*RECITAL 2009, Senlis, 24–26 juin 2009*

# Méta-moteur de traduction automatique : proposition d'une métrique pour le classement de traductions

Marion Potet
Laboratoire d'informatique de Grenoble, équipe GETALP
UJF - BP 53, 38041 Grenoble Cedex 9
Marion.Potet@imag.fr

**Résumé.** Compte tenu de l'essor du Web et du développement des documents multilingues, le besoin de traductions "à la volée" est devenu une évidence. Cet article présente un système qui propose, pour une phrase donnée, non pas une unique traduction, mais une liste de N hypothèses de traductions en faisant appel à plusieurs moteurs de traduction pré-existants. Neufs moteurs de traduction automatique gratuits et disponibles sur le Web ont été sélectionnés pour soumettre un texte à traduire et réceptionner sa traduction. Les traductions obtenues sont classées selon une métrique reposant sur l'utilisation d'un modèle de langage. Les expériences conduites ont montré que ce méta-moteur de traduction se révèle plus pertinent que l'utilisation d'un seul système de traduction.

**Abstract.** Considering the Web and multilingual documents development expansion, the need of fast translation has become an evidence. This paper presents a system that proposes, for a given sentence, a list of N translation hypotheses instead of a single translation, using several machine translation systems already existing. Nine free and available (on the Internet) automatic translation engines have been chosen to submit a text to be translated and to receive its translation. The translations obtained are evaluated individually with a language model adapted and a metric elaborated by us, and in this way classified by relevance order. The experiment have pointed out that this meta-translation engine is more useful than the use of one system for translation.

**Mots-cl** traduction automatique, web, modèle de langage, méta-moteur de traduction.

**Keywords:** automatic translation, web, language model, meta-translator.

# 1 Introduction

Le première approche utilisée pour traduire des textes était basée sur des règles linguistiques qui visent à formaliser toutes les connaissances nécessaires à la traduction. Celle-ci nécessite beaucoup de travail de la part des linguistes pour définir le vocabulaire et la grammaire. Cette méthode a donné naissance, par exemple, au célèbre système de traduction Systran. D'autres méthodes existent, comme les méthodes empiriques et parmi elles, l'approche statistique, dont les bases théoriques ont été posées par (Brown *et al.*, 1993) puis (Koehn *et al.*, 2003), qui fait en sorte que toute la connaissance translingue soit acquise de manière automatique à partir de corpus.

Plusieurs travaux se sont intéressés à la comparaison des performances des systèmes issus des différentes approches. Il en ressort que, pour des performances équivalentes, les approches expertes à base de règle et les approches empiriques font des erreurs différentes lors de la traduction. Dans (Dugast *et al.*, 2008), par exemple, on remarque que les oublis de mots et la génération de mots inconnus sont des types d'erreurs spécifiques aux systèmes statistiques alors que les erreurs dans l'ordre des mots, tout comme les erreurs de vocabulaire, sont spécifiques aux systèmes à base de règles. D'autre part, (Rayner & Bouillon, 1995) précisent que les règles semblent encoder les informations grammaticales alors que les statistiques encodent les informations liées au domaine. Par ailleurs, la combinaison d'un système basé sur les règles et d'un module statistique, par exemple au moyen de la post-édition comme dans (Simard *et al.*, 2007) et (Koehn *et al.*, 2007), montre une amélioration significative de la qualité de traduction.

Cet article propose de combiner les résultats de plusieurs moteurs de traduction automatique issus de systèmes de natures différentes. Son originalité est qu'il fait appel à différents moteurs de traduction automatique du Web afin d'obtenir plusieurs traductions qui sont ensuite classées par ordre de pertinence à l'aide d'une métrique basée sur la fluidité de la traduction. L'idée est de tirer parti de cette variabilité inter-systèmes en regroupant, pour une même phrase, les résultats obtenus des différents moteurs afin de les mettre en concurrence. Le système utilise plusieurs moteurs de traduction simultanément ; par définition ; nous pouvons lui attribuer l'appellation de *méta-moteur* de traduction.

Dans un premier temps, des interfaces de traduction sont référencées, sélectionnées puis testées. Il en résulte une liste de N moteurs de traduction qui sont ensuite interrogés par programme afin d'obtenir N traductions d'une phrase source initiale.

Une fois les résultats des différents moteurs de traduction obtenus, il s'agit de scorer chaque traduction, en vue de les classer. Pour une phrase source, il en résulte une liste ordonnée d'hypothèses de traduction. La section 3, décrit la métrique utilisée pour classer ces traductions. La démarche est validée par une évaluation automatique (score BLEU) et par une évaluation subjective dont les résultats sont décrits dans la section suivante. Enfin, le système a été implémenté à travers une interface graphique, accessible sur le Web, présentée dans la section 5.

# 2 Récupération de traductions

Le cadre applicatif est restreint au domaine des informations journalistiques ou dépêches, plus communément appelé "news", en vue de valoriser ce travail par des d'applications directes comme la traduction automatique d'articles journalistiques de la langue anglaise vers la langue

française.

Une sélection manuelle des moteurs de traduction a été réalisée sur la base des caractéristiques suivantes : disponible sur le net ; gratuit ; permettre la traduction de phrases ou de textes ; traiter la traduction de la langue anglaise vers la langue française ; ne pas pratiquer de "blacklistage"[1]. Peuvent également être prises en compte la performance des résultats, la rapidité de réponse et une bonne utilisabilité. Selon ces critères, 22 interfaces ont ainsi été répertoriées.

Lors de l'analyse et des tests des systèmes de traduction présents sur le web, plusieurs interfaces différentes présentaient systématiquement, pour une même phrase à traduire, des résultats identiques. Ces interfaces ont été regroupées et neuf moteurs de traduction distincts ont été identifiés. Comme plusieurs interfaces utilisent le même outil traductif, le choix se porte sur celles qui présentent un temps de réponse court et une bonne utilisabilité (pas d'identification auprès du serveur, méthode d'envoi du formulaire, etc). Par soucis d'anonymat, dans la suite de l'article, les neuf interfaces finalement retenues seront désignées par : $MT_G$, $MT_S$, $MT_R$, $MT_A$, $MT_E$, $MT_F$, $MT_L$, $MT_P$, $MT_W$.

# 3 Classement des traductions

## 3.1 Description des corpus

Etant donné le cadre applicatif, les corpus doivent nécessairement comporter des données journalistiques de type "news" et ils doivent atteindre une taille suffisante pour permettre des traitements statistiques fiables. Etant donné le peu de ressources disponibles, dans le domaine spécifique des données journalistiques, nous avons, pour le besoin, constitué nos propres corpus à partir de textes écrits collectés sur le World Wide Web. Nous avons constitué, d'une part, un corpus monolingue en français pour le modèle de langage sur lequel d'appuie la métrique ainsi qu'un corpus bilingue français/anglais pour tester le système.

L'apprentissage du modèle de langage, utilisé pour classer les traductions de chaque moteurs, nécessite un ou des corpus représentatifs des conditions d'utilisation et de l'application envisagée : des corpus monolingues français de données contemporaines et journalistiques et de taille suffisante pour une estimation fiable des probabilités.

Trois modèles de langage ont été entraînés à partir des corpus d'apprentissage ( décrits dans le tableau 1). Ils sont ensuite combinés par interpolation linéaire. Le corpus de développement nécessaire à l'estimation des pondérations de chaque modèle de langage, est composé de données journalistiques du site *France24*[2] du 1er au 14 mai 2008.

Par ailleurs, l'évaluation des systèmes nécessite un corpus de test, bilingue et aligné, de phrases issues du domaine journalistique qui vont servir de référence pour mesurer la qualité des traducteurs. Pour cela, nous avons donc constitué un corpus bilingue à partir des versions anglaises et françaises du site Web de $France24$. Les phrases alignées ont ensuite été sélectionnées et extraites manuellement, directement d'après la mise en correspondance automatique des do-

[1] Le nombre de requêtes permises pour une même machine sur un moteur de traduction est suceptible d'être limité ou contrôlé par le serveur de certains moteurs de traduction

[2] http ://www.france24.com

[5] Association Européenne pour les ressources linguistiques.

| Source | Description | Nombre de mots | Période |
|---|---|---|---|
| France24 | www.France24.com | 4 M | février/avril 2008 |
| Web | données du web | 72 M | juin 2003/avril 2008 |
| Le Monde | CDRom de ELRA[5] | 23 M | janvier/décembre 2003 |

TAB. 1 – Description des corpus d'apprentissage du modèle de langage

cuments. Il est à noter que les textes relataient le même événement sans être pour autant la traduction l'un de l'autre. Les phrases traduites prélevées ont donc dû être vérifiées et, au besoin, corrigées une par une. Le résultat est un corpus de 300 phrases, bilingue et aligné de données journalistiques récentes.

## 3.2 Description du modèle de langage

Avant tout traitement, les corpus d'apprentissage ont été normalisés avec la boîte à outils *CLIPS-Text-Toolkit-2.5* (Bigi & Le, 2008). Les modèles de langage ont été entraînés à partir des corpus à l'aide de l'outil SRILM (Stolcke, 2002). Les modèles utilisent des trigrammes de mots, ils ont étés lissés avec la méthode de repli de Kneser-Ney modifiée et un pruning à $10^{-9}$ leur été appliqué.

Nous avons donc créé trois modèles de langage que nous nommerons, en rappel à leur corpus d'apprentissage respectif (vus dans la table 1), *LeMonde*, *Web* et *France24*. Le modèle de langage final est le résultat de l'interpolation linéaire du modèle *France24* avec un coefficient de 0,41, du modèle *Web* avec un coefficient de 0,42 et du modèle *LeMonde* avec un coefficient de 0,17. Son vocabulaire est de 60 147 mots et le taux de mots inconnus est de 1,36 % dans le corpus de test.

Nous avons choisi un apprentissage à vocabulaire fermé, un vocabulaire ouvert étant, à notre avis, inadapté dans le cas de données journalistiques fortement dépendante de l'actualité car le nombre de mots inconnus est très important et la distribution des probabilités en est fortement affectée.

## 3.3 Définition d'une métrique pour classer les traductions

Un modèle de langage analyse une phrase et lui applique plusieurs métriques. Celles qui nous interressent sont : le nombre de mots de la phrase, le nombre de mots de la phrase ne figurant pas dans le modèle de langage ou mots inconnus (notés $OOVs$[3]) et le logarithme de la probabilité de la phrase $W^K = w_1 w_2 \cdots w_K$, avec $w_k$ les $K$ mots de la phrase. Cette dernière notée $logprob$ s'estime comme suit :

$$logprob(W^K) = \sum_{k=2}^{K} log\ P(w_k|w_{k-1}, w_{k-2})$$

avec $P(w_k|w_{k-1}, w_{k-2})$ la probabilité du trigramme $w_{k-2}, w_{k-1}, w_k$.

[3]OOV= Out Of Vocabulary.

En pratique, l'outil SRILM possède une commande $ngram$ qui, à partir d'un modèle de langage, applique à toute phrase les métriques précédemment citées. Cette commande appliquée à la phrase *"je fais un essai"* donnera le résultat suivant :

```
<s> je fais un essai </s>
1 sentence, 4 words, 0 00Vs, logprob= -10.1085
```

Dans notre application, la métrique recherchée sera utilisée pour comparer les hypothèses de traduction issues des différents moteurs en ligne.

Pour l'attribution d'un score aux différentes traductions, il est inutile d'introduire un paramètre destiné à normaliser la longueur des phrases car les phrases, étant la traduction d'une seule et même phrase source, ont nécessairement approximativement le même nombre de mots. Le critère $logprob$, bien que fortement dépendant du nombre de mots dans la phrase, semble donc en parti approprié à notre application. Le problème est, qu'avec le modèle de langage utilisé, ce critère sur-évalue les phrases comportant un ou plusieurs mots inconnus.

Dans le cas des données d'actualité, le vocabulaire du corpus d'apprentissage est très large et le modèle de langage associe une probabilité trop importante aux mots hors vocabulaire. Ainsi, certaines suites de mots comme "Je fais un MOTINC" peuvent se voir attribuer une probabilité d'apparition beaucoup plus forte que la phrase "Je fais un essai". Pour traiter ce problème, une technique consiste à assigner une probabilité que l'on définira aux mots n'apparaissant pas dans le corpus d'apprentissage. Notre métrique de classement doit donc tenir compte du nombre de mots inconnus dans la phrase, en leur attribuant une pondération adéquate. Il est alors nécessaire d'introduire une pénalité $\epsilon$ appliquée à chacun des mots hors vocabulaire rencontrés dans les phrases. Pour cela nous avons déterminé empiriquement une valeur adéquate en vérifiant qu'elle soit inférieure au plus petit des logarithmes de probabilité attribué à un mot connu. Celle-ci a été fixée à $\epsilon = -8$. Finalement, le score $LP_{OOV}$ attribué à une phrase traduite $W^K$ s'écrit :

$$LP_{OOV}(W^K) = \sum_{k=2}^{K} logP(w_k \mid w_{k-1}w_{k-2}) + OOVs \times \epsilon \quad (1)$$

où $OOVs$ est le nombre de mots hors vocabulaire de la phrase.

Par la suite, nous classerons les phrases par maximisation du score $LP_{OOV}$ de l'équation 1.

# 4 Expérimentations et résultats

Le *méta-moteur* a l'intéret de présenter, pour une seule et même phrase à traduire, plusieurs traductions provenant de différents systèmes de traduction, d'une part, et de les présenter classées selon un critère de qualité, d'autre part. C'est ce classement qui va être jugé lors des évaluations. Le système de classement est donc évalué, avec les méthodes usuelles, en considérant uniquement la traduction qu'il classe première ($1stBest$). Il présente, en effet, un intérêt particulier s'il apporte plus d'information que l'utilisation d'un seul des moteurs de traduction, pris séparément, en l'occurrence le meilleur. On prévoit donc que les performances de notre système soient meilleures que celles du meilleur moteur de traduction utilisé. Pour vérifier cette hypothèse, nous procéderons à une évaluation automatique que l'on confirmera par une évaluation subjective.

## 4.1 Evaluation automatique

La qualité des traductions est estimée automatiquement par la métrique existante la plus utilisée dans le domaine, le score BLEU (Papineni *et al.*, 2002), complété avec le score NIST (Doddington, 2002) et le score METEOR (Lavie & Agarwal., 2005). La métrique BLEU repose sur la comparaison de la sortie du traducteur avec les traductions dites de référence. Cette métrique mesure le recouvrement lexical de la phrase traduite avec une ou plusieurs phrases données comme références de la traduction. Le score BLEU varie de 0 à 1 et, étant un score de précision, il est d'autant meilleur qu'il est grand. Néanmoins difficile à interpréter dans l'absolu, nous comparerons le score BLEU de notre système au score Bleu des moteurs de traduction en ligne utilisés. Les scores NIST et METEOR, quant à eux, reprennent le principe du score BLEU et l'adaptent légèrement.

Le tableau 2 montre les résultats de l'évaluation des neufs moteurs de traduction utilisés. Ils sont classés par ordre de performance, sur notre corpus de test aligné de 300 phrases. Le traducteur $MT_G$ est en tête avec un score BLEU de 0,311. On constate également que les scores NIST et METEOR sont corrélés avec l'évaluation BLEU.

| | METEOR | NIST | BLEU |
|---|---|---|---|
| $MT_G$ | 0,165 | 6,68 | 0,311 |
| $MT_R$ | 0,145 | 6,20 | 0,258 |
| $MT_S$ | 0,141 | 6,07 | 0,252 |
| $MT_P$ | 0,142 | 6,06 | 0,251 |
| $MT_A$ | 0,135 | 5,89 | 0,234 |
| $MT_E$ | 0,122 | 5,83 | 0,216 |
| $MT_W$ | 0,130 | 5,78 | 0,230 |
| $MT_F$ | 0,122 | 5,78 | 0,216 |
| $MT_L$ | 0,120 | 5,51 | 0,206 |

TAB. 2 – Scores BLEU, NIST et METEOR des 9 moteurs de traduction sur notre corpus journalistique de 300 phrases

Le tableau 3 présente les résultats de l'évaluation du classement proposé par la métrique de l'équation 1. L'ensemble des traductions classées premières par notre système, que l'on nommera par la suite $1stBest$, obtient un score BLEU de 0,317. De la même façon, on appelle respectivement $2ndBest$ et $3rdBest$ l'ensemble des traductions classées par le système en deuxième et troisième position.

| | METEOR | NIST | BLEU |
|---|---|---|---|
| 1stBest | 0,170 | 6,82 | 0,317 |
| 2ndBest | 0,154 | 6,48 | 0,285 |
| 3rdBest | 0,144 | 6,21 | 0,261 |

TAB. 3 – Scores BLEU, NIST et METEOR des trois premiers résultats de notre méta-traducteur

Le tableau 4 détaille la proportion des différents moteurs de traduction présents dans les trois premiers résultats. Bien que $MT_G$ soit présent à 80,7 % dans les trois premières réponses données par le système, les autres traducteurs apparaissent également en tête des résultats du clas-

sement. Les neufs moteurs de traduction sélectionnés apportent donc tous leur contribution aux performances des trois meilleurs sélectionnés.

| % | $MT_G$ | $MT_R$ | $MT_P$ | $MT_A$ | $MT_S$ | $MT_E$ | $MT_L$ | $MT_F$ | $MT_W$ | total |
|---|---|---|---|---|---|---|---|---|---|---|
| 1stBest | 54,88 | 21,44 | 5,4 | 5,7 | 2,8 | 4,3 | 3,4 | 2,0 | 0 | 100 |
| 2ndBest | 15,3 | 46,4 | 13,0 | 6,0 | 5,8 | 6,0 | 4,3 | 4,3 | 0,9 | 100 |
| 3rdBest | 46,8 | 8,5 | 8,8 | 8,0 | 9,0 | 6,3 | 4,8 | 4,5 | 3,0 | 100 |
| Total | 39,0 | 25,4 | 9,0 | 5,5 | 5,9 | 5,5 | 4,1 | 3,6 | 1,3 | 100 |

TAB. 4 – Proportions des moteurs dans les trois premiers résultats du méta-traducteur

En sélectionnant automatiquement la meilleure phrase, au sens de la métrique, proposée par les neuf moteurs de traduction, notre système présente un score BLEU de 0,317 alors que le meilleur des moteurs de traduction utilisé ($MT_G$) obtient un score de 0,311. Les scores NIST et METEOR sont eux aussi nettement améliorés. Néanmoins, il est difficile de savoir si ce gain est réellement significatif. De ce fait, il paraît utile d'avoir recours à une évaluation humaine subjective pour vérifier le résultat obtenu de façon automatique.

## 4.2 Evaluation subjective

Le but de cette évaluation subjective est de faire appel à des annotateurs pour confirmer les résultats obtenus avec l'évaluation automatique. Vérifier le classement complet des traductions par le système serait très complexe à mettre en œuvre, on préfèrera évaluer le premier choix établi par notre système comparativement au meilleur des moteurs de traduction utilisé, en l'occurrence $MT_G$.

On évalue donc la meilleure traduction sélectionnée par notre système comparativement à la traduction que donne $MT_G$. La consigne donnée aux participants est de choisir, parmi les deux phrases, celle qui leur semble la meilleure, ou de n'effectuer aucun choix s'ils jugent celles-ci équivalentes. Il y a donc trois réponses possibles : la phrase donnée par notre système est meilleure, la phrase donnée par $MT_G$ est meilleure, aucune n'est meilleure que l'autre. Il est à noter que nous ne précisons évidemment pas la provenance des phrases.

Seize volontaires ont participé à l'expérimentation et chaque phrase a été évaluée par six annotateurs différents. Nous avons vérifié la cohérence des réponses à l'aide de l'indice de Bayes. Plus l'erreur de Bayes est petite, plus les annotateurs sont cohérents dans leurs réponses. Nous avons donc éliminé les phrases pour lesquelles les annotateurs étaient trop partagés, pour obtenir 78 phrases de test restantes (indice de Bayes < 0,9).

Sur ces phrases, on peut donc dire que les participants préfèrent dans 55 % des cas les réponses renvoyées par notre système contre 33 % pour $MT_G$.

L'évaluation subjective confirme l'évaluation automatique : il est préférable d'utiliser une traduction choisie parmi plusieurs, provenant de moteurs différents, que de n'utiliser systématiquement que celui qui obtient le meilleur score moyen. On peut en déduire que ce concept de méta-traducteur est une solution pour améliorer la qualité des traductions et que le classement proposé par notre système est pertinent.

# 5 Création d'une interface graphique en ligne

Pour mettre ce service de méta-traducteur à disposition d'utilisateurs potentiels, nous avons créé une interface graphique qui permet d'interagir avec le système à partir du Web.

L'interface permet à l'utilisateur d'entrer un texte à traduire et de choisir les moteurs de traduction à utiliser parmi les 9 disponibles ( figure 1). Une fois le texte traduit entré et les moteurs de traduction à utiliser sélectionnés, l'utilisateur a le choix de l'affichage des résultats. Ceux-ci peuvent apparaître classés par ordre de pertinence ou non. L'option d'affichage classé enverra la requête aux moteurs de traduction sélectionnés et, après réception des réponses, les évaluera à l'aide du modèle de langage et les affichera classées (figure 2). L'option d'affichage non classé déroule le processus jusqu'à réception des requêtes et affiche les traductions selon leur ordre d'arrivée sur le serveur. L'interface est accessible, à partir de l'URL suivante :
$http://www-clips.imag.fr/geod/User/marion.potet/GUI/metatranslator.php$.

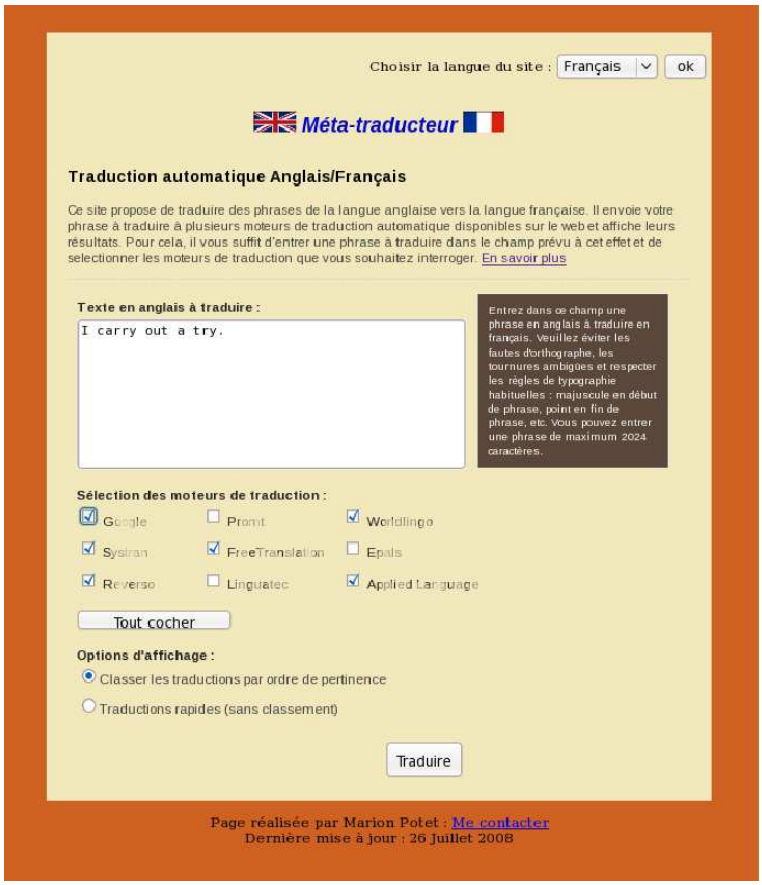


FIG. 1 – Interface graphique : page d'accueil

# 6 Conclusion

Au terme de ce travail, il ressort que, une solution possible pour améliorer la qualité de la traduction automatique, est de tirer parti de la variabilité des résultats de plusieurs systèmes de traduction. C'est l'intéret et l'originalité de l'outil traductif proposé dans cet article. Notre

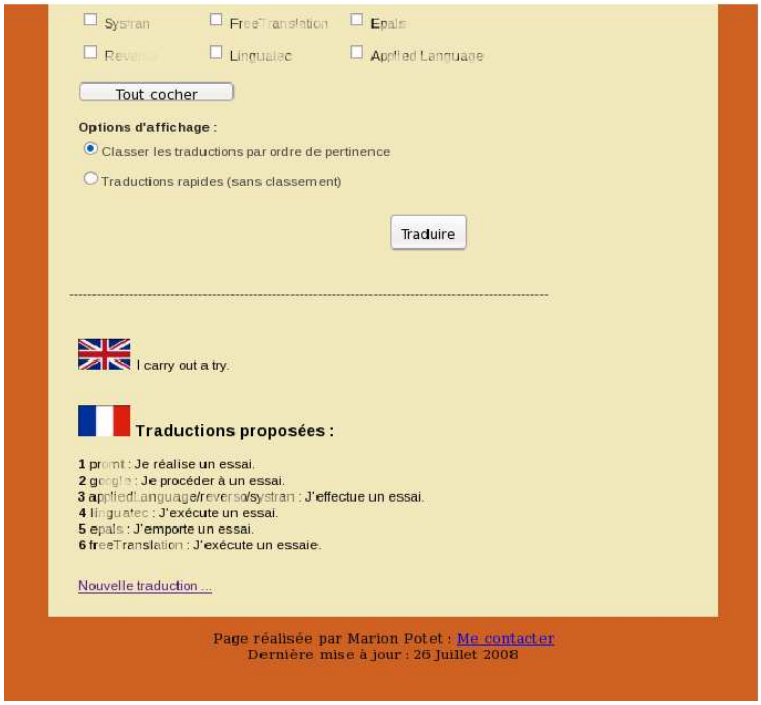


FIG. 2 – Interface graphique : résultat d'une traduction

méta-traducteur utilise, en effet, plusieurs moteurs de traduction existants et met à profit, pour chaque situation, ceux qui semblent répondre le mieux. La validation des solutions proposées a été réalisée dans la cadre d'une traduction de phrases anglaises vers la langue française de données de type journalistique. Pour chaque phrase donnée à traduire, les traductions obtenues sont classées selon un critère qualitatif évalué par une métrique reposant sur un modèle de langage. Dans l'échantillon de phrases testé, il s'est révélé être plus pertinent de proposer la meilleure traduction, sélectionnée automatiquement par notre système parmi les neufs hypothèses de traduction, que de proposer systématiquement un seul des moteurs de traduction utilisé. Ceci a été confirmé aussi bien lors de l'évaluation automatique que lors de tests subjectifs auprès d'utilisateurs potentiels. Au vu de ces résultats, l'outil a été mis à disposition sur le web à travers une interface graphique. Il est ainsi possible de soumettre un texte à traduire, de sélectionner des traducteurs automatiques à interroger et de visualiser la liste des traductions proposées classées par ordre de pertinence.

# Références

BIGI B. & LE V. B. (2008). Normalisation et alignement de corpus français et vietnamiens : Format et logiciels. In *9es journées internationales d'analyse statistique des données textuelles, France, Lyon*.

BROWN P. E., PIETRA V. J. D., PIETRA S. A. D. & MERCER R. L. (1993). The mathematics of machine translation : Parameter estimation. *Computational linguistics*, **19**, 263–311.

DODDINGTON G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Human Language Technology Conference archive Proceedings of*

*the second international conference on Human Language Technology Research, San Diego, California*, p. 138–145.

DUGAST L., SENELLART J. & KOEHN P. (2008). Can we relearn an rbmt system ? In *The Third Workshop on statistical Machine Translation, Columbus, Ohio, USA*, volume 3, p. 175–179.

KOEHN P., DUGAST L. & SEMELLARD J. (2007). Statistical post-editing on systran's rule based translation system. In *The Second Workshop on Statistical Machine Translation, Prague, Czech Republic*, volume 23, p. 220–223.

KOEHN P., OCH F. J. & MARCU D. (2003). Statistical phrase-based translation. In *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Edmonton, Canada*, volume 1, p. 48–54.

LAVIE & AGARWAL. (2005). Meteor : An automatic metric for mt evaluation with improved correlation with human judgments. In *Workshop on Statistical Machine Translation, Michigan, USA*, p. 65–72.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : A method for automatic evaluation of machine translation. In *the 40th Annual Meeting on Association for Computational Linguistics SESSION : Machine translation and evaluation, Philadelphia, Pennsylvania*, p. 311–318.

RAYNER M. & BOUILLON P. (1995). Hybrid transfer in an english-french spoken language translator. In *Journées internationales : Language engineering, Montpellier, FRANCE*, volume 15, p. 153–162.

SIMARD M., GOUTTE C. & ISABELLE P. (2007). statistical phrase-based post-editing. In *The Conference of the North American Chapter of the Association for Computational Linguistics, pages , Rochester, USA*, p. 508–515.

STOLCKE A. (2002). Srilm, an extensible language modeling toolkit. In *The International Conference on Spoken Language Processing, Denver, Colorado*, volume 3, p. 901–904.