

# White Paper Lingway



**Speeding up the building of  
multilingual terminologies**

**JUNE 2005**

**>lingway**

Solutions in language processing

*New Lingway technology accelerates the building of high quality multilingual terminologies by automatically extracting terms from non-aligned corpus. This innovative capability resolves the problems of aligned corpus scarcity, especially in emerging fields and evolving terminology.*

## Contents

---

- Contents..... 2
- Foreword ..... 3
- 1. Methodology ..... 4
  - Presentation ..... 4
  - Procedures ..... 4
- 2. Results ..... 9
  - Presentation ..... 9
  - Evaluation..... 9
    - Quality ..... 9
    - Examples of results ..... 10
- 3. Conclusion..... 12

## Foreword

---

The global demand for translation keeps growing (both in volume and the quality required), bringing with it the urgent need for more efficient multilingual computer tools to assist translation specialists, as well as to enhance machine translation performance.

One of the most critical and costly tasks in any multilingual project is the **building and enrichment of terminologies**. Up to now the process has been the **real stumbling block to quality translation**, whether carried out by professional translators or by automatic translation systems.

Lingway has introduced the **Lingway Terminology Builder**, a platform that allows users to quickly build terminologies for multilingual applications. The solution, developed within the Lingway KM software, uses a combination of statistical and linguistic tools to extract bilingual terminology from corpus.

**Lingway Terminology Builder** speeds up the process of building pertinent terminologies with attested source and target terms that are both field-specific and relevant to the customer's context.

**Lingway Terminology Builder** substantially reduces the time required to build a terminology (by a factor of up 2 to 3).

The originality and advantage of the Lingway approach lies in its exceptional efficiency and the high quality of the results. Terms (both source and target) are extracted from the corpus most likely to provide the information required. Lingway's bilingual terminology extraction system does not require that corpus be aligned, unlike other solutions in the market.

Below is a description of the procedures and tools we use to build terminologies, as well as some remarks on the results obtained.

## 1. Methodology

---

### **Presentation**

Terminology knowledge bases are of key importance when it comes to translation, as they are directly responsible for the quality of the results. These databases determine the terms that will be used by professional translators, as well by automatic translation systems such as Systran.

A lexical database for a given language can usually be broken down into three levels:

- A hard core (which would consist of the approximately 3,000 commonly used terms),
- A generic or standard language (several tens of thousands of terms),
- Extensions by fields (terminology layer) that can number up to several tens of thousands, and even several hundred thousand terms).

Lingway has generic dictionaries that cover the first two categories, and additional terminology extensions in a number of key technical fields.

Lingway has developed and validated a method for automatically building and enriching terminologies based on these dictionaries and on Lingway KM software.

### **Procedures**

Lingway uses the following procedures to build its terminologies:

- Constitution of a corpus (source and target languages) representative a given field,
- Extraction of terms from the source language corpus,
- Use of the Lingway KM cross-language search engine to identify potential translation options in the target language corpus.

The resulting bilingual terminology base can then be consulted by professional translators or imported to machine translation software (notably the Systran format).

### **Constituting a field-specific corpus**

This is an extremely important step as quality of the translation options depends on the relevance of the corpus selected (both source and target). We estimate that you need a minimum of 40 Mb of text in the source language and approximately twice that amount in the target language to attain an acceptable terminology volume and level of performance. The corpus must be varied enough to produce a terminology truly representative of the field.

### **Exploiting “customer” resources**

In most cases, customers already have corpus related to the field they wish to cover, whether it be internal corporate documentation (patents, articles, company publications) or websites (online publications, meetings, specialized or company websites, blogs, etc.) Corporate-related single language or bilingual terminologies can also be used.

### **Exploiting “Web” resources**

If a customer does not have a corpus, or if the corpus available is insufficient, Lingway offers the option of developing a customized corpus using a proven methodology.

The first step in the procedure is to constitute a fairly limited set of field-specific terms in the source language. A first set of documents or web pages containing these terms (or, more exactly, significant sub-sets of the above sets) is extracted: a statistical tool enables you to identify other terms in the field and by process iteration, obtain a relatively large corpus in a given field. There are also lists of textual corpora, publications (political, news, etc), multilingual websites, and other sources that may also be exploited.

Specialists in each of the foreign languages evaluate certain key features for each corpus, such as the field, language level (technical, general, common, formal, etc.), coding, format and the relevance of the corpus to the project in consideration.

## Extraction of source terms

The Lingway KM software performs the task of automatically extracting terms in the source language.

Lingway KM's terminology extractor carries out a linguistic analysis of each sentence in the corpus, and identifies typical or "model" linguistic structures.

The table below shows some examples:

Model	Example
"Adjective + Noun"	<i>laptop computer</i>
"Noun + Preposition + Noun"	<i>state of the art</i>
"Noun + Preposition + Adjective + Noun"	<i>increase of the minimum wage</i>

The potential translation options are filtered on the basis of complementary statistical criteria so that only the most appropriate terms are retained in the final extraction.

One of the elements taken into account is the "in-width" coverage of each term in the corpus. If the corpus is constituted by a few hundred websites and a term, even a frequently used one, appears in only one of the websites, it will not be retained.

## Obtaining translations

This step is the most critical phase of the process. It can be broken down into three sub-steps:

- Production of a large number of potential translation options.
- Calculation of the number of occurrences of the translation option in the target corpus.
- Selection of the best translation option (or, in a few exceptional cases, options).

### Production of a large number of potential translation options

This sub-step exploits the Lingway KM "cross-language" search engine to produce all the different combinations of translations using the knowledge contained in the existing dictionary and terminologies.

If there is the following knowledge, for example:

Source term (English)	Potential Translation (French)
term	condition
term	terme
term	vocable
use	utilisation
use	exercice
use	usage
use	fonction

The Lingway KM cross-language search engine will produce:

Source term (English)	Translation option calculation (French)	Readable version
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

In fact, the English term itself may be listed among potential translation options, as well as translation proposals for words that are not listed in the dictionary, but potential options nonetheless. The latter is particularly true in highly technical fields that employ many neologisms (FR "productisation", "browseur", "déboguer" and "débugger", etc.).

Please note that the EN model <Noun + "of" + Noun> resulted in FR translation options <Noun + "de" + Noun>. In other cases, several target prepositions may be attested, such as EN <Noun + Noun> which may produce FR <Noun + "de" + Noun> as well as <Noun + "en" + Noun>.

These transformations are used to obtain correct terms, such as (Medical field):

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

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

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

### Counting of translation option occurrences in the target corpus

This means counting the number of occurrences of a translation option.

For instance, the previous example might result in:

Translation Option (French)	Number of occurrences in the target corpus (French)
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

The following example in the Medical field is particularly informative because the translation options include those drawn from existing terminologies and others found by the Lingway Terminology Builder. Listed are the various translation options for the EN term "hearing loss":

Translation Option (French)	Number of occurrences in the target corpus (French)
perte auditive	192
perte d'audition/perde de l'audition	155
surdité partielle	7
baisse de l'acuité auditive	4
perte de l'ouïe	1
perte audition	1
déperdition auditive	0
déperdition de l'audition	0
perte de la faculté d'entendre	0
déperdition de la faculté d'entendre	0
déperdition de l'ouïe	0
perte pour l'audition	0
perte en audition	0
(...)	0

### Selection of the best translation option or options

Several elements are taken into consideration in the selection process:

- Relative frequency of the translation options (the most frequent options are retained and the less frequent eliminated).
- Absolute frequency of the translation options (options that are hardly ever used are eliminated).
- Frequency of the translation options in relation to the frequency of the source term (options whose frequency is abnormally lower in the target language than that of its source term in the source corpus are eliminated).

In the first example, the only translation retained will be:

Source term (english)	Translation option selected (French)
term of use	<i>conditions d'utilisation</i>

In other cases, several translation options may eventually be retained, as in the next example:

Source term (English)	Translation option selected (French)	Target frequency
hearing loss	perte auditive	<b>192</b>
term of use	perte d'audition/perte de l'audition	<b>155</b>

The second term comes from an existing terminology, but not the most frequent. Please note two other translation options, "surdité partielle" and "baisse de l'acuité auditive" also came from existing terminologies but as they appeared far less frequently than the first two, were not validated by the target corpus.

## 2. Results

---

### **Presentation**

Below is a description of the results obtained for a project carried out by Lingway for Systran, the automatic translation company. The project involved building three large terminology bases for the fields of Computer Science (18,000 terms), Medicine (27,000 terms) and Finance (6,000 terms). These terminologies were automatically exported to the Systran format and integrated into the **Systran 5.0 Professional Premium** solution, with quality control ensured by Systran teams.

### **Evaluation**

#### **Quality**

Systran integrated the terminology bases produced by Lingway in its translation software and compared its performance to that of its existing system with the help of its SLP (**Systran Linguistic Platform**). The platform displayed the following elements for the corpus under evaluation:

- Each source sentence.
- The machine translation with the new terminology base.
- The machine translation without the new terminology base.
- A comparison of the two translations.

The SLP system also provided a professional translator with the opportunity to offer his opinion on the improvement, regression or equivalency of each sentence produced by the new terminology base.

The SLP platform then calculated the percentage of improvements, regressions or equivalencies for the whole corpus.

The initial results obtained by integrating the **Lingway Terminology Builder's** base showed substantial improvement in the quality of the translations: +49%.

A small regression percentage was noted in the first phase of the analysis. These regressions are often due to the evaluation method, which privileges terms listed in traditional dictionaries, while Lingway privileges the terms attested in the target corpus: the term "*driver Linux*" proposed by Lingway is badly graded compared to "*module de gestion de périphérique de Linux*» for instance. The latter is a more cumbersome translation but, even more important, was not a translation option attested in the corpus.

Another reason for some of the regressions is linked to the side effects of the translation software (added pairs can block some of the software rules).

Another difficulty encountered, and since resolved, was connected to the problem of surface form, in particular pairs that did not agree in number (singular/plural).

Once the side effects were eliminated, the results were as follows:

Category	Percentage
Improvement	53%
Regression	3%
Equivalence	44%

The balance sheet is definitely positive and the quality of the translations using the terminology bases unquestionably superior.

## Examples of results

Below are some examples of the results obtained in the field of computer science. The FR terms provided by the **Lingway Terminology Builder** are indicated in **green**

Source	[Site Info] [Privacy Policy] [Advertise] (Freq: 1153)
Systran Translation + Lingway Terminology	[Infos du site] [Politique de confidentialité] [annoncez]
"Plain" Systran Translation	[Information de site] [politique d'intimité] [annoncez]
Comparison	[Information de [Infos du site] [politique d'intimité] [Politique de confidentialité] [annoncez]

In this first example, you can note the value of the approach in the choice of the word "confidentialité," which fits the context (computer science) far better than "intimité."

Source	Product Submission Form : (Freq: 633)
Systran Translation + Lingway Terminology	Formulaire de soumission de produit :
"Plain" Systran Translation	Forme de soumission de produit :
Comparison	Forme Formulaire de soumission de produit :

Below, Lingway uses the word "Formulaire» to translate the English word "Form," rather than "Forme" which is less appropriate to the context.

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

In this very interesting example, it's not the literal translation that's in question. Les words "Windows" and "XP" are recognized as belonging to the same term "Windows XP," and remain linked in the translation.

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

Same as the above, but this time for "**Windows XP Professionnel**".

<b>Source</b>	<b>content creation (Freq: 176)</b>
<b>Systran Translation + Lingway Terminology</b>	<b>création de contenu</b>
<b>"Plain" Systran Translation</b>	création contente
<b>Comparison</b>	création <i>contente de contenu</i>

No remarks needed.

<b>Source</b>	<b>Company Info (Freq: 176)</b>
<b>Systran Translation + Lingway Terminology</b>	<b>Infos société</b>
<b>"Plain" Systran Translation</b>	Information de compagnie
<b>Comparison</b>	<i>Information de compagnie</i> <i>Infos société</i>

The translation "**Infos société**" is definitely easier to understand than "information de compagnie."

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

The translation provided by the **Lingway Terminology Builder** base is superior to the literal translation.

### 3. Conclusion

---

**Lingway Terminology Builder's** methodology and tools allow users to automatically or semi-automatically build terminologies, with the resulting advantages:

- Gain in Time**      **Lingway Terminology Builder's** ability to construct terminologies automatically or semi-automatically means that your terminology resources are rapidly operational.
- Lower Costs**      As a result, the cost of building terminologies is substantially reduced, by at least 50% and up to 75% in certain cases.
- Better Quality**      The corpus approach to translation guarantees the pertinence of attested source and target terms. In addition, the solution offers a tool to measure the level of confidence in the quality of the translation. Evaluations already carried out show a marked improvement in translation quality.
- Greater Relevance**      The use of Web corpus enables users to regularly scan the Internet for updated content and to acquire the latest terminology, which is particularly important in fields such as technology.