

White Paper Lingway



**Comment construire rapidement
des bases terminologiques multilingues**

JUIN 2005

>lingway

Solutions in language processing

La technologie Lingway permet dorénavant la construction rapide de bases terminologiques multilingues de qualité, par extraction automatique à partir de corpus non alignés, ce qui élimine le problème de la rareté des corpus alignés par rapport aux domaines émergents et aux terminologies en évolution.

Sommaire

Sommaire	2
Avant-propos	3
1. Méthodologie	4
Présentation	4
Les étapes proposées par la méthodologie	4
Constitution de corpus dans un domaine.....	4
Extraction de termes source	5
Obtention des traductions.....	5
2. Résultats obtenus.....	9
Présentation	9
Evaluation.....	9
Qualité obtenue	9
Exemples de résultats	10
3. Conclusion.....	12

Avant-propos

Les demandes dans le monde de la traduction étant de plus en plus importantes (en volume et en qualité exigée), il y a aujourd'hui un besoin d'outils informatiques multilingues de plus en plus performants, non seulement d'aide aux traducteurs humains, mais également d'aide à l'amélioration de la traduction automatique.

L'un des postes actuellement les plus coûteux et cruciaux dans un projet multilingue est **la construction ou l'enrichissement des bases terminologiques, véritable goulot d'étranglement de la qualité de la traduction**, qu'elles soient destinées aux traducteurs ou aux logiciels de traduction automatique.

Lingway propose la solution **Lingway Terminology Builder**, permettant la construction rapide de bases terminologiques de qualité pour le développement d'applications multilingues. Cette solution a été développée à partir du logiciel Lingway KM. Mariant linguistique et statistiques, elle permet de réaliser une extraction semi-automatique de terminologie bilingue sur corpus.

Lingway Terminology Builder permet de construire très rapidement des terminologies particulièrement pertinentes, avec des termes source et cible, validés dans le domaine et dans le contexte fixés par le client.

Lingway Terminology Builder permet de réduire considérablement les coûts de réalisation d'une terminologie (pouvant aller jusqu'à les diviser par 2 ou 3).

L'originalité et l'intérêt de l'approche de Lingway résident dans la rapidité et la qualité du résultat (à la fois source et cible), les termes étant extraits des corpus les plus pertinents par rapport au problème à résoudre. Contrairement aux méthodes classiquement proposées, les corpus bilingues d'extraction ne doivent pas obligatoirement être alignés.

Dans ce document nous décrivons d'abord nos méthodes et outils pour mettre au point des bases terminologiques, et ensuite nous commentons les résultats obtenus

1. Méthodologie

Présentation

Les bases de connaissance terminologiques ont une importance cruciale en traduction, car de la qualité de ces bases dépend directement la qualité des traductions obtenues. Ces bases sont en effet utilisées aussi bien par un traducteur humain qui est aidé au niveau de la terminologie, que par un logiciel de traduction automatique du type de Systran.

Une base lexicale pour une langue donnée peut généralement être divisée en trois couches :

- un noyau dur composé des 3 000 mots les plus courants,
- une langue générale ou standard (quelques dizaines de milliers de mots),
- des extensions par domaine (couche terminologique) pouvant atteindre des dizaines de milliers, voire des centaines de milliers de termes.

Lingway possède des dictionnaires de langue générale couvrant bien entendu les deux premiers niveaux, mais également des extensions terminologiques dans certains domaines technologiques.

En se basant sur ces dictionnaires et sur le logiciel Lingway KM, une méthode permettant de construire et d'enrichir automatiquement des bases terminologiques bilingues a été mise au point et validée.

Les étapes proposées par la méthodologie

Les étapes de la méthodologie de Lingway pour la construction de bases terminologiques sont les suivantes :

- Constitution de corpus (langues source et cible) représentatifs du domaine souhaité,
- Extraction de termes à partir du corpus en langue source,
- Utilisation du moteur de recherche cross-langage de Lingway KM pour identifier des candidats traductions dans le corpus cible.

La base terminologique bilingue qui en résulte peut ensuite être utilisée en simple consultation par les traducteurs humains, ou bien être importée par des logiciels (format Systran notamment).

Constitution de corpus dans un domaine

Il s'agit d'une étape très importante car de la pertinence des corpus (source et cible) choisis, dépend la qualité des termes proposés. Pour que la méthode fonctionne de manière optimale, et pour atteindre une volumétrie terminologique satisfaisante, nous estimons qu'il faut un minimum de 40 Mo de texte en corpus source et environ le double en cible. Les corpus devront être suffisamment variés pour que les ressources constituées soient représentatives du domaine.

Exploitation des ressources "client"

Dans un grand nombre de cas, le client dispose déjà de corpus associés au domaine qu'il veut couvrir, qu'il s'agisse de documentation interne à l'entreprise (brevets, articles, documents d'entreprise), ou de sites Web (revues en ligne, conférences, sites spécialisés, sociétés, blogs, ...). Les terminologies mono ou bilingues propres à l'entreprise peuvent également être exploitées.

Exploitation des ressources "Web"

Si le client ne dispose pas de corpus ou si les corpus disponibles sont insuffisants, Lingway peut en option mettre au point des corpus, en utilisant une méthodologie bien testée.

Le point de départ de cette méthode est de choisir un ensemble relativement réduit de mots du domaine dans la langue source. Ensuite, un premier ensemble de documents ou pages Web contenant ces mots (ou plus exactement des sous-ensembles significatifs de l'ensemble de ces mots) est extrait ; un calcul statistique permet de déterminer d'autres mots du domaine, et l'itération de ce processus permet l'obtention d'un corpus relativement large dans un domaine donné. Par ailleurs il existe des listes de corpus de langues, des listes de journaux (politiques, d'actualité, etc.), des sites multilingues, etc. qui peuvent de même être exploités.

Un expert de chaque langue étrangère porte un jugement sur certaines caractéristiques de chaque corpus, comme par exemple le domaine, le niveau de langue (technique, général, familier, soutenu, etc.), codage, format, et une indication sur la pertinence par rapport au projet.

Extraction de termes source

L'extraction automatique de termes en langue source est prise en charge par le logiciel Lingway KM.

Dans Lingway KM, l'extracteur de termes se base sur l'analyse linguistique de chaque phrase du corpus, et la reconnaissance de structures linguistiques typiques ou "patrons". Le tableau suivant en montre quelques exemples :

Patron	Exemple
« Nom + Adjectif »	<i>Ordinateur portable</i>
« Nom + Préposition + Nom »	<i>Employé de banque</i>
« Nom + Préposition + Nom + Adjectif »	<i>Augmentation du salaire minimum</i>

Chaque candidat terme sera retenu ou pas dans l'extraction finale, sur la base de critères statistiques complémentaires afin de ne retenir que les termes effectivement représentatifs.

Par exemple, un élément pris en compte est la couverture de chaque terme "en largeur" dans le corpus. Si par exemple le corpus est constitué d'une centaine de sites Web, et qu'un terme, même fréquent, n'apparaît que dans un seul de ces sites, il ne sera pas retenu.

Obtention des traductions

Cette étape est la plus cruciale du processus.

Elle peut être divisée en trois sous-étapes :

- production d'un grand nombre de candidats traductions,
- comptage des occurrences des candidats traductions dans le corpus cible,
- sélection de la (ou, exceptionnellement, des) meilleure(s) traduction(s).

Production d'un grand nombre de candidats traductions

Cette sous-étape utilise le moteur « cross-langage » de Lingway KM, qui va produire toutes les différentes combinaisons des traductions par rapport aux connaissances du dictionnaire et des terminologies existantes.

Par exemple, si dans les connaissances il y a :

Mot source (EN)	Traduction possible (FR)
term	condition
term	terme
term	vocable
use	utilisation
use	exercice
use	usage
use	fonction

Le moteur cross-langage de Lingway KM produit :

Terme source (EN)	Calcul candidat traduction (FR)	Version lisible
term of use	condition + de + utilisation	condition d'utilisation
term of use	condition + de + exercice	condition d'exercice
term of use	condition + de + usage	condition d'usage
term of use	condition + de + fonction	condition de fonction
term of use	terme + de + utilisation	terme d'utilisation
term of use	terme + de + exercice	terme d'exercice
term of use	terme + de + usage	terme d'usage
term of use	terme + de + fonction	terme de fonction
term of use	vocable + de + utilisation	vocable d'utilisation
term of use	vocable + de + exercice	vocable d'exercice
term of use	vocable + de + usage	vocable d'usage
term of use	vocable + de + fonction	vocable de fonction

En fait, on peut ajouter à cette liste des candidats traduction le terme anglais lui même, ainsi que des propositions de traductions de mots non répertoriées dans le dictionnaire, mais en principe possibles. Ce dernier point est surtout visible dans les domaines très techniques où il y a un grand nombre de néologismes (FR "productisation", "browser", "déboguer" ou "debugger", etc.).

Notons que le patron EN <Nom + "of" + Nom> a donné des candidats FR <Nom + "de" + Nom>. Dans d'autres cas plusieurs prépositions cible peuvent être testées, comme EN <Nom + Nom> qui peut donner FR <Nom + "de" + Nom> mais également <Nom + "en" + Nom>, <Nom + "pour" + Nom>, <Nom + "contre" + Nom>, etc.

Ces transformations sont utilisées par exemple dans l'obtention des termes corrects suivants (domaine Médecine) :

EN "cancer mortality" : FR "mortalité *par* cancer"

EN "children's hospital" : FR "hôpital *pour* enfant"

EN "flu shot" : FR "vaccin *contre* la grippe"

Comptage des traductions dans le corpus cible

Il s'agit ici de compter les occurrences, modulo certaines variantes acceptées, des candidats traduction.

Ainsi pour l'exemple précédent on aura :

Candidat traduction (FR)	Nombre d'occurrences dans le corpus cible (FR)
condition d'utilisation	228
condition d'exercice	6
condition d'usage	3
condition de fonction	0
terme d'utilisation	1
terme d'exercice	0
terme d'usage	1
terme de fonction	1
vocable d'utilisation	0
vocable d'exercice	0
vocable d'usage	0
vocable de fonction	0
<i>term of use</i>	3

L'exemple suivant, dans le domaine Médecine, est très intéressant parce que parmi les candidats traductions, certains proviennent de bases terminologiques existantes et d'autres ont été trouvés par [Lingway Terminology Builder](#).

Il s'agit des candidats traduction du terme EN "hearing loss" :

Candidat traduction (FR)	Nombre d'occurrences dans le corpus cible (FR)
perte auditive	192
perte d'audition/ perte de l'audition	155
surdit� partielle	7
baisse de l'acuit� auditive	4
perte de l'ou�e	1
perte audition	1
d�perdition auditive	0
d�perdition de l'audition	0
perte de la facult� d'entendre	0
d�perdition de la facult� d'entendre	0
d�perdition de l'ou�e	0
perte pour l'audition	0
perte en audition	0
[...]	0

Sélection de la ou des meilleures traductions

Pour effectuer cette sélection, on tient compte de plusieurs éléments :

- fréquence relative des candidats traductions (on ne retiendra pas les traductions moins fréquentes par rapport aux autres),
- fréquence absolue des candidats traductions (on ne retiendra pas les traductions très peu fréquentes),
- fréquence des candidats traductions par rapport à la fréquence du terme source (on ne retiendra pas les traductions dont la fréquence serait anormalement moins élevée en langue cible, que celle de son terme source dans le corpus source).

Dans notre premier exemple, la seule traduction retenue sera évidemment :

<i>Terme source (EN)</i>	<i>Traduction retenue (FR)</i>
term of use	condition d'utilisation

Dans d'autres cas, plusieurs traductions pourront être retenues éventuellement. C'est le cas de notre deuxième exemple :

<i>Terme source (EN)</i>	<i>Traduction retenue (FR)</i>	<i>Fréquence cible</i>
hearing loss	perte auditive	192
hearing loss	perte d'audition/ perte de l'audition	155

Le deuxième terme provient d'une terminologie existante, mais pas le plus fréquent. A noter que les deux candidats traductions suivants, "surdité partielle" et "baisse de l'acuité auditive" venaient eux aussi de terminologies existantes mais ils ne sont pas validés par le corpus cible car beaucoup moins fréquents que les deux premiers.

2. Résultats obtenus

Présentation

Cette section décrit les résultats obtenus dans le cadre d'un projet réalisé par Lingway pour le compte de la société Systran. Ce projet a consisté à construire trois grandes bases terminologiques dans les domaines de l'Informatique (18.000 termes), la Médecine (27.000 termes) et l'Economie (6.000 termes). Ces terminologies, exportées automatiquement dans le format Systran, ont été intégrées dans le logiciel **Systran 5.0 Professional Premium** et une évaluation de qualité a été menée par les équipes de Systran.

Evaluation

Qualité obtenue

La société Systran a donc intégré les bases terminologiques produites par Lingway dans son logiciel de traduction, et comparé l'évolution des résultats avec la SLP (**Systran Linguistic Platform**) qui permet de visualiser pour un corpus d'évaluation :

- chaque phrase source.
- la traduction automatique avec la nouvelle base terminologique.
- la traduction automatique sans la nouvelle base terminologique.
- une comparaison entre les deux traductions.

La SLP permet également à un traducteur humain de donner son avis pour chaque phrase sur l'amélioration, la régression ou l'équivalence apportée par la base terminologique introduite.

La plate-forme SLP calcule alors les pourcentages d'amélioration, régression ou équivalence sur tout un corpus.

Les premiers résultats obtenus avec l'intégration de la base terminologique produite par la solution **Lingway Terminology Builder** ont montré une amélioration considérable de la qualité des traductions : +49%.

Dans la première phase d'analyse, un faible pourcentage de régression a été noté. Les régressions sont souvent dues à la méthode d'évaluation, qui privilégie les termes connus dans un dictionnaire classique, alors que Lingway privilégie les termes attestés dans le corpus cible : par exemple le terme proposé par Lingway "*driver Linux*" est mal noté contre "*module de gestion de périphérique de Linux*", version un peu plus lourde mais surtout pas attestée dans les corpus.

Une deuxième raison à l'origine de certaines régressions est liée à des effets de bord du logiciel de traduction (les paires ajoutées peuvent bloquer certaines règles du logiciel).

Un autre problème rencontré, résolu depuis, était lié à des problèmes de forme de surface, et notamment de paires non accordées en nombre (singulier/pluriel).

Après élimination des effets de bord, les résultats estimés sont les suivants :

Catégorie	Pourcentage
Amélioration	53%
Régression	3%
Equivalence	44%

Le bilan est donc plus que positif, et la qualité de la traduction utilisant les bases terminologiques est incontestablement meilleure.

Exemples de résultats

Nous présentons ici quelques exemples de résultats du domaine Informatique. Sont marqués en **violet** les termes FR apportés par la terminologie construite par Lingway.

Source	[Site Info] [Privacy Policy] [Advertise] (Freq: 1153)
Traduction Systran + Lingway Terminology	[Infos du site] [Politique de confidentialité] [annoncez]
Traduction Systran "nu"	[Information de site] [politique d'intimité] [annoncez]
Comparaison	<i>[Information de [Infos du site] [politique d'intimité]</i> [Politique de confidentialité] [annoncez]

Dans ce premier exemple, on remarque la pertinence de l'approche par rapport au choix du mot "confidentialité", meilleur naturellement que "intimité" dans le contexte Informatique.

Source	Product Submission Form : (Freq: 633)
Traduction Systran + Lingway Terminology	Formulaire de soumission de produit :
Traduction Systran "nu"	Forme de soumission de produit :
Comparaison	<i>Forme</i> Formulaire de soumission de produit :

Ici, c'est le mot anglais "Form" qui est mieux traduit par "Formulaire" que par "Forme" dans le contexte.

Source	Oct 27, 2003, Connecting Windows XP to Multiple Networks: (Freq: 256)
Traduction Systran + Lingway Terminology	27 oct. 2003, Windows XP se reliant aux réseaux multiples :
Traduction Systran "nu"	27 oct. 2003, Windows se reliant XP aux réseaux multiples :
Comparaison	27 oct. 2003, Windows XP se reliant XP aux réseaux multiples :

Dans cet exemple très intéressant, ce n'est pas la traduction mot par mot qui est en cause. Les mots "Windows" et "XP" étant reconnus comme faisant partie d'un même terme "**Windows XP**", ils restent solidaires dans la traduction.

Source	Dec 01, 2003, About the new Windows XP Professional Logon Screen : (Freq: 256)
Traduction Systran + Lingway Terminology	1er déc. 2003, au sujet du nouvel écran de connexion de Windows XP Professionnel :
Traduction Systran "nu"	1er déc. 2003, au sujet du nouvel écran professionnel de procédure de connexion de Windows XP :
Comparaison	1er déc. 2003, au sujet du nouvel écran <i>professionnel de procédure</i> de connexion de Windows XP Professionnel :

Idem que le précédent, pour "**Windows XP Professionnel**".

Source	content creation (Freq: 176)
Traduction Systran + Lingway Terminology	création de contenu
Traduction Systran "nu"	création contente
Comparaison	création <i>contente de contenu</i>

Tout commentaire semble inutile.

Source	Company Info (Freq: 176)
Traduction Systran + Lingway Terminology	Infos société
Traduction Systran "nu"	Information de compagnie
Comparaison	<i>Information de compagnie</i> Infos société

"**Infos société**" est sans doute une traduction plus compréhensible que "information de compagnie".

Source	StarTech.com is the professionals' source for hard-to-find computer parts. (Freq: 150)
Traduction Systran + Lingway Terminology	StarTech.com est la source des professionnels pour les pièces difficiles à trouver d'ordinateur.
Traduction Systran "nu"	StarTech.com est la source des professionnels pour dur-à-trouve des pièces d'ordinateur.
Comparaison	StarTech.com est la source des professionnels pour <i>dur-à-trouve des les</i> pièces difficiles à trouver d'ordinateur.

La traduction de la base construite par **Lingway Terminology Builder** est meilleure que la traduction mot par mot.

3. Conclusion

La méthodologie et les outils **Lingway Terminology Builder** permettent une construction automatique ou semi-automatique de bases terminologiques qui offre les caractéristiques suivantes :

- Gain de temps** La construction automatique, ou semi-automatique de bases par **Lingway Terminology Builder** permet de disposer des ressources terminologiques très rapidement
- Prix** En conséquence coût de réalisation d'une terminologie est très sensiblement réduit, d'au moins 50% et jusqu'à 75% dans certains cas.
- Qualité** L'approche de traduction par corpus garantit des termes source et cible particulièrement pertinents. De plus une métrique de la confiance en la qualité des traductions est proposée par la solution. Les évaluations réalisées montrent une amélioration extrêmement visible en qualité de traduction
- Evolutivité** L'utilisation de corpus du Web permet un balayage périodique des textes les plus nouveaux de façon à acquérir la terminologie nouvelle, qui pour certains domaines (technologie...) est particulièrement importante