Normalisation automatique du vocabulaire source pour traduire depuis une langue à morphologie riche

Franck Burlot François Yvon
LIMSI, CNRS, Univ. Paris-Sud, Université Paris Saclay, 91 403 Orsay, France
nom.prénom@limsi.fr

RÉSUMÉ

Lorsqu'ils sont traduits depuis une langue à morphologie riche vers l'anglais, les mots-formes sources contiennent des marques d'informations grammaticales pouvant être jugées redondantes par rapport à l'anglais, causant une variabilité formelle qui nuit à l'estimation des modèles probabilistes. Un moyen bien documenté pour atténuer ce problème consiste à supprimer l'information non pertinente de la source en la normalisant. Ce pré-traitement est généralement effectué de manière déterministe, à l'aide de règles produites manuellement. Une telle normalisation est, par essence, sous-optimale et doit être adaptée pour chaque paire de langues. Nous présentons, dans cet article, une méthode simple pour rechercher automatiquement une normalisation optimale de la morphologie source par rapport à la langue cible et montrons que celle-ci peut améliorer la traduction automatique.

ABSTRACT

Learning Morphological Normalization for Translation from Morphologically Rich Languages

When translating from a morphologically rich language into English, source side word forms encode grammatical information that can be considered as redundant with respect to English, leading to data sparsity issues. A well-known way to mitigate this problem is to remove irrelevant information from the source through normalization. This pre-processing is usually performed in a deterministic fashion, using hand-crafted rules. This normalization is, in essence, suboptimal and needs to be adapted for each new language pair. We introduce here a simple way to automatically search for an optimal normalization of the source morphology with respect to the target-side language and show that it can improve machine translation.

MOTS-CLÉS : traduction automatique, langue morphologiquement riche, classification.

KEYWORDS: machine translation, morphology-rich language, clustering.

1 Introduction

Traduire depuis une langue source à morphologie riche comme le tchèque ou le russe vers une langue plus analytique comme l'anglais conduit à de nombreuses difficultés dues à d'importantes divergences dans les systèmes linguistiques de ces paires de langues.

Les langues morphologiquement riches considérées dans cet article ont des tendances synthétiques, ce qui signifie qu'elles marquent généralement de l'information grammaticale dans des terminaisons de mots, comme le cas qui signale la fonction grammaticale du mot dans la phrase. Un tel phénomène est inexistant en anglais, où la fonction du mot est le plus souvent marquée par un ordre des mots

particulier (le sujet est situé à gauche du verbe) ou par une préposition. Ces divergences témoignent d'un manque de symétrie manifeste entre ces deux types de langues. Ainsi, alors que du côté source les adjectifs varient en genre, nombre et cas, leur traduction anglaise est invariable.

De telles différences affectent négativement la qualité de la traduction de différentes manières :

— La multiplicité des mots-formes sources implique que chacune de ces formes a une fréquence inférieure à son équivalent anglais, ce qui rend difficile l'estimation fiable de paramètres, surtout pour les lemmes rares ;

— Un cas extrême survient lorsqu'il s'agit de traduire un mot-forme qui n'a pas été observé dans les données d'entraînement. Même lorsque d'autres formes relevant du même lemme ont été observées, un système de traduction qui ne manipulerait que des mots serait incapable de déduire ce type de rapprochement et produira une sortie erronée.

Un moyen bien connu pour atténuer ce problème consiste à enlever l'information considérée comme non pertinente par rapport à l'anglais. Par exemple, le genre, le nombre et le cas des adjectifs sont couramment éliminés puisque toutes les formes sources se traduisent par un seul mot anglais. Cette solution a été abondamment étudiée (ex. Ney & Popovic (2004); Durgar El-Kahlout & Yvon (2010) pour la paire allemand-anglais, Goldwater & McClosky (2005) pour la paire tchèque-anglais) et utilisée par de nombreux systèmes participant aux campagnes d'évaluation WMT [1] (ex. Lo *et al.* (2016); Marie *et al.* (2015) pour la paire russe-anglais). Le même type de solution est par ailleurs employé pour traduire dans la direction opposée (Minkov *et al.*, 2007; Toutanova *et al.*, 2008; Fraser *et al.*, 2012) où la sortie du système de traduction normalisée doit additionnellement être réinfléchie. Ces procédures comportent de multiples limitations : elles dépendent de la paire de langues étudiée et reposent sur des ensembles de règles qui doivent être adaptées à chaque direction de traduction. Il est également probable que de telles méthodes sont sous-optimales par rapport à la tâche, puisqu'elles ignorent les particularités des données utilisées pour entraîner le système de traduction. Talbot & Osborne (2006) proposent une méthode automatique pour regrouper les mots qui partagent la même traduction en employant des méthodes de sélection de modèle ; toutefois, contrairement à notre modèle, cette méthode n'est pas spécifiquement conçue pour traiter des problèmes de morphologie et ne se base pas sur une analyse morpho-syntaxique.

Nous présentons (section 2) un modèle simple et indépendant de la langue étudiée qui permet de réaliser une telle normalisation en regroupant automatiquement les formes sources qui tendent à être traduites par les mêmes mots cibles [2]. Cette similarité de traduction est mesurée par l'entropie de la distribution des mots cibles $\mathbf{E}$ alignés à un mot source f : $H(\mathbf{E}|f)$. Les résultats expérimentaux obtenus pour la traduction du tchèque et du russe vers l'anglais et le français (section 3.2) montrent que cette procédure de classification améliore la qualité de la traduction. Nous proposons enfin à la section 4 une description détaillée des classes obtenues avec notre modèle.

1. `http://statmt.org/`

2. Notre implémentation est disponible sur `https://github.com/franckbrl/bilingual_morph_normalizer`.

2 Classification de la source

2.1 Gain d'information

L'objectif est de réaliser une classification des formes sources en fusionnant celles qui se traduisent par le(s) même(s) mot(s). Nous supposons que chaque forme source f est la combinaison d'un lemme, d'une partie du discours (PdD) et d'une séquence d'étiquettes morphologiques [3] et que le corpus parallèle a été aligné mot-à-mot. Ces alignements permettent d'estimer des probabilités de traduction lexicale $p(e|f)$, ainsi que des probabilités unigrammes $p(f)$ qui constituent l'entrée de notre algorithme.

Nous explicitons dans un premier temps l'intuition de notre méthode dans le cas simple où le corpus ne contient qu'un lemme pour chaque PdD. Nous notons respectivement $\mathbf{f}$ l'ensemble des mots-formes (ou des positions dans le paradigme) pour ce lemme et $\mathbf{E}$ la totalité du vocabulaire anglais. L'entropie conditionnelle du modèle de traduction est alors :

$$\begin{aligned} H(\mathbf{E}|\mathbf{f}) &= \sum_{f\in\mathbf{f}} p(f)H(\mathbf{E}|f) \\ &= \sum_{f\in\mathbf{f}} \frac{p(f)}{\log_2|\mathbf{E}_{a_f}|} \sum_{e\in\mathbf{E}_{a_f}} -p(e|f)\log_2 p(e|f), \end{aligned} \tag{1}$$

où $\mathbf{E}_{a_f}$ est l'ensemble des mots anglais alignés au moins une fois au mot source f. Le terme de normalisation ($\log_2|\mathbf{E}_{a_f}|$) garantit que toutes les valeurs d'entropie sont comparables et ne dépendent pas de la quantité de mots cibles alignés à f.

Partant d'un état initial, où chaque mot-forme f correspond à une classe singleton, et en procédant de manière ascendante, nous recherchons les paires de classes (f_1, f_2) dont la fusion réduit l'entropie conditionnelle. Dans ce but, nous calculons le gain d'information (GI) [4] issu de l'opération de fusion :

$$\begin{aligned} \mathrm{GI}(f_1,f_2) = {} & p(f_1)H(\mathbf{E}|f_1) \\ & + p(f_2)H(\mathbf{E}|f_2) \\ & - p(f')H(\mathbf{E}|f') \end{aligned} \tag{2}$$

où f_1 et f_2 sont des classes candidates à la fusion et f' est la classe qui résulte de cette fusion.

Le gain d'information correspond à la différence entre la combinaison des entropies des classes f_1 et f_2 avant et après la fusion en f'. Si les mots sources correspondants ont une distribution semblable sur les mots cibles, le gain d'information est positif, tandis que quand leurs traductions sont différentes, il est négatif et leur fusion conduit à une perte d'information.

Notons que l'entropie totale $H(\mathbf{E}|\mathbf{f})$ du modèle de traduction peut être recalculée de manière incrémentale après la fusion de la paire (f_1, f_2) par :

3. Ainsi, le mot tchèque *autem* (en voiture) est représenté par : *auto+Nom+Neutre+Singulier+Instrumental*.
4. Le gain d'information prend une valeur entre -1 et 1.

$$H(\mathbf{E}|\mathbf{f}) \leftarrow H(\mathbf{E}|\mathbf{f}) - \mathrm{GI}(f_1, f_2) \tag{3}$$

Nous pouvons interpréter le gain d'information comme une mesure de similarité entre deux mots-formes, qui pourrait être employé dans le cadre d'un modèle probabiliste de classification comme le *partitionnement en k-moyennes*. La difficulté réside ici dans le fait que nous ne sommes pas en mesure de décider en avance et de manière satisfaisante quel nombre de classes on souhaite obtenir. L'objectif de cette classification est donc double et nous recherchons :

— des classes cohérentes, qui réunissent des formes dont la traduction en langue cible est proche ;
— un nombre de classes optimal correspondant au niveau adéquat de granularité dans la normalisation, avec un espace de recherche délimité d'un côté par la représentation des mots selon leur forme (aucune normalisation) et de l'autre par une représentation en lemmes (niveau maximal de normalisation dans nos conditions).

Ainsi, étant dans l'incapacité de paramétrer manuellement le nombre de classes *a priori*, il convient de trouver un optimum en fonction des données observées. La procédure que nous proposons pour atteindre ces objectifs est décrite à la section 2.2.

2.2 Classifier les cellules du paradigme

En pratique, notre algorithme est appliqué au niveau des PdD plutôt qu'individuellement sur les lemmes : nous supposons ainsi que pour une PdD p donnée, tous les lemmes ont le même nombre n_p de variantes morphologiques (ou de cellules dans leur paradigme). Ainsi, bien que restant basée sur des statistiques individuelles collectées au niveau des lemmes, la valeur du gain d'information sera cumulée sur l'ensemble des lemmes d'une PdD donnée.

Comme expliqué plus haut, pour chaque lemme d'une PdD donnée, le point de départ est une matrice de gains d'information $L_{\mathbf{l}} \in [-1:1]^{n_p \times n_p}$, où $L_{\mathbf{l}}(i,j)$ est le gain d'information obtenu après la fusion des formes l_i et l_j du lemme $\mathbf{l}$. L'agrégation de ces matrices produit *la matrice pour les PdD* $M_p \in [-1:1]^{n_p \times n_p}$ qui contient la moyenne des gains d'information issus de la fusion de deux cellules pour la partie du discours p.

$C(\mathrm{p}) \leftarrow \{1, ..., n_p\}$
$i, j \leftarrow \arg\max_{i',j' \in C(\mathrm{p})^2} M_p(i', j')$
répéter
  fusionner i et j dans $C(p)$
  pour chaque $l \in V_{lem}$ **faire**
    supprimer $L_{\mathbf{l}}(i,j)$, créer $L_{\mathbf{l}}(ij)$
    calculer $p(ij)$, $p(\mathbf{E}|ij)$ et $H(\mathbf{E}|ij)$
    calculer $L_{\mathbf{l}}(ij,k)$ **pour** $k \in C(\mathrm{p})$
  $M_p \leftarrow \sum_{\mathbf{l} \in V_{\mathrm{lem}}} L_{\mathbf{l}}$
  $i, j \leftarrow \arg\max_{i',j' \in C(\mathrm{p})^2} M_p(i', j')$
jusqu'à $M_p(i,j) < m$ *or* $|C(p)| = 1$

Algorithme 1 – Un algorithme de classification ascendant

La procédure complète est décrite dans l'algorithme 1. Elle commence avec n_p classes pour chaque PdD et accomplit des opérations de fusion tant que le gain d'information obtenu pour la fusion dépasse un seuil minimal m. À chaque acceptation de fusion, les paramètres de la nouvelle classe (probabilité unigramme, probabilité de traduction et entropie) sont recalculés *pour tous les lemmes* du vocabulaire ($\mathbf{l} \in V_{lem}$) et utilisés pour actualiser les matrices de gains d'information des PdD M_p.

Lorsque cette procédure se termine, on obtient pour chaque PdD p une classification $C(p)$ qui peut être employée pour normaliser les données sources de diverses manières (voir section 3.2).

3 Résultats expérimentaux

Nous évaluons le modèle de normalisation de la morphologie sur une tâche de traduction pour trois paires de langues : tchèque-anglais, russe-anglais et tchèque-français. Notons que cette dernière paire comprend deux langues à morphologie riche.

3.1 Conditions expérimentales

Les systèmes de traduction automatique sont entraînés avec Moses (Koehn *et al.*, 2007) et optimisés avec KB-MIRA (Cherry & Foster, 2012). Les alignements sont obtenus avec Fast_align (Dyer *et al.*, 2013). Tous les systèmes traduisant vers l'anglais emploient le même modèle de langue de 4-grammes entraîné avec KenLM (Heafield, 2011) sur les données anglaises de presse distribuées à WMT2016 [5], ainsi que sur le côté cible des données parallèles utilisées pour le grand modèle tchèque-anglais (voir ci-dessous), pour un total d'environ 150 millions de phrases. Le modèle de langue français a, quant à lui, été entraîné sur le corpus monolingue News-2014 de WMT et sur le côté cible des données parallèles. La tokenisation des textes anglais et français repose sur nos propres outils de pré-traitement des textes (Déchelotte *et al.*, 2008).

La normalisation du côté source est opérée indépendamment pour chaque corpus introduit plus bas, en employant les données parallèles d'entraînement du système de traduction. Les lemmes et les étiquettes morpho-syntaxiques ont été obtenus avec Morphodita (Straková *et al.*, 2014) pour le tchèque et TreeTagger (Schmid, 1994; Sharoff & Nivre, 2011) pour le russe. Une légère pré-sélection des données sources à traiter tend à fournir de meilleurs résultats et nous décidons de ne pas considérer lors de la classification les lemmes apparaissant moins de 100 fois, ainsi que les mots-formes dont la fréquence est inférieure à 10 dans les données d'entraînement, afin de limiter le bruit engendré par les alignements initiaux. Lors de la classification des cellules du paradigme (section 2.2), nous fixons le gain d'information minimum m à 0.

En pratique, nous avons remarqué que nous obtenions des résultats sensiblement meilleurs et un temps de traitement plus court que le calcul exact de l'algorithme 1, avec un régime d'actualisation alternatif pour la matrice de gains d'information M. Une fois initialisée de la manière décrite ci-dessus comme la somme des matrices de gains d'information $L_{\mathbf{l}}$, nous traitons M comme une matrice de similarité et employons une actualisation proche de l'algorithme de « linkage clustering ». Après la création des clusters c_1 et c_2, la cellule de matrice correspondant à la nouvelle classe est calculée par :

5. www.statmt.org/wmt16

$$M(c_1, c_2) = \frac{\sum_{f_1 \in c_1} \sum_{f_2 \in c_2} M(f_1, f_2)}{|c_1| \times |c_2|}, \quad (4)$$

ce qui évite d'avoir à actualiser toutes les matrices L_1.

Les expériences de traduction automatique qui suivent sont réalisées pour la paire russe-anglais avec pour données parallèles le corpus News-Commentary fourni à WMT 2016 (190 000 phrases). Nous discutons également des résultats obtenus avec deux corpus tchèque-anglais : un petit système est entraîné sur News-Commentary (190k phrases fournies à WMT 2016) et un système plus grand qui ajoute au premier Europarl (Koehn, 2005) et un sous-ensemble du corpus CzEng (Bojar *et al.*, 2016) identifié comme relevant du domaine de la presse d'actualités (total d'un million de phrases). Ces systèmes sont enfin optimisés sur les corpus de WMT Newstest-2015 et évalués sur Newstest-2016. Le système tchèque-français a été entraîné sur le corpus Europarl (622k phrases parallèles), optimisé sur Newstest-2014 et évalué sur Newstest-2013.

3.2 Résultats de traduction automatique

Nous présentons ici les effets de la normalisation du vocabulaire source sur deux types de systèmes de traduction automatique : statistique et neuronal.

3.2.1 Traduction automatique statistique

La classification apprise sur le petit corpus d'entraînement tchèque-anglais a permis de grandement réduire le vocabulaire source initial. Nous avions au départ 158 914 chaînes de caractères distinctes, correspondant ensuite à 237 378 formes entièrement désambiguïsées (représentées par un lemme et de l'information morpho-syntaxique). En appliquant le modèle de classification, nous avons finalement obtenu un vocabulaire de 90 170 entrées normalisées. Cette réduction du vocabulaire source correspond à une réduction du nombre de mots hors-vocabulaire (MHV) au moment de l'évaluation de ce petit système dont les résultats sont présentés dans le tableau 1. Ceci montre que notre modèle apporte une réponse efficace au problème de la dispersion des données.

L'application du modèle aux données d'entraînement peut se faire de différentes manières selon son utilisation pour les alignements et/ou la traduction :

- Les alignements appris sur la source infléchie, initialement employés pour apprendre la normalisation de la source, sont également utilisés dans le système de traduction (ali cs pour le tchèque et ali ru pour le russe). [6] La source normalisée n'intervient donc que dans le modèle de traduction pour l'extraction des segments et le modèle de réordonnancement (cx-en, cx-fr et rx-en).
- De nouveaux alignements mot-à-mot sont appris sur la source normalisée (ali cx et ali rx). Le système de traduction qui emploie ces alignements effectue une traduction de la source infléchie (non normalisée) vers la cible (cs-en, cs-fr et ru-en).
- La source normalisée sert à entraîner de nouveaux alignements (ali cx et ali rx), ainsi que le modèle de traduction (cx-en, cx-fr et rx-en).

6. Par convention, nous dénotons cs ou cx (resp. ru ou rx) les versions brutes et normalisées du tchèque (resp. du russe). Comme il est d'usage, en et fr désignent respectivement les données anglaises et françaises.

— Les cas ci-dessus sont comparés à un système de base où la traduction (cs-en, cs-fr et ru-en) et les alignements (ali cs et ali ru) sont entraînés sur une source infléchie.

La comparaison de ces différentes configurations permet de mesurer plus précisément l'apport de la normalisation pour l'amélioration des alignements et de modèles de traduction.

TABLE 1 – Scores BLEU pour le tchèque (petites données)

Système	BLEU	MHV
cs-en (ali cs)	21,26	2189
cx-en (ali cx)	22,62 (+1,36)	1888
cx-en (ali cs)	22,34 (+1,08)	1914
cs-en (ali cx)	22,19 (+0,93)	2152
cx-en (100 plus fréq)	**22,82 (+1,56)**	1893

TABLE 2 – Scores BLEU pour le tchèque (grandes données)

Système	BLEU	MHV
cs-en (ali cs)	23,85	1878
cx-en (ali cx)	24,57 (+0,72)	1610
cx-en (ali cs)	24,36 (+0,51)	1627
cs-en (ali cx)	24,14 (+0,29)	1832
cx-en (100 plus fréq)	**24,85 (+1,00)**	1614
cx-en ($m = -10^{-4}$)	24,44 (+0,59)	1604
cx-en ($m = 10^{-4}$)	24,05 (+0,20)	1761
cx-en (manuel)	24,46 (+0,61)	1623

TABLE 3 – Scores BLEU pour le russe

Système	BLEU	MHV
ru-en (ali ru)	19,76	2260
rx-en (ali rx)	**21,02 (+1,26)**	2033
rx-en (ali ru)	20,92 (+1,16)	2033
ru-en (ali rx)	20,53 (+0,77)	2048
rx-en (100 plus fréq)	20,89 (+1,13)	2026

TABLE 4 – Scores BLEU pour le tchèque-français

Système	BLEU	MHV
cs-fr (ali cs)	19,57	1845
cx-fr (ali cx)	**20,19 (+0,62)**	1592

Au tableau 1, l'utilisation du modèle de normalisation à la fois pour les alignements (ali cx) et le système de traduction (cx-en) donne une amélioration de 1,36 points BLEU. L'utilisation de la classification uniquement pour les alignements ou pour le système de traduction conduit à des résultats inférieurs, qui restent toutefois nettement meilleurs que ceux du système de base (cs-en). Ceci tend à démontrer que les deux modèles bénéficient de la normalisation de la source. Nous présentons enfin une autre façon d'appliquer la normalisation, qui consiste à conserver les mots-formes initiaux pour

les 100 lemmes les plus fréquents (100 plus fréq), ce qui conduit au meilleur résultat que nous ayons obtenu pour la paire de langues tchèque-anglais, avec une amélioration de 1,56 points BLEU par rapport au système de base.

Nous observons la même tendance pour le plus grand système tchèque-anglais (voir tableau 2), même si les contrastes en score BLEU sont légèrement moins nets, en raison de la plus grande quantité de données employée. L'apprentissage du système de traduction observe dans ce cas plus de mots-formes et souffre donc moins du problème de dispersion des données. Avec ce système, nous avons également expérimenté différentes valeurs de gain d'information minimum m pour l'acceptation d'une fusion introduit en section 2.2, ce qui laisse penser que la valeur optimale pour m est proche de 0. Nous observons ici une propriété de notre algorithme : un m élevé produit plus de classes, ce qui a pour effet d'augmenter le nombre de MHV. Lorsque m est fixé à 10^{-4}, le système de traduction compte 1761 MHV, soit 157 de plus qu'avec $m = -10^{-4}$.

Des résultats avec une normalisation manuelle (manuel) sont donnés dans le tableau 2. Les règles de normalisation employées sont proches de celles qui sont présentées dans (Burlot *et al.*, 2016), où les noms se distinguent par leur nombre et leur polarité (affirmatifs et négatifs), les adjectifs par leur polarité et leur degré de comparaison, etc. Nous avons en plus appliqué des règles aux classes de verbes qui se distinguent par leur temps et leur polarité, à l'exception de la troisième personne du singulier au présent qui est conservée. Cette normalisation manuelle donne une amélioration (+0,61) qui est presque deux fois inférieure à celle de notre meilleur système (+1,00).

Les résultats pour la paire russe-anglais suivent la même tendance que la paire tchèque-anglais, à l'exception du fait que la conservation des 100 mots-formes les plus fréquents ne fournit pas d'amélioration sur la normalisation complète des données d'entraînement. Nous soupçonnons que cette différence est partiellement due au verbe tchèque très fréquent *být* (être) au présent qui a une flexion riche, tandis que dans beaucoup de cas, le russe a pour équivalent un zéro. Ainsi, le fait de conserver dans les données toutes les formes de ce verbe tchèque permet d'outrepasser efficacement le caractère délexicalisé de notre modèle, puisque le nombre de formes dans les deux langues source et cible est proche. Le russe ne possédant pas de verbe être, la même méthode consiste à conserver des formes qui n'ont pas d'équivalent en anglais.

TABLE 5 – Réduction des MHV dans la traduction vers le français

source (cs)	**Nasbírali** jsme 79 bodů.
cs-fr	Nous avons **nasbírali** 79 points.
cx-fr	Nous avons **accumulé** 79 points.
référence	On a terminé avec 79 points.

Enfin, nous constatons au tableau 4 que la normalisation du tchèque optimisée par rapport au français permet également d'améliorer la traduction, notamment en réduisant les MHV au moment de l'évaluation (tableau 5), bien qu'une telle langue cible soit déjà morphologiquement plus riche que l'anglais. [7] L'amélioration que nous observons est toutefois moins grande que dans le cas de la traduction vers l'anglais. Nous posons que cela est dû à un degré de normalisation du tchèque moins avancé lorsqu'il partage certaines propriétés avec la cible, comme la flexion de l'adjectif, ce

7. Dans les données d'entraînement tchèque-français, le mot-forme *nasbírali* (tableau 5) n'apparaît qu'une seule fois et le modèle d'alignement ne l'a relié à aucun mot français, si bien que le système de traduction n'est pas capable d'en fournir une traduction. Après la normalisation des données, ce mot se retrouve dans la même classe que d'autres formes à différents genres et nombres, par exemple *nasbíral* (5 occurrences) et *nasbíralo* (1 occurrence), et le modèle d'alignement trouve des traductions françaises pour cette nouvelle classe plus fréquente, notamment le mot *accumulé* qui apparaît dans la sortie.

qui conduit notre modèle à créer moins de classes. C'est ce type de question que nous proposons d'étudier plus en profondeur dans la section 4.

3.2.2 Traduction automatique neuronale

Nous présentons ici des systèmes neuronaux tchèques-anglais entraînés sur les grandes données introduites en section 3.1. La normalisation du tchèque est opérée au moyen du modèle appris sur ces mêmes données.

Ces systèmes ont été entraînés avec la boîte à outils Nematus (Sennrich *et al.*, 2017). Les modèles ont été validés sur newstest-2015 avec une fréquence de 10 000 mises-à-jour. La patience a été paramétrée à 10 validations, ce qui a conduit les systèmes à apprendre pendant 600 000 mises-à-jour en moyenne (deux à trois semaines). Du dropout a été appliqué à toutes les couches. L'algorithme d'optimisation utilisé est « adadelta ». Tous les systèmes sont enfin testés sur Newstest-2016.

Les résultats présentés proviennent de deux systèmes à base de mots infléchis et trois à base de mots normalisés. Ces mots sont segmentés selon l'algorithme « byte pair encoding » (BPE) en cible comme en source (Sennrich *et al.*, 2016). Outre des systèmes de type BPE-à-BPE, nous proposons également des systèmes à la source factorisée (Sennrich & Haddow, 2016), où la représentation des mots sources est concaténée à celle de caractéristiques des mots, comme les PdD. [8] Ces systèmes sont les suivants :

- **cs-en** : les mots tchèques infléchis sont segmentés en BPE et traduits vers l'anglais (BPE) ;
- **cx-en** : les mots tchèques normalisés sont segmentés en BPE et traduits vers l'anglais (BPE) ;
- **cx-en factorisé (lemmes, classes)** : la source est représentée par des lemmes segmentés et les identifiants des classes appris lors de la normalisation du tchèque.
- **cs-en factorisé (mots, PdD)** : la source est représentée par des mots tchèques infléchis segmentés et des PdD ;
- **cx-en factorisé (lemmes, classes, PdD)** : la source est représentée par des lemmes segmentés, des classes issues de la normalisation et des PdD.

Lorsque un mot est segmenté, la PdD ou la classe qui lui correspond doit être dupliquée afin que nous obtenions le même nombre d'éléments dans toutes les séquences de facteurs correspondant à une même phrase. Pour tous les systèmes factorisés, nous ajoutons un facteur comportant des tokens qui signalent si la PdD ou la classe à la position courante correspond à un mot non segmenté, ou s'il s'agit d'un début, milieu ou d'une fin de mot (Sennrich & Haddow, 2016).

TABLE 6 – Scores BLEU pour le tchèque-anglais (traduction neuronale)

cs-en	cx-en	cx-en factorisé (lemmes, classes)	cs-en factorisé (mots, PdD)	cx-en factorisé (lemmes, classes, PdD)
21,45	21,14	21,75	21,89	**22,42**

Les résultats pour ces différents systèmes sont présentés au tableau 6. Nous observons tout d'abord que les systèmes non factorisés favorisent sensiblement la représentation des mots infléchis par rapport aux mots normalisés (-0,31). Lorsque les identifiants de classes sont séparés des lemmes dans un système à la source factorisée, la performance remonte (+0,61) et surpasse les mots infléchis (+0,30). Notons que le système factorisé à base de lemmes et de classes obtient un résultat semblable

8. Par PdD, nous entendons désormais la séquence d'étiquettes comprenant la catégorie et les information morphologiques fines (genre, nombre, cas, temps, etc.).

à celui basé sur des mots et des PdD (-0.14). Ceci porte à croire que la normalisation a permis de sélectionner correctement dans les PdD l'information grammaticale pertinente à la prédiction des mots anglais. Ainsi, pour le système, les mots infléchis et les PdD semblent comporter certaines redondances qui ne permettent pas d'améliorer grandement la traduction.

Le meilleur de ces systèmes comprend des lemmes segmentés, des identifiants de classe et des PdD. Il dépasse de 0.53 points le système factorisé à base de mots infléchis et de 0.97 le premier système à base de mots infléchis. Les PdD semblent donc plus efficaces lorsqu'elles sont associées aux mots normalisés. Nous posons que l'avantage de ce système réside dans le fait qu'il représente deux types d'information grammaticale clairement distingués. D'une part, les classes comportent l'information qui doit être traduite avec le mot (comme le nombre des noms), et d'autre part, les PdD représentent une information d'ordre plutôt syntaxique, comme le cas (sujet, objet), seul indicateur de la fonction du mot anglais à prédire, et donc de sa position dans la phrase cible (à gauche, à droite du verbe).

4 Évaluation qualitative du modèle

Nous proposons dans cette dernière section une analyse de la classification obtenue à l'aide de notre modèle, en mettant notamment en évidence l'influence exercée par la langue cible sur ce processus.

4.1 Normalisation du tchèque par rapport à l'anglais

Les classes obtenues lors de la normalisation du tchèque par rapport à l'anglais confirment certaines intuitions linguistiques. En effet, le tableau 7 montre que la normalisation des noms a permis de regrouper dans une même classe des formes qui auparavant se distinguaient par leur cas, phénomène grammatical absent de l'anglais. En revanche, ces classes reflètent clairement la distinction du nombre, qui est une propriété marquée en anglais. Quelques singletons ont par ailleurs été créés, notamment pour le cas instrumental au nombre duel (classe 0). Ce genre ne s'applique en tchèque qu'aux parties du corps qui constituent une paire (les mains, les pieds, les yeux et les oreilles) et n'est marqué qu'à l'instrumental. Il correspond généralement en anglais (et en français) à une construction prépositionnelle dans *rukama* → *avec [les] mains*), et la présence d'une préposition anglaise (ici *with*) dans la traduction empêche alors le rattachement de la cellule au reste du paradigme. Ces formes se trouvent parallèlement souvent incluses dans des expressions idiomatiques, comme *mezi čtyřma* ***očima*** (entre quatre yeux), qui correspondent dans les données d'entraînement à des traductions non littérales : *in private* (en privé). C'est ainsi que cette forme n'a pas été rapprochée d'autres membres du paradigme qui se traduisent plus couramment par *yeux*.

TABLE 7 – Quelques classes nominales tchèques optimisées pour l'anglais (grand système)

NOMS CS-EN				
Classe 0	Classe 1	Classe 13	Classe 16	Classe 12
		Fém+Sing+Nominatif	Masc+Sing+Nominatif	Neut+Plur+Nominatif
	Fém+Sing+Vocatif		Masc+Sing+Vocatif	
		Fém+Sing+Accusatif	Masc+Sing+Accusatif	Neut+Plur+Accusatif
		Fém+Sing+Génitif	Masc+Sing+Génitif	Neut+Plur+Génitif
		Fém+Sing+Datif	Masc+Sing+Datif	Neut+Plur+Datif
		Fém+Sing+Prépos	Masc+Sing+Prépos	Neut+Plur+Prépos
Fém+Duel+Instru		Fém+Sing+Instru	Masc+Sing+Instru	Neut+Plur+Instru

Les formes du vocatif féminin sont également à part (classe 1), à cause de mots très fréquents comme *paní* (madame) que le modèle d'alignement a trop souvent relié à des mots anglais situés autour de la traduction, comme le nom de famille qui suit (*paní Ashton*). Nous constatons ici un premier impact de la qualité des alignements sur le résultat de notre modèle.

Le genre, enfin, est une caractéristique qui distingue clairement les noms dans les classes obtenues. En effet, il constitue une partie intrinsèque du nom et l'étiqueteur que nous avons employé pour le tchèque ne regroupe pas par un même lemme des paires comme *maître* - *maîtresse*. Nous notons toutefois qu'avec un seuil de gain d'information minimum *m* inférieur à zéro (voir section 2.2), les genres masculin animé et masculin inanimé tendent à se regrouper dans les mêmes classes. Ce regroupement devient possible lorsque l'étiqueteur analyse un même lemme comme animé ou non selon le contexte.

TABLE 8 – Quelques classes verbales tchèques optimisées pour l'anglais (petit système)

VERBES CS-EN		
Classe 6	Classe 9	Classe 11
Pers3+Sing+Prés	Pers2+Plur+Cond	Pers1+Plur+Cond
Pers3+Sing+Impér	Pers2+Plur+Fut	Pers1+Plur+Fut+Affirm
	Pers2+Plur+Prés	Pers1+Plur+Fut+Négat
	Pers2+Sing+Prés	

Nous présentons quelques classes verbales obtenues avec les petites données pour la paire tchèque-anglais dans le tableau 8. Nous constatons d'abord que la troisième personne du singulier au présent est à part (classe 6), ce qui correspond à la seule forme marquée de la plupart des verbes anglais conjugués au présent : *I cluster, he cluster**s*** (*je classifie, il classifie*). [9] Dans la classe 9 sont regroupées des formes de verbes à la deuxième personne. D'une part, cette classe regroupe différents temps (conditionnel, futur et présent), ce qui est une fois de plus probablement dû à des erreurs d'alignement où un verbe tchèque n'a pas été associé à l'auxiliaire anglais (*will, would*) marquant un temps ou un mode. D'autre part, contrairement aux auxiliaires, il semble que les pronoms personnels de deuxième personne aient généralement été correctement rattachées aux verbes tchèques. [10] La classe regroupe le singulier et le pluriel, qui ont à la fois un pronom identique en anglais (*you*) et une forme verbale indifférenciée dans la plupart des cas. Cette classe est néanmoins distinguée de la classe 11, qui ne regroupe que des formes de la première personne du pluriel vraisemblablement alignées au pronom anglais *we* (différent du singulier *I*). Dans cette dernière classe, nous notons de nouveau une confusion entre le futur et le conditionnel, mais aussi au niveau de la polarité. En effet, la négation tchèque est marquée par un préverbe (*ne-*), si bien que la particule négative anglaise (*not*) doit être alignée au verbe tchèque. En réalité, ces alignements sont très souvent manqués et il s'agit ici d'un problème bien connu pour cette paire de langues en traduction automatique (Rosa, 2013).

La normalisation des adjectifs tchèques par rapport à l'anglais présente moins de bruit que pour les verbes (tableau 9). En effet, la classe 25 regroupe toutes les formes de l'adjectif qualificatif sans considération du genre, du nombre ou du cas, faisant écho à l'adjectif anglais qui est invariable. Ces adjectifs sont ensuite distingués de leurs formes comparatives (classe 29) et des formes superlatives (classe 39). Enfin, nous constatons que les adjectifs négatifs sont réunis dans une classe cohérente

9. Nous observons également dans cette classe une forme impérative à la troisième personne, ce que l'étiqueteur n'applique qu'au mot *budiž* (*soit* dans *que la lumière* ***soit***) relevant du lemme *být* (être). Notre modèle a ainsi rapproché cette forme de celle du présent car toutes deux partagent fréquemment la même traduction dans les données d'entraînement.

10. Ceci est attendu dans le cas très courant où le pronom sujet d'un verbe tchèque est omis, si bien que le pronom anglais est rattaché au verbe tchèque marquant la personne dans sa conjugaison.

TABLE 9 – Quelques classes adjectivales tchèques optimisées pour l'anglais (grand système)

ADJECTIFS CS-EN			
Classe 25	Classe 29	Classe 39	Classe 31
Masc+Sing+Nomin	Masc+Sing+Nomin+Comparat	Masc+Sing+Nomin+Superlat	Masc+Sing+Nomin+Négat
Masc+Sing+Accus	Masc+Sing+Accus+Comparat	Masc+Sing+Accus+Superlat	Masc+Sing+Accus+Négat
Masc+Plur+Instru	Masc+Plur+Instru+Comparat	Masc+Plur+Instru+Superlat	Masc+Plur+Instru+Négat
Fém+Plur+Prépos	Fém+Plur+Prépos+Comparat	Fém+Plur+Prépos+Superlat	Fém+Plur+Prépos+Négat
Neut+Sing+Datif	Neut+Sing+Datif+Comparat	Neut+Sing+Datif+Superlat	Neut+Sing+Datif+Négat
...	...	...	...

(classe 31), contrairement à ce que nous avons précédemment observé pour les verbes. Ici, la négation tchèque correspond généralement à une négation lexicalisée en anglais (*possible*, ***im****possible*), si bien que les traductions d'adjectifs affirmatifs et négatifs se traduisent souvent par deux mots différents, ce à quoi notre modèle est sensible.

À titre de comparaison, lors de la normalisation du russe par rapport à l'anglais, notre modèle a créé deux classes adjectivales : la première pour l'unique forme invariable du comparatif (singleton) et la suivante pour toutes les autres formes (réunissant tous les genres, nombres et cas). Quant au superlatif, il n'est pas marqué morphologiquement en russe et est composé avec le déterminant *samyj* (*samyj krasivyj*, **le plus** beau), qui doit s'aligner avec l'article et l'adverbe anglais *the most* (le plus).

TABLE 10 – Des classes de pronoms personnels tchèques optimisées pour l'anglais (grand système)

PRONOMS PERSONNELS CS-EN	
Classe 7	Classe 32
Sing+Pers1+Nomin	Sing+Pers1+Accus
	Sing+Pers1+Datif
	Sing+Pers1+Prépos
	Sing+Pers1+Génitif
	Sing+Pers1+Instru

La normalisation des pronoms personnels tchèques par rapport à l'anglais (tableau 10) distingue le nominatif des autres cas à la première personne du singulier. En effet, en anglais, le pronom sujet ***I*** (cas nominatif) a une forme différente du pronom objet *me* qui correspond à la fois à l'objet direct (accusatif), indirect (datif) et à toutes sortes de compléments introduits par une préposition (autres cas) : *with* ***me*** (*avec moi*). [11]

4.2 Normalisation du tchèque par rapport au français

La normalisation du tchèque par rapport au français comporte certains points communs avec celle opérée par rapport à l'anglais. C'est le cas notamment des noms (tableau 11) qui ne se distinguent, en français comme en anglais, que par le nombre. Notons que notre modèle a ici aussi laissé les formes du duel féminin et du vocatif féminin au singulier dans deux classes à part (0 et 1).

Les adjectifs, en revanche, ont été classifiés différemment, comme on peut le voir dans le tableau 12. Cette fois, les classes sont toujours cohérentes du point de vue du nombre, ce qui s'explique par le fait que les adjectifs français varient également en nombre. Nous notons toutefois une certaine

11. Par souci de clarté, nous n'avons disposé dans cette classe que les pronoms tchèques de forme courte (*mě*), alors qu'elle contient également les formes longues (*mne*) distinguées par notre étiqueteur.

TABLE 11 – Quelques classes nominales tchèques optimisées pour le français

NOMS CS-FR				
Classe 0	Classe 1	Classe 8	Classe 11	Classe 12
		Fém+Sing+Nominatif	Masc+Plur+Nominatif	Masc+Sing+Nominatif
	Fém+Sing+Vocatif		Masc+Plur+Vocatif	Masc+Sing+Vocatif
		Fém+Sing+Accusatif	Masc+Plur+Accusatif	Masc+Sing+Accusatif
		Fém+Sing+Génitif	Masc+Plur+Génitif	Masc+Sing+Génitif
		Fém+Sing+Datif	Masc+Plur+Datif	Masc+Sing+Datif
		Fém+Sing+Prépos	Masc+Plur+Prépos	Masc+Sing+Prépos
Fém+Duel+Instru		Fém+Sing+Instru	Masc+Plur+Instru	Masc+Sing+Instru

confusion avec la catégorie du genre, dont plusieurs valeurs cohabitent parfois au sein d'une même classe. Cette confusion provient du fait que le genre de l'adjectif dépend souvent de l'accord, par exemple, avec le nom dans le cadre du syntagme nominal. Le genre des noms ne correspondant pas nécessairement entre le français et le tchèque, un adjectif masculin tchèque se retrouve aligné avec des formes masculines et féminines françaises. *A priori*, une telle situation ne devrait pas mener à un gain d'information élevé lorsque l'on considère la fusion de deux adjectifs tchèques de genres différents. Toutefois, si le hasard des données sur-représente un genre français particulier pour deux adjectifs tchèques de genre différent, notre modèle opère une fusion. C'est ce qui s'est produit avec les classes 16 et 18 qui regroupent des formes féminines et neutres.

TABLE 12 – Quelques classes adjectivales tchèques optimisées pour le français

ADJECTIFS CS-FR			
Classe 16	Classe 18	Classe 24	Classe 25
Fém+Sing+Nomin	Fém+Plur+Nomin	Fém+Plur+Nomin+Négat	Fém+Plur+Nomin+Comparat Fém+Plur+Nomin+Superlat
Fém+Sing+Accus	Fém+Plur+Accus	Fém+Plur+Accus+Négat	Fém+Plur+Accus+Comparat Fém+Plur+Accus+Comparat
...	...	...	...
Classe 16	Classe 18	Classe 29	Classe 32
Neut+Sing+Datif	Neut+Plur+Datif	Neut+Plur+Datif+Négat	Neut+Plur+Datif+Comparat
...	...	...	...

Nous constatons enfin que notre modèle regroupe souvent les formes comparatives et superlatives. Ceci est encore une fois dû à des erreurs du modèle d'alignement qui associe visiblement correctement un comparatif tchèque (*brzčejší*) à un adverbe et un adjectif français (*plus tôt*), mais qui a plus de difficulté à rattacher également l'article formant le superlatif français (*nejbrzčejší*, ***le*** *plus tôt*), si bien que les deux constructions françaises ne sont plus distinguées, ce qui a pour effet de regrouper les deux mots sources tchèques dans la même classe.

Les verbes tchèques optimisés par rapport au français sont divisés en 19 classes, dont l'une rassemble les formes du passé qui se distinguaient auparavant par le genre, probablement car l'accord du participe passé français est bien plus rare que celui du tchèque. Ce regroupement a permis au système de traduction d'estimer des paramètres plus fiables sur toutes les formes d'un verbe passé tchèque. Le tableau 13 présente l'un des effets de la normalisation sur la qualité de la traduction, où le verbe *zaznamenali* est traduit par une forme française au présent (*note*) par le système cs-fr, puis finalement par un passé (*ont enregistré*) avec le système cx-fr (section 3.2).

L'un des inconvénients de notre modèle réside donc dans sa grande dépendance envers la qualité des alignements. Or, lorsqu'il s'agit d'aligner une langue synthétique à une langue analytique, de

TABLE 13 – Meilleure traduction du temps (cx-fr)

source (cs)	Republikánští advokáti ostatně za posledních deset let **zaznamenali** pouze 300 případů volebních podvodů v USA.
cs-fr	Républicains avocats après ces dix dernières années, **note** seulement 300 cas de fraude électorale aux États-Unis.
cx-fr	D'ailleurs, les avocats républicains au cours des dix dernières années **ont enregistré** seulement 300 cas de fraude électorale aux États-Unis.
référence	D'ailleurs, les avocats républicains n'**ont recensé** que 300 cas de fraude électorale aux États-Unis en dix ans.

nombreux mots grammaticaux en cible, comme des auxiliaires ou des prépositions, restent non alignés (Burlot & Yvon, 2015). Néanmoins, notre modèle de normalisation est capable de prendre en compte les particularités morphologiques de la langue cible, comme l'absence de déclinaison casuelle nominale en français et en anglais, et peut distinguer le cas sujet et objet des pronoms. Notons enfin que les phénomènes linguistiques perçus par ce modèle nécessiteraient un grand nombre de règles coûteuses si la normalisation devait être fait manuellement.

5 Conclusion

Nous avons présenté une méthode simple et indépendante de la paire de langue étudiée pour déduire la normalisation d'une langue à morphologie riche par rapport à une langue cible qui consiste à regrouper dans une même classe des mots qui partagent la même traduction. Nous avons montré que cette méthode améliore la qualité de la traduction automatique.

À l'avenir, nous projetons d'aborder le sujet de la traduction dans la direction opposée, vers la langue à morphologie riche, ce qui exige une étape de post-traitement succédant à la traduction vers une langue normalisée. En effet, il convient dans ce cas de prédire les mots-formes en sélectionnant la bonne cellule dans la classe prédite par le système de traduction.

Remerciements

Ce travail a été partiellement financé par le programme de recherche et d'innovation de l'Union européenne Horizon 2020, dans le cadre de l'accord de subvention No. 645452 (QT21).

Références

BOJAR O., DUŠEK O., KOCMI T., LIBOVICKÝ J., NOVÁK M., POPEL M., SUDARIKOV R. & VARIŠ D. (2016). CzEng 1.6 : Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech and Dialogue : 19th International Conference, TSD* : Springer Verlag.

BURLOT F., KNYAZEVA E., LAVERGNE T. & YVON F. (2016). Two-step mt : Predicting target morphology. In *Proceedings of the International Workshop on Spoken Language Translation*, IWSLT'16, Seattle, USA.

BURLOT F. & YVON F. (2015). Morphology-aware alignments for translation to and from a synthetic language. In *Proceedings of the International Workshop on Spoken Language Translation*, IWSLT'15, p. 188–195, Da Nang, Vietnam.

CHERRY C. & FOSTER G. (2012). Batch tuning strategies for statistical machine translation. In *Proceedings of the NAACL-HLT*, p. 427–436, Montreal, Canada.

DÉCHELOTTE D., ADDA G., ALLAUZEN A., GALIBERT O., GAUVAIN J.-L., MAYNARD H. & YVON F. (2008). LIMSI's statistical translation systems for WMT'08. In *Proceedings of NAACL-HLT Statistical Machine Translation Workshop*, Columbus, Ohio.

DURGAR EL-KAHLOUT I. & YVON F. (2010). The pay-offs of preprocessing for German-English Statistical Machine Translation. In M. FEDERICO, I. LANE, M. PAUL & F. YVON, Eds., *Proceedings of the International Workshop on Spoken Language Translation*, p. 251–258.

DYER C., CHAHUNEAU V. & SMITH N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proc. NAACL*, p. 644–648, Atlanta, Georgia.

FRASER A., WELLER M., CAHILL A. & CAP F. (2012). Modeling inflection and word-formation in SMT. In *Proc. EACL*, p. 664–674, Avignon, France.

GOLDWATER S. & MCCLOSKY D. (2005). Improving statistical MT through morphological analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, p. 676–683.

HEAFIELD K. (2011). KenLM : Faster and Smaller Language Model Queries. In *Proc. WMT*, p. 187–197, Edinburgh, Scotland.

KOEHN P. (2005). A parallel corpus for statistical machine translation. In *Proc. MT-Summit*, Phuket, Thailand.

KOEHN P., HOANG H., BIRCH A., CALLISON-BURCH C., FEDERICO M., BERTOLDI N., COWAN B., SHEN W., MORAN C., ZENS R., DYER C., BOJAR O., CONSTANTIN A. & HERBST E. (2007). Moses : Open source toolkit for statistical machine translation. In *Proc. ACL :Systems Demos*, Prague, Czech Republic.

LO C.-K., CHERRY C., FOSTER G., STEWART D., ISLAM R., KAZANTSEVA A. & KUHN R. (2016). NRC Russian-English Machine Translation System for WMT 2016. In *Proceedings of the First Conference on Machine Translation*, p. 326–332, Berlin, Germany.

MARIE B., ALLAUZEN A., BURLOT F., DO Q.-K., IVE J., KNYAZEVA E., LABEAU M., LAVERGNE T., LÖSER K., PÉCHEUX N. & YVON F. (2015). LIMSI@WMT'15 : Translation task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, p. 145–151, Lisbon, Portugal.

MINKOV E., TOUTANOVA K. & SUZUKI H. (2007). Generating complex morphology for machine translation. In *Proc. ACL*, Prague, Czech Republic : Association for Computational Linguistics.

NEY H. & POPOVIC M. (2004). Improving word alignment quality using morpho-syntactic information. In *Proc. COLING*, p. 310–314, Geneva, Switzerland.

ROSA R. (2013). Automatic post-editing of phrase-based machine translation outputs. Master's thesis, Institute of Formal and Applied Linguistics, Charles University.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

SENNRICH R., FIRAT O., CHO K., BIRCH-MAYNE A., HADDOW B., HITSCHLER J., JUNCZYS-DOWMUNT M., LÄUBLI S., MICELI BARONE A., MOKRY J. & NADEJDE M. (2017). *Nematus : a Toolkit for Neural Machine Translation*, In *Proc. EACL 2017 Software Demonstrations*, p. 65–68. ACL.

SENNRICH R. & HADDOW B. (2016). Linguistic input features improve neural machine translation. In *Proc. WMT 2016*, p. 83–91.

SENNRICH R., HADDOW B. & BIRCH A. (2016). Neural machine translation of rare words with subword units. In *Proc. ACL 2016*.

SHAROFF S. & NIVRE J. (2011). The proper place of men and machines in language technology processing Russian without any linguistic knowledge. In *Russian Conference on Computational Linguistics*.

STRAKOVÁ J., STRAKA M. & HAJIČ J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proc. ACL : System Demos*, p. 13–18, Baltimore, Maryland.

TALBOT D. & OSBORNE M. (2006). Modelling lexical redundancy for machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, p. 969–976.

TOUTANOVA K., SUZUKI H. & RUOPP A. (2008). Applying morphology generation models to machine translation. In *Proc. ACL-08 : HLT*, p. 514–522, Columbus, Ohio.