La recherche d'unités de traduction sur le Web par maximisation de co-occurrence

Giselle SANCHEZ[1] et Daniel GAUTHERET

Media Langues, 502 Rue Paradis, 13008 Marseille

[1] sanchez@media-langues.com

1. Introduction

Ce document présente un principe simple permettant de répondre avec un moteur de recherche à des problèmes de traduction de termes techniques ou d'expressions non triviales. Les procédures proposées ne prétendent pas se substituer à une véritable recherche terminologique au moyen de dictionnaires spécialisés ou de documentation, mais il arrive très souvent que ces ressources ne soient pas accessibles au traducteur : d'une part, les dictionnaires techniques sont chers, spécialisés dans des domaines étroits et rarement au fait des développements récents ou des expressions toutes faites appréciées de nombreux rédacteurs, d'autre part, les contraintes de temps permettent rarement d'effectuer une recherche documentaire approfondie, même sur Internet. Par conséquent, même si une véritable recherche terminologique reste la panacée, il faut admettre qu'elle est de moins en moins accessible et que les traducteurs doivent souvent recourir à des solutions plus expéditives.

Les procédures que nous présentons ici exploitent les possibilités des moteurs de recherche sur le Web en s'appuyant sur le principe très simple de **maximisation de co-occurrence**. Ce principe consiste à rechercher des pages web contenant à la fois le mot à traduire et sa traduction. La traduction est d'autant plus juste que le nombre de pages contenant les deux termes est élevé. Observer de telles co-occurrences peut sembler assez improbable a priori, mais avec le nombre gigantesque de documents indexés par les moteurs de recherche (près de 10 milliards de pages par Google en 2005) comprenant un très grand nombre de sites multilingues, des co-occurrences apparaissent pour la majorité des termes techniques et se prêtent même le plus souvent à une analyse statistique. Nous présenterons également un moyen de contrôle simple, le **test d'usage**, permettant de valider les résultats d'une recherche de co-occurrence et d'éviter les très nombreux pièges posés par les mauvaises traductions déjà présentes sur Internet.

A l'aide d'exemple concrets présentant chacun des problèmes de traduction spécifiques, nous montrerons comment la maximisation de co-occurrence peut apporter des solutions rapides et efficaces aux traducteurs. Nos exemples utilisent le plus populaire des moteurs de recherche, Google, mais fonctionneraient également avec tout autre moteur générique comme AltaVista, MSN search, Yahoo, etc.

2. Quand le traducteur croit connaître la traduction correcte

Exemple : **« Smart card »** EN>FR

Il s'agit du cas le plus simple. Nous devons traduire un terme et nous avons quelques idées (pas forcément justes) de traductions possibles. La recherche de co-occurrence va nous permettre de valider très rapidement nos propositions.

Nous devons traduire en français le terme « smart cards ». La première traduction qui nous vient à l'esprit est « cartes intelligentes ». Appliquons donc pour la première fois une **recherche de co-occurrence** sur ces deux termes. Nous entrons dans Google les deux expressions:

```
« smart cards » « cartes intelligentes »
```

Ne surtout pas oublier les guillemets pour délimiter chaque expression. Après avoir lancé la recherche, nous obtenons une page de réponse commençant par:

```
Résultats 1 - 10 sur un total d'environ 132
```

Il y a donc 132 pages ou documents contenant à la fois les termes « smart cards » et « cartes intelligentes ». C'est un résultat encourageant, mais avant d'accepter cette traduction, il est indispensable d'effectuer le **test d'usage**. Ceci consiste à vérifier que le terme apparaît à une fréquence proportionnelle à la représentation de sa langue sur le Web. Pour comprendre cela, considérez le **Tableau 1** qui présente le nombre moyen de pages trouvées par un moteur de recherche avec un terme donné dans différentes langues, lorsque ce terme est observé 100 fois en anglais.

Tableau 1 : Représentation des langues sur le Web (anglais = 100). Cette table donne le nombre d'occurrences moyen sur Google d'un échantillon de termes ou d'expressions en français, espagnol et allemand, lorsque le terme ou l'expression équivalent est trouvé 100 fois en anglais.

	EN	FR	ES	DE
Termes ou expressions courants (moyenne sur 8 termes)	100	9	6	16
Termes ou expressions techniques (moyenne sur 12 termes)	100	7	4	6

On voit notamment que pour un terme courant observé 100 fois en anglais, on ne trouve son équivalent français que 9 fois ou son équivalent espagnol que 6 fois. C'est à dire qu'un terme français se rencontre environ 10 fois moins que sa traduction anglaise. Il ne s'agit que d'une moyenne, et nous rencontrerons quelques cas particuliers dans cette étude, mais il reste que toute déviation importante par rapport à cette moyenne doit être interprétée par le traducteur comme un signal d'alarme: la traduction choisie risque d'être erronée.

Voyons donc comment se comporte l'expression « cartes intelligentes » dans le test d'usage :

```
« cartes intelligentes »
```

```
Résultats 1 - 10 sur un total d'environ 632
```

```
« smart Cards »
```

```
Résultats 1 - 10 sur un total d'environ 1 040 000
```

Il y a deux indices alarmants dans ce résultat. Premièrement, le nombre d'occurrence en français n'est pas dix fois inférieur à l'anglais comme on s'y attendait, mais 1500 fois inférieur ! Deuxièmement, le nombre de co-occurrences (132) n'est pas très éloigné du nombre d'occurrences en français (632). Ceci indique que les rédacteurs ayant sous les yeux le terme anglais « smart card » tendent à employer la traduction « carte intelligente ». On peut donc soupçonner une erreur de traduction.

Il s'agit bien évidemment d'un anglicisme et, dans le cadre de cet exercice, où le traducteur cherche plutôt à confirmer ses intuitions, nous admettrons maintenant que la traduction plus probable de « carte à puce » nous vienne à l'esprit. Passons donc au test d'usage :

```
« cartes à puce »
```

```
Résultats 1 - 10 sur un total d'environ 137 000
```

Voici un résultat beaucoup plus satisfaisant. La fréquence du terme français représente environ 13% de celle du terme anglais, ce qui est conforme aux 10% attendus. Réalisons une ultime vérification des co-occurrences :

```
« cartes à puce » « smart cards »
```

```
Résultats 1 - 10 sur un total d'environ 1 330
```

Soit dix fois plus que la co-occurrence « cartes intelligentes » + « smart cards ». Ceci indique fortement que le terme de « carte à puce » est le plus approprié.

3. De l'intérêt de connaître la représentation des langues sur le Web

Exemple : **« Informatique »** FR>EN

Dans ce deuxième exemple, nous souhaitons traduire en anglais le terme « informatique ». Le premier terme qui nous vient à l'esprit est « informatics ». Passons à la recherche de co-occurrence :

```
informatique informatics
```

```
Résultats 1 - 10 sur un total d'environ 86 000
```

Encore une fois, voilà un résultat encourageant. Mais que donne le test d'usage ?

```
informatique
```

```
Résultats 1 - 10 sur un total d'environ 50 000 000
```

```
informatics
```

```
Résultats 1 - 10 sur un total d'environ 19 500 000
```

Ce résultat est suspect : le terme anglais est moins utilisé que le terme français, alors qu'il devrait l'être 10 fois plus. Effectivement, il nous revient à l'esprit que le mot « informatique » est une invention française qui n'a été que récemment introduite en anglais. Le terme anglais courant est « computer science ». Vérifions la co-occurrence :

```
informatique « computer science »
```

```
Résultats 1 - 10 sur un total d'environ 232 000
```

Voilà qui est mieux que les 86000 co-occurrences observées avec « informatics ». On remarquera néanmoins que 86000 n'est pas un nombre négligeable, ce qui montre que le terme « informatics » est incontestablement passé à l'usage courant en anglais. Passons enfin au test d'usage :

```
« computer science »
```

```
Résultats 1 - 10 sur un total d'environ 69 300 000
```

L'usage de « computer science » est donc plus élevé que celui de « informatics ». Curieusement cependant, le terme n'est pas trouvé 10 fois plus que son équivalent français « informatique », mais pratiquement autant. Cette fois, l'explication n'est pas une nouvelle erreur de traduction. Il s'avère simplement que le terme informatique désigne en français deux concepts : la discipline scientifique et le qualificatif équivalent de « computer » en anglais. L'arbitrage du traducteur reste indispensable pour trancher dans de tels cas.

4. Ajouter du contexte et fouiller dans un site multilingue

Exemple : **« Conserves de fruits »** FR>EN

Dans les exemples précédents, nous avons présenté des cas relativement simples où le traducteur avait une intuition raisonnable de la bonne traduction. Voyons maintenant comment la recherche de co-occurrence peut aider le traducteur dans le cas où il n'a aucune idée de la traduction, mais connaît toutefois le contexte d'application du terme ou de l'expression.

Nous devons traduire en anglais un catalogue français de produits alimentaires. Comment traduire « conserves de fruits » ? Tentons la première expression qui nous vient à l'esprit :

```
« conserves de fruits » « fruit conserves »
```

```
Résultats 1 - 10 sur un total d'environ 13
```

Ce faible nombre d'occurrence est inquiétant. Passons au test d'usage :

« conserves de fruits »

Résultats **1 - 10** sur un total d'environ **3720**

« fruit conserves »

Résultats **1 - 10** sur un total d'environ **3220**

Nous attendrions plutôt de l'ordre de 30000 réponses en anglais. Nous voilà convaincu qu'il s'agit d'une mauvaise traduction. Imaginons maintenant que nous n'avons pas d'autre idée de traduction pour ce terme. Que faire dans ce cas? Nous allons passer à une **recherche par contexte**.

Associons le terme recherché à un mot que nous espérons lui trouver associé dans la langue cible, par exemple un nom de fruit: strawberry.

« conserves de fruits » strawberry

Voyons cette fois à quoi ressemble la page de résultats :

Notre but est maintenant d'identifier parmi les sites proposés un texte en français accompagné de sa traduction en anglais. Le cinquième document, titré «Agricultural analyses and design » est justement un document bilingue, ce qui apparaît dès la lecture du titre bilingue à l'ouverture.

Ouvrons le document en mode HTML et recherchons « conserves de fruits » à l'aide de la commande « CTRL-F ». Nous arrivons à la phrase suivante :

```
L'industrie de préparation de produits horticoles est plus
orientée à la production d'une variété de jus de fruits (fruits
de la passion, ananas …) et les conserves de fruits (confiture,
pommades).
```

Si la traduction anglaise de cette phrase est bien présente dans ce document, on devrait retrouver la traduction des autres mots de cette phrase également. Cherchons par exemple la traduction de « confiture », que nous connaissons bien : « jam ». Nous trouvons effectivement « jam » un peu plus loin, dans le contexte :

```
The horticulture processing industry is mostly confined to the
production of several fruit juices (passion fruit, pineapple) and
fruit preserves (jams, jellies).
```

Et cette phrase nous amène à la traduction de « conserves de fruits » par «fruit preserves ». Vérifions cette traduction par un test d'usage :

```
« fruit preserves »
```

```
Résultats 1 - 10 sur un total d'environ 71 800
```

Soit vingt fois plus que « conserves de fruits » : voilà qui est plus satisfaisant que « fruit conserves ».

Il faut noter toutefois qu'une autre traduction valide, « canned fruits », est observée 99.700 fois. Elle nous a échappé parce que nous nous sommes contentés de la première occurrence trouvée, mais nous aurions pu trouver cette autre solution en analysant quelques documents supplémentaires. Comme toujours, la qualité de la traduction dépendra du temps que le traducteur sera prêt à consacrer à ce travail et de ses connaissances intrinsèques.

5. Un piège déjoué par le test d'usage

Exemple : **« Produits bio »** FR>EN

« Produits bio », voici à nouveau un terme français bien connu dont la traduction la plus évidente « bio products » est absolument fausse. En voici la preuve :

```
« produits bio »
```

```
Résultats 1 - 10 sur un total d'environ 89 000
```

```
« bio products »
```

```
Résultats 1 - 10 sur un total d'environ 75 400
```

De toute évidence, il n'y a pas assez de « bio products », et le doute devient une certitude lorsque l'on regarde la page de résultats :

Tous ces sites font référence à l'industrie biomédicale ou chimique : rien à voir avec la notion française de produit « bio » ! La recherche de co-occurrence « bio products » et « produits bio » donne pourtant 27 solutions. Soit probablement autant de mauvaises traductions !

Pour parvenir à la bonne traduction, aidons-nous encore une fois du contexte pour trouver un texte bilingue qui contiendra le terme correct. Essayons par exemple le mot « bread » :

```
« produits bio » bread
```

```
Résultats 1 - 10 sur un total d'environ 525
```

Voici un extrait de la page de solutions :

Il faut maintenant repérer un site susceptible de comporter le terme français et sa traduction. Par mesure de sécurité, étant donné que nous avons constaté l'utilisation d'un faux ami, nous conseillons d'éliminer tous les sites susceptibles d'être réalisés par des francophones. Oublions donc tous les sites en « .fr » ou « .be ». Le dernier site de la première page attire notre attention. En effet, son adresse indique :

```
http://www.royle-int.co.uk/fr/products.html
```

D'après son nom de domaine en « .uk », ce site est anglais. Notons que son adresse contient « `/fr` », suggérant la présence de versions en d'autres langues. Ouvrons d'abord la page trouvée par le moteur de recherche. A l'aide de la fonction Rechercher (CTRL-F), nous trouvons immédiatement l'expression « produits bio » dans le contexte suivant :

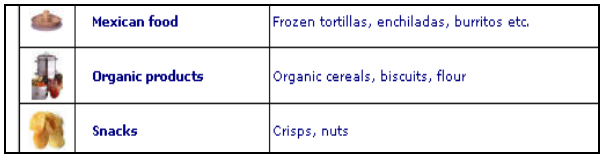
Curieusement, le mot « bread », qui avait été demandé au moteur de recherche, ne se trouve nulle part dans cette page. C'est un événement fréquent avec certains moteurs et il ne doit pas nous inquiéter. Comment trouver maintenant la page traduite en anglais ? De nombreux sites multilingues sont organisés selon une hiérarchie rationnelle, avec un répertoire différent pour chaque langue, noté par « `/fr` » pour le français, « `/uk` », « `/en` » ou « `/eng` » pour l'anglais, « `/es` » pour l'espagnol , etc.. Toutefois, on trouve également des sites dont les pages rédigées dans la langue principale ne se trouvent dans aucun répertoire particulier, et seules les pages traduites dans d'autres langues utilisent la notation en répertoire. Il n'y donc pas de règle absolue et il convient d'essayer différentes possibilités. Dans le cas qui nous intéresse, l'adresse avec un répertoire « `/uk` » …

```
http://www.royle-int.co.uk/uk/products.html
```

… nous dirige vers une page d'erreur. Nous essayons donc avec une adresse sans répertoire :

```
http://www.royle-int.co.uk/products.html
```

… et nous tombons sur la version anglaise du site. Très rapidement, nous identifions la partie de la page équivalente à celle lue en français :



	Mexican food	Frozen tortillas, enchiladas, burritos etc.
	Organic products	Organic cereals, biscuits, flour
	Snacks	Crisps, nuts

La traduction recherchée serait donc « organic products ». Vérifions son usage :

```
« organic products »
```

```
Résultats 1 - 10 sur un total d'environ 734 000
```

… soit environ 8 fois plus que “produits bio” en français, ce que confirme la co-occurrence :

```
« organic products » « produits bio »
```

```
Résultats 1 - 10 sur un total d'environ 171
```

Nous tenons la bonne traduction !

6. La recherche directe du terme source dans la langue cible

Exemple : **« Le cachet de la poste faisant foi »** FR>EN

« le cachet de la poste faisant foi » est une expression très appréciée de l’administration française. Pour mesurer son usage, il faut la considérer comme un bloc :

```
« le cachet de la poste faisant foi »
```

```
Résultats 1 - 10 sur un total d'environ 48 000
```

Si l’on considère qu’il s’agit d’une expression de 7 mots, le nombre d’occurrences est impressionnant ! Ceci nous indique que nous devons trouver une expression anglaise utilisée environ 500.000 fois ! Hélas, nous n’avons aucune notion de cette expression et nous devons donc entreprendre une recherche par contexte. La première idée serait d’utiliser le mot anglais « post », mais il est vraiment trop peu spécifique (Washington Post, post-doc…) et il n’est pas réservé à l’anglais (post-opératoire). Utilisons plutôt le mot « stamp » qui pourrait être associé à « cachet ».

```
« le cachet de la poste faisant foi » stamp
```

Résultats **1 - 10** sur un total d'environ **95**

En fouillant les premières pages proposées, on trouve bon nombre de « post marked stamp » , ou « according to post stamp », mais ce sont souvent des sites français et nous devons donc être vigilants. Effectivement :

« post marked stamp »

Résultats **1 - 10** sur un total d'environ **31**

Nous sommes loin des 500.000 ! Une fouille plus approfondie des sites bilingues nous apporte « postmarked stamp ». Le test d'usage est bien meilleur :

« postmarked stamp »

Résultats **1 - 10** sur un total d'environ **64 600**

Avec une telle fréquence d'utilisation, il s'agit probablement d'une expression acceptable en anglais, mais nous sommes toujours loin du compte attendu. La co-occurrence est d'ailleurs très décevante :

« postmarked stamp » « le cachet de la poste faisant foi »

Résultats **1 - 1** sur **1**

Mais comment faire si les sites bilingues ne semblent pas apporter de réponse satisfaisante ? Essayons de trouver de nouveaux sites **en effectuant une recherche directe en langue cible**. Nous devons régler le moteur de recherche pour qu'il ne trouve que des sites en langue cible, ici l'anglais, au moyen des options de recherche avancée :

Nous recherchons ensuite l'expression française seule :

« le cachet de la poste faisant foi »

Ainsi, nous forçons le moteur à nous montrer des pages identifiées comme « en anglais » qui comprennent l'expression française (ou ayant un trafic important en provenance de pages contenant cette expression). Le moteur nous propose ainsi 189 sites :

Le premier site est un site bilingue canadien. Après l'avoir ouvert en HTML, nous trouvons directement l'expression « must be post-marked ». Effectuons un test d'usage :

```
« must be post-marked »
```

```
Résultats 1 - 10 sur un total d'environ 21 800
```

C'est toujours acceptable, mais il manque encore quelque chose. Il suffit en fait de se rappeler que le terme « postmarked » en un seul mot est plus utilisé que « post-marked » en deux mots (voir le test d'usage plus haut). Essayons donc :

```
« must be postmarked »
```

```
Résultats 1 - 10 sur un total d'environ 570 000
```

Et voilà ! La co-occurrence est également améliorée :

```
« must be postmarked » « le cachet de la poste faisant foi »
```

```
Résultats 1 - 10 sur un total d'environ 81
```

… et l'on pourra même utiliser le moteur de recherche pour connaître la préposition la mieux appropriée :

« must be postmarked by »

Résultats **1 - 10** sur un total d'environ **257 000**

Ou bien :

« must be postmarked no later than »

Résultats **1 - 10** sur un total d'environ **91 200**

L'expression « must be postmarked by » est donc la meilleure traduction que nous puissions trouver ici pour « le cachet de la poste faisant foi ». Bien qu'elle soit très éloignée de l'expression française et qu'elle demande une reformulation globale de la phrase, la recherche de co-occurrence s'est avéré suffisante pour trouver sa traduction.

7. Résumé

Selon le concept de **maximisation de co-occurrence**, la meilleure traduction d'un terme est celle qui se trouve le plus fréquemment associée à ce terme lors d'une recherche sur le Web. Lorsque le traducteur envisage plusieurs traductions, il doit choisir celle qui présente le plus grand nombre de co-occurrences.

La **fréquence d'usage** est indispensable pour contrôler la validité des traductions. Les langues sont représentées sur le Web de façon très inégale et, par conséquent, un terme et sa traduction ne se rencontrent pas avec la même fréquence. En se référant au Tableau 1, un traducteur peut vérifier si les fréquences d'usage observées sont conformes aux valeurs attendues. Dans le cas d'une déviation importante, la traduction doit être considérée comme potentiellement incorrecte.

Lorsqu'aucune des traductions envisagées ne semble satisfaisante, le traducteur peut effectuer une **recherche par contexte** dans des sites bilingues. Dans ce cas, l'expression à traduire est associée à un terme en langue cible dont on a toutes les raisons de croire qu'il sera associé à l'expression à traduire. Une telle recherche identifie dans la plupart des cas de nombreux sites bilingues, dans lesquels on recherchera l'expression en langue source puis, son équivalent en langue cible ailleurs dans la même page.

Lorsque la recherche par contexte n'identifie aucun site pertinent, ou ne permet d'identifier que des traductions non satisfaisantes en termes de co-occurrence ou de fréquence d'usage, on peut effectuer une **recherche directe dans la langue cible**. A l'aide des options avancées du moteur de recherche, l'expression source est directement recherchée dans les pages en langue cible. Cette méthode identifie fréquemment de nouvelles pages bilingues qui peuvent contenir la traduction recherchée.