Enrichissement d'un lexique bilingue par apprentissage analogique

Philippe Langlais — **Alexandre Patry**

Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
CP. 6128 Succ. Centre-Ville
H3C 3J7 Montréal, Canada

felipe@iro.umontreal.ca

RÉSUMÉ. *La présence de mots inconnus dans les applications langagières représente un défi de taille bien connu auquel n'échappe pas la traduction automatique. Les systèmes professionnels de traduction offrent à cet effet à leurs utilisateurs la possibilité d'enrichir un lexique de base avec de nouvelles entrées. Récemment, (Stroppa et Yvon, 2005) démontraient l'intérêt de l'apprentissage analogique pour l'analyse morphologique d'une langue. Dans cette étude, nous montrons qu'il offre également une réponse adaptée au problème de l'enrichissement d'un lexique bilingue et à la traduction d'entrées lexicales inconnues en particulier.*

ABSTRACT. *Unknown words are a well-known hindrance to natural language applications. In particular, they drastically impact machine translation quality. An easy way out commercial translation systems usually offer their users is the possibility to add unknown words and their translations into a dedicated lexicon. Recently, (Stroppa et Yvon, 2005) shown how analogical learning alone deals nicely with morphology in different languages. In this study we show that analogical learning offers as well an elegant and efficient solution to the problem of identifying potential translations of unknown words.*

MOTS-CLÉS : *analogie formelle, enrichissement de lexiques bilingues, traduction automatique.*

KEYWORDS: *formal analogy, bilingual lexicon projection, machine translation.*

1. Introduction

Une analogie proportionnelle ou *analogie* met en relation quatre entités, ce que l'on dénote habituellement [x : y :: z : t] et se lit : « x est à y ce que z est à t ». Pour ne prendre qu'un exemple, lors de l'exposition EXPO NANO sur les nanotechnologies qui s'est tenue à la Cité des sciences et de l'industrie de la ville de Paris[1], on pouvait lire la description suivante pour familiariser les élèves du collège à la notion de nanomètre :

> *Il y a la même différence de taille entre un atome et une balle de tennis,*
> *qu'entre cette même balle et notre planète.*

Même si le deuxième et le troisième terme sont identiques dans cette relation, il s'agit bien là d'une analogie[2] que nous pourrions noter [atome : balle :: balle : planète]. Le raisonnement par analogie, qui manipule ce type de relations, est un principe bien connu en sciences cognitives et en intelligence artificielle (Gentner *et al.*, 2001). L'aptitude à raisonner par analogie a notamment longtemps fait l'objet de questions dans les tests SAT (*Scholastic Assessment Test*) aux États-Unis. Ces tests, introduits en 1926, sont destinés aux étudiants désirant s'inscrire au lycée. Ils incluaient jusqu'en 2004[3] des questions où l'on présentait une paire de mots stimuli (ex. : *maçon* : *pierre*) à un sujet qui devait identifier parmi plusieurs autres paires celle correspondant au stimuli (ex. : *charpentier* : *bois*). Comme le discute (Turney, 2006), ce type d'analogie consiste à rapprocher des relations impliquées entre deux paires d'éléments :

> *Analogy is a high degree of relational similarity*
> *(L'analogie est un haut degré de similarité relationnelle)*

(Turney et Littman, 2005) proposent une approche basée sur le modèle de l'espace vectoriel, populaire en recherche d'information, qui permet de répondre correctement à 47 % de 374 questions typiquement posées dans ces tests. Une amélioration de ce système est décrite par (Turney, 2006).

(Pirrelli et Yvon, 1999) décrivent les fondements théoriques de l'*apprentissage analogique* qui s'applique potentiellement à tout problème où une correspondance doit être établie entre deux niveaux de représentation linguistique. Ils remarquent qu'une approche classique consiste à décomposer le problème à un niveau de granularité où une correspondance entre les unités des deux représentations linguistiques peut être établie avec suffisamment de fiabilité. Le passage d'un niveau de représentation à un autre s'effectue alors en composant chacune des correspondances. Dans beaucoup de problèmes d'intérêt en traitement des langues, cette correspondance n'est pas biunivoque et met en difficulté cette approche compositionnelle. Par exemple, dans le problème de la conversion d'une forme orthographique en une séquence de phonèmes, (Pirrelli et Yvon, 1999) rappellent que certains phénomènes sont difficiles à capturer, comme le relâchement trisyllabique : le premier *i* se prononce différemment dans

1. http://www.cite-sciences.fr/csmedia/storage/PM_Document/idv_nano.pdf.
2. (Lepage, 2003, pp. 40-42) cite Euclide qui distingue les proportions continues impliquant trois termes des proportions discrètes en impliquant quatre.
3. http://www.collegeboard.com/about/newstat/newstat.html.

*host**il**e* (/aj/) et dans *host**i**lity* (/i/), alors qu'ils ont le même voisinage orthographique immédiat. (Yvon, 1997) montre que l'analogie offre une solution adaptée au problème, sans pour autant qu'un découpage explicite des chaînes orthographiques (par exemple en graphèmes) ou des chaînes phonémiques (par exemple en syllabes) ne soit nécessaire.

(Lepage, 2003) propose un traitement particulièrement riche du rôle de l'analogie dans la langue, tant d'un point de vue formel, algorithmique qu'historique. L'auteur décrit différentes expériences, notamment en analyse morphologique, qui attestent du bien-fondé applicatif de l'analogie. Des principes dégagés dans ce travail, (Lepage et Denoual, 2005) présentaient récemment le système ALEPH, un système de traduction par l'exemple basé sur le principe de préservation d'*analogies formelles* (analogies s'identifiant à partir de la forme graphique des termes en présence) entre les phrases d'un bitexte. Ce système obtenait des performances honorables dans la tâche partagée IWSLT'05 (Eck et Hori, 2005). Une description plus précise du système est disponible (Lepage et Denoual, 2006). Une variante de ce système (Lepage et Lardilleux, 2007), faisant notamment usage d'une méthode originale d'alignement sous-phrastique, participait récemment à la campagne d'évaluation IWSLT'07 (Fordyce, 2007) avec des résultats moins probants lorsqu'on la compare aux autres systèmes. Il faut cependant souligner que cette année-là, les participants étaient invités à utiliser des ressources externes en sus des 40 000 paires de phrases distribuées par direction de traduction, ce que (Lepage et Lardilleux, 2007) n'ont pas fait.

(Stroppa et Yvon, 2005) proposent une formalisation algébrique à la fois concise et accessible de l'analogie formelle. Ils démontrent expérimentalement l'élégance et la puissance de l'apprentissage analogique dans deux tâches d'étiquetage morphologique ; la première consistait à étiqueter morphosyntaxiquement à l'aide d'un jeu d'étiquettes fines les mots inconnus d'une langue ; la seconde visait à prédire l'arbre d'analyse morphologique d'un lemme inconnu (lire également (Yvon *et al.*, 2004)).

D'autres auteurs se sont intéressés, ces dernières années, au potentiel applicatif des analogies formelles. (Claveau et L'Homme, 2005) montraient notamment qu'un type particulier d'analogies formelles très simples à calculer permettait de structurer les termes d'un domaine. (Moreau et Claveau, 2006) montrent également le bénéfice du raisonnement par analogie pour l'extension de requêtes dans un système monolingue de recherche d'information.

Dans cette étude, nous montrons que le raisonnement par analogie offre également une réponse adéquate au problème concret de la traduction d'entrées lexicales inconnues pour différentes directions de traduction. La plupart des systèmes de traduction ignorent ces mots, soit en les retirant du texte à traduire, soit en s'assurant qu'ils apparaissent non modifiés dans la traduction. Nous montrons que notre approche permet de traduire automatiquement environ 60 % des mots inconnus d'une application.

Cet article est organisé comme suit. Nous rappelons en section 2 le principe général de l'apprentissage analogique. Nous montrons en section 3 qu'il s'applique naturellement à l'enrichissement d'un lexique bilingue. Le protocole expérimental que nous

avons suivi est décrit en section 4. Nous évaluons notre approche en section 5 en la comparant à deux systèmes de base. Nous dressons en section 6 un bilan de ce travail, et proposons des perspectives de recherche.

2. Raisonnement analogique

2.1. *Apprentissage analogique*

L'approche mise en place dans cette étude pour l'enrichissement de lexiques bilingues s'inscrit dans le cadre théorique de l'apprentissage par analogie discuté par (Pirrelli et Yvon, 1999). Nous disposons d'un ensemble $\mathcal{L}$ d'observations représentées chacune par un ensemble de traits. On en distingue deux sous-ensembles qui définissent respectivement les *espaces* d'*entrée* $\mathcal{I}$ et de *sortie* $\mathcal{O}$. Étant donnée une observation incomplète où seuls les traits d'entrée sont connus, la tâche d'inférence consiste à prédire les valeurs des traits de sortie.

On trouve dans (Pirrelli et Yvon, 1999) plusieurs exemples d'encodage d'une observation en traits, chacun adapté à une tâche particulière du traitement automatique des langues. Dans leurs travaux, les traits sont soit des symboles, soit des chaînes de caractères, soit un ensemble des deux. Par exemple, `(walk,WALK+Verb+Pres)` et `(walked,WALK+Verb+Past)` sont deux observations où les traits d'entrée (le premier terme) sont ici des chaînes de caractères et les traits de sortie sont constitués de trois symboles (séparés par `+`). (Stroppa, 2005) étend cette panoplie à des traits structurés comme les arbres souvent rencontrés dans les applications langagières. Dans cette étude, nous souhaitons mettre au point un traducteur. Une observation est tout simplement encodée par une paire de chaînes de caractères, où une forme source est associée à sa traduction, comme dans `(analogie,analogy)`.

Sans perte de généricité, nous définissons formellement l'ensemble d'observations connues par :

$$\mathcal{L} = \{(i, o) \,|\, i \in \mathcal{I}, o \in \mathcal{O}\}$$

où $\mathcal{I}$ et $\mathcal{O}$ sont respectivement l'ensemble des formes des systèmes linguistiques d'entrée et de sortie de l'application. On désigne par $I(u)$ et $O(u)$ les *projections* respectives dans l'espace d'entrée et de sortie de l'observation u ; c'est-à-dire que si $u \equiv (i, o)$, alors $I(u) \equiv i$ et $O(u) \equiv o$ respectivement.

Pour une observation incomplète u, la procédure d'inférence met en œuvre trois étapes :

1) construire l'ensemble de triplets analogiques ou *stems* de u :

$$\mathcal{E}_{\mathcal{I}}(u) = \{(s, v, w) \in \mathcal{L}^3 \,|\, [I(s) : I(v) :: I(w) : I(u)]\}$$

2) construire l'ensemble des solutions trouvées aux équations analogiques formées par projection (ou transfert) des stems de u dans l'espace de sortie :

$$\mathcal{E}_{\mathcal{O}}(u) = \{t \in \mathcal{O} \,|\, [O(s) : O(v) :: O(w) : t] \,, \forall (s, v, w) \in \mathcal{E}_{\mathcal{I}}(u)\}$$

3) sélectionner les éléments solutions parmi $\mathcal{E}_{\mathcal{O}}(u)$.

La première étape identifie des relations analogiques dans l'espace d'entrée. La deuxième résout des équations analogiques dans l'espace de sortie avec comme *biais inductif* qu'une analogie dans l'espace d'entrée correspond à une analogie dans l'espace de sortie[4]. À l'issue des deux premières étapes, un ensemble de solutions est généré. Nous appelons donc cette partie de l'apprentissage analogique le *générateur*. Comme nous le verrons, nombre de ces formes ne sont pas désirables (le plus souvent parce qu'elles n'appartiennent pas à l'espace de sortie). Il convient donc de choisir parmi elles, la ou les solutions candidates, ce qui est le rôle de la troisième et dernière étape que nous appelons le *sélectionneur*.

L'apprentissage analogique possède des ressemblances avec l'approche des k plus proches voisins (k-ppv). Il s'agit en effet d'une approche passive qui n'effectue aucune généralisation explicite à partir du corpus d'entraînement qui doit être conservé au moment des tests. Les deux approches recherchent également des exemples « analogues » dans l'espace d'entrée. L'approche des k-ppv requiert la définition d'une mesure de similarité entre deux formes de l'espace d'entrée alors que l'apprentissage analogique nécessite la définition d'une relation analogique entre quatre objets (chaînes, symboles, arbres, etc.). Comme le discutent (Pirrelli et Yvon, 1999 ; Stroppa, 2005), il existe de nombreux avantages à la seconde approche. Le plus important selon nous est que des relations entre les objets sont établies séparément dans chaque espace de représentation et seul le biais analogique (à une analogie dans l'espace d'entrée correspond une analogie dans l'espace de sortie) permet de mettre ces niveaux en correspondance.

Ainsi si notre base d'apprentissage est constituée des trois exemples suivants où l'espace d'entrée contient des phrases arborées de la langue française et l'espace de sortie contient des phrases anglaises (traductions des phrases françaises) :

$$\mathcal{L} \equiv \left\{ \begin{array}{l} (\texttt{S(Pr(elle).Vb(mange))}, \texttt{ she eats}) \\ (\texttt{S(Pr(elle).Vb(parle))}, \texttt{ she talks}), \\ (\texttt{S(Pr(elles).Vb(mangent))}, \texttt{ they eat}) \end{array} \right\}$$

alors la traduction de la phrase arborée `S(Pr(elles).Vb(parlent))` est possible car l'analogie $[\texttt{S(Pr(elle).Vb(parle))} : \texttt{S(Pr(elle).Vb(mange))} :: \texttt{S(Pr(elles).Vb(mangent))} : \texttt{S(Pr(elles).Vb(parlent))}]$ peut être établie dans l'espace d'entrée, produisant dans l'espace de sortie, grâce au biais inductif, l'équation $[\texttt{she eats} : \texttt{she talks} :: \texttt{they eat} : \texttt{?}]$ dont la forme *they talk* est la solution. Le point important est qu'en aucun moment dans le processus, n'est explicité à quelle sous-chaîne de sortie correspond une sous-chaîne d'entrée, pas plus d'ailleurs que n'est spécifié le rôle des symboles grammaticaux dans l'espace d'entrée.

En particulier, nous pouvons, avec l'apprentissage analogique, inverser simplement l'espace d'entrée et de sortie de l'application. Sans aucune autre modification, nous aurions ainsi un système convertissant une chaîne anglaise en sa traduction ar-

4. (Stroppa, 2005) discute en profondeur les implications de ce biais inductif et le compare à celui d'autres méthodes d'apprentissage.

borée française. Cette réversibilité est loin d'être possible dans les approches comme les k-ppv pour lesquelles les classes sont essentiellement des symboles parmi un ensemble fini (et souvent réduit) comme les étiquettes morphosyntaxiques d'une langue.

Il convient cependant de souligner que la recherche d'exemples proches dans l'apprentissage analogique est une opération de complexité *a priori* cubique en la dimension de l'espace d'entrée, alors qu'elle est seulement linéaire dans le cas des k-ppv. Dans de nombreuses applications incluant celle présentée dans cette étude, cette recherche est trop coûteuse pour être effectuée exhaustivement et des heuristiques (pouvant impliquer des distances) doivent être appliquées pour réduire l'espace de recherche (voir section 3.2).

L'exemple que nous venons de présenter ainsi que l'application que nous visons dans cette étude s'accommodent toutes les deux de la même définition d'analogie que nous précisons dans la section suivante.

2.2. *Analogie formelle et équation analogique*

Différents liens peuvent unir quatre objets en relation analogique. Dans le cas des tests SAT, par exemple, ces liens sont de nature sémantique (ex. :[`hache` : `bûcheron` :: `roulette` : `dentiste`], [`aluminium` : `métal` :: `nouvelle` : `livre`]). Des analogies dites formelles avec lesquelles nous travaillons dans cette étude dénotent des relations graphiques entre les formes mises en relation. En français [`fournit` : `fleurit` :: `fournie` : `fleurie`] et en anglais [`abandoning` : `abandonment` :: `amending` : `amendment`] sont deux exemples de relations de nature morphologique. Le lecteur intéressé trouvera dans (Lepage, 2003) de nombreux exemples de relations analogiques (formelles) dans des langues très différentes.

Plusieurs définitions de l'analogie formelle ont été proposées (Lepage, 1998 ; Pirrelli et Yvon, 1999). (Stroppa et Yvon, 2005) en donnent une définition à la fois générale et intuitive qui aide à clarifier l'algorithme de résolution que nous présentons dans la section suivante.

Définition (analogie formelle) pour tout $(x, y, z, t) \in \Sigma^{\star^4}$, $[x : y :: z : t]$ **ssi** il existe une factorisation $(f_x, f_y, f_z, f_t) \in (\Sigma^{\star^d})^4$ telle que, $\forall i \in [1, d]$:

$$(f_y^{(i)}, f_z^{(i)}) \in \left\{ (f_x^{(i)}, f_t^{(i)}), (f_t^{(i)}, f_x^{(i)}) \right\}$$

où Σ représente l'alphabet. La plus petite valeur de d vérifiant cette définition est appelée le *degré* de l'analogie ; $f_x^{(i)}, f_y^{(i)}, f_z^{(i)}$ et $f_t^{(i)}$ sont appelés les *facteurs* de x, y, z et t respectivement. Par construction, nous avons $|f_x| = |f_y| = |f_z| = |f_t| = d$.

Par exemple, [`believer` : `unbelievable` :: `dreamer` : `undreamable`] est une proportion analogique formelle puisque nous pouvons facilement vérifier que la factorisation suivante répond à la définition :

$$\begin{array}{lcl} f_{believer} & \equiv & (\epsilon, believ, er) \\ f_{unbelievable} & \equiv & (un, believ, able) \\ f_{dreamer} & \equiv & (\epsilon, dream, er) \\ f_{undreamable} & \equiv & (un, dream, able) \end{array}$$

On peut démontrer qu'il n'existe pas de factorisation plus petite (en terme du nombre de facteurs impliqués). Par conséquent, le degré[5] de cette analogie est 3.

Il est important de souligner, dès à présent, que toute relation analogique formelle ne revêt pas nécessairement un caractère linguistique. (Lepage, 2003) parle à ce sujet du caractère aveugle de l'analogie. Il suffit pour s'en convaincre de considérer la relation [`abstention` : `mention` :: `absorption` : `morpion`] qui est bien une analogie formelle puisque :

$$\begin{array}{lcl} f_{abstention} & \equiv & (abs, \epsilon, t, ention) \\ f_{mention} & \equiv & (m, \epsilon, \epsilon, ention) \\ f_{absorption} & \equiv & (abs, orp, t, ion) \\ f_{morpion} & \equiv & (m, orp, \epsilon, ion) \end{array}$$

mais que nous qualifions de fortuite. (Lepage et Lardilleux, 2007) observent sur corpus, que la majeur partie des analogies de forme entre *chunks* possède un statut linguistique.

Lorsque l'un des termes (généralement le dernier) d'une relation analogique est absent, nous parlons d'*équation analogique*, notée $[x : y :: z :?]$. Une telle équation dénote l'ensemble des formes qui sont en relation analogique avec le stem $\langle x, y, z \rangle$:

$$[x : y :: z :?] = \{t \mid [x : y :: z : t]\}$$

Dans la suite de l'exposé, lorsque nous mentionnons les termes d'analogie ou d'équation analogique, nous faisons référence – sauf indication contraire – à une analogie ou une équation analogique formelle.

2.3. *Solveur d'équation analogique*

(Stroppa et Yvon, 2005) montrent qu'il est possible de calculer les solutions d'une équation analogique $[x : y :: z :?]$ à l'aide d'un transducteur à états finis construit à partir d'automates reconnaissant x, y et z. Cette approche généralise l'algorithme proposé initialement par (Lepage, 1998) que nous avons implémenté dans cette étude.

5. Le degré d'une analogie n'est pas une notion dont notre approche fait usage. Nous la mentionnons cependant en raison d'un point que nous discutons en fin de section 2.3.

Ce dernier présente l'avantage pratique de générer beaucoup moins de solutions que l'algorithme de (Stroppa et Yvon, 2005) (nous verrons que bon nombre des solutions générées par notre solveur sont indésirables). En contrepartie, il arrive que notre implémentation ne produise pas certaines solutions qui devraient l'être.

Afin de résoudre l'équation $[x : y :: z :?]$, l'algorithme décrit par (Lepage, 1998) synchronise deux tables d'édition, l'une entre x et y, l'autre entre x et z, sous l'intuition que les alignements de coût minimal dans ces deux tables révèlent les facteurs impliqués dans la définition que nous avons donnée d'une analogie formelle. Considérons, par exemple, l'équation $[\texttt{even} : \texttt{evenly} :: \texttt{uneven} : \texttt{?}]$. L'alignement entre *even* et *evenly* révèle les facteurs *even* et *ly*, tandis que de l'alignement entre *even* et *uneven* révèle les facteurs *un* et *even*. L'algorithme « compose alors dans le bon ordre » (ou synchronise) ces facteurs pour former ici la solution *un + even + ly*[6].

Cette synchronisation est au cœur même de l'algorithme proposé par (Lepage, 1998) et il convient de saluer l'intuition qu'il lui a fallu pour la mettre au point. En effet, si dans l'exemple précédent, les trois chaînes en présence contiennent toutes la sous-chaîne *even*, ce qui simplifie la synchronisation, il n'en va pas toujours de même. Ainsi, pour l'équation $[\texttt{even} : \texttt{usual} :: \texttt{unevenly} : \texttt{?}]$ que nous prenons comme exemple pour illustrer l'algorithme, la synchronisation doit rendre compte du fait que *even* et *usual* commutent. Comme nous allons le voir, identifier les sous-chaînes qui commutent à partir d'une table d'édition n'est pas un problème simple. En fait, l'algorithme que nous présentons n'y parvient pas toujours[7].

Nous reproduisons en figure 2 notre implémentation de l'algorithme de synchronisation décrit de manière (beaucoup) plus compacte dans (Lepage, 1998). Il requiert le calcul préalable de deux tables d'édition. L'algorithme standard (Wagner et Fisher, 1974) peut être appliqué avec comme particularité que le coût associé à l'insertion d'un caractère dans y ou z est nul plutôt qu'unitaire, précisément car ces caractères sont constitutifs des solutions recherchées. Les tables calculées pour notre exemple sont présentées en figure 1.

Avant d'illustrer le processus de synchronisation, nous souhaitons souligner que l'algorithme présenté en figure 2 prend pour acquis que l'opération d'édition (la valeur r dans l'algorithme) de chaque case d'une table d'édition est fixée. En pratique, cependant, plusieurs opérations (au maximum trois) peuvent mener à un minimum (local) de la distance d'édition (e) dans une case donnée de la table ; auquel cas l'algorithme doit être appliqué de manière concurrentielle. Nous avons adopté la stratégie d'appliquer l'algorithme séquentiellement sur différents chemins de plus faible coût échantillonnés aléatoirement dans chaque table. Comme il existe un nombre potentiellement exponentiel (en la taille des chaînes en présence) de chemins de coût minimal, nous fixons de manière arbitraire le nombre maximal `s` de chemins considérés dans

6. Le symbole *+* ne fait bien sûr pas partie de la solution.
7. (Langlais et Yvon, 2008) identifient une situation qui pose problème à l'algorithme présenté ici. Y. Lepage applique une version non déterministe de cet algorithme pour contourner ce problème (communication personnelle).

4	4	4	4	4	4	n	4	4	3	3	2	1	0	0	0
3	***3***	3	3	3	3	e	3	3	3	2	1	0	0	0	0
2	***2***	2	2	2	2	v	2	2	2	1	0	0	0	0	0
1	***1***	1	1	1	1	e	1	1	1	0	0	0	0	0	0
0	***0***	***0***	***0***	***0***	***0***		0	0	0	0	0	0	0	0	0
l	a	u	s	u	◁		▷	u	n	e	v	e	n	l	y

Figure 1. *Table de distance d'édition entre even et usual (partie gauche), et even et unevenly (partie droite). La signification des chemins marqués par des valeurs en gras italique ou encadrées est expliquée dans la légende de la figure 3*

chaque table d'édition. Nous appliquons donc (au plus) $\mathtt{s}^2$ fois l'algorithme de la figure 2. Notons qu'une même forme peut être générée par plusieurs synchronisations.

Cette façon de procéder n'est pas des plus efficaces, puisque de nombreux tests conduits lors des exécutions séquentielles de l'algorithme de synchronisation sont effectués plusieurs fois. De même, nous calculons les tables d'édition au complet alors qu'il est possible de n'en calculer qu'une sous-partie (Ukkonen, 1985). Quoi qu'il en soit, notre implémentation permet de résoudre en une seconde plus d'une centaine d'équations ayant au moins une solution ou plus de mille équations pour des stems échantillonnés aléatoirement[8]. Ce niveau de rapidité suffit à la réalisation de notre étude ; le problème n'étant pas tellement le temps mis pour résoudre une équation que le nombre d'équations envisagées (voir la section 3.2).

La figure 3 illustre deux synchronisations produites en échantillonnant deux des six cent quatre-vingt-un chemins de coût minimal entre *usual* et *even.* Nous observons qu'aucun de ces alignements ne révèle clairement la commutation de *even* avec *usual.* La table d'édition entre *unevenly* et *even* n'abrite, quant à elle, qu'un seul chemin de coût minimal (représenté à droite). La première solution, obtenue grâce au chemin en gras italique dans la figure 1, est *usuaunlly* qui n'est de manière évidente pas une solution souhaitable. La seconde, *unusually*, est obtenue en considérant plutôt le chemin encadré. C'est la solution qu'une personne familière de la langue anglaise fournirait spontanément à l'équation [`even : usual :: unevenly : ?`]. Les deux formes sont cependant bien des solutions légitimes de l'équation au sens de la définition que nous avons rappelée.

Nous avons vu qu'une équation analogique peut admettre zéro, une ou plusieurs solutions qui ne sont pas nécessairement des formes légitimes de la langue étudiée. Le tableau 1 présente les solutions trouvées par notre solveur à l'équation [`even : usual :: unevenly : ?`] en fonction du nombre maximal (`s`) de meilleurs chemins échantillonnés. Il existe un total de soixante-douze solutions à cette équation (selon

8. Temps mesuré sur un ordinateur Pentium cadencé à 3GHz pour `s=10`. Lorsqu'une équation admet une solution, elle en admet habituellement plusieurs, ce qui ralentit la résolution.

In : AB et AC, les tables d'édition entre A et B et entre A et C
Out : sol, une solution à l'équation $[A : B :: C :?]$ ou `échec`

$sol \leftarrow \phi$; $a \leftarrow |A|$; $b \leftarrow |B|$; $c \leftarrow |C|$

while (not $a = b = c = 0$) **do**
  if ($AB[a, b].r = AC[a, c].r$) **then**
    if ($AB[a, b].r =$ S) **then**
      if ($A_a = B_b = C_c$) **then**
        copie(A_a)
      else if ($A_a \neq B_b \neq C_c$) **then**
        return`échec`
      else
        if ($C_c = A_a$) **then** copie(B_b) **else** copie(C_c)
      $a \leftarrow a - 1$
      $b \leftarrow b - 1$
      $c \leftarrow c - 1$
    else if ($AB[a, b].r =$ I) **then**
      if ($AB[a, b].e > AC[a, c].e$) **then** copie(B, b) **else** copie(C, c)
    else
      $a \leftarrow a - 1$
  else
    if ($AB[a, b].r =$ I) **then**
      copie(B, b)
    else if ($AB[a, b].r =$ D) **then**
      $a \leftarrow a - 1$
      $c \leftarrow c - 1$
    else if ($AC[a, c].r =$ I) **then**
      copie(C, c)
    else
      $a \leftarrow a - 1$
      $b \leftarrow b - 1$

Avec les macro-opérateurs :
copie(x,i) $\equiv$ $sol \leftarrow$ x$_i$.sol ; $i \leftarrow i - 1$
copie(c) $\equiv$ $sol \leftarrow$ c.sol

Figure 2. *Algorithme de synchronisation des deux tables d'édition AB et AC entre A et B, et A et C respectivement. Une case dans l'une de ces tables est une paire $< e, r >$ qui contient la valeur de la distance d'édition (e) et l'opération ayant permis d'y arriver (r pour retour arrière) qui prend l'une des trois valeurs D(élétion), S(ubstitution) ou I(nsertion). L'algorithme décrit dans (Lepage,1998) contient des tests supplémentaires qui permettent de détecter plus rapidement un échec dans la synchronisation*

a) usua + un + l + ly

				e	v	e	n
u	**s**	**u**	**a**				**l**
I	I	I	I	D	D	D	S

		e	v	e	n		
u	**n**	e	v	e	n	**l**	**y**
I	I	D	D	D	D	I	I

b) un + usu + a + l + ly

e				v		e	n
	u	**s**	**u**		**a**		**l**
D	I	I	I	D	I	D	S

		e	v	e	n		
u	**n**	e	v	e	n	**l**	**y**
I	I	D	D	D	D	I	I

Figure 3. *Illustration de deux synchronisations mise en œuvre lors de la résolution de* $[even : usual :: unevenly :?]$ *et les solutions produites par l'algorithme de la figure 2 pour chacune d'elles. La synchronisation représentée en **a)** correspond au chemin en gras italique dans la figure 1 ; la synchronisation présentée en **b)** à celui qui est encadré*

la définition d'analogie que nous avons fournie), bien qu'une seule ici soit d'intérêt (*unusually*).

s	nb	(solution, fréquence)
5	5	(usuunally, 1) (usualunly, 1) (usunually, 1) (usuaunlly, 1) (uunsually, 1)
10	6	(usuunally, 2) (usunually, 2) (**unusually**, 2) (uunsually, 2) …
100	6	(**unusually**, 36) (uunsually, 27) (usunually, 19) (usuunally, 12) …
200	6	(**unusually**, 88) (uunsually, 56) (usunually, 32) (usuunally, 17) …
400	15	(**unusually**, 102) (unusualyl, 66) (uunsually, 66) (uunsualyl, 38) …
1000	21	(**unusually**, 168) (unusualyl, 104) (uunsually, 104) (unusulyal, 56) …

Tableau 1. *Solutions les plus fréquemment générées par différentes variantes de notre solveur à l'équation* $[\texttt{even} : \texttt{usual} :: \texttt{unevenly} : \texttt{?}]$. nb *indique le nombre de solutions trouvées*

Que la forme attendue (*unusually*) soit la plus fréquemment générée, lorsque l'échantillonnage est important n'est pas étonnant. Intuitivement, lorsqu'il existe une factorisation démontrant la relation analogique entre quatre termes, il en existe plusieurs qui font toutes intervenir les mêmes commutations (c'est-à-dire, la même factorisation minimale). Voici par exemple deux des factorisations démontrant que $[\texttt{voir} : \texttt{revoir} :: \texttt{faire} : \texttt{refaire}]$ est une analogie (la première est de degré 2, la seconde de degré 4), les deux factorisations capturant passivement qu'une opération de préfixation lie les deux paires de mots en présence :

$$\begin{array}{lcl} f_{voir} & \equiv & (\epsilon, voir) \\ f_{revoir} & \equiv & (re, voir) \\ f_{faire} & \equiv & (\epsilon, faire) \\ f_{refaire} & \equiv & (re, faire) \end{array} \qquad \begin{array}{lcl} f_{voir} & \equiv & (\epsilon, \epsilon, vo, ir, \epsilon) \\ f_{revoir} & \equiv & (r, e, vo, ir, \epsilon) \\ f_{faire} & \equiv & (\epsilon, \epsilon, fa, ir, e) \\ f_{refaire} & \equiv & (r, e, fa, ir, e) \end{array}$$

En d'autres termes, plus le degré d'une analogie est faible, plus les chances sont grandes que deux alignements échantillonnés correspondent à la même factorisation minimale. Nous référons le lecteur intéressé à (Stroppa, 2005) pour de plus amples informations à ce sujet.

3. Application à l'enrichissement d'un lexique bilingue

Les principes décrits dans la section précédente peuvent être appliqués au problème de l'enrichissement d'un lexique bilingue. Cette opération qui consiste à étendre un lexique existant à de nouvelles entrées présente de nombreux intérêts pratiques, notamment dans le cas de paires de langues faiblement dotées.

La couverture d'un lexique, aussi grande soit-elle, n'est pas garante de son utilité. Ainsi, même pour une paire de langues largement dotée comme le français et l'anglais, il n'existe pas de lexique bilingue couvrant les termes de tous les domaines de spécialités. Ceci justifie notamment l'intérêt de travaux visant à identifier la traduction de termes dans des ressources bilingues comparables (Fung, 1995 ; Déjean et Gaussier, 2002 ; Morin *et al.*, 2007) ou de manière plus proactive à apprendre automatiquement à traduire les termes d'un domaine particulier, comme ceux du domaine médical (Claveau et Zweigenbaum, 2005 ; Claveau, 2007).

Nous tenons à souligner que l'approche que nous préconisons ici ne permet d'enrichir un lexique donné qu'avec des formes morphologiquement apparentées à celles connues de ce lexique. Ceci est dû à notre choix d'utiliser des analogies formelles. Il est à ce titre intéressant de s'interroger sur la nature des phénomènes morphologiques que nous sommes en mesure de capturer (passivement). Nous voyons plusieurs problèmes à cette entreprise. Le premier est d'ordre pratique : l'apprentissage analogique peut être appliqué à plusieurs langues et aucun des auteurs n'a les compétences nécessaires pour effectuer cette analyse sur les quatre langues concernées par cette étude. Le deuxième argument est plus fondamental et reprend un point important discuté par (Lepage, 2003, p. 339) : l'apprentissage analogique ne fait pas usage d'information morphologique. Il serait certes très intéressant de vérifier dans quelle mesure les facteurs utilisés par notre système (ce que (Pirrelli et Yvon, 1999) appellent en anglais les *cores*) correspondent à des morphèmes, mais ce travail dépasse de loin le cadre volontairement expérimental que nous nous sommes fixé. (Lepage, 2003, p. 339) souligne également que certains phénomènes morphologiques sont difficiles à délimiter. Nous invitons le lecteur à consulter les nombreux exemples d'opérations d'infixation, de suffixation et de préfixation dans différentes langues que donne (Lepage, 2003, pp. 340–341) et dont l'analogie permet de rendre compte.

3.1. *Principe*

Notre approche peut-être illustrée par l'exemple de la figure 4 où nous cherchons à traduire en anglais le mot français (inconnu) `futilité`. Nous identifions pour cela

Étape 1	recherche de stems sources (français)
[têtes : tête :: futilités : futilité]	
[hostilités : hostilité :: futilités : futilité]	
[activités : activité :: futilités : futilité]	
...	
Étape 2	transfert + résolution cible
têtes↔*heads*	*hostilité*↔*hostility*
tête↔*head*	*hostilités*↔*hostilities*
futilités↔*trivialities*	*activités*↔*hobbies*
futilités↔*gimmicks*	*activité*↔*hobby*
...	
⇓	
[heads : head :: gimmicks : ?]	⇒ gimmick
[hostilities : hostility :: trivialities : ?]	⇒ triviality
[hobbies : hobby :: trivialities : ?]	⇒ triviality
...	
Étape 3	sélection des candidats
⟨triviality, 2⟩, ⟨~~gimmick~~, 1⟩, ...	

Figure 4. *Illustration de l'approche que nous préconisons à la traduction du mot français (inconnu) : futilité. Le générateur propose entre autres solutions* triviality *et* gimmick*. Le sélectionneur écarte la deuxième d'entre elles*

(étape 1) des relations analogiques dans la langue source comme : [activités : activité :: futilités : futilité]. Nous projetons (étape 2) ces relations en langue cible de manière à définir des équations analogiques (cibles) comme : [heads : head :: gimmicks : ?] ou [hostilities : hostility :: trivialities : ?] que nous résolvons. La dernière étape consiste à choisir parmi les solutions ainsi colligées celles que nous retenons.

L'application de l'apprentissage analogique à notre problème est immédiate. Le corpus d'observations $\mathcal{L}$, que nous désignons par *lexique amorce* dans la suite, est un ensemble de paires de mots (s, t) en relation de traduction (ex. : hostilités, hostilities). Il est important de bien réaliser qu'aucun aménagement particulier à l'apprentissage analogique décrit en section 2.1 n'est nécessaire. Il serait en particulier trompeur de penser qu'il existe une version monolingue puis une version bilingue de l'apprentissage analogique.

Que notre application soit de nature bilingue est uniquement lié au choix d'encodage des observations de $\mathcal{L}$. Le fait de choisir pour traits d'entrée et de sortie des chaînes dans la langue source et la langue cible, respectivement, nous permet d'obtenir un traducteur. Sans aucune modification de nos outils, nous pourrions réaliser un étiqueteur morphosyntaxique (monolingue) en prenant pour trait de sortie les étiquettes morphosyntaxiques de chaque mot en entrée, comme dans *(il mange, pronom + verbe)*.

3.2. *Implémentation*

La simplicité apparente de l'apprentissage analogique dissimule toutefois des problèmes pratiques que nous décrivons maintenant.

3.2.1. *Générateur*

L'étape 1 d'identification des triplets analogiques dans l'espace d'entrée est une opération trop coûteuse en temps (cubique en la dimension de l'espace d'entrée). Nous utilisons deux techniques pour réduire cette complexité. La première consiste à utiliser les équations analogiques dans l'espace d'entrée en mode génératif : plutôt que de vérifier tous les triplets $\langle s, v, w\rangle$ entretenant une relation analogique avec u, nous cherchons les solutions à $[I(v) : I(s) :: I(u) :?]$. Celles qui appartiennent à l'espace d'entrée définissent avec les deux premiers termes de la proportion les triplets que nous recherchons. Il s'agit d'une méthode exacte qui repose sur la propriété d'inversion des rapports (Stroppa, 2005) :

$$[x : y :: z : t] \equiv [y : x :: t : z]$$

Cette méthode, qui réduit la construction de $\mathcal{E}_{\mathcal{I}}(u)$ à une opération de complexité quadratique en la dimension de l'espace d'entrée, est encore trop coûteuse. Nous appliquons donc une seconde méthode, cette fois-ci heuristique, qui consiste à ne calculer les équations analogiques que sur les seuls mots proches de $I(u)$; formellement, nous construisons $\mathcal{E}_{\mathcal{I}}(u)$ selon (étape 1) :

$$\mathcal{E}_{\mathcal{I}}(u) = \{t \mid \ [x : y :: I(u) : t] \ , \forall x \in v_{\mathrm{n}}(I(u)) \text{ et } y \in v_{\mathrm{n}}(x)\}$$

où $v_{\mathrm{n}}(x)$ est une fonction de voisinage d'une lexie x qui réunit les $\mathtt{n}$ formes les plus proches de x selon une fonction f :

$$v_{\mathrm{n}}(x) = \underset{y}{\mathrm{argmin}_{\mathrm{n}}}\{f(x, y)\}$$

où f dans cette étude est la distance d'édition avec coût unitaire (Levenshtein, 1966)[9].

3.2.2. *Sélectionneur*

Nous avons mentionné qu'une équation analogique peut générer plusieurs solutions, certaines n'étant pas des formes légitimes de l'espace de sortie. L'étape 3 du processus d'inférence consiste dans notre cas à ne retenir que les solutions analogiques qui sont présentes dans un (grand) lexique monolingue cible $\mathcal{V}$, que nous appelons le *lexique de validation*. Nous avons compilé à cet effet à partir de textes

9. Le lecteur attentif aura noté que nous avons souligné en section 2.1 que contrairement à l'approche des k plus proches voisins, le raisonnement par analogie ne requiert pas de distance. La distance que nous utilisons ici n'est pas constitutive de l'approche (comme c'est le cas dans les k-ppv) mais répond seulement à des considérations pratiques : nous pourrions, par exemple, nous en affranchir en tirant aléatoirement des triplets dans l'espace d'entrée.

variés un lexique monolingue totalisant 466 439 formes différentes. (Lepage et Lardilleux, 2007) proposent une approche moins rigide où sont éliminées les formes qui contiennent des m-grammes (pour m suffisamment grand) non vus dans le corpus d'apprentissage. (Langlais *et al.*, 2008) étudient la possibilité d'entraîner un classificateur à séparer les formes désirables des autres.

3.2.3. ANALOG

Notre implémentation de l'inférence analogique, que nous baptisons ANALOG, est contrôlée par deux métaparamètres : le seuil d'échantillonnage `s`, c'est-à-dire le nombre maximal d'alignements de coût minimal considérés dans une table d'édition au moment de la résolution d'une équation et le seuil de voisinage `n` qui contrôle le nombre de formes voisines considérées dans l'étape 1 de recherche de relations analogiques. L'influence de ces paramètres est étudiée en section 5.

Lorsque ANALOG produit plusieurs traductions d'une forme source, elles sont proposées en ordre décroissant de fréquence avec laquelle elles sont générées. Il ne s'agit pas ici de la fréquence avec laquelle notre solveur génère une forme mais du nombre de fois où une forme est la solution d'une équation analogique cible résolue pendant l'étape 2. Des exemples de traductions produites par notre système sont présentés dans le tableau 2. La traduction *reincorporated* produite par ANALOG pour le mot *réintégrés* est, par exemple, la solution de l'équation [`insert : reinsert :: incorporated : ?`] générée par projection de l'analogie [`inséré : réinséré :: intégrés : réintégrés`], mais est aussi la solution des équations [`integrating : reintegrating :: incorporated : ?`] et [`integrate : reintegrate :: incorporated : ?`], toutes les deux projections de l'analogie [`intégration : réintégration :: intégrés : réintégrés`]. Quatorze autres équations cibles ont également pour solution cette même traduction.

4. Protocole expérimental

4.1. *Corpus*

Nous avons réalisé nos expériences en utilisant les corpus de la campagne d'évaluation des systèmes de traduction qui s'est tenue lors de l'atelier WMT'06 (Koehn et Monz, 2006). Dans cette tâche, les bitextes d'entraînement étaient constitués d'environ 700 000 paires de phrases extraites de textes parlementaires européens. Un corpus disjoint de 2 000 phrases extraites de cette même source (`in-domain`) était à traduire par les systèmes participants, ainsi que 1 064 phrases[10] hors domaine (`out-domain`) en provenance du site Internet de `Project Syndicate` (`http://www.project-syndicate.com`), une organisation sans but lucratif qui distribue des articles de revue sur des thèmes variés (politique, économie, science, etc.).

10. Nous avons retiré 30 de ces phrases qui contenaient des problèmes d'encodage.

source	cand	$\mathcal{V}$	(candidat, fréquence)
anti-agricole	296	5	(anti-farm, 5) (anti-agricultural, 3) (anti-farming, 3) (anti-rural, 3) (anti-farmer, 3)
concentrerait	2947	7	(concentrat, 11) (concentrate, 4) (summarized, 4) (summarizing, 4) (concentrating, 3) (focuss, 3) (focus, 3)
écrirait	156	4	(writs, 1) (write, 1) (writes, 1) (writ, 1)
réintégrés	2686	18	(reinstated, 20) (reintegrated, 17) (re-integrated, 13) (re-entered, 10) (reincluded, 8) (reinvolved, 8) (reincorporated, 8) (reinserted, 7) (reinstate, 7) (reintegrate, 6) (reinstating, 4) (accomplished, 3) (rebuilt, 3) (reinclude, 3) (rejoined, 3) (reverte, 2) (reintegration, 2) (reintegrating, 2)
galette	218	1	(pancake, 13)

Tableau 2. *Exemples de traductions obtenues avec un lexique amorce entraîné sur 100 000 paires de phrases.* `cand` *indique le nombre de solutions analogiques cibles générées (étape 2) ; la colonne* $\mathcal{V}$ *dénombre les traductions candidates retenues une fois validées par le lexique de validation (étape 3)*

Plusieurs problèmes surviennent lors de la traduction de textes ; nous nous intéressons ici au problème de la traduction des mots inconnus, c'est-à-dire dans notre cas, des mots des corpus de test absents du corpus d'entraînement. La plupart des systèmes de traduction ignorent ces mots, soit en les retirant du texte à traduire, soit en s'assurant qu'ils apparaissent non modifiés dans la traduction.

La distribution des mots inconnus dans la tâche WMT'06 est caractérisée dans le tableau 3. On remarque que les taux de mots inconnus dans cette tâche sont relativement bas et sans surprise et plus importants pour le corpus de test hors domaine. Également, de manière attendue, les corpus de la langue allemande contiennent plus de mots inconnus que les autres ; cette langue comportant notamment de nombreux mots composés.

Une analyse grossière des 441[11] mots inconnus relevés pour le français révèle que 83 d'entre eux (20 %) sont des noms propres, 54 (12 %) contiennent au moins un chiffre (numéros de page ou de lois, années, etc.), 37 (8 %) sont des mots composés, 18 (4 %) sont des mots d'emprunt (souvent des mots latins ou grecs), 7 sont des acronymes et 4 sont simplement dus à des problèmes de découpage en mots (tokenisation). Les 238 mots restants (54 %) sont ce que nous appelons des « mots ordinaires » qui n'ont simplement pas été rencontrés au moment de l'entraînement. Ce sont ces mots qui sont susceptibles de recevoir par ANALOG des traductions.

11. Quatre mots inconnus seulement sont en commun entre `in-domain` et `out-domain` pour la langue française.

	français		espagnol		allemand	
$\|\text{voc}\|_{train}$	80 343		100 435		186 231	
test-	in	out	in	out	in	out
$\|\text{voc}\|_{test}$	7 230	5 263	7 719	5 322	8 812	6 067
$\|\text{unk}\|$	180	265	233	292	469	599
oov %	0,26	1,22	0,38	1,37	0,84	2,87

Tableau 3. *Nombre de formes sources différentes des corpus d'entraînement et de test, $|unk|$ désigne le nombre de formes sources du corpus de test inconnues du corpus d'entraînement et oov % indique le pourcentage d'occurrences de formes sources inconnues dans le matériel de test*

4.2. *Protocole*

Nos expérimentations simulent une situation typique du développement d'un système de traduction basé sur l'exemple : nous disposons d'un bitexte à partir duquel est appris un lexique bilingue (probabiliste dans notre cas). Nous disposons de plus d'une (grande) collection de textes en langue cible que nous utilisons ici pour compiler le lexique de validation ($\mathcal{V}$). Notre but est de prédire des traductions de mots inconnus du corpus d'entraînement. Nous avons éliminé de notre étude les formes numériques que nous pouvons traiter de manière plus simple.

Évaluer la qualité de différentes variantes de notre approche nécessite le parcours de plusieurs listes de traductions. En plus de se révéler fastidieuse, cette entreprise est délicate : beaucoup de traductions produites par ANALOG ne sont valides que dans certains contextes. Il suffit pour cela de consulter les exemples du tableau 2 pour s'apercevoir de la difficulté de la tâche. Nous avons donc, dans un premier temps, procédé à une évaluation automatique en utilisant un lexique bilingue de référence. Ce lexique, dénommé $\mathcal{L}_{ref}$, est obtenu par entraînement sur la totalité du bitexte d'entraînement d'un modèle lexical à l'aide de la trousse à outils GIZA++ (Och et Ney, 2000)[12].

Par le même procédé, nous avons entraîné différents lexiques amorce ($\mathcal{L}_T$) sur des portions de T paires de phrases du corpus d'entraînement ; ce qui nous permet, en faisant varier T, d'étudier le comportement d'ANALOG en fonction de la taille du lexique amorce. Les valeurs de T étudiées sont 5 000, 10 000, 100 000 et 200 000 paires de phrases.

Nous traduisons avec ANALOG les mots sources du corpus de test de WMT'06 absents du lexique amorce ($\mathcal{L}_T$) mais présents dans le lexique de référence ($\mathcal{L}_{ref}$). Une traduction candidate est considérée correcte si elle est présente dans $\mathcal{L}_{ref}$.

12. En pratique, pour éliminer une partie du bruit d'un lexique appris automatiquement, nous le croisons (intersection) avec un lexique résultant de l'entraînement d'un modèle lexical entraîné dans la direction opposée (anglais-français *vs* français-anglais).

4.3. *Évaluation automatique*

La qualité des traductions produites par ANALOG est mesurée à l'aide de deux taux : le taux de réponse `r %` indique le pourcentage de mots inconnus qui reçoivent par ANALOG au moins une traduction (qu'elle soit bonne ou pas) et le taux de justesse `p %` qui mesure parmi les mots ayant reçu (au moins) une traduction le pourcentage de ceux qui reçoivent une traduction valide parmi celles proposées. Ces taux ne mesurent pas la précision et le rappel des traductions produites au regard de la référence, mais nous montrons plus loin qu'ils sont des bornes inférieures de la qualité de la première traduction produite pour un mot inconnu par ANALOG.

4.4. *Approches de base*

À des fins de comparaison, deux approches de base (*baseline*) sont testées sur la même tâche dans les mêmes conditions. La première approche (BASE1) est fondée sur les cognats et consiste à proposer comme traduction d'un mot source inconnu x, les n mots cibles de $\mathcal{V}$ les plus similaires (au sens de la distance d'édition) :

$$\left\{ \underset{y \in \mathcal{V}}{\operatorname{argmin}_{\mathrm{n}}}\, ed(y, x) \right\}$$

Cette approche marchera d'autant mieux que les langues sont proches (ex. *docteur* → *doctor*). La deuxième approche (BASE2) ressemble davantage à l'approche analogique et consiste à identifier les n formes sources du lexique amorce $\mathcal{L}_T$ les plus proches du mot inconnu x, puis à proposer leurs traductions telles qu'indiquées par ce lexique (ex. *demanda* → *demande* → *request*) :

$$\left\{ \cup_{v \in \text{vois}(x)} O(v) \right\} \quad \text{avec vois}(x) \equiv \underset{w \in \mathcal{L}_\mathcal{T}}{\operatorname{argmin}_{\mathrm{n}}}\, ed(I(w), x)$$

Chacune de ces approches est testée selon deux variantes. La première ($_{id}$) propose le même nombre de traductions qu'ANALOG fournit dans les mêmes conditions expérimentales (les approches sont donc directement comparables) ; la seconde variante ($_{10}$) propose dix traductions pour chaque mot inconnu.

5. Expériences

5.1. *Évaluation automatique*

Les performances du système ANALOG en fonction de la taille du bitexte sur lequel le lexique amorce est entraîné sont consignées dans le tableau 4 pour la direction de traduction français-anglais. Le seuil d'échantillonnage `s` est fixé à 20, le nombre de voisins `n` à 30. La première case du tableau indique, par exemple, que sur le corpus de test `in-domain`, notre système propose (au moins) une traduction pour 30,7 %

des mots du jeu de test inconnus du lexique amorce appris sur 5 000 paires de phrases ($\mathcal{L}_{5000}$). Pour la moitié (50,8 %) des mots recevant (au moins) une traduction par ANALOG, au moins l'une des traductions proposées est valide (selon $\mathcal{L}_{ref}$).

Le tableau 4 appelle plusieurs commentaires. Il convient tout d'abord de garder à l'esprit que ce que nous mesurons ici est davantage l'aptitude d'une approche à reconstruire un lexique de référence à partir d'un lexique amorce. En particulier, nous ne mesurons pas ici les traductions produites pour les mots inconnus du lexique de référence. Globalement, les performances d'ANALOG augmentent avec la taille du lexique amorce. Ceci est normal car plus ce lexique est grand, plus il contient de formes sources qui peuvent entrer en relation analogique avec un mot inconnu et plus les traductions de ces mots sont nombreuses, ce qui permet également de créer davantage de relations analogiques dans la langue cible.

Nous observons qu'ANALOG se comporte de manière similaire lorsqu'il traduit des mots du domaine des textes parlementaires européens et des mots hors domaine. Ceci est encourageant dans une perspective applicative et souligne l'une des forces de l'apprentissage analogique qui ne s'appuie pas sur des occurrences de formes, comme le font le plus souvent les approches statistiques, mais sur des relations paradigmatiques les liant (ex. : masculin/féminin).

Les approches BASE1 et BASE2 sont inférieures en qualité à l'approche ANALOG (BASE1 est, sans surprise, la moins bonne des trois). Il est possible avec ces approches

$\mathcal{L}_T$	5 000		10 000		50 000		100 000		200 000	
	p %	r %	p %	r %	p %	r %	p %	r %	p %	r %
	in-domain									
ANALOG	50,8	30,7	54,4	44,3	57,9	63,9	57,0	63,8	57,7	64,4
BASE1$_{id}$	31,6	30,7	32,3	44,3	24,7	63,9	20,3	63,8	20,9	64,4
BASE2$_{id}$	34,5	30,7	37,1	44,3	39,0	63,9	37,8	63,8	34,4	64,4
BASE1$_{10}$	26,7	100,0	28,3	100,0	23,9	100,0	20,0	100,0	16,6	100,0
BASE2$_{10}$	26,3	100,0	30,8	100,0	29,3	100,0	27,6	100,0	24,9	100,0
unk	[3 171 , 6,8]		[2 245 , 6,1]		[754 , 3,7]		[456 , 2,8]		[253 , 2,0]	
	out-domain									
ANALOG	52,4	28,9	54,4	42,4	51,7	68,0	53,6	73,4	55,3	79,2
BASE1$_{id}$	28,0	28,9	29,0	42,4	27,3	68,0	23,1	73,4	26,8	79,2
BASE2$_{id}$	32,9	28,9	35,0	42,4	32,5	68,0	35,9	73,4	40,8	79,2
BASE1$_{10}$	24,7	100,0	25,9	100,0	25,1	100,0	20,9	100,0	25,2	100,0
BASE2$_{10}$	21,7	100,0	26,4	100,0	27,2	100,0	29,4	100,0	33,6	100,0
unk	[2 270 , 6,2]		[1 701 , 5,5]		[621 , 3,2]		[402 , 2,4]		[226 , 1,7]	

Tableau 4. *Résultats d'*ANALOG *en fonction de la taille du lexique amorce ($\mathcal{L}_T$). Les lignes préfixées de unk indiquent le nombre de mots à traduire ainsi que le nombre moyen de traductions dans $\mathcal{L}_{ref}$*

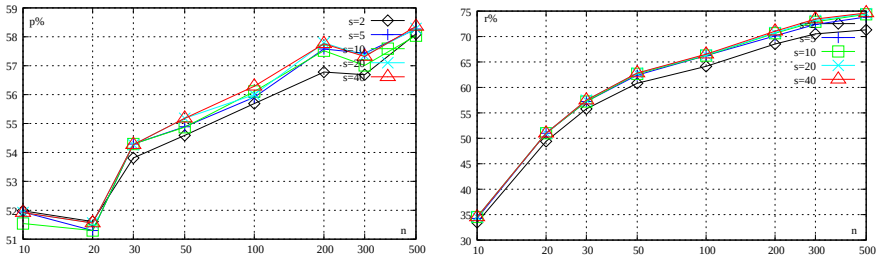

Figure 5. *Justesse (à gauche) et réponse (à droite) de la configuration $\mathcal{L}_{50\,000}$ en fonction des seuils d'échantillonnage $\mathtt{s}$ et de voisinage $\mathtt{n}$ pour la direction français-anglais*

d'obtenir un taux de réponse à 100 %, au prix d'une détérioration de la qualité des traductions produites, ce qui suggère qu'une combinaison des deux approches (comme par exemple écouter BASE2 lorsque ANALOG est silencieux) améliorerait les performances globales.

Nous estimons que pour la moitié des entrées non traduites, c'est une absence de relation analogique identifiée dans l'espace de sortie qui aboutit à l'absence de candidat. En moyenne pour le lexique amorce $\mathcal{L}_{100\,000}$, une entrée inconnue entre en relation analogique 988 fois du côté source, ce qui génère 52 formes sources qui appartiennent au lexique amorce. Du côté cible, une moyenne de 99 solutions analogiques sont proposées (par forme inconnue source) ; une moyenne de 5 d'entre elles seulement sont validées par le lexique de validation $\mathcal{V}$ et donc considérées ici.

Les performances d'ANALOG en fonction des seuils d'échantillonage `s` et de voisinage `n` sont étudiées pour le lexique amorce $\mathcal{L}_{50\,000}$[13]. Les taux de justesse et de réponse obtenus en faisant varier la valeur des deux seuils sont rapportés en figure 5. On observe très nettement l'influence que `n` a sur les performances. Pour des valeurs faibles de `n`, la justesse est aux alentours de 52 % et le taux de réponse de 30 % ; pour les valeurs élevées de `n`, ces taux augmentent à 58 % et 75 % respectivement. L'influence du nombre `s` d'échantillonnages effectués par notre solveur est beaucoup moins marquée ; les meilleures performances sont mesurées pour les valeurs les plus grandes.

Dans le tableau 5, sont présentés les résultats d'ANALOG pour trois directions de traduction : français, espagnol et allemand vers l'anglais. Les performances pour la direction français-anglais sont celles qui sont présentées dans le tableau 4 et ne sont rappelées qu'à des fins de comparaison. Nous nous concentrons sur la langue cible an-

13. Certaines sessions de traduction, notamment pour des seuils élevés d'échantillonnage et de voisinage requièrent de nombreuses heures de calcul. Nous avons donc opté pour un lexique amorce offrant un bon compromis entre le nombre de mots à traduire (754) et sa représentativité.

glaise car c'est celle qui est la plus communément étudiée au sein de la communauté. Nous observons une baisse d'environ 10 % de la justesse et du taux de réponse pour la direction allemand-anglais. Cela est probablement dû à l'heuristique sur la distance d'édition que nous avons introduite dans le générateur et qui n'est pas particulièrement appropriée à la traduction des mots composés qui sont nombreux en allemand. Nous observons également que pour les deux nouvelles directions étudiées, le taux de justesse tend à diminuer pour les lexiques amorces les plus grands. Du bruit dans la référence serait bien sûr une explication, puisqu'elle est obtenue automatiquement. Une autre explication serait qu'ANALOG introduit d'autant plus de bruit que de nombreuses analogies sont identifiées.

	Français		Espagnol		Allemand	
$\mathcal{L}_T$	p %	r %	p %	r %	p %	r %
5 000	51,4	30,7	52,8	30,3	49,3	23,1
10 000	55,3	44,4	52,0	45,2	47,6	33,3
50 000	58,8	64,3	54,0	66,5	44,6	53,2
100 000	58,2	65,1	53,9	69,1	45,8	55,6
200 000	59,4	65,2	46,4	71,8	43,0	59,2

Tableau 5. *Performances d'*ANALOG *mesurées sur le corpus* `in-domain` *pour différentes langues sources et différentes tailles de lexiques amorces. Les traductions sont effectuées vers l'anglais.* $\mathcal{L}_T$ *indique le nombre de paires de phrases utilisées pour entraîner le lexique amorce*

5.2. *Évaluation manuelle*

Une inspection des traductions proposées par ANALOG révèle certains problèmes dont cette évaluation ne rend pas compte. En particulier, certaines traductions correctes sont rejetées à cause d'un lexique de référence erroné ou incomplet. C'est notamment le cas des exemples de la figure 6 où, par exemple, `circumventing` et `fellow` sont des traductions légitimes de `contournant` et de `concitoyen` respectivement.

Nous avons donc entrepris une évaluation manuelle des traductions produites à partir du lexique amorce $\mathcal{L}_{100\,000}$ par les approches ANALOG et BASE2 pour les 127 mots du jeu de test `in-domain` inconnus de $\mathcal{L}_{ref}$ et ne contenant aucun chiffre. Nous avons décidé, non sans arbitraire, d'identifier comme valide une traduction candidate dès lors qu'elle est synonyme d'une traduction possible du mot source. *derisory*, *ridiculous*, ou *laughable* sont par exemple trois traductions proposées par ANALOG pour le mot français *dérisoires* que nous considérons toutes valides.

De cette évaluation il résulte que 75 (60 %) des mots inconnus reçoivent au moins une traduction valide par ANALOG, contre 63 (50 %) avec BASE2. 81 % des traductions produites en tête par ANALOG sont bonnes contre 22 (35 %) pour BASE2. Des 52 mots

contournant (*49 candidats*)
ANALOG ◇ (circumventing, 55) (undermining, 20) (evading, 19) (circumvented, 17) (overturning, 16) (circumvent, 15) (circumvention, 15) (bypass, 13) (evade, 13) (skirt, 12) ...
$\mathcal{L}_{ref}$ ◇ **skirting**, **bypassing**, by-pass, overcoming
concitoyen (*24 candidats*)
ANALOG ◇ (citizens, 26) (fellow, 26) (**fellow-citizens**, 26) (people, 26) (citizen, 23) (fellow-citizen, 21) (fellows, 5) (peoples, 3) (civils, 3) (fellowship, 2) ...
$\mathcal{L}_{ref}$ ◇ **fellow-citizens**

Figure 6. *Les dix meilleures traductions produites par* ANALOG *à partir du lexique amorce* $\mathcal{L}_{200\,000}$ *pour deux mots inconnus et leurs traductions dans* $\mathcal{L}_{ref}$ *sont montrées ici. Les traductions candidates en gras sont présentes dans la liste de référence*

n'ayant pas reçu par ANALOG de traduction satisfaisante, 38 (73 %) n'ont en fait reçu aucune traduction. Ces mots sont en majorité des noms propres, des mots d'une autre langue (latin, grec ou anglais) ainsi que des mots composés.

Nous concluons de cette évaluation manuelle à une échelle certes modeste, mais que nous croyons représentative, que notre approche permet pour la direction français-anglais de proposer une traduction valide en tête pour 80 % des entrées inconnues *ordinaires* (voir la section 4.1) de notre jeu de tests. Nous notons aussi que les taux de justesse et de réponse que nous mesurons de manière automatique sont des bornes inférieures des taux mesurés manuellement.

5.3. ANALOG *versus* GIZA++

Nous l'avons souligné, nos lexiques de référence contiennent des erreurs inhérentes au fait que nous les avons obtenus automatiquement. Nous avons notamment remarqué des erreurs régulières sur les mots peu fréquents. Il nous a donc semblé intéressant de comparer les traductions produites par ANALOG et par GIZA++ pour ces mots peu fréquents. Plus précisément, nous avons comparé la première proposition produite par ANALOG et par les modèles IBM 2 (Brown *et al.*, 1993) tels qu'implémentés par la boîte à outils GIZA++ sur un échantillon de 142 mots apparus de une à trois fois dans le bitexte de 200 000 paires de phrases utilisées pour entraîner $\mathcal{L}_{200\,000}$. L'évaluation consistait à décider du statut valide ou incorrect de chaque traduction. Il est à noter que pour les mots peu fréquents, cette distinction ne pose pas de grand problème, dans la mesure où les traductions sont plus souvent clairement mauvaises que potentiellement bonnes.

Les résultats de cette évaluation sont consignés dans le tableau 6. Il en ressort qu'environ 42 % des traductions produites par GIZA++ sont mauvaises, contre 4 %

		ANALOG			
		BONNE	MAUVAISE	SILENCE	
GIZA++	BONNE	59	2	22	83
	MAUVAISE	41	3	15	59
		100	5	37	142

Tableau 6. *Comparaison de la première traduction produite par* ANALOG *et par* GIZA++ *(modèle IBM 2) sur un échantillon aléatoire de 142 mots peu fréquents du corpus* $\mathcal{L}_{200\,000}$

seulement de celles produites par ANALOG. En revanche, ce dernier est silencieux pour 37 des 142 mots évalués.

Il ne faut bien sûr pas s'étonner des mauvais résultats de GIZA++ puisque les mots considérés sont peu fréquents, donc difficiles à modéliser à l'aide d'une approche statistique. Comme ANALOG est indifférent à la fréquence des mots, il est plus apte à les traduire ; ce qui est intéressant, car comme la loi de Zipf nous le rappelle, les corpus contiennent beaucoup de mots peu fréquents.

5.4. *Impact en traduction*

Nous avons mesuré l'apport d'ANALOG à un système de traduction statistique à l'état de l'art basé sur les modèles à segments contigus (*phrase-based models*). Le système décrit par (Patry *et al.*, 2006) a été utilisé à cette fin. Nous avons pour cela simplement ajouté dans la table de transfert du système, des paires associant un mot inconnu à la première traduction proposée par ANALOG ou par BASE2. Nous avons pris pour lexique amorce, celui obtenu à partir de la totalité du matériel d'entraînement de la tâche WMT'06, c'est-à-dire le lexique $\mathcal{L}_{ref}$.

La qualité des traductions est évaluée par rapport à une unique référence en terme du taux d'erreur au niveau des mots (WER pour *word-error rate*) et du score BLEU (Papineni *et al.*, 2002). Les performances de différentes variantes du système de traduction sont consignées en table 7 pour les seules phrases du corpus de test qui contiennent au moins un mot inconnu.

Des gains modestes, mais stables, pour les différentes directions de traduction, sont mesurés par les deux métriques d'évaluation automatique pour le système enrichi des meilleures traductions produites par ANALOG. Ce résultat n'est pas surprenant puisque dans le système de base (ligne BASE), les mots inconnus ne sont pas traduits et apparaissent tels quels dans la traduction. On observe cependant que BASE2 dégrade légèrement les performances. La raison est que BASE2 produit parfois des traductions erronées pour des invariants de traduction comme des noms propres, alors qu'ANALOG ne parvient généralement pas à les traduire. (Pirrelli et Yvon, 1999) avancent l'idée

que le silence lié à l'apprentissage analogique peut être bénéfique à certaines applications. Cela semble être le cas de la nôtre. En forçant BASE2 à proposer une traduction seulement dans les cas où ANALOG en produit une (ligne BASE2$_{id}$), les performances sont légèrement meilleures que le système de base, mais toujours inférieures à celles d'ANALOG.

	Français		Espagnol		Allemand	
	WER	BLEU	WER	BLEU	WER	BLEU
BASE	61,8	22,74	54,0	27,00	69,9	18,15
+BASE2	61,8	22,72	54,2	26,89	70,3	18,05
+BASE2$_{id}$	61,7	22,81	54,1	27,01	70,1	18,14
+ANALOG	61,6	22,90	53,7	27,27	69,7	18,30
nb. phrases	387		452		814	

Tableau 7. *Performances d'un système de traduction statistique utilisant ou non la première traduction proposée par* BASE2, BASE2$_{id}$ *ou* ANALOG *pour chaque mot inconnu du corpus de test*

6. Discussion et perspectives

Dans cette étude, nous avons étudié l'adéquation de l'apprentissage analogique à enrichir un lexique amorce. Nous nous sommes plus particulièrement intéressés au problème de la traduction de mots inconnus. À l'aide d'une référence construite automatiquement, nous avons mesuré, pour des lexiques amorces de taille suffisante, que plus de 60 % des mots inconnus ordinaires reçoivent par ANALOG un ensemble de traductions qui contient, dans environ la moitié des cas, une traduction de référence. Nous avons observé des performances comparables pour la direction espagnol-anglais mais une perte de l'ordre de 10 % pour la direction allemand-anglais que nous attribuons à l'heuristique que nous utilisons pour réduire la complexité de notre approche.

Une évaluation manuelle fait ressortir, pour la direction français-anglais, la bonne qualité des traductions produites. Nous observons en particulier sur un échantillon donné qu'environ 80 % des mots ordinaires inconnus reçoivent des traductions valides et que la même proportion reçoit en tête une traduction correcte.

Nous avons également mis en lumière qu'ANALOG était potentiellement utile à la traduction de mots peu fréquents, souvent mal appréhendés par les approches statistiques. Nous avons enfin validé notre approche en montrant que la première traduction produite par ANALOG pour un mot inconnu permettait d'améliorer un système de traduction statistique en l'état de l'art.

Une idée similaire à celle que nous avons présentée initialement dans (Langlais et Patry, 2007a) a fait l'objet d'une étude publiée peu de temps après par (Denoual, 2007). L'auteur y montre – dans le cadre de la tâche IWSLT'06 (Paul, 2006) de traduction du japonais vers l'anglais – que la traduction par analogie des mots inconnus

rencontrés dans le jeu de tests de cette tâche améliore les performances du système de traduction sous-jacent. Ceci confirme le bien-fondé de notre approche et renforce son universalité, le japonais et l'anglais appartenant à des familles linguistiques bien distinctes.

Plusieurs auteurs se sont intéressés à l'identification de traductions dans des corpus comparables, soit pour des mots simples (Fung et Yee, 1998 ; Rapp, 1999 ; Takaaki et Matsuo, 1999 ; Koehn et Knight, 2002), soit pour des termes de spécialité (Morin et Daille, 2004). Les techniques proposées dans ces travaux peuvent être employées également à l'enrichissement d'un lexique bilingue. Il convient cependant de souligner que, contrairement à ces approches, les traductions proposées par ANALOG émergent du seul principe de l'analogie et ne nécessitent aucune autre ressource externe qu'un lexique amorce. Nous ne sommes donc pas soumis au problème non trivial de l'acquisition de corpus dédiés (parallèles ou non) qui doivent contenir les mots que nous avons à traduire ainsi que leurs traductions. (Morin *et al.*, 2007) montraient récemment, dans le cadre d'une étude sur la traduction de termes de spécialité, que l'identification de corpus comparables sur Internet n'était pas une tâche aussi facile que ce que l'on pourrait croire.

La majeure partie des analogies que nous identifions au niveau des mots capturent des informations de nature morphologique. L'utilisation d'informations morphologiques en traduction statistique fait l'objet de plusieurs études ; le travail de (Nießen, 2002) étant à notre connaissance l'un des premiers. Selon les paires de langues considérées, des gains ont été mesurés lorsque les données d'entraînement sont de taille modeste (Lee, 2004 ; Popovic et Ney, 2004 ; Goldwater et McClosky, 2005).

Contrairement à ces études, ANALOG ne requiert pas d'analyse morphologique préalable, que ce soit du langage source, du langage cible ou des deux. Il serait intéressant de comparer notre approche à des approches inférant ces informations de manière non supervisée (Freitag, 2005). Une extension au cas bilingue du travail de (Hathout, 2001) sur les signatures analogiques serait également une piste de recherche intéressante.

Nous travaillons actuellement sur plusieurs améliorations de l'idée que nous avons présentée. Tout d'abord, l'analogie ne fait pas loi. Comme le souligne (Lepage, 2003), l'analogie revêt un caractère aveugle et nous avons montré qu'il était possible de mettre en relation analogique des formes qui ne sont pas reliées linguistiquement. Nous pensons qu'il serait utile, voir nécessaire, de distinguer les analogies fortuites des autres. Dans (Langlais *et al.*, 2008), nous abordons ce problème comme une tâche de classification.

Ensuite, même si dans cette étude nous nous sommes restreints à identifier des analogies entre mots, rien dans notre approche ne nous y contraint. Aussi, projetons-nous d'appliquer ANALOG à l'enrichissement de tables de séquences de mots couramment utilisées dans les systèmes de traduction en l'état de l'art. Une étude que nous avons menée récemment (Langlais et Patry, 2007b) suggère que cela est possible.

Enfin, comme nous l'avons mentionné, l'apprentissage analogique abrite certains problèmes pratiques non triviaux, dont celui d'identifier dans un corpus les triplets analogiques. Nous avons utilisé ici une approche heuristique basée sur la distance d'édition pour réduire l'espace de recherche d'ANALOG. Nous pensons qu'il est possible d'utiliser certaines propriétés algébriques des analogies formelles pour réduire cet espace, sans approximation. Des résultats prometteurs sont rapportés par (Langlais et Yvon, 2008).

Remerciements

Cette étude a grandement profité de discussions que nous avons eues avec Nicolas Stroppa et François Yvon ainsi que du tutoriel donné à TALN'06 par Yves Lepage. Nous remercions les relecteurs de cet article pour leurs commentaires avisés.

7. Bibliographie

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.

Claveau V., « Traduction automatique de termes biomédicaux pour la recherche d'information interlingue », *4th CORIA*, Saint-Étienne, France, 2007.

Claveau V., L'Homme M.-C., « Apprentissage par analogie pour la structuration de terminologie - Utilisation comparée de ressources endogènes et exogènes », *6e rencontre de Terminologie et Intelligence Artificielle (TIA'05)*, Rouen, France, avril, 2005.

Claveau V., Zweigenbaum P., « Automatic Translation of Biomedical Terms by Supervised Transducer Inference », *10th Conference on Artificial Intelligence in Medicine (AIME'05)*, Aberdeen, Écosse, juillet, 2005.

Denoual E., « Analogical Translation of Unknown Words in a Statistical Machine Translation Framework », *Machine Translation Summit, XI*, Copenhagen, Sept. 10-14, 2007.

Déjean H., Gaussier r., « Une nouvelle approche à l'extraction de lexiques bilingues à partir de corpus comparables », *Lexicometrica*, 2002. Numéro spécial : Alignement lexical dans les corpus multilingues.

Eck M., Hori C., « Overview of the IWSLT 2005 Evaluation Campaign », *2nd International Workshop on Spoken Language Translation (IWSLT'05)*, Pittsburgh, Pennsylvania, USA, p. 1-22, 2005.

Fordyce C. S., « Overview of the IWSLT 2007 Evaluation Campaign », *4th International Workshop on Spoken Language Translation (IWSLT'07)*, Trento, Italy, p. 1-12, 2007.

Freitag D., « Morphology Induction from Term Clusters », *Proc. of the 9th CoNLL*, Ann Arbor, Michigan, USA, p. 128-135, 2005.

Fung P., « A pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora », *33rd ACL*, p. 236-243, 1995.

Fung P., Yee L. Y., « An IR Approach for Translating New Words from Nonparallel, Comparable Texts », *Proc. of the 36th ACL*, San Francisco, California, p. 414-420, 1998.

Gentner D., Holyoak K. J., Konikov B. N., *The Analogical Mind*, MIT Press, Cambridge, MA, 2001.

Goldwater S., McClosky D., « Improving statistical MT through morphological analysis », *Proc.of HLT-EMNLP*, Vancouver, British Columbia, Canada, p. 676-683, 2005.

Hathout N., « Analogies morpho-synonymiques. Une méthode d'acquisition automatique de liens morphologiques à partir d'un dictionnaire de synonymes », *8e conférence TALN*, Tours, France, 2001.

Koehn P., Knight K., « Learning a Translation Lexicon from Monolingual Corpora », *Proc. of the ACL Workshop on Unsupervised Lexical Acquisition*, Philadelphia, Pennsylvania, USA, p. 9-16, 2002.

Koehn P., Monz C., « Manual and Automatic Evaluation of Machine Translation between European Languages », *Proceedings on the Workshop on Statistical Machine Translation*, Association for Computational Linguistics, New York City, p. 102-121, June, 2006.

Langlais P., Patry A., « Enrichissement d'un lexique bilingue par analogie », *14e conférence sur le Traitement Automatique des Langues Naturelles (TALN'07)*, Toulouse, France, p. 101-110, June, 2007a.

Langlais P., Patry A., « Translating Unknown Words by Analogical Learning », *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic, p. 877-886, June, 2007b.

Langlais P., Yvon F., Algorithms for Analogical Learning with Formal Analogies, Technical report, ENST, Paris, 2008.

Langlais P., Yvon F., Zweigenbaum P., An Analogical Learning Approach for Translating Terms, Technical report, ENST, Paris, 2008.

Lee Y.-S., « Morphological Analysis for Statistical Machine Translation », *Proc. of HLT-NAACL*, Boston, Massachusetts, USA, 2004.

Lepage Y., « Solving Analogies on Words : an Algorithm », *COLING-ACL*, p. 728-734, Montreal, Canada, 1998.

Lepage Y., « De l'analogie rendant compte de la commutation en linguistique », mai, 2003, Mémoire d'habilitation à diriger des recherches, Université Joseph Fourier, Grenoble I.

Lepage Y., Denoual E., « ALEPH : an EBMT system based on the preservation of proportionnal analogies between sentences across languages », *International Workshop on Statistical Language Translation (IWSLT)*, Pittsburgh, PA, oct., 2005.

Lepage Y., Denoual E., « Purest ever example-based machine translation : Detailed presentation and assessment », *Machine Translation*, vol. 19(3), p. 251-282, Dec., 2006.

Lepage Y., Lardilleux A., « The GREYC Machine Translation System for the IWSLT 2007 Evaluation Campaign », *4th International Workshop on Spoken Language Translation (IWSLT'07)*, Trento, Italy, p. 49-54, 2007.

Levenshtein V. I., « Binary codes capable of correcting deletions, insertions and reversals », *Sov. Phys. Dokl.*, vol. 6, p. 707-710, 1966.

Moreau F., Claveau V., « Extensions de requêtes par relations morpho-syntaxiques apprises automatiquement », *3e Conférence en Recherche d'Informations et Applications (CORIA'06)*, Lyon, France, mars, 2006.

Morin E., Daille B., « Extraction de terminologies bilingues à partir de corpus comparables d'un domaine spécialisé », *TAL*, vol. 45(3), p. 103-122, 2004.

Morin E., Daille B., Takeuchi K., Kageura K., « Bilingual Terminology Mining - Using Brain, not brawn comparable corpora », *45th ACL*, Prague, Czech Republic, p. 664-671, June, 2007.

Nießen S., Improving Statistical Machine Translation using Morpho-syntactic Information, PhD thesis, RWTH, Aachen, Germany, 2002.

Och F., Ney H., « Improved Statistical Alignment Models », *38th annual meeting of the Association for Computational Linguistics (ACL'00)*, Hongkong, China, p. 440-447, 2000.

Papineni K., Roukos S., Ward T., Zhu W.-J., « BLEU : a Method for Automatic Evaluation of Machine Translation », *Proc. of the 40th ACL*, Philadelphia, Pennsylvania, USA, p. 311-318, 2002.

Patry A., Gotti F., Langlais P., « Mood at work : Ramses versus Pharaoh », *Proc. of the HLT-NAACL Workshop on Statistical Machine Translation*, New York City, USA, p. 126-129, 2006.

Paul M., « Overview of the IWSLT 2007 Evaluation Campaign », *3rd International Workshop on Spoken Language Translation (IWSLT'07)*, Kyoto, Japan, p. 1-15, 2006.

Pirrelli V., Yvon F., « The hidden dimension : a paradigmatic view of data-driven NLP », *Journal of Experimental & Theroretical Artifical Intelligence*, vol. 11, p. 391-408, 1999.

Popovic M., Ney H., « Towards the Use of Word Stems and Suffixes for Statistical Machine Translation », *Proc. of the 4th LREC*, Lisbon, Portugal, p. 1585-1588, 2004.

Rapp R., « Automatic Identification of Word Translations from Unrelated English and German Corpora », *37th annual meeting of the Association for Computational Linguistics (ACL'99)*, College Park, Maryland, p. 519-526, 1999.

Stroppa N., Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles, PhD thesis, École nationale supérieure des télécommunications, Paris, France, nov., 2005.

Stroppa N., Yvon F., « An analogical learner for morphological analysis », *9th Conf. on Computational Natural Language Learning (CoNLL)*, Ann Arbor, MI, p. 120-127, June, 2005.

Takaaki T., Matsuo Y., « Extraction of Translation Equivalents from Non-Parallel Corpora », *Proc. of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'99)*, Chester, England, p. 109-119, 1999.

Turney P. D., « Similarity of Semantic Relations », *Computational Linguistics*, vol. 32, n° 3, p. 379-416, Sept., 2006.

Turney P., Littman M., « Corpus-based Learning of Analogies and Semantic Relations », *Machine Learning Journal*, vol. 60, n° 1-3, p. 251-278, sep., 2005.

Ukkonen E., « Algorithms for approximate string matching », *Information and Control*, vol. 64, p. 100-118, 1985.

Wagner R. A., Fisher M. J., « The String-to-String Correction Problem », *Journal of ACM*, vol. 21(1), p. 168-173, 1974.

Yvon F., « Paradigmatic cascades : a lingusitically sound model of pronunciation by analogy », *35th Annual Meeting fo the ACL*, Madrid, Spain, July, 1997.

Yvon F., Stroppa N., Delhay A., Miclet L., Solving analogical equations on words, Technical report, École Nationale Supérieure des Télécommunications, Paris, France, Jul., 2004.