>>> revista tradumàtica
Traducció i Tecnologies de la Informació i la Comunicació
Número 06: Terminologia i traducció : desembre 2008 : http://www.fti.uab.cat/tradumatica/revista :
ISSN: 1578-7559

# Terminologia i traducció automàtica

**Fiona Bell, Mathias Lemke**
**Translendium SL**

**Resum:** Aquest article descriu les característiques de funcionament del sistema de traducció automàtica Translendium. També detalla l'estructura del mòdul terminològic dins del sistema així com el procés per a introduir-hi nova terminologia i validar-la. Aquest procés s'il·lustra amb un cas real.
**Paraules clau:** Terminologia, traducció automàtica, gestió terminològica, control de qualitat.

**Resumen:** Este artículo describe las características de funcionamiento del sistema de traducción automática Translendium. También detalla la estructura del módulo terminológico dentro del sistema, así como el proceso que se sigue para introducir nueva terminología y validarla. Este proceso se ilustra con un caso real.
**Palabras clave:** Terminología, traducción automática, gestión terminológica, control de calidad.

**Abstract:** This article describes the characteristics of the automatic translation programme Translendium. It also gives a detailed description of the programme's terminology module as well as the process for entering and validating new terminology. This process is illustrated with a real case study.
**Key words:** Terminology, machine translation, terminology management, quality assignment.

## 1. Introducció

Sembla evident que una major cobertura lèxica d'un programa de traducció automàtica ha de tenir necessàriament com a conseqüència l'augment de la qualitat de traducció. Per cobertura lèxica es pot entendre la cobertura del vocabulari general però també la cobertura de termes fets servir en un o més camps d'especialitat concrets.
Així, si alimentem un programa de traducció automàtica amb molts termes de diferents especialitats, en principi la qualitat de la traducció automàtica s'hauria de veure afectada positivament. Aquest no és sempre el cas, però.
El cas que presentem en aquest article representa un cas òptim d'importació i la posterior utilització terminològica en un sistema de TA.

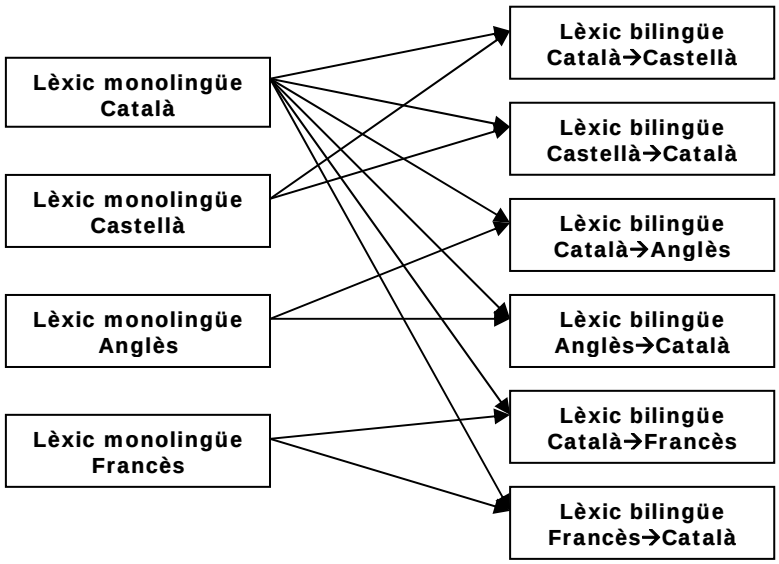
## 2. L'arquitectura del sistema de TA de Translendium

El traductor *de Translendium* és un sistema de traducció modular basat en regles i amb un enfocament de transferència que es pot desglossar en subtasques d'anàlisi, transferència i generació. Cada una d'aquestes subtasques fa servir mòduls de coneixement lingüístic diferents.

Una direcció de traducció utilitza dos tipus de mòduls lingüístics (els diccionaris i les gramàtiques) i un mòdul de processament (l'anomenat nucli del sistema o kernel). Tant els diccionaris com les gramàtiques són modulars. És a dir, cada direcció de traducció consta de mòduls lèxics monolingües i bilingües i de mòduls gramaticals d'anàlisi, de transferència i de

generació. Aquesta modularitat del sistema de traducció permet el manteniment i el desenvolupament eficients dels diferents components lingüístics perquè són compartits per les corresponents direccions de traducció.

**Mòduls lèxics objecte de la importació terminològica:**




Cada parell de llengües conté dos diccionaris monolingües i dos diccionaris bilingües, un per cada direcció de traducció.

Els lèxics monolingües de *Translendium* estan formats per un conjunt d'entrades amb la corresponent informació morfològica, sintàctica i semàntica formalitzada en forma de trets i valors. Un dels lèxics, el lèxic de la llengua d'origen, és el lèxic utilitzat pel sistema per analitzar el text d'entrada que s'ha de traduir. L'altre, el lèxic de la llengua de destinació, s'utilitza per generar correctament les paraules en la traducció. Un mateix lèxic monolingüe serveix tant per a la fase d'anàlisi com per a la fase de generació.

Els lèxics bilingües contenen les d'entrades bilingües per a una direcció de traducció amb la possibilitat de tenir informació de selecció lèxica contextual i/o de transformacions <estructurals associades. Les entrades s'identifiquen a partir del lema de la llengua d'origen (SLCAN), el lema de la llengua de destinació (TLCAN) i les categories gramaticals de la llengua d'origen i de destinació (SLCAT i TLCAT). A més dels lemes i les categories gramaticals de la llengua d'origen i de destinació, un altre dels trets lèxics obligatori és l'àrea temàtica (TAG).

L'àrea temàtica (o TAG) és la indicació del camp d'especialitat en el qual s'aplica l'entrada bilingüe. Alguns mots es poden traduir de forma diferent segons el context del camp d'especialitat en el qual s'utilitzen. Els lèxics de *Translendium* tenen una llista predefinida d'àrees temàtiques organitzades en forma de jerarquia. L'usuari pot fer servir les àrees temàtiques ja existents però pot també crear àrees temàtiques noves i assignar-hi les entrades bilingües corresponents.

**3. Presentació d'un cas d'èxit**

**Importació de la terminologia de salut al programa de TA de la Generalitat de Catalunya.**

A finals del 2006 *Translendium* va realitzar una importació massiva de terminologia del camp de la medicina per incorporar al traductor de la *Generalitat* de Catalunya. La terminologia lliurada per a la importació provenia del *TermCat*, el Centre Terminologia de Catalunya, i incloïa equivalències bilingües de termes en les sis direccions de traducció: català⇔castellà, català⇔anglès i català⇔francès.
Les equivalències dels sistemes català⇔castellà i català⇔anglès eren de prop de 20.000 termes per direcció. Les equivalències del català⇔francès eren d'aproximadament la meitat, 10.000 equivalències terminològiques.
En tots els casos, la ingent quantitat de termes feia que la importació pogués tenir grans repercussions sobre el lèxic de sistema, que té aproximadament entre 40.000 i 50.000 entrades per direcció de traducció, i les gramàtiques del traductor, amb la qual cosa la importació necessitava un procés d'avaluació posterior per assegurar-ne la validesa i comprovar realment l'augment de la qualitat de la traducció.

**4. Importació de terminologia**

El procés d'importació de terminologia mèdica en el sistema de *Translendium*, clarament determinat per l'arquitectura lèxica abans esmentada, va implicar:

1. Importació d'equivalències de termes bilingües en un mòdul terminològic de medicina en els tres parells de llengües català⇔castellà, català⇔anglès i català⇔francès.
   - P. ex. la llista d'equivalències català-anglès facilitada pel TermCat era del tipus:

| | | | |
|---|---|---|---|
| *dineïna* | *f* | *dynein* | *n* |
| *diniprofil·lina* | *f* | *diniprofylline* | *n* |
| *dinitolmida* | *f* | *dinitolmide* | *n* |
| *dinitrat d'isosorbida* | *m* | *isosorbide dinitrate* | *n* |

2. Creació de les corresponents entrades monolingües en quatre diccionaris: català, castellà, francès i anglès. A causa de la modularitat dels lèxics del sistema de *Translendium*, la importació al lèxic monolingüe del català va suposar una sola importació de tots els termes catalans que apareixien en les equivalències bilingües de les sis direccions de traducció.

   El procés de creació de les entrades monolingües en el sistema de *Translendium* és en si mateix un procés gairebé automàtic en el qual les eines lingüístiques específiques creen per defecte les entrades tant simples com complexes (o multimots) a partir de l'anàlisi morfològica i una sèrie d'assumpcions i generalitats aplicades al lèxic. En aquest cas el procés de creació quasi no va necessitar una revisió posterior perquè es complien uns requisits òptims per a la importació:

   - Els llistats d'equivalències originals contenien, a més de l'entrada del terme, informació morfològica (categoria lèxica, gènere i nombre), amb la qual cosa era

segur i fiable generar els trets morfològics i sintàctics de les entrades sense incórrer en errors.

- La gran majoria de termes eren de categoria nominal, la categoria que menys complexitat sintàctica i morfològica té..
- Els termes eren correctes des del punt de vista formal i no necessitaven revisió. Els termes no contenien marques de puntuació ni es tractava de cadenes de caràcters massa llargues ni de sintaxi confusa.

Les entrades monolingües simples i compostes es van crear de forma separada:

- Per crear per defecte els trets morfològics, sintàctics i semàntics es va adaptar l'eina lèxica de creació de les noves entrades simples definint els trets més probables segons els sufixos que contenien els termes. La sufixació en terminologia mèdica és freqüent i molt regular, amb la qual cosa les generalitats que se'n podien derivar són molt fiables i en general necessiten poca revisió posterior.

    - P. ex. les paraules catalanes acabades en "*-dòncia*" són paraules femenines que flexionen en "*-dòncies*" al plural i denoten un procés (PRO) → "*eritrodòncia*", "*oclusodòncia*".
    - P. ex. les paraules castellanes acabades en "*-caína*" denoten un material (MAT) i són també femenines i flexionen en "*-caínas*" en plural → "*benzocaína*", "*cincocaína*".

- La creació dels mots compostos, o multimots en terminologia de *Translendium*, es va poder dur a terme directament gràcies a la informació morfològica facilitada amb les entrades i també gràcies al fet que els models de mots compostos coincidien amb els models que preveu la base de dades lèxica de *Translendium*.

    - P. ex. "*oclusión en tijera de Brodie*" → nom femení al qual se li pot assignar el model NST-STRING (nom+part invariable), on "*oclusión*" és el nucli del multimot i "*en tijera de Brodie*" és una cadena de caràcters invariable.
    - P. ex. "*bifurcació aòrtica*" → nom femení al qual se li pot assignar el model de multimot NST-VAR (nom+part variable), on "*bifurcació*" és el nucli del multimot i "*aòrtic*" és l'adjectiu que el qualifica.

El procés de creació de les entrades bilingües va ser molt més senzill, ja que solament calia afegir la marca d'àrea temàtica (TAG) en la qual s'aplicava el terme, en aquest cas el de medicina, i en alguns casos, pocs, s'havien d'establir certes restriccions en certes entrades per aconseguir controlar millor les traduccions:

- P. ex. Una equivalència terminològica entre català i francès com

    part m mise bas m

    havia de fer constar la restricció de gènere del masculí per diferenciar-la de l'equivalència ja existent restringida al femení de "*part*" → "*partie*" per evitar traduccions del tipus:

    La **part** final del trajecte circumval·la les granges, travessa el torrent i acaba tornant al punt de partida

La ***mise basse** finale du trajet entoure les fermes, <A\[il|elle\]> traverse le torrent et <A\[il|elle\]> finit par rendre au point de départ

Similarment la traducció del mot "urgència" podia ser doble en funció del nombre gramatical que traduïa i això havia de quedar reflectit en el lèxic bilingüe del sistema. Introduint una restricció de nombre segons si el terme apareixia en singular o plural en el text d'origen, es podia destriar la traducció que s'havia de generar en el text de sortida.

| | | | |
|---|---|---|---|
| urgència | f | urgence | f |
| urgències | f pl | service d'urgence | m |

Tots els hospitals tenen **urgències**
Tous les hôpitaux ont **service d'urgence**

La **urgència** és deguda a l'escassetat de pluja
L'**urgence** est due au manque de pluie

## 5. El procés d'avaluació i validació de la importació terminològica

La ingent quantitat de termes importats en el lèxic feia necessari un procés de validació exhaustiu que d'una banda assegurés que la terminologia no desestabilitzava el sistema interferint en l'anàlisi dels textos i en les traduccions del lèxic de vocabulari general i, per l'altra, que asseverés l'augment de qualitat de traducció de textos de l'àrea temàtica mèdica gràcies als nous termes importats. Així el test va implicar:

1. La traducció dels textos generals de referència utilitzats a *Translendium* i la comparació amb la versió instal·lada en el traductor de la *Generalitat* per cada parell de llengües. Durant el procés de desenvolupament dels lèxics de *Translendium* i cada cop que es lliura una nova versió, es tradueixen un nombre determinat d'unitats de traducció i es comparen amb la versió immediatament anterior. D'aquesta manera es pot garantir la millora de la qualitat de traducció.
   El resultats de traducció obtinguts no mostraven cap interferència entre la terminologia i el lèxic de sistema. La gran quantitat de terminologia importada en el sistema no afectava les gramàtiques amb possibles casos nous d'homografia lèxica degut a la poca ambigüitat dels termes mèdics.

2. La traducció de textos d'especialitat mèdica lliurats pel mateix *Departament de Salut*. Aquests textos eren un exemple real dels textos que serien objecte de la traducció automàtica amb la terminologia. Les millores que van sorgir en aquest procés són de dos tipus:

   - Termes que apareixien en els textos i que quedaven sense traduir primer perquè no estaven coberts en el lèxic de sistema, ara es traduïen.
     - P. ex.
       *La malaltia de Kawasaki és una vasculitis sistèmica aguda;*
       *La maladie de Kawasaki est une ***vasculitis** systémique aiguë;*
       *La maladie de Kawasaki est une **vascularite** systémique aiguë;*

*Los pacientes en los que se identifican micobacterias no tuberculosas no se consideran casos de TB.*
*Els pacients en qui s'identifiquen* ****micobacterias*** *no tuberculoses no es consideren casos de TB.*
*Els pacients en qui s'identifiquen* ***micobacteris*** *no tuberculosos no es consideren casos de TB.*

- Algunes paraules que podrien tenir més d'una traducció ara es traduïen pel terme més adient en un context mèdic.

  - P. ex.
    *La mortalidad habitual de la gripe oscila entre el 0,06-0,18% de las personas que enferman, aunque depende de la cepa gripal causante del proceso epidémico.*
    *La mortalitat habitual de la grip oscil·la entre el 0,06-0,18%-0,18 de les persones que emmalalteixen, encara que depèn del* ***<A\[*cep|soca\]> gripal*** *causant del procés epidèmic.*
    *La mortalitat habitual de la grip oscil·la entre el 0,06-0,18%-0,18 de les persones que emmalalteixen, encara que depèn de la* ***soca gripal*** *causant del procés epidèmic.*

    *Hi ha més de cent tipus diferents de VPH, dels quals més de trenta poden infectar les mucoses del tracte genital dels dos sexes.*
    *Hay más de cien tipos diferentes de VPH, de los cuales más de treinta pueden infectar las mucosas del* ****trato*** *genital de los dos sexos.*
    *Hay más de cien tipos diferentes de VPH, de los cuales más de treinta pueden infectar las mucosas del tracto genital de los dos sexos.*

    *Evolució de la presa d'estròg*ens locals i generals.
    *Evolución de la* ***<A\[*presa|toma\]>*** *de estrógenos locales y generales.*
    *Evolución de la* ***toma*** *de estrógenos locales y generales.*

## 6. Conclusions

La importació massiva de terminologia en els lèxics d'un sistema de traducció automàtica pot donar resultats molt positius sempre i quan aquesta importació compleixi certs requisits:

- La terminologia no pot ser vaga ni imprecisa. Com més específica sigui més unívoca serà i menys distorsió crearà en la traducció de vocabulari general. Per unívoca s'entén terminologia tan específica d'un camp especialitat que és poc probable trobar-la en usos generals o d'altres camps i també, per tant, monosèmica, és a dir terminologia que remet a un sol concepte. La terminologia mèdica és un bon exemple d'especificitat i univocitat terminològica.
- La terminologia en un sistema de TA ha d'estar ben delimitada i marcada per etiquetes d'àrea temàtica o camp d'especialitat. Aquesta modularitat és necessària per poder escollir la terminologia adient quan es tradueixen textos d'un camp d'especialitat concret. En el cas del sistema de traducció automàtica de *Translendium*, el programa pot dur a

terme aquesta elecció seleccionant del lèxic les traduccions corresponents a un camp d'especialitat concret.

- La terminologia de categoria nominal és la més freqüent però també la que respon millor quan s'utilitza en traducció automàtica pel fet que les seves implicacions pel que fa a les gramàtiques són mínimes i incideixen poc en l'anàlisi i generació sintàctica de les llengües de les quals i a les quals es tradueix.
- La importació de la terminologia en un sistema de traducció automàtica complex com és el de *Translendium* és més ràpida i efectiva i necessita menys revisió si els llistats terminològics inclouen informació morfològica dels termes.
- La terminologia ha de ser plenament correcta i estable pel que fa a la forma, és a dir, no ha de presentar errors ortogràfics ni errors de concordança.
- La terminologia ha de ser d'ús general dins del camp d'especialitat i tenir objectius clars de traducció. Ha de ser, doncs, terminologia activa i actualitzada que es pugui donar en textos del camp d'especialitat que cobreix.