La traduction assistée par ordinateur

»http://www.fxm.ch/Fr/Langues-Traduction/TraductionOrdinateur.fr.htm

La Traduction Assistée par Ordinateur (TAO) est un ensemble de technologies visant à aider les humains à traduire des textes à l'aide des ordinateurs.

La traduction assistée par ordinateur est une technologie différente de la traduction automatique (ex. : Google Translate). En traduction automatique, l'ordinateur traduit lui-même le texte, l'humain n'est pas censé intervenir. En traduction assistée par ordinateur, c'est l'humain qui traduit en s'aidant de logiciels informatiques.

Ces logiciels permettent par exemple de créer automatiquement des lexiques (dictionnaires) bilingues à partir d'anciennes traductions. Ceci est particulièrement utile dans le cas de la traduction technique : même si un traducteur professionnel maîtrise bien une langue, il est souvent difficile pour lui/elle de traduire les termes techniques. Par exemple, les termes *thérapie hormonale* et *pré-ménopause* sont des expressions que l'on ne trouve pas dans les dictionnaires bilingues généralistes qui ne contiennent que des mots non techniques (Larousse, Petit Robert, etc.).

Les logiciels d'extraction automatique de lexiques bilingues

Pour aider les traducteurs, des chercheurs ont mis au point des programmes informatiques qui analysent les traductions passées et identifient automatiquement les paires de termes qui sont des traductions. Le résultat de cette analyse est un lexique bilingue de termes techniques, comme montré dans le schéma ci-contre.

De cette façon, un traducteur peut utiliser les anciennes traductions de ses confrères pour l'aider à traduire des textes qui traitent d'un domaine dont il ne connait pas les termes techniques. Le logiciel se charge pour lui d'aller retrouver dans les textes les traductions des termes qu'il ne connaît pas.

Les recherches en « extraction automatique de lexiques bilingues » ont démarré dans les années 70 et aujourd'hui, cette fonctionnalité est présente dans de nombreux logiciels commerciaux de TAO. Pourtant, cette technologie présente une limite majeure : que faire lorsque l'on aborde un nouveau domaine technique pour lequel il n'existe encore aucune traduction ? C'est dans ce cadre de recherches que s'inscrit mon travail de thèse.

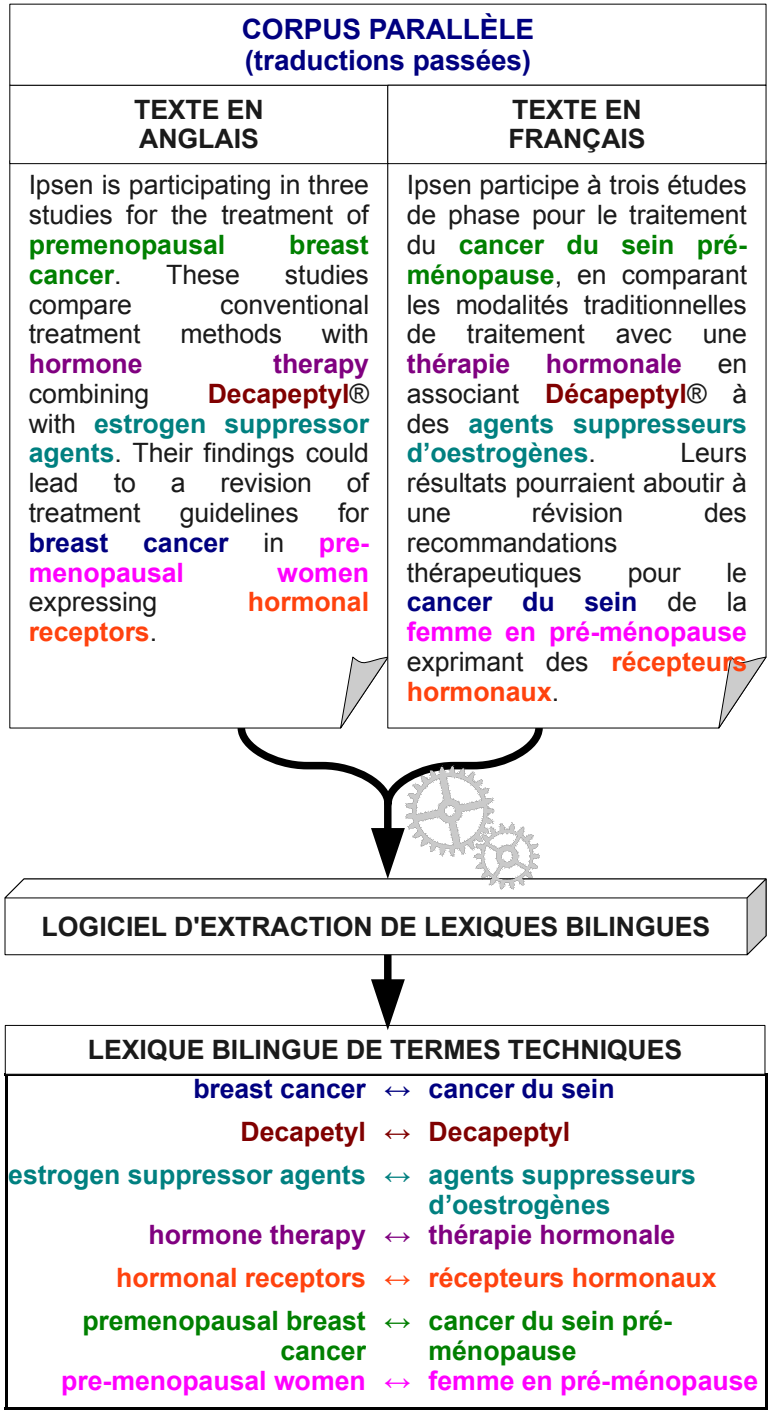


Fig.1 : Processus de création automatique d'un lexique bilingue

Extraire des lexiques bilingues sans traductions passées grâce aux corpus comparables

Lorsqu'il n'existe aucune traduction passée, on a recours à ce que l'on appelle des « corpus comparables ». Les corpus comparables sont un ensemble de textes dans deux langues différentes, traitant du même sujet sans pour autant être des traductions les uns des autres. Comme on peut le voir dans le schéma de droite, les corpus comparables sont plus difficiles à traiter pour le logiciel.

En effet, les mots des textes anglais et français ne suivent pas le même ordre ; certaines traductions n'existent pas (ex : *Decapeptyl*) ; certains termes sont traduits de façon plus libre (ex : *hormone therapy* est traduit par *hormonothérapie* au lieu de *thérapie hormonale*).

Toutes ces raisons font que le lexique extrait est de moins bonne qualité : **plusieurs traductions sont proposées** (au traducteur de retrouver la bonne) et parfois **aucune bonne traduction n'a été trouvée**. Bref, les algorithmes doivent encore être améliorés avant d'être utilisés dans des logiciels commerciaux !

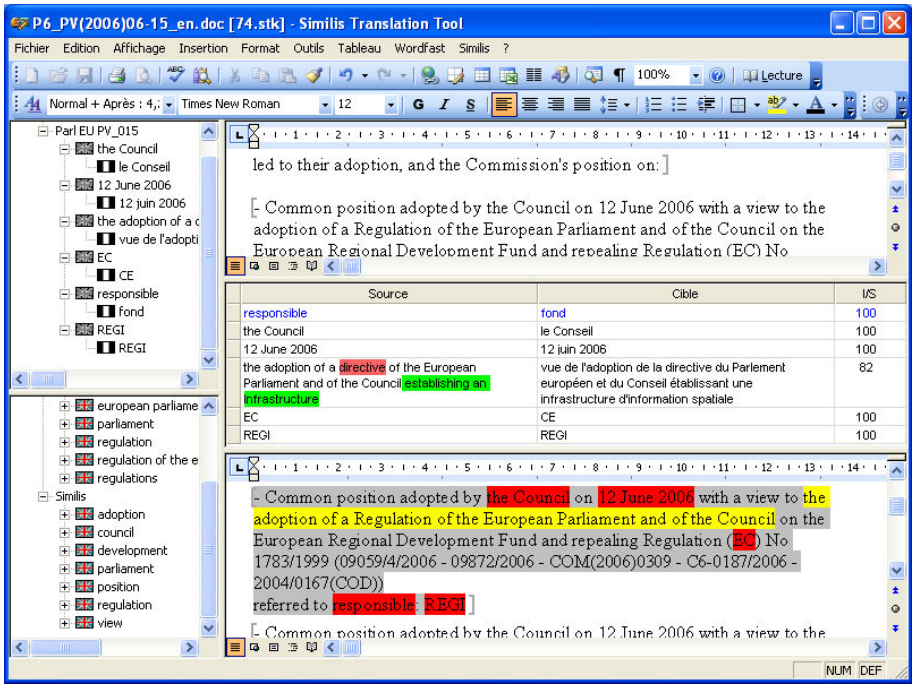


Fig. 2 : Logiciel d'aide à la traduction Similis

CORPUS COMPARABLE	
TEXTE ANGLAIS SUR LE CANCER DU SEIN	**TEXTE FRANÇAIS SUR LE CANCER DU SEIN**
Ipsen is participating in three phase III studies conducted under the auspices of the International Breast Cancer Study Group for the treatment of **premenopausal breast cancer**. These studies compare conventional treatment methods with **hormone therapy** combining **Decapeptyl**® with **estrogen suppressor agents**. They are scheduled to continue through to 2015. Their findings could lead to a revision of treatment guidelines for **breast cancer** in **pre-menopausal women** expressing **hormonal receptors**.	**Hormonothérapie** Dans environ deux tiers des cancers du sein, les cellules cancéreuses présentent des **récepteurs hormonaux** en excès. La tumeur est alors dite hormono sensible car les œstrogènes stimulent la prolifération cancéreuse par l'intermédiaire de ces récepteurs. Dans le **cancer du sein** les traitements hormonaux agiront soit en diminuant le taux d'œstrogènes dans le sang et donc la stimulation des **récepteurs hormonaux** (castration, anti-aromatases), soit en bloquant les **récepteurs hormonaux**.

LOGICIEL D'EXTRACTION DE LEXIQUES BILINGUES

LEXIQUE BILINGUE DE TERMES TECHNIQUES

breast cancer	→	**cancer du sein**
	→	sein
	→	cancer
	→	cellules cancéreuses
Decapetyl	→	?
estrogen suppressor agents	→	?
hormone therapy	→	thérapie
	→	récepteurs hormonaux
	→	hormono
hormonal receptors	→	hormonothérapie
	→	récepteurs
	→	**récepteurs hormonaux**
premenopausal breast cancer	→	?
pre-menopausal women	→	?

Comment fonctionnent les logiciels d'extraction de lexiques bilingues à partir de corpus comparables ?

Il existe deux sortes méthodes :
- » la méthode **distributionnelle**, utilisée depuis les années 90 ;
- » la méthode **compositionnelle**, appliquée aux corpus comparables depuis 2009.

La méthode distributionnelle

Dans cette méthode, on considère que deux termes sont des traductions l'un de l'autre s'ils apparaissent souvent avec les mêmes mots (ils ont la même *distribution*).

Dans l'illustration ci-dessous, on voit que le terme anglais *hormone therapy* et le terme français *hormonothérapie* sont entourés de mots qui ont le même sens : *woman* ↔ *femmes* ; *ovaries*↔ *ovaire* ; *chemotherapy* ↔ *chimiothérapie* ; *radiotherapy* ↔ *radiothérapie* ; t*reatment* ↔ *traitement*. Les mots de contextes qui ont le même sens sont identifiés grâce à un dictionnaire bilingue généraliste qui permet de traduire tout les mots non techniques.

Hormone therapy et *hormonothérapie* ont des distributions similaires : on considère alors qu'il est fort possible que *hormone therapy* et *hormonothérapie* soient des traductions.

TEXTES ANGLAIS

It can also occur if a woman has had surgery to remove her ovaries or if chemotherapy, radiotherapy, or **hormone therapy** causes ovarian failure (….)
Hormone therapy for cancer offers a better tolerated option than traditional chemotherapy and is particularly suitable for long-term treatment.

TEXTES FRANÇAIS

Quatre techniques sont principalement utilisées pour le traitement du cancer de l'ovaire : la chirurgie, la radiothérapie, la chimiothérapie et l'**hormonothérapie** (…)
Maintenant, il est important que les femmes parlent à leur médecin pour déterminer si l'**hormonothérapie** est une possibilité pour elles.

La méthode compositionnelle

Dans cette méthode, on considère que deux termes sont des traductions s'ils sont composés des mêmes éléments (ils ont la même *composition*).

Dans l'illustration plus bas, on voit que le terme *hormonal receptors* est composé des mêmes éléments que le terme *récepteurs hormonaux* : *hormonal* correspond à *hormonaux* et *receptors* correspond à *récepteurs.* La traduction des éléments a pu être obtenue grâce à un dictionnaire bilingue généraliste.

Comme *hormonal receptors* et *récepteurs hormonaux* contiennent les mêmes éléments, on considère qu'il est fort possible qu'ils soient des traductions l'un de l'autre.

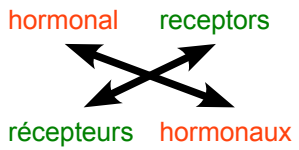
Le principe de cette méthode est simple. Toutefois, les langues ont des *morphologies* différentes, c'est-à-dire qu'elles ne construisent pas les mots exactement de la même façon :

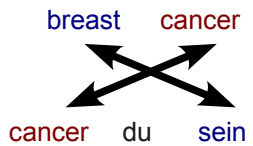
» l'ordre des éléments est différent : *hormonal receptors* vs. *récepteurs hormonaux*

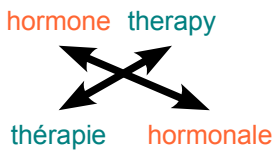
» des petits mots comme des prépositions ou des déterminants peuvent venir s'insérer : *breast cancer* vs. *cancer du sein*

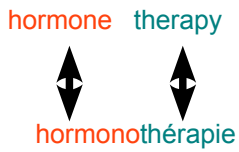
» la traduction des éléments n'est pas exacte : pour traduire *hormone therapy*, il ne faut pas traduire *hormone* par sa traduction exacte en français (le nom commun *hormone*) mais plutôt par une variante : l'adjectif *hormonal* ou le morphème lié *hormono-*.

Les éléments dont sont composés les mots sont appelés « morphèmes ». Les **différences morphologiques entre les langues sont autant de difficultés qu'ils convient de résoudre**. C'est en grande partie ce à quoi je me suis attelée pendant mon travail de thèse.









Mon travail de thèse

Mon travail de thèse est composé de deux parties :

» Dans un premier temps, j'ai développé un logiciel d'extraction de lexiques bilingues à partir de corpus comparables qui utilise la méthode distributionnelle. Cette méthode est connue depuis les années 90. Elle a été évaluée dans diverses applications (traduction automatique, moteurs de recherche multilingues) mais pas dans le cadre de la traduction assistée par ordinateur. J'ai donc évalué mon logiciel dans ce cadre : j'ai demandé à des traducteurs de traduire des textes techniques en utilisant les lexiques extraits par le logiciel et observé leurs réactions et la qualité des traductions qu'ils avaient produites. Cette évaluation a montré que la méthode distributionnelle n'est pas satisfaisante.

Aux yeux des traducteurs, le lexique n'est pas d'assez bonne qualité : plusieurs traductions sont proposées (au traducteur de retrouver la bonne) et parfois aucune bonne traduction n'a été trouvée. La solution proposée consiste alors employer la méthode compositionnelle. Cette approche avait déjà été testée avec succès par mes directeurs de thèse en 2009. Toutefois, elle peut encore être améliorée.

» La deuxième partie de mon travail de thèse a consisté à essayer d'améliorer la méthode compositionnelle. Je me suis penchée sur deux aspects :

1) J'ai proposé des solutions pour gérer les problèmes liés aux **différences morphologiques entres langues**. Ceci permet d'identifier plus de traductions dans les textes.

2) J'ai également expérimenté des techniques permettant d'aider le traducteur à trouver la bonne traduction parmi les traductions proposées par l'algorithme. Ces techniques consistent à **ordonner les traductions** de la plus à la moins plausible.

Malgré des résultats encourageants, je reste assez pessimiste dans ma conclusion quant à la possibilité d'utiliser les corpus comparables pour extraire des lexiques qui soient *directement* utilisables par les traducteurs.

PETIT LEXIQUE

Traduction Automatique
Technologie dont le but est de traduire automatiquement un texte, sans recours à un humain.

Lexique bilingue spécialisé
Liste de traductions de termes techniques

Corpus parallèle
Ensemble de textes dans une langue source accompagnés de leur traduction dans une langue cible.

Morphologie des langues
La morphologie est la branche de la linguistique qui étudie la façon dont sont formés les mots.

Traduction assistée par ordinateur
Technologie dont le but est d'aider l'humain à traduire un texte, sans chercher à le remplacer.

Extraction de lexiques bilingues
Technologie permettant de créer automatiquement des dictionnaires bilingues.

Corpus comparable
Ensemble de textes dans deux langues, qui traitent du même sujet mais ne sont pas des traductions les uns des autres.

Morphèmes
Élément composant un mot (racine, préfixe, suffixe). Par exemple, *fillette* est composé de deux morphèmes : la racine *fill-* et le suffixe *-ette*.