auxiliaires n'ont pas une contribution uniforme relativement aux performances de nos systèmes. Par exemple, l'italien s'est avéré utile pour améliorer la qualité de la traduction des noms dans le système fr $\rightarrow$ en, mais pas pour les autres catégories morphosyntaxiques. D'autres sources, comme par exemple un module de TAL spécialisé, pourraient de même avoir des contributions variables en fonction des éléments traduits. Il sera donc important de pouvoir prendre cela en compte.

Nous avons volontairement limité la discussion à l'utilisation de prédictions sous forme de mots ou de couples de mots *via* des modèles auxiliaires bigrammes : il s'avère que les modèles d'ordre supérieur, lorsqu'ils sont estimés sur des traductions obtenues par triangulation, sont en effet trop bruités pour contribuer positivement. D'autres types de prédictions, pour autant qu'elles puissent être produites de manière fiable, pourraient facilement être prises en compte à travers des modèles d'ordre supérieur. Par exemple, il semble donc intéressant d'envisager d'utiliser des prédictions provenant de terminologies bilingues, voire de mémoires de traduction comme dans (Simard et Isabelle, 2009), ou encore, comme nous l'avons déjà évoqué, directement d'un traducteur humain.

En outre, pour mieux cerner les phénomènes auxquels nous nous sommes confrontés, il nous a paru préférable de faire opérer l'adaptation lexicale à un niveau local. Or, comme certains travaux l'ont montré (Carpuat, 2009), les traductions automatiques peuvent bénéficier fortement de la prise en compte d'un niveau discursif. Une telle *macro-adapation* exigera des corpus de développement et de test plus importants en taille.

Finalement, de nombreuses applications décrites dans le cadre présenté figure 2 présentent, à nos yeux, un véritable intérêt, à la fois en recherche et d'un point de vue applicatif. En particulier, nous avons commencé à réfléchir aux problèmes de combinaison de systèmes, d'utilisation de textes paraphrasés et de traduction assistée par l'humain.

Remerciements

Ce travail a été partiellement financé par l'OSEO dans le cadre du programme Quaero.

7. Bibliographie

Bellagarda J. R., « An overview of statistical language model adaptation », *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, Sophia Antipolis, France, p. 165-174, 2001.

Bertoldi N., Federico M., « Domain Adaptation for Statistical Machine Translation with Monolingual Resources », *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, p. 182-189, 2009.

Brown P. F., Cocke J., Pietra S. D., Pietra V. J. D., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S., « A Statistical Approach to Machine Translation », *Computational Linguistics*, vol. 16, n° 2, p. 79-85, 1990.

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.

Callison-Burch C., Fordyce C. S., Koehn P., Monz C., Schroeder J., « Further Meta-Evaluation of Machine Translation », *Proceedings of the Third Workshop on Statistical Machine Translation*, Columbus, Ohio, p. 70-106, 2008.

Carpuat M., « One Translation Per Discourse », *Proceedings of the NAACL-HLT Workshop on Semantic Evaluations*, Boulder, United States, 2009.

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

Chen S. F., Goodman J. T., « An Empirical study of smoothing techniques for language modeling », *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, p. 310-318, 1996.

Crego J. M., Marino J. B., « Improving SMT by coupling reordering and decoding », *Machine Translation*, vol. 20, n° 3, p. 199-215, 2007.

Crego J. M., Max A., Yvon F., « Plusieurs langues (bien choisies) valent mieux qu'une : traduction statistique multisource par renforcement lexical », *Proceedings of TALN*, Senlis, France, 2009.

Crego J. M., Max A., Yvon F., « Local lexical adaptation in Machine Translation through triangulation : SMT helping SMT », *Proceedings of COLING*, Beijing, China, 2010.

Dymetman M., Max A., Yamada K., « Towards Interactive Text Understanding », *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, p. 109-112, 2003.

Fiscus J. G., « A post-processing system to yield reduced error word rates : Recognizer output voting error reduction (ROVER) », *in* I. S. P. Society (ed.), *IEEE Workshop on Automatic Speech Recognition and Understanding*, Piscataway, NJ, p. 347-354, 1997.

Hildebrand A. S., Vogel S., « Combination of machine translation systems via hypothesis selection from combined n-best lists », *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, Waikiki, Hawaï, p. 254-261, 2008.

Ignat C., Improving Statistical Alignment and Translation using highly multilingual corpora, PhD thesis, Université de Strasbourg, 2009.

Kay M., « Triangulation in translation », *Keynote at the MT 2000 Conference*, University of Exeter, 2000.

Knight K., « Decoding Complexity in Word-Replacement Translation Models », *Computational Linguistics*, vol. 25, n° 4, p. 607-615, 1999.

Koehn P., « Europarl : A Parallel Corpus for Statistical Machine Translation », *Proceedings of MT Summit*, Phuket, Thailand, 2005.

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

Koehn P., Birch A., Steinberger R., « 462 Machine Translation Systems for Europe », *Proceedings of the 12th MT Summit*, Ottawa, Canada, p. 65-72, 2009.

Koehn P., Och F. J., Marcu D., « Statistical Phrase-Based Translation », *Proceedings of NAACL/HLT*, Edmonton, Canada, 2003.

Kumar S., Och F. J., Macherey W., « Improving Word Alignment with Bridge Languages », *Proceedings of EMNLP-CoNLL*, Prague, Czech Republic, 2007.

Leusch G., Matusov E., Ney H., « The RWTH System Combination System for WMT 2009 », *Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athens, Greece, p. 51-55, 2009.

Lopez A., « Statistical Machine Translation », *ACM Computing Surveys*, 2008.

Mariño J., Banchs R. E., Crego J. M., de Gispert A., Lambert P., Fonollosa J., Costa-jussà M., « N-gram Based Machine Translation », *Computational Linguistics*, vol. 32, n° 4, p. 527-549, 2006.

Matusov E., Leusch G., Banchs R. E., Bertoldi N., Dechelotte D., Federico M., Kolss M., Lee Y.-S., Mariño J., Paulik M., Roukos S., Schwenk H., Ney H., « System Combination for Machine Translation of Spoken and Written Language », *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, n° 7, p. 1222-1237, September, 2008.

Max A., Crego J. M., Yvon F., « Contrastive Lexical Evaluation of Machine Translation », *Proceedings of LREC*, Valletta, Malta, 2010.

Nomoto T., « Multi-Engine Machine Translation with Voted Language Model », *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, Barcelona, Spain, p. 494-501, 2004.

Och F. J., « Minimum Error Rate Training for Statistical Machine Translation », *Proceedings of ACL*, Sapporo, Japan, 2003.

Och F. J., Ney H., « Statistical Multi-Source Translation », *Proceedings of MT Summit*, Santiago de Compostela, Spain, 2001.

Papineni K., Roukos S., Ward T., Zhu W.-J., « BLEU : a method for automatic evaluation of machine translation », *Proceedings of ACL*, Philadelphia, USA, 2002.

Paulik M., Fügen C., Schaaf T., Schultz T., Stüker S., Waibel A., « Document Driven Machine Translation Enhanced Automatic Speech Recognition », *Proceedings of the International Conference on Speech Language Techonology (InterSpeech)*, Lisbon, Portugal, 2005.

Placeway P., Lafferty J., « Cheating with imperfect transcripts », *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, PA, p. 2115-2118, 1996.

Rosti A.-V., Ayan N. F., Xiang B., Matsoukas S., Schwatz R., Dorr B. J., « Combining Outputs from Multiple Machine Translation Systems », *Proceedings of NAACL-HTL*, Rochester, USA, 2007.

Schroeder J., Cohn T., Koehn P., « Word Lattices for Multi-Source Translation », *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, Athens, Greece, p. 719-727, 2009.

Schwartz L., « Multi-source translation methods », *MT at work : Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, Waikiki, Hawaï, p. 279-288, 2008.

Schwenk H., « Investigations on large-scale lightly-supervised training for statistical machine translation », *Proceedings of the International Workshop on Spoken Language Translation*, Hawaï, USA, p. 182-189, 2008.

Simard M., Isabelle P., « Phrase-based Machine Translation in a Computer-assisted Translation Environment », *Proceedings of Machine Translation Summit XII*, Ottawa, Canada, 2009.

Stolcke A., « SRILM – An Extensible Language Modeling Toolkit », *Proceedings of the International Conference on Spoken Langage Processing (ICSLP)*, vol. 2, Denver, CO, p. 901-904, 2002.

Tillmann C., « A Unigram Orientation Model for Statistical Machine Translation », *Proceedings of ACL'04*, Boston, MA, p. 101-104, May, 2004.

Utiyama M., Isahara H., « A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation », *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Proceedings of the Main Conference*, Association for Computational Linguistics, Rochester, New York, p. 484-491, 2007.

Vilar D., Xu J., d'Haro L. F., Ney H., « Error Analysis of Statistical Machine Translation Output », *Proceedings of LREC*, Genoa, Italy, 2006.

Wu H., Wang H., « Pivot Language Approach for Phrase-Based Statistical Machine Translation », *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic, p. 856-863, June, 2007.

Wu H., Wang H., « Revisiting Pivot Language Approach for Machine Translation », *Proceedings of ACL*, Suntec, Singapore, p. 154-162, August, 2009.

Zens R., Och F. J., Ney H., « Phrase-based statistical machine translation », *in* M. Jarke, J. Koehler, G. Lakemeyer (eds), *KI-2002 : Advances in artificial intelligence*, vol. 2479 of *LNAI*, Springer Verlag, p. 18-32, 2002.