Décompte des vraies analogies entre énoncés et entre *chunks* sur un corpus de plusieurs dizaines de milliers d'énoncés

Yves Lepage

GREYC, Université de Caen Basse-Normandie
Tél. : 02-31-56-74-82, Fax : 02-31-25-73-30
Yves.Lepage@info.unicaen.fr

27 novembre 2007

1 Introduction et motivation

Le travail rapporté ici trouve son origine dans la conception et la réalisation d'un système de traduction automatique de phrases courtes qui repose sur l'utilisation de la structure des langues et vise à traiter les divergences entre langues lors de la traduction. Ce système, baptisé ALEPH,[1] a été décrit dans [Lepage and Denoual2005] et [Lepage and Lardilleux2007]. C'est un système par l'exemple composé d'un moteur exploitant des données consistant en phrases exemples alignées dans les deux langues, source et cible. Les figures 1 et 2 illustre en quatre étapes le processus fondamental de traduction dans ce système.

Le fonctionnement du système est comme suit. Une phrase est proposée au système à la traduction, par exemple :

紅茶一杯下さい。

Dans un premier temps, la phrase à traduire est placée au sein de la structure formée par l'ensemble des phrases exemples. Pour illustrer cela, sur la figure 1,

[1] Acronyme de Analogie en Langues Et Procédés par Homomorphisme.

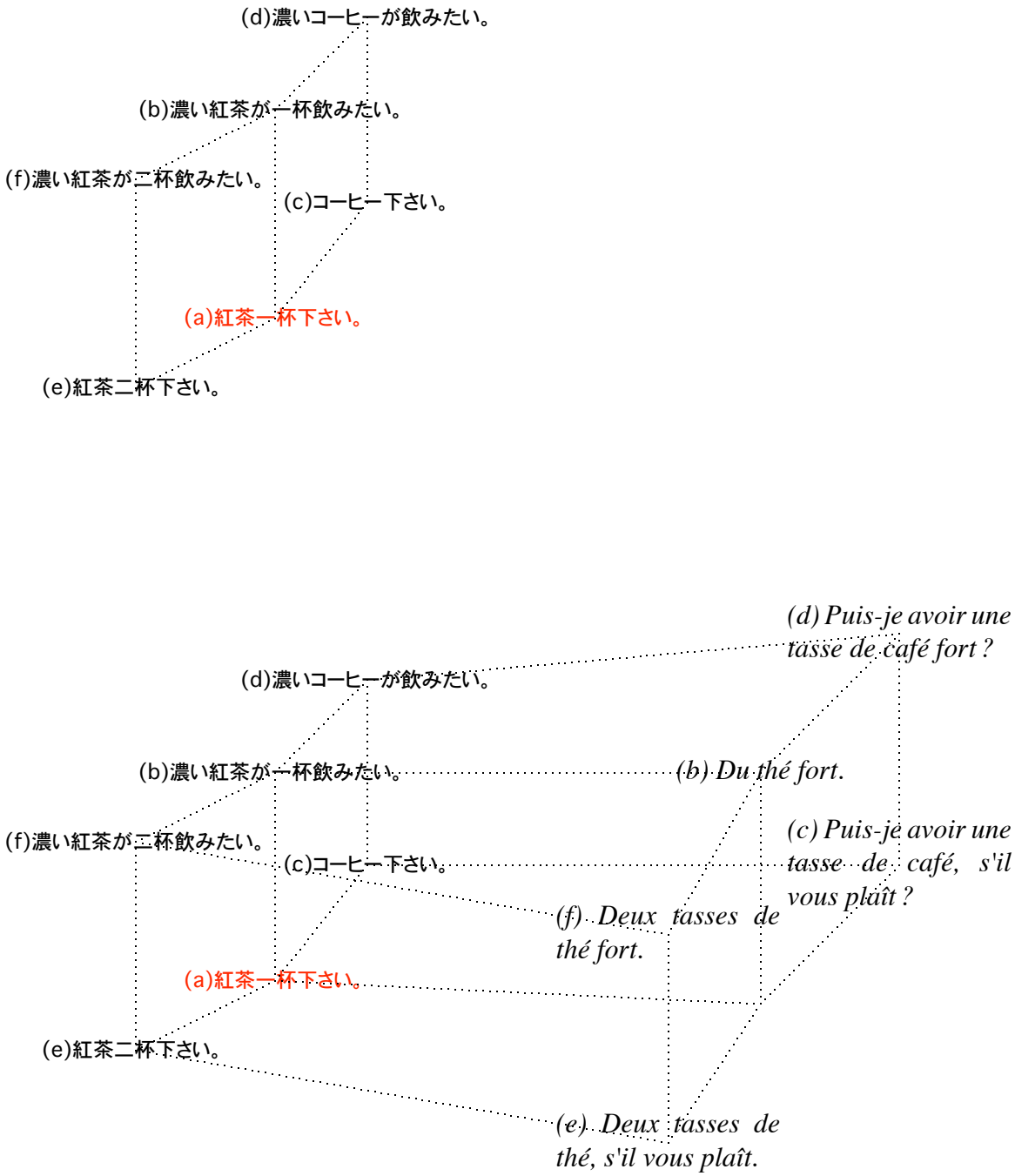

Fig. 1 – Première et deuxième étapes du processus de traduction : le système place la phrase à traduire dans la structure des exemples en langue source ; il peut alors déterminer les phrases cibles correspondant aux phrases entourant la phrase à traduire.

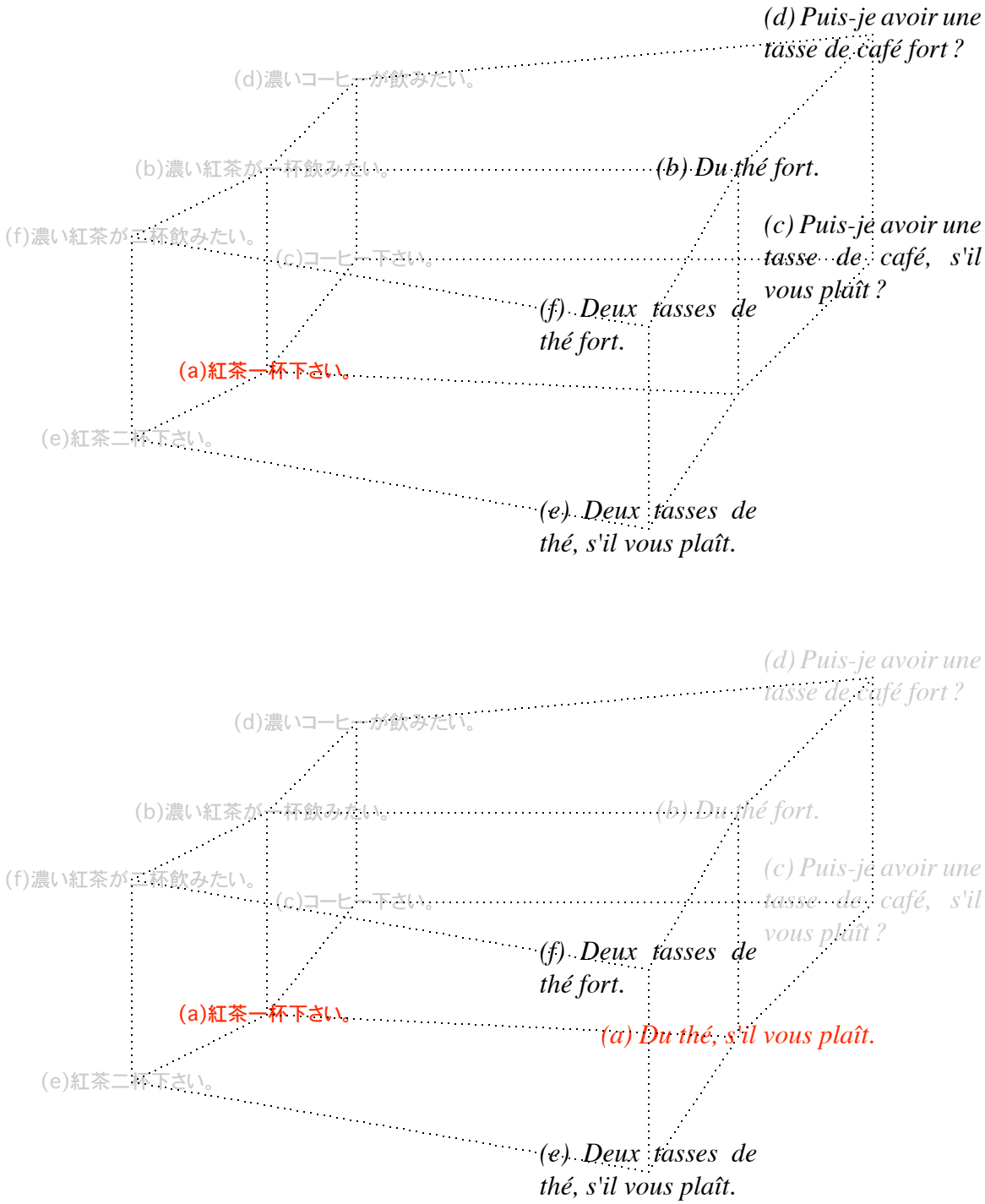

Fig. 2 – Troisième et quatrième étapes de traduction : le système essaie par calculs de combler le trou dans la structure en langue cible ; s'il y parvient, il propose la phrase obtenue comme traduction de la phrase source à traduire.

deux parallélogrammes enserrent la phrase à traduire. Dans un deuxième temps, le système détermine les phrases exemples en langue cible correspondant aux phrases en langue source précédentes. Cette opération laisse apparaître un trou dans la structure de la langue cible. Dans un troisième temps, le système essaie, par calculs, de combler ce trou par une phrase. S'il y parvient, il peut dans un quatrième temps proposer cette phrase comme résultat de traduction. Ici donc,

Du thé, s'il vous plaît.

est proposé comme traduction de

紅茶一杯下さい。

En résumé le processus de traduction repose sur l'idée :

- qu'il existe une cohérence monolingue intrinsèque qui doit être mise en évidence par des opérations fondamentales et universelles d'ordre symbolique (et peut-être pas forcément d'ordre strictement linguistique) ;
- que l'hypothèse de correspondance entre langues passe nécessairement par la mise en correspondance des structures de la langue.

Les structures dont il est question ici ne sont pas des arbres d'analyse de phrases. Il s'agit de la structure des langues au sens où les phrases s'y enchâssent. La structure n'est donc pas à chercher dans les phrases, mais dans le rapport entre les phrases. Ce qui permet la traduction, c'est le fait que chaque énoncé de langue est au cœur d'un réseau qui le situe dans le système de sa langue. Notre système de traduction est donc un système structuraliste. Cela implique que la méthode n'impose pas de savoir à quel mot correspond quel mot pour effectuer une traduction, ni même de savoir explicitement à quelle structure correspond quelle structure. Par exemple, dans la traduction de la phrase espagnole :

Atravesó el río flotando.

en anglais :

It floated across the river.

on peut dire que le mot *Atravesó* est traduit par *across* (un verbe traduit par une préposition). Mais cela est une vision incomplète. Pour être complet, il faut dire que ce même mot *Atravesó* est en fait traduit non seulement par l'adverbe *across*, mais aussi par le pronom personnel *It* (le *ó* de *Atravesó* indique une troisième personne singulier) et par la terminaison verbale *-ed* (le passé est aussi exprimé par le *ó* de *Atravesó*). Par opposition, dans la méthode proposée, le détail de

cette correspondance entre les deux phrases (au niveau interne) reste implicite car il résulte de l'ensemble des relations analogiques monolingues (externes aux phrases) dans lesquelles chacune de ces deux phrases intervient. Il est à souligner que le calcul des ces relations est effectué entièrement au niveau des caractères dans notre système, sans aucune segmentation en mots. Cela constitue un avantage pour le traitement des langues ne marquant pas la segmentation en mots comme le chinois ou le japonais.

2 L'analogie proportionnelle

2.1 Cadre linguistique

Les structures sur lesquelles notre système travaille sont les parallélogrammes visibles dans les figures 1 et 2. Il s'agit de relations entre quatre phrases établissant qu'une première phrase est à une deuxième ce qu'une troisième est à une quatrième. Par exemple, on peut dire que

Du thé, s'il vous plaît.

est à

Du thé fort.

ce que

Deux tasses de thé, s'il vous plaît.

est à

Deux tasses de thé fort.

ce type de relation entre quatre objets A, B, C et D se note habituellement $A : B :: C : D$ et constitue ce que l'on appelle des analogies proportionnelles.

Les analogies existent entre mots, comme dans l'exemple latin tiré de [de Saussure1995, p. 221-222] :

oratorem : *orator* :: *honorem* : *honor*

Cet exemple est celui par lequel Saussure montre que l'analogie proportionnelle est une opération synchronique qui explique la mise en concurrence de deux formes à une même époque. Dans cet exemple, une première forme héritée d'un état antérieur de la langue, *honos*, et une deuxième forme produite par analogie, *honor*. L'analogie proportionnelle succède aux changements phonétiques, en l'occurrence le rhotacisme qui avait provoqué le changement de *honosem* en *honorem*[2] et elle rétablit la régularité troublée par le changement phonétique. Ici, la régularité :

oratorem : *orator* :: *honosem* : *honos*

détruite par le changement de *honosem* en *honorem*, est rétablie par :[3]

$$\textit{oratorem} : \textit{orator} :: \textit{honorem} : x \quad \Rightarrow \quad x = \textit{honor}$$

Cet exemple latin introduit la problématique que nous nous proposons d'étudier : celle du jeu entre la forme et le sens dans le cadre de l'analogie proportionnelle. On peut en effet distinguer les analogies selon qu'elles sont de forme ou de sens.

Voici premièrement un exemple d'analogie de sens mais pas de forme :

nageoires : *poisson* :: *ailes* : *oiseau*

En conjugaison, de telles analogies sont appelées anomalies selon la tradition grammaticale latine :

manger : *je mangerai* :: *aller* : *j'irai*

Voici deuxièmement un exemple d'analogie de forme qui n'est pas une analogie de sens :

plume : *porte-plume* :: *fenêtre* : *porte-fenêtre*

Les analogies de ce genre sont appelées fausses analogies.

Troisièmement, les seules analogies vraies sont celles pour lesquelles la forme et le sens sont respectés comme dans les exemples de conjugaison appartenant à un même paradigme :

donner : *ils donneront* :: *marcher* : *ils marcheront*

[2]Le rhotacisme est le changement d'un /s/ intervocalique en /R/.

[3]Les Anglo-saxons appellent loi de Sturtevant le fait que les lois phonétiques, d'application régulière, produisent en définitive du désordre, mais que l'analogie, d'application irrégulière, rétablirait l'ordre.

Hermann Paul constate que c'est le cadre des vraies analogies qui permet la production, à la volée, des formes de conjugaison ou de déclinaison [Paul1920, p. 110] :

> On construit les formes supplémentaires dans l'instant où on en a besoin d'après les paradigmes, c'est-à-dire, par analogie.

Pour ce qui est de l'application de l'analogie hors de la morphologie, il semble lui revenir d'avoir vu le premier son application au-delà des mots, puisqu'on peut lire sous sa plume :

> Un grand nombre de formes morphologiques et de *relations syntaxiques* [c'est nous qui soulignons] est obtenu à l'aide des groupes de proportions sans que le locuteur ait jamais le sentiment de quitter le terrain sûr de l'acquis.

Léonard Bloomfield étend cette application aux énoncés complexes en écrivant [Bloomfield1933, p. 276] :

> Lorsqu'un locuteur prononce un énoncé complexe, nous sommes incapables de dire s'il l'a déjà entendu ou s'il vient de le créer par analogie.

et il illustre son propos par l'exemple suivant (que nous traduisons en français) :

> *Bébé a faim* : *Annie a faim*
> *pauvre Bébé* : *pauvre Annie* :: *Donne une orange à Bébé* : x
> *L'orange de Bébé* : *L'orange d'Annie*

2.2 L'argument de la pauvreté du stimulus

L'analogie a cependant mauvaise réputation en syntaxe à cause du fameux argument, controversé, de la pauvreté du stimulus.[4] Cet argument est apporté en soutien à l'idée qu'il n'y aurait aucune procédure d'induction à l'œuvre lors de l'apprentissage de la langue :

- parce que les enfants produisent des énoncés qu'ils n'ont jamais entendus auparavant ;
- parce qu'ils ne produiraient jamais certains énoncés agrammaticaux.

Un tel exemple d'énoncés est le cas d'*auxiliary fronting* en anglais. Des enfants pourraient produire des énoncés comme :

Is the student who is in the garden hungry ?

[4]La volume 9 datant de 2002 de la revue *Lingusitic Review* contient quatre articles en forme de réponses les uns aux autres d'adversaires et de défenseurs de l'argument de la pauvreté du stimulus.

dont ils n'auraient jamais entendu auparavant la structure. En plus, ils ne produiraient jamais des énoncés du type :

$$^{*}\textit{Is the student who in the garden is hungry ?}$$

Pourtant, formellement, ces deux énoncés sont les deux solutions possibles de l'équation analogique suivante :

$$\textit{The student in the garden is hungry.} : \textit{Is the student in the garden hungry ?} :: \textit{The student who is in the garden is hungry.} : x$$

dont ils peuvent fort bien avoir entendu auparavant les trois termes. L'argument induirait donc à conclure en faveur du caractère inné de la grammaire pour lequel Chomsky donne un autre exemple, basé, lui aussi, sur l'utilisation d'une analogie. Cette analogie est une fausse analogie, c'est-à-dire une analogie de forme qui n'est pas une analogie de sens.[5]

$$\textit{Abby is baking vegan pies.} : \textit{Abby is baking.} :: \textit{Abby is too tasteful to pour gravy on vegan pies.} : \textit{Abby is too tasteful to pour gravy on.}$$

Contre l'argument de la pauvreté du stimulus, [Pullum and Scholtz2002] montre que la structure d'*auxiliary fronting* apparaît en fait dans des livres pour enfants et même dans le corpus CHILDES. Les auteurs effectuent même des comptages que contestent [Legate and Yang2002], ce qui laisse le débat ouvert. Cependant, [Pullum1999] admet que les analogies proportionnelles ne sauraient franchir certaines frontières syntaxiques. L'exemple anglais cité par l'auteur est le suivant, qui ne fait pas sens :

$$\begin{array}{rcl} \textit{white skirt} & : & \textit{green blouse} \\ & :: & \\ \textit{Often commentators who are white skirt the problem of institutional racism.} & : & \textit{*Often commentators who are green blouse the problem of institutional racism.} \end{array}$$

[5]Noam Chomsky, Conference at the university of Michigan, *1998, A report by Aaron Stark*. Dans la troisième phrase, la sauce est mise sur les petits pâtés végétariens, alors qu'elle l'est sur Abby dans la quatrième phrase. La différence de structure entre la troisième et la quatrième phrases n'est donc pas celle qui existe entre la première et la deuxième phrases.

2.3 Décompte des vraies analogies

Nous nous posons donc la question de savoir dans quelle mesure les analogies de forme peuvent être des analogies de sens, c'est-à-dire dans quelle mesure les analogies de forme sont des vraies analogies. Le résultat que nous établissons est que, sur le corpus que nous avons étudié, les analogies de forme sont dans plus de 95% des cas des vraies analogies pour deux "morceaux de langue" ne franchissant pas de frontières syntaxiques : les énoncés (en fait, des phrases) et les *chunks*.

3 Expériences et résultats

3.1 Le corpus utilisé

Nous travaillons avec le BTEC, *Basic Traveller's Expression Corpus*, qui est un corpus multilingue de 160 000 énoncés dans le domaine du tourisme. Ce corpus a été développé dans le cadre du consortium C-STAR pour la recherche en traduction automatique de la parole. Les énoncés que l'on y trouve sont isolés de leur contexte, relativement courts et beaucoup sont en fait semblables.

La figure **??** donne quelques exemples d'énoncés contenus de ce corpus. Les tableaux 1 et 2 donnent des statistiques sur la caractérisation des situations d'énonciation et sur le nombre de mots ou d'énoncés.

Dans le cadre d'une première expérience sur les énoncés, nous avons utilisé la globalité du BTEC en trois langues, le chinois, l'anglais et le japonais. Dans le cadre d'une seconde expérience de mesure sur les *chunks*, nous avons utilisé seulement un sous-ensemble de 20 000 énoncés qui correspondent aux données de la campagne d'évaluation IWSLT 2005 [Eck and Hori2005]. Ces 20 000 énoncés constituent en fait un échantillonnage des 160 000 énoncés du BTEC.

How about at one p.m. ?	午後一時はどうですか。
What does that mean ?	あれはどういう意味ですか。
Where's the nearest underground station ?	もよりの地下鉄の駅 はどこにありますか。
Uh-huh.	なるほど。
Where can I take a taxi ?	タクシ乗り場どこですか。

Fig. 3 – Quelques exemples d'énoncés en anglais et en japonais contenus dans le BTEC.

Tab. 1 – Proportion des différentes situations d'énonciation dans le BTEC.

tourisme	7,7%	dans l'avion	3,6%
au restaurant	7,3%	communication	6,4%
parler affaires	5,3%	à l'aéroport	5,5%
prendre contact	4,0%	chez quelqu'un	2,3%
etc.	...		

Tab. 2 – Statistiques sur les énoncés et leur longueur en mots ou en caractères dans le BTEC.

Langue	Enoncés différents	Mots / énoncé	Caractères / énoncé
anglais	97 769	$7{,}94 \pm 3{,}86$	$16{,}21 \pm 7{,}84$
japonais	103 274	$5{,}85 \pm 3{,}11$	$35{,}14 \pm 18{,}81$
chinois	96 234	$11{,}00 \pm 5{,}77$	non calculé

3.2 Décompte des analogies entre énoncés

Avec un Pentium 4 à 2,8 GHz et 2 gigaoctets de mémoire, il nous a fallu quelques dix jours pour récupérer l'ensemble des analogies de forme contenues dans notre corpus d'environ 100 000 phrases. Ce décompte se fait en se basant sur la définition de l'analogie donnée dans [Lepage2004] qui fait intervenir des calculs de distance d'édition entre chaînes de caractères et une propriété sur les caractères. Les exemples de la figure 4 sont des exemples d'analogies obtenues. Le tableau 3 donne les chiffres obtenus pour chacune des langues. Pour l'anglais, nous trouvons plus de deux millions d'analogies de forme impliquant plus de 50 000 phrases. Autrement dit, la moitié des phrases du corpus sont en analogie immédiate avec d'autres phrases du corpus. Sur ces phrases uniquement, on obtient une moyenne de 35 analogies par phrase, mais il faut souligner que la distribution n'est pas gaussienne.

Parmi les analogies de forme, certaines peuvent ne pas être des analogies en sens. C'est la cas de la dernière analogie donnée dans la figure 4.

Une estimation par échantillonnage confirme l'impression que presque toutes les analogies de forme sont en fait des analogies de sens. Aussi bien en anglais

I need it : *I don't need it.* :: *You look well.* : *You don't look well.*

There is no towel. : *There are no towels.* :: *How much is the best seat?* : *How much are the best seats?*

It's long. : *Do you have one that's a little bit longer?* :: *It's small.* : *Do you have one that's a little bit smaller?*

Can you cash this traveler's check? : *Could you cash this traveler's check?* :: *Can you cash some traveler's checks?* : *Could you cash some traveler's checks?*

He had a gun. : *She had a gun.* :: *Hit.* : *Shit.*

Fig. 4 – Exemples d'analogies obtenues en anglais.

Tab. 3 – Nombre d'analogies de forme entre énoncés.

Langue	nombre d'analogies de forme	nombre d'énoncés impliqués	nombre d'analogies en moyenne par énoncés impliqués
anglais	2 384 202	53 250	44,77
japonais	1 910 062	53 572	35,65
chinois	1 639 068	49 675	33,00
Intersection par traduction			
anglais ∩ chinois	238 135	25 554	9,32
anglais ∩ japonais	336 287	24 674	13,63
japonais ∩ chinois	329 429	25 527	13,11
anglais ∩ japonais ∩ chinois	68 164	13 602	5,01
Intersection par traduction après production d'énoncés par analogie			
anglais ∩ chinois	1 536 298	49 297	31,16
anglais ∩ japonais	1 910 689	50 536	37,63
japonais ∩ chinois	1 569 037	51 442	30,50
anglais ∩ japonais ∩ chinois	1 507 380	49 052	30,71

qu'en japonais, l'inspection de 666 analogies par des vérificateurs humains donnent après test statistique qu'au moins 96% des analogies de forme seraient des analogies de sens (seuil de rejet de 0,1%).

Dans une expérience tendant à augmenter ce score, nous nous sommes basés sur l'idée que la traduction préservait le sens et que, donc, si des énoncés en correspondance dans les trois langues donnaient lieu à des analogies dans les trois langues, alors, on devrait avoir une analogie de sens presque à coup sûr. Le tableau 3 donne le nombre d'analogies dans chaque langue qui correspondent à une analogie dans une autre langue, ainsi que l'intersection de telles analogies dans les trois langues. Le nombre de telles analogies diminue sérieusement pour tomber à presque 70 000. Mais le même test statistique que précédemment montre que plus de 99% de telles analogies sont des analogies de sens (seuil de rejet de 0,1%). Un exemple de fausse analogie trouvée lors de ce test est :

Could you tell me how to fill this from? : *Could you tell me how to fill this form?* :: *Where is the conference centre?* : *Where is the conference center?*

Dans une troisième expérience de comptage, nous avons essayé d'imposer la correspondance entre analogies au travers des langues en produisant les phrases lorsqu'elles n'existaient pas. Par exemple, le quadruplet d'énoncés suivants en anglais forme une analogie de forme :

I prefer French food.
I'd prefer an aisle seat.
I like French food.
I'd like an aisle seat.

Mais si l'on prend son correspondant en japonais :

フランス料理の方がいい。
通路側の方がいい。
フランス料理がいいんですが。
通路側がいいです。

il n'y a pas d'analogie de forme, mais il y a bien analogie de sens. En résolvant l'équation analogique suivante :

フランス料理の方がいい。：通路側の方がいい。:: フランス料理がいいんですが。：**x**

on force la production de l'énoncé 通路側がいいんですが。qui forme avec les trois énoncés précédents une analogie de forme et de sens, c'est-à-dire une vraie analogie. Nous avons effectué cette production d'énoncés nouveaux (certains peuvent être agrammaticaux ou mal formés). Les chiffres ont été portés dans le tableau 3. Nous avons de nouveau appliqué notre test statistique sur un échantillon de 666 énoncés en anglais et en japonais. Le résultat est que plus de 97% des analogies de forme seraient des analogies de sens (seuil de rejet de 0,1%). Un exemple d'analogie de forme qui n'a pas été jugé analogie de sens est :

$$\textit{Can I eat?} : \begin{matrix}\textit{Can I eat on}\\ \textit{the train?}\end{matrix} :: \begin{matrix}\textit{I'm afraid you're}\\ \textit{wrong.}\end{matrix} : \begin{matrix}\textit{I'm afraid you're on}\\ \textit{the wrong train.}\end{matrix}$$

3.3 Décompte des analogies entre *chunks*

Dans la perspective de l'application de notre système à la traduction sous-phrastique, et donc pour les mêmes raisons méthodologiques que pour les énoncés, nous avons entrepris la même estimation du nombre de vraies analogies entre chunks sur le corpus BTEC. Dans cette étude, nous nous limitons au japonais car il présente l'avantage d'être une langue à marqueurs de cas, c'est-à-dire que les groupes verbaux ou nominaux sont balisés par des marqueurs spéciaux généralement distingués dans l'écriture et qui apparaissent systématiquement en fin de groupes. Par exemple, dans la phrase suivante, les marquers ont été repérés dans des rectangles.

わたしの場合[、]たいてい仕事[で][、]めった[に]遊び[で][は]行き[ません][。]
/watasi no baai[,] taitei sigoto [de][,] metta [ni] asobi [de] [ha] iki[masen][.]/

Tab. 4 – Statistiques sur les données japonaises utilisées Pour compter le nombre de mots en japonais, nous avons utilisé ChaSen. Un *chunk* apparaît en moyenne 4,1 fois dans le corpus.

Type de données	Taille des données			Nombre moyen en mots
		en mots	en caractères	
Enoncés	20 000	173 091	339 579	8,7
Chunks	99 719	693 526	718 819	6,9

Nous avons listé a priori de tels marqueurs et avons segmenté le corpus à notre disposition afin que le nombre d'analogies de forme produites soit maximal. La méthode est la suivante : nous segmentons le texte avec chaque marqueur séparément et calculons, sur un échantillon, le nombre d'analogies de forme obtenues. Nous retenons alors le marqueur pour lequel ce nombre est maximal, découpons le texte avec ce marqueur, et réappliquons alors la méthode au texte découpé. A chaque étape, un nouveau marqueur est retenu, jsuqu'à ce que nous observions un plateau dans le progression du nombre d'analogies. La figure 5 donne la progression du nombre d'analogies avec, en regard, la liste des marqueurs retenus à chaque étape. la liste des marqueurs obtenus est la suivante :

- le point final ;
- la virgule ;
- la particule de génitif の /no/ ;
- la particule de lieu ou d'instrumental で /de/ ;
- la particule de direction へ /e/ ;
- la particule de lieu ou de datif に /ni/ ;
- la particule d'accusatif ou de deuxième actant を /wo/ ;
- la particule du thème ou du sujet が /ga/ ;
- la particule d'origine から /kara/ ;
- enfin, la terminaison verbale du passé ou d'accompli ました /-masita/.

Nous avons aussi observé que la croissance du nombre de *chunks* était à peu près linéaire en la taille du nombre d'énoncés, ce qui peut se résumer en disant qu'un énoncé est constitué en moyenne de cinq *chunks*. On calcule aussi qu'en moyenne, un *chunk* apparaît 4,1 fois dans le corpus.

De la même manière que pour les énoncés, nous avons récolté automatiquement toutes les analogies de forme entre *chunks*. Nous en dénombrons presque deux millions pour 40 000 chunks. Nous avons alors estimé par test statistique sur 500 *chunks* le nombre d'analogies de forme qui seraient des analogies de sens. Le

Tab. 5 – Estimation du nombre de vraies analogies entre *chunks* c'est-à-dire du nombre d'analogies de forme et de sens avec une hypothèse nulle de 96%.

Type de données		Nombre d'analogies de forme	Nombre d'analogies vraies	
			% observées	p-value
Enoncés	20 000	4 428	100%	n.p.
Chunks	99 719	2 131 269	96%	0,005

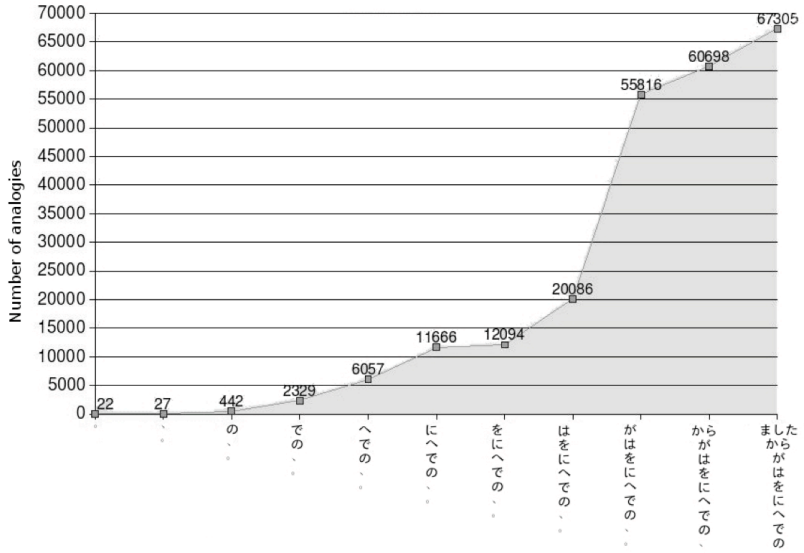

Fig. 5 – Progression du nombre d'analogies sur un échantillon de même taille de *chunks*. Les marqueurs retenus à chaque étape sont visibles en abscisse.

résultat est que au moins 96% des analogies de forme entre *chunks* constitueraient des analogies de sens (seuil de rejet de 0,05%).

Il est particulièrement intéressant d'estimer l'accroissement du nombre d'analogies entre énoncés et entre *chunks* extraits de ces mêmes énoncés. Nous avons pu déterminer que le nombre d'analogies entre *chunks* croît en la puissance 1,7 du nombre d'analogies entre énoncés dont sont extraits ces *chunks*. Cet exposant de 1,7 est la valeur donnant la meilleure corrélation de Pearson possible par rapport à tous les autres exposants possibles (entre 1,5 et 2,5). On peut donc conclure que cet accroissement est approximativement quadratique.

4 Conclusion

Nous avons décrit des expériences de comptage de vraies analogies, c'est-à-dire d'analogies de forme et de sens. Notre but était de savoir si, majoritairement, les analogies de forme, que l'on peut détecter automatiquement sont des analogies de sens ou non. Cette étude se justifie pour étayer l'utilisation de l'analogie proportionnelle dans le cadre d'un système de traduction automatique par l'exemple

dont nous avons rappelé le principe.

Sur un corpus d'environ 100 000 énoncés différents dans plusieurs langues, nous avons détecté environ deux millions d'analogies de forme dans chacune de langues. Ces analogies font intervenir presque 50 000 phrases, c'est-à-dire presque la moité du corpus. Des test statistiques ont permis d'estimer que plus de 96% des analogies de forme sont bien des analogies de sens. De telles chiffres peuvent être obtenus en prenant aussi bien une estimation basse par intersection entre le chinois, l'anglais et le japonais, qu'en prenant une estimation haute en imposant les analogies de forme par production automatique d'énoncés.

Sur un corpus plus petit de 20 000 énoncés en japonais, une expérience similaire de décompte de vraies analogies entre *chunks* japonais a permis de trouver la même proportion d'analogies de forme étant analogies de sens, c'est-à-dire plus de 96%. Par comparaison avec les analogies entre énoncés, le nombre d'analogies entre *chunks* croît de façon presque quadratique en le nombre d'analogies entre énoncés.

La conclusion de tout ce travail est donc que, contrairement à l'a priori négatif qui prévaut en syntaxe à l'encontre de l'analogie, l'analogie de forme est fiable dans l'immense majorité des cas : les analogies de forme sont le plus souvent des analogies de sens entre énoncés ou entre chunks.

Références

[Bloomfield1933] Leonard Bloomfield. 1933. *Language*. Holt, New York.

[de Saussure1995] Ferdinand de Saussure. 1995. *Cours de linguistique générale*. Payot, Lausanne et Paris. [1^{e} éd. 1916].

[Eck and Hori2005] Thomas Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 evaluation campaign. In Carnegie Mellon University, editor, *Proc. of the International Workshop on Spoken Language Translation*, pages 1--22.

[Legate and Yang2002] Julie Anne Legate and Charles D. Yang. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, 19 :151--162.

[Lepage and Denoual2005] Yves Lepage and Etienne Denoual. 2005. Purest ever example-based machine translation : detailed presentation and assessment. *Machine Translation Journal*, 19 :251--282.

[Lepage and Lardilleux2007] Yves Lepage and Adrien Lardilleux. 2007. The greycmachine translation system for the iwslt2007 evaluation campaign. In

Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2007), Trento.

[Lepage2004] Yves Lepage. 2004. Lower and higher estimates of the number of ``true analogies'' between sentences contained in a large multilingual corpus. In *Proceedings of COLING-2004*, volume 1, pages 736--742, Genève, August.

[Paul1920] Hermann Paul. 1920. *Prinzipien der Sprachgeschichte*. Niemayer, Tübingen. 5^e éd., [1^e éd. 1880].

[Pullum and Scholtz2002] Geoffrey K. Pullum and Barbara C. Scholtz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19 :9--50.

[Pullum1999] Geoffrey K. Pullum, 1999. *Generative grammar*, pages 340--343. The MIT Press, Cambridge.