

1. La traducció assistida per ordinador

Índex

1.1. Introducció

La informàtica s'ha introduït plenament en tots els sectors productius i activitats professionals. La traducció no és una excepció i des de fa ja anys el procés de traducció es du a terme gairebé sempre fent servir un ordinador. Molt sovint l'ús de l'ordinador es limita a aplicacions ofimàtiques estàndard, especialment processadors de textos. Les consultes a fonts d'informació habituals en la traducció (diccionaris generals i terminològics, enciclopèdies, etc.) es fa ja també de manera generalitzada a través d'un ordinador, ja sigui mitjançant la consulta de recursos instal·lats a la màquina o mitjançant la consulta a llocs webs específics. L'ordinador ha esdevingut també el mitjà principal de comunicació entre el traductor i els seus clients, especialment mitjançant el correu electrònic. Però rellevar a l'ordinador a ser un simple substitut d'una màquina d'escriure, un diccionari i un servei de correus eficient és menysprear les seves enormes possibilitats.

Els *programes de traducció assistida per ordinador* engloben un seguit d'aplicacions informàtiques especialment dissenyades per a assistir de manera eficient al traductor en la seva tasca.

En un sentit ampli, els *sistemes de traducció assistida* engloben totes les aplicacions informàtiques dissenyades per a tasques específiques del procés de traducció.

En un sentit més específic, generalment es parla de sistemes de traducció assistida quan ens referim als programes que ajuden a traduir a partir de consultes a una o diverses memòries de traducció i de manera opcional a un o més glossaris terminològics. Tanta importància té el concepte de memòria de traducció que sovint s'ha anomenat als sistemes de traducció com a *memòries de traducció* o bé *sistemes de gestió de memòries de traducció*.

Lippmann (1971) feia una descripció pionera del concepte de traducció assistida per ordinador:

> *Computer-aided translation (CAT) is a storage and retrieval operation carried out on line with a computer during the time in which a translation is produced. A system of dictionary access and updating routines, text-processing facilities, and on-line utilities is designed to telescope the delay between the initiation of a translation and its finished print out. The system does not attempt to simulate the human translator by producing an autonomous translation via programmed algorithms; rather, it serves as an extension of the capabilities of the user, who is able to call on the resources of the computer as needed in the process of translation and get an immediate response. Under the system described, users communicate over ordinary telephone lines with the computer by means of remote terminals. In employing the system, the user can switch back and forth as many times as required among human translation, direct dictionary look up, editing, printing, and system interrogation, and thereby achieve rapid interaction toward the desired goal, i.e. a finished translation.*

Així, un dels objectius principals d'un sistema de traducció assistida és posar a l'abast del traductor de manera automàtica i ràpida tots els recursos que el puguin resultar d'utilitat. Per fer la seva tasca un traductor habitualment consulta diccionaris generals, terminològics i enciclopèdics. Aquests diccionaris poden estar en paper i poden constituir diversos volums. Això fa que la consulta manual d'aquests recursos pugui resultar molt costosa en termes de temps. Un recurs molt interessant, però a la vegada difícil de gestionar de manera manual, són els exemples de traduccions anteriors, ja sigui fetes pel mateix traductor com per un altre professional. Sovint, quan es tradueix una determinada oració es té la sensació de ja haver-la traduït anteriorment. El fet de disposar d'un registre de traduccions accessible de manera fàcil i ràpida pot suposar un estalvi de temps important. Els sistemes de traducció assistida, com veurem més endavant, ens donen un accés automàtic i immediat a tots aquest recursos.

El procés de traducció amb un sistema de traducció assistida es divideix de manera genèrica en els següents passos:

- **Tractament de format**: el sistema de traducció assistida ens permetrà crear projectes de traducció a partir d'arxius en diferents formats. D'aquesta manera, amb una única eina

podem traduir arxius en diferents formats sense la necessitat de disposar un gran conjunt d'eines específiques.

- **Segmentació**: La traducció del text original es fa en petites unitats. Aquestes unitats s'anomenen *segments* i en general es pot assimilar un segment a una única oració (com veurem més endavant no sempre coincideix un segment amb una oració). La finalitat d'aquesta divisió en segments o *segmentació* és bàsicament la consulta a memòries de traducció. La probabilitat de trobar una unitat semblant a la unitat que estem traduint és inversament proporcional a la mida d'aquesta unitat. Per tant, és més probable trobar una oració similar a la memòria que no pas un paràgraf.
- **Assignació de recursos**: un cop creat el projecte de traducció, i per tal de treure profit de totes les possibilitats de l'eina de traducció assistida, és imprescindible assignar-li una sèrie de recursos, principalment glossaris terminològics i memòries de traducció.
- **Traducció**: aquest pas es fa també segment a segment i l'eina mostra la informació rellevant (coincidències exactes i parcials a les memòries de traducció, entrades dels glossaris presents al segment a traduir, etc.) d'una manera clara.
- **Creació de recursos a mesura que es va traduint***:* durant la traducció d'un projecte és un bon moment per crear nous recursos. Les memòries de traducció es creen de manera totalment automàtica durant la traducció. Mentre anem traduint podem incloure noves entrades als glossaris terminològics. D'aquesta manera ens estalviarem noves consultes en el futur.
- **Revisió dins de l'eina de traducció assistida**: Aquestes eines disposen de diferents funcions per assegurar la qualitat de la traducció. Des de correctors ortogràfics i gramaticals fins la comprovació d'etiquetes i marques de format. És important fer una primera revisió de la traducció dins de l'eina de traducció assistida, abans de la creació dels documents traduïts.
- **Creació dels documents traduïts**: D'igual manera que l'eina és capaç de tractar diversos formats per introduir-los en un projecte de traducció, també és capaç de generar els documents traduïts en el mateix format original, i mantenint en un bon grau la maquetació de l'arxiu.
- **Revisió dels documents traduïts en el seu format original**: És molt important fer una darrera revisió de la traducció un cop exportada al seu format original.

1.2. Components bàsics d'un sistema de traducció assistida

1.2.a Entorn de treball

Entenem per entorn de treball d'un sistema de traducció assistida la seva interfície gràfica. Un dels aspectes més importants de l'entorn de treball és la disposició de la informació en la pantalla. La majoria d'aspectes d'aquesta disposició són personalitzables, però es poden distingir alguns grups generals. La primera distinció es pot fer entre:

- Programes que disposen d'una interfície de treball pròpia. En aquest grup podem classificar un gran nombre d'eines de traducció assistida: OmegaT, Virtaal, SDL, Déja Vu, WordFast Professional, etc. A la figura 1 podem observar la interfície gràfica de Virtaal.

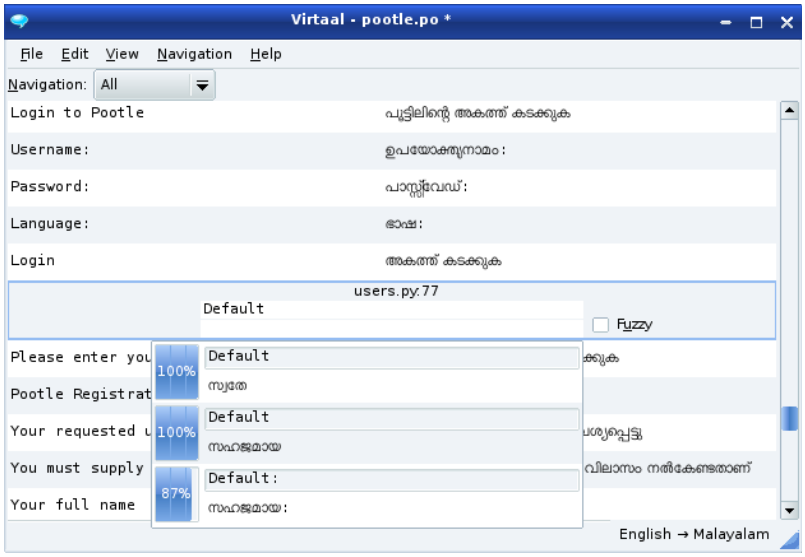


Figura 1: Virtaal

- Programes que s'integren en un altre programa, generalment un processador de textos: Anaphraseus (que s'integra dins d'OpenOffice), WordFast Classic i Trados WorkBench (que s'integren en Microsoft Word). A la figura 2 podem veure Anaphraseus integrat a l'OpenOffice.

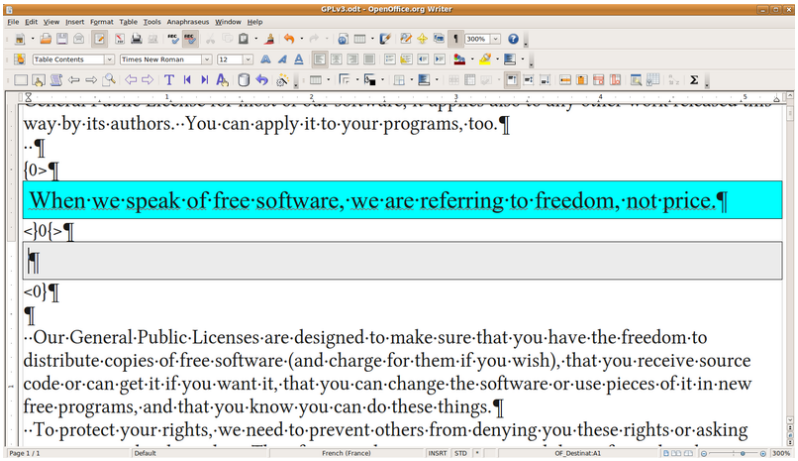


Figura 2: Anaphraseus

Dins del primer grup, és a dir, dels programes que disposen d'una interfície pròpia, podem distingir diferents subgrups atenent a la disposició dels segments originals i traduïts. Així, podem trobar els següents subgrups:

- Segment traduït sota l'original: com per exemple, en OmegaT. Molt sovint, en aquesta disposició, els segments no actius només es mostren en un idioma (generalment l'original si el segment encara no està traduït i la traducció en cas d'haver-se traduït). El segment que estem traduint es mostra en les dues versions. A la figura 3 podem veure la interfície d'OmegaT.

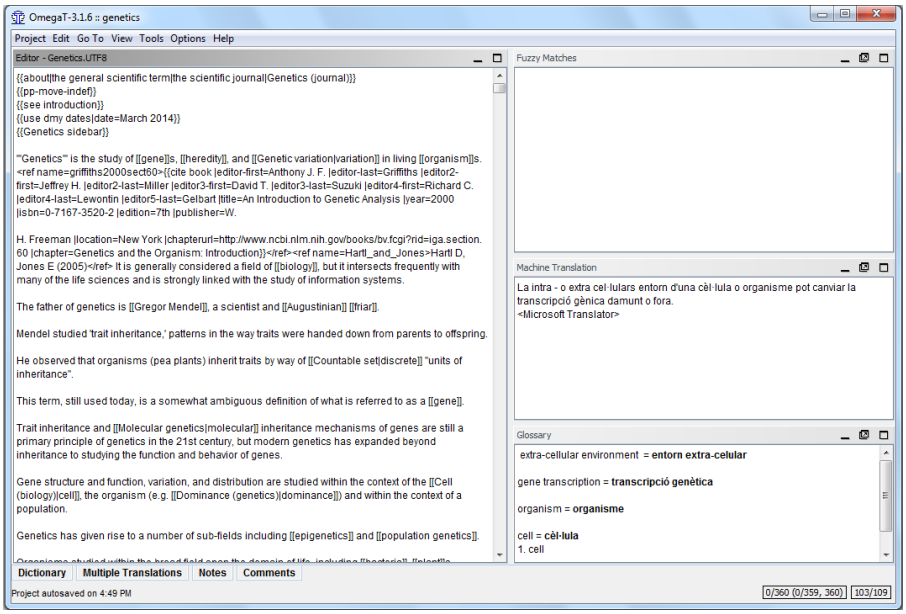


Figura 3: OmegaT

- Disposició en dues columnes: on generalment l'esquerra correspon a l'original i la dreta a la traducció. A la figura 4 podem veure la interfície de Déjà Vu X2.

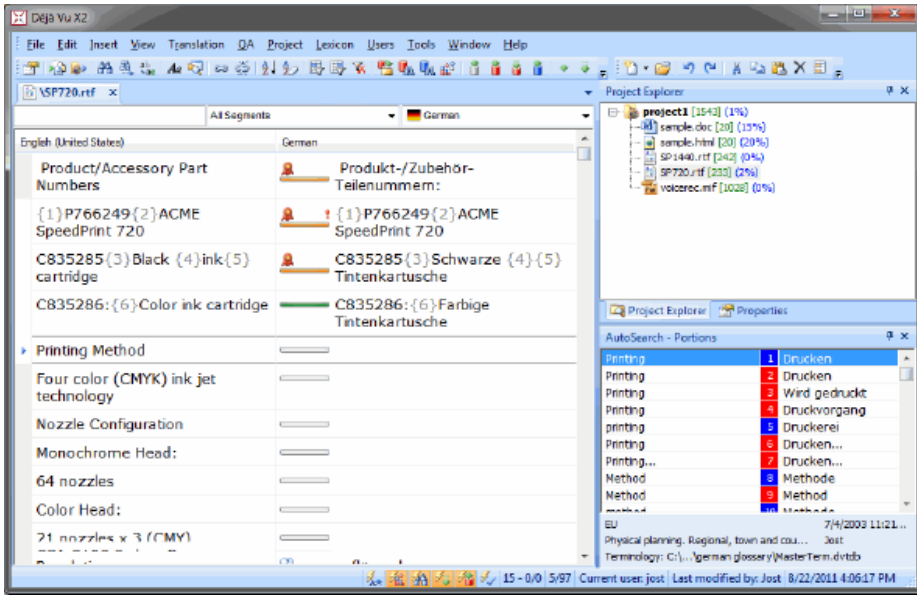


Figura 4: Déjà Vu X2

A més de la disposició del segment original i traduït també és molt important la disposició de les finestres de consulta a memòries de traducció i a bases de dades terminològiques. En general, la posició i mida d'aquestes pantalles és totalment configurable. L'usuari pot modificar la mida i la posició d'aquestes finestres per ajustar-les a les seves necessitats i preferències.

1.2.b. Consulta a memòries de traducció

El sistema de traducció assistida ha de ser capaç de fer consultes a una o més memòries de traducció. L'objectiu de la consulta és trobar segments de la memòria que siguin iguals o similars al segment que estem traduint. Si troba un segment igual estem parlant d'una *coincidència exacta* (*exact match*) i si el segment és semblant parlem de *coincidències parcials* (*fuzzy match*).

Aquesta cerca de segments semblants es fa tant amb els segments emmagatzemats a la memòria de traducció com en els segments traduïts anteriorment dins del mateix projecte. En aquest cas parlem de *repeticions internes*.

L'usuari pot ajustar la similitud mínima que ha de tenir un segment per a que aparegui com a coincidència parcial. El sistema mostrarà totes les coincidències que tinguin una similitud igual o superior a la mínima establerta per l'usuari, i en ordre descendent de similitud, és a dir, començant per la més semblant. Els sistemes de traducció assistida, a més, mostraran en colors les diferències entre el segment que estem traduint i el que apareix a la memòria (Figura 5). Alguns sistemes, a més, intentaran combinar fragments més petits presents a la memòria per a fer una proposta de traducció a partir de les traduccions d'aquests segments.

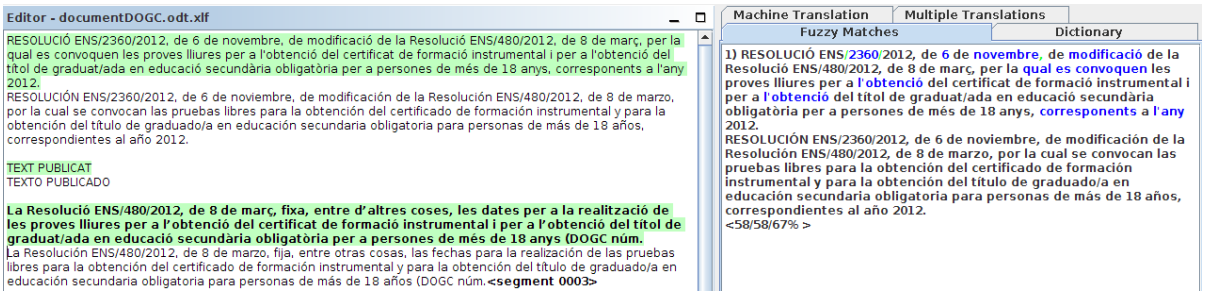

Editor - documentDOGC.odt.xlf
RESOLUCIÓ ENS/2360/2012, de 6 de novembre, de modificació de la Resolució ENS/480/2012, de 8 de març, per la qual es convoquen les proves lliures per a l'obtenció del certificat de formació instrumental i per a l'obtenció del títol de graduat/ada en educació secundària obligatòria per a persones de més de 18 anys, corresponents a l'any 2012.
RESOLUCIÓN ENS/2360/2012, de 6 de noviembre, de modificación de la Resolución ENS/480/2012, de 8 de marzo, por la cual se convocan las pruebas libres para la obtención del certificado de formación instrumental y para la obtención del título de graduado/a en educación secundaria obligatoria para personas de más de 18 años, correspondientes al año 2012.
TEXT PUBLICAT
TEXTO PUBLICADO
La Resolució ENS/480/2012, de 8 de març, fixa, entre d'altres coses, les dates per a la realització de les proves lliures per a l'obtenció del certificat de formació instrumental i per a l'obtenció del títol de graduat/ada en educació secundària obligatòria per a persones de més de 18 anys (DOGC núm.
La Resolución ENS/480/2012, de 8 de marzo, fija, entre otras cosas, las fechas para la realización de las pruebas libres para la obtención del certificado de formación instrumental y para la obtención del título de graduado/a en educación secundaria obligatoria para personas de más de 18 años (DOGC núm.<segment 0003>
Machine Translation
Multiple Translations
Fuzzy Matches
Dictionary
1) RESOLUCIÓ ENS/2360/2012, de 6 de novembre, de modificació de la Resolució ENS/480/2012, de 8 de març, per la qual es convoquen les proves lliures per a l'obtenció del certificat de formació instrumental i per a l'obtenció del títol de graduat/ada en educació secundària obligatòria per a persones de més de 18 anys, corresponents a l'any 2012.
RESOLUCIÓN ENS/2360/2012, de 6 de noviembre, de modificación de la Resolución ENS/480/2012, de 8 de marzo, por la cual se convocan las pruebas libres para la obtención del certificado de formación instrumental y para la obtención del título de graduado/a en educación secundaria obligatoria para personas de más de 18 años, correspondientes al año 2012.
<58/58/67% >

Figura 5. Coincidències parcials a OmegaT

En el següent capítol, dedicat íntegrament a les memòries de traducció veurem a fons tots aquests aspectes i molts altres, com el concepte d'indexació de les memòries per a una cerca ràpida.

1.2.c. Consulta a glossaris i diccionaris

El sistema de traducció assistida també ha de ser capaç de fer consultes a glossaris i diccionaris terminològics. En aquest cas, cada segment que estem traduint pot tenir més d'una entrada del glossari i s'han de mostrar tots els resultats. La principal dificultat per a la cerca a glossaris és fer que sigui capaç de trobar les entrades de manera independent a la forma morfològica en la que es trobi al segment. És a dir, ha de ser capaç de trobar l'entrada *allergic reaction* tant si al text es troba en singular com en plural (*allergic reactions*). Per a llengües morfològicament més riques, com per exemple el català, la tasca es complica, ja que a vegades el plural implicar pluralitzar tots els elements (*reacció al·lèrgica – reaccions al·lèrgiques*) i en altres casos només un d'ells (*lletra de canvi – lletres de canvi*). En llengües amb una morfologia molt més rica encara el terme pot aparèixer en el segment en moltíssimes formes (механическое напряжение, механического напряжения, механические напряжения, механических напряжений...) i totes haurien de mostrar l'entrada corresponent del glossari. Les diferents eines de traducció assistida poden tenir més o menys habilitat en tractar les variants morfològiques dels termes. En el capítol 3, dedicat a les bases de dades terminològiques aprofundirem en aquest tema. A la Figura 6 podem veure la pantalla de terminologia d'OmegaT. Si ens fixem en els resultats, tot i disposar del terme *prova lliure* no s'ha detectat automàticament ja que en l'original apareix en plural.

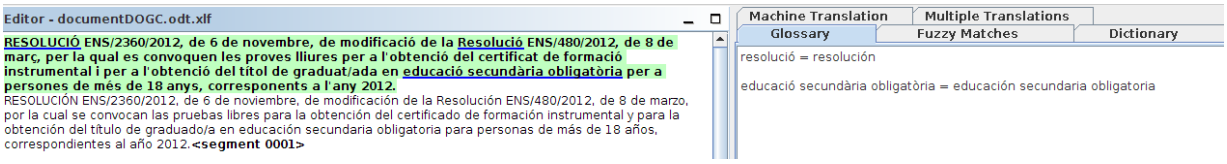

Editor - documentDOGC.odt.xlf
RESOLUCIÓ ENS/2360/2012, de 6 de novembre, de modificació de la Resolució ENS/480/2012, de 8 de març, per la qual es convoquen les proves lliures per a l'obtenció del certificat de formació instrumental i per a l'obtenció del títol de graduat/ada en educació secundària obligatòria per a persones de més de 18 anys, corresponents a l'any 2012.
RESOLUCIÓN ENS/2360/2012, de 6 de noviembre, de modificación de la Resolución ENS/480/2012, de 8 de marzo, por la cual se convocan las pruebas libres para la obtención del certificado de formación instrumental y para la obtención del título de graduado/a en educación secundaria obligatoria para personas de más de 18 años, correspondientes al año 2012.<segment 0001>
Machine Translation
Multiple Translations
Glossary
Fuzzy Matches
Dictionary
resolució = resolución
educació secundària obligatòria = educación secundaria obligatoria

Figura 6. Pantalla de terminologia de l'eina OmegaT

Algunes eines de traducció assistida permeten també la consulta automàtica a diccionaris generals. Aquesta funcionalitat no és tan interessant per a un traductor professional però en certs casos pot resultar d'utilitat. Com que la consulta és totalment automàtica, si es disposa d'aquesta funcionalitat i d'un diccionari adient pot ser bona idea activar aquesta opció en els nostres projectes. A la figura 7 podem observar aquesta funcionalitat a l'OmegaT. Donat que la informació que apareix és molt

extensa, si fem doble clic en una determinada paraula, OmegaT et mostra directament la informació associada a aquella paraula.

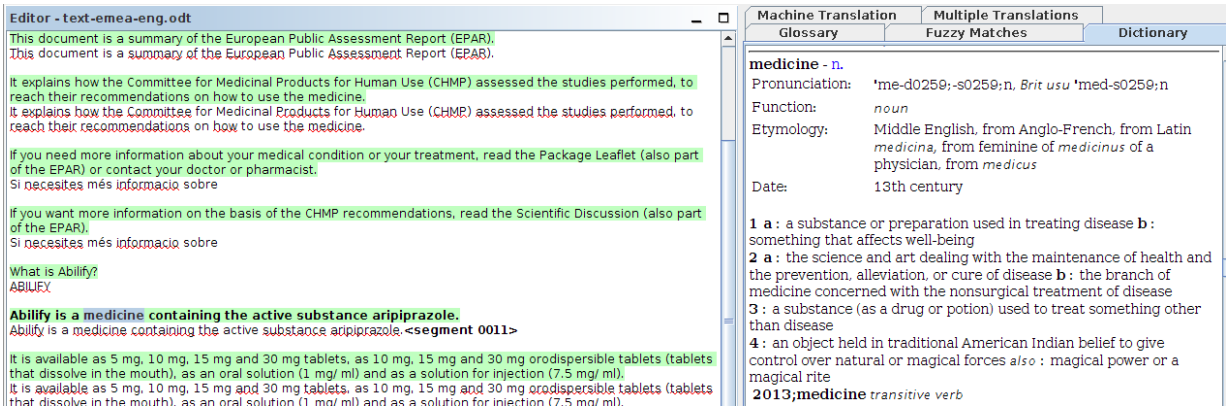

Editor - text-emea-eng.odt
This document is a summary of the European Public Assessment Report (EPAR).
This document is a summary of the European Public Assessment Report (EPAR).
It explains how the Committee for Medicinal Products for Human Use (CHMP) assessed the studies performed, to reach their recommendations on how to use the medicine.
It explains how the Committee for Medicinal Products for Human Use (CHMP) assessed the studies performed, to reach their recommendations on how to use the medicine.
If you need more information about your medical condition or your treatment, read the Package Leaflet (also part of the EPAR) or contact your doctor or pharmacist.
Si necesites més informacio sobre
If you want more information on the basis of the CHMP recommendations, read the Scientific Discussion (also part of the EPAR).
Si necesites més informacio sobre
What is Abilify?
ABILIFY
Abilify is a medicine containing the active substance aripiprazole.
Abilify is a medicine containing the active substance aripiprazole.<segment 0011>
It is available as 5 mg, 10 mg, 15 mg and 30 mg tablets, as 10 mg, 15 mg and 30 mg orodispersible tablets (tablets that dissolve in the mouth), as an oral solution (1 mg/ ml) and as a solution for injection (7.5 mg/ ml).
It is available as 5 mg, 10 mg, 15 mg and 30 mg tablets, as 10 mg, 15 mg and 30 mg orodispersible tablets (tablets
Machine Translation
Multiple Translations
Glossary
Fuzzy Matches
Dictionary
medicine - n.
Pronunciation: 'me-d0259;-s0259;n, Brit usu 'med-s0259;n
Function: noun
Etymology: Middle English, from Anglo-French, from Latin medicina, from feminine of medicinus of a physician, from medicus
Date: 13th century
1 a : a substance or preparation used in treating disease b : something that affects well-being
2 a : the science and art dealing with the maintenance of health and the prevention, alleviation, or cure of disease b : the branch of medicine concerned with the nonsurgical treatment of disease
3 : a substance (as a drug or potion) used to treat something other than disease
4 : an object held in traditional American Indian belief to give control over natural or magical forces also : magical power or a magical rite
2013;medicine transitive verb

Figura 7. Consulta automàtica a un diccionari en OmegaT

1.2.d. Combinació de traducció assistida i traducció automàtica

Molts sistemes de traducció assistida permeten fer consultes a sistemes de traducció automàtica, de manera que a més de presentar els resultats provinents d'una memòria de traducció presenten també el resultat de traduir el segment amb un sistema de traducció automàtica. Aquesta consulta a sistemes de traducció automàtica pot arribar a ser molt productiva per alguns parells de llengües, ja que bona part de les propostes es poden aprofitar fent alguns canvis mínims. Això acostuma a succeir per a parells de llengües prou properes, com per exemple català-castellà, català-francès, etc. Si es fa servir aquesta opció cal anar molt en compte, ja que per les presses tindrem tendència a acceptar com a totalment bones algunes propostes de traducció que no són del tot correctes.

El capítol 4 d'aquest llibre el dedicarem enterament a la traducció automàtica i tornarem a parlar amb més detall d'aquesta combinació. A la figura 8 podem observar la pantalla de Traducció automàtica d'OmegaT en funcionament.

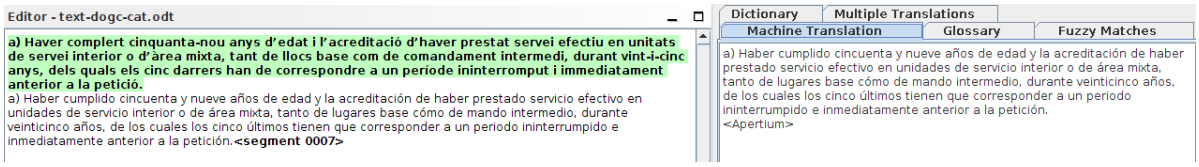

Editor - text-dogc-cat.odt
a) Haver complert cinquanta-nou anys d'edat i l'acreditació d'haver prestat servei efectiu en unitats de servei interior o d'àrea mixta, tant de llocs base com de comandament intermedi, durant vint-i-cinc anys, dels quals els cinc darrers han de correspondre a un període ininterromput i immediatament anterior a la petició.
a) Haber cumplido cincuenta y nueve años de edad y la acreditación de haber prestado servicio efectivo en unidades de servicio interior o de área mixta, tanto de lugares base cómo de mando intermedio, durante veinticinco años, de los cuales los cinco últimos tienen que corresponder a un periodo ininterrumpido e inmediatamente anterior a la petición.<segment 0007>
Dictionary
Multiple Translations
Machine Translation
Glossary
Fuzzy Matches
a) Haber cumplido cincuenta y nueve años de edad y la acreditación de haber prestado servicio efectivo en unidades de servicio interior o de área mixta, tanto de lugares base cómo de mando intermedio, durante veinticinco años, de los cuales los cinco últimos tienen que corresponder a un periodo ininterrumpido e inmediatamente anterior a la petición.
<Apertium>

Figura 8. Traducció automàtica a OmegaT

1.2.e. Tractament de formats

Els sistemes de traducció assistida han de ser capaços de tractar un conjunt variat de formats d'arxiu. Així, han de ser capaços d'importar arxius corresponents a processadors de text (Word, OpenOffice-LibreOffice), html, fitxers propis de llenguatges de programació, etc. El procés d'importació consistirà en seleccionar i mostrar només el text que cal traduir i amagar el corresponent a marques de format. Un cop traduït i revisat el projecte, el programa ha de ser capaç d'exportar la traducció, és a dir, crear els arxius traduïts que han d'estar en el mateix format que l'original.

Així doncs, l'eina de traducció assistida ens permetrà treballar en un mateix entorn amb diversos formats que requereixen habitualment programes d'edició específics. A la taula 1 oferim alguns exemples d'eines i formats que poden tractar[1]

Eina	**Formats**
OmegaT	Plain text, HTML, XHTML, StarOffice, OpenOffice.org, OpenDocument (ODF), MS Office Open XML, Help & Manual, HTML Help Compiler (HCC), LaTeX, DokuWiki, QuarkXPress CopyFlow Gold, DocBook, Android Resource, Java Properties, Typo3 LocManager, Mozilla DTD, Windows RC, WiX, ResX, INI files, XLIFF , PO , SubRip Subtitles, SVG Images
Open Language Tools	XLIFF , HTML/XHTML, XML, DocBook SGML, ASCII, StarOffice/OpenOffice/ODF, PO , .properties, .java (ResourceBundle), .msg/.tmsg (catgets)
SDL Trados	Features four translation environments: dedicated TagEditor, MSWord Interface, SDLX, the integrated interface SDL Trados Studio 2011. Additional filters for translating with TagEditor available: Word, Excel, PowerPoint, OpenOffice , InDesign, QuarkXPress, PageMaker, Interleaf, Framemaker, HTML, SGML, XML, SVG, ... Includes SDL MultiTerm for terminology management and Project Management Dashboard for automating tasks and tracking.
Virtaal	XLIFF, PO and MO, TMX, TBX, Wordfast TM, Qt ts Many others via converters in the Translate Toolkit
WordFast Classic	MS Word, Excel, PowerPoint (for Windows and Mac); tagged documents

Taula 1. Formats tractats per diversos sistemes de traducció assistida

El capítol 5 d'aquest llibre el dediquem íntegrament a aspectes tècnics relacionats amb el tractament de formats.

1 La informació s'ha extret de la Vikipèdia [http://en.wikipedia.org/wiki/Computer-assisted_translation] i pot no estar del tot actualitzada.

1.3. El procés de traducció amb un sistema de traducció assistida

En aquest capítol explicarem el procés genèric de traducció d'un document o projecte de traducció amb un sistema de traducció assistida. Aprofitarem en aquest capítol per parlar en profunditat d'alguns aspectes que no queden recollits en altres capítols: la segmentació i el format SRX, el format XLIFF per a l'intercanvi de projectes de traducció i localització i alguns aspectes relacionats amb els correctors ortogràfics i gramaticals. Tot i que aquests aspectes no es tracten en altres capítols del llibre, es pot trobar informació addicional en els diferents enunciats de les pràctiques associades a aquest llibre.

1.3.a. Tractament del format

En el nostre projecte de traducció haurem de tractar un o més fitxers que estaran en un o més formats informàtics. En la majoria dels casos, l'eina de traducció assistida serà capaç de tractar els formats requerits sense problemes, així que aquest pas acostuma a ser totalment transparent per al traductor i el du a terme sense més preocupacions.

En algunes ocasions, la nostra eina de traducció assistida no serà compatible amb el format del fitxer que hem de traduir. Un cas clar és el tractament del format doc de Microsoft Word. No totes les eines de traducció assistida el poden importar. Per exemple, OmegaT no és compatible amb aquest format. Cal recordar que és un format propietari i que a més les eines que sí que són compatibles amb doc requereixen que tinguis el Microsoft Word instal·lat al sistema, condició que no es dóna sempre. En casos com aquests, la solució és senzilla i funciona en la majoria d'ocasions: transformar l'arxiu doc a un arxiu compatible (com el format ODF emprat per LibreOffice i OpenOffice i considerat com a estàndard per la I.S.O.; o bé, el docx de Microsoft). La conversió acostuma a funcionar sense entrebancs. Un cop finalitzat el procés de traducció es podrà fer la conversió inversa sense més inconvenients.

En altres casos més complexos, és possible que no hi hagi un format intermedi disponible per a la nostra eina de traducció assistida. En molts d'aquests casos la solució estarà en la utilització del format XLIFF que s'explica en aquest mateix capítol a la secció 1.3.c. Molt probablement hi haurà algun programa que pugui transformar el format en qüestió en aquest format estàndard, l'XLIFF, que és compatible amb la gran majoria d'eines de traducció assistida.

1.3.b. Segmentació: el format SRX

El procés de segmentació consisteix a dividir el text d'entrada en unitats d'una mida adequada para poder presentar-les una darrera l'altra al traductor. Les consultes als diferents recursos, com per exemple les memòries de traducció es faran en aquestes unitats de text. No convé que aquestes unitats siguin massa grans (per exemple un paràgraf sencer) perquè la probabilitat de trobar fragments iguals o semblants a la memòria de traducció és menor si el fragment és gran. Tampoc convé que les unitats siguin massa petites, per exemple, una o dues paraules, ja que la unitat i coherència del text es trencaria i faria impossible la seva traducció. Així, la mida ideal acostuma a ser quelcom semblant a una frase o oració. A cada una d'aquestes unitats se'ls anomena *segment* i al

procés de crear-los *segmentació*. Ara bé, les eines de traducció assistida no acostumen a incorporar gaire coneixement lingüístic i per aquest motiu la segmentació es fa a partir d'elements textuals, com signes de puntuació, presència de caràcters en majúscules, etc. El primer que ens ve al cap és que segmentarem per punts ".", però això no sempre funciona. Veiem aquest exemple:

El Sr. Martínez vindrà amb l'A.V.E. de les 15.30 h. Després es reunirà amb el Dr. Pérez en el nostre despatx de l'Av. Diagonal.

Com podem observar, una segmentació basada únicament en punts "." produiria una gran quantitat de segments.

El Sr.

Martínez vindrà amb l'A.

V.

E.

de les 15.

30 h.

Després es reunirà amb el Dr.

Pérez en el nostre despatx de l'Av.

Diagonal.

Les regles de segmentació s'acostumen a definir mitjançant *expressions regulars* que defineixen punts de possibles talls de segments i especifiquen si s'ha de produir el tall o no.

> Una *expressió regular* (o col·loquialment anomenades *regexp*, acrònim de l'anglès *regular expression*) és una representació, segons unes regles sintàctiques d'un llenguatge formal, d'una porció de text genèric a buscar dins d'un altre text, com per exemple uns caràcters, paraules o patrons de text concrets. [font Vikipèdia]

Hi ha un llenguatge XML estàndard per a la definició de regles de segmentació: el format SRX (*Segmentation Rule eXchange*). A continuació mostrem un SRX molt simple, que només defineix un parell de regles.

```
<?xml version="1.0" encoding="UTF-8"?>
<srx xmlns="http://www.lisa.org/srx20" xmlns:okpsrx="http://okapi.sf.net/srx-ext
ensions" version="2.0">
<body>
<languagerules>
<languagerule languagerulename="default">
<rule break="no">
<beforebreak>([A-Z]\.){2,}</beforebreak>
<afterbreak>\s</afterbreak>
</rule>
<rule break="yes">
<beforebreak>\.</beforebreak>
<afterbreak>\s</afterbreak>
</rule>
</languagerule>
</languagerules>
</body>
</srx>
```

La primera de les regles:

```
<rule break="no">
<beforebreak>([A-Z]\.){2,}</beforebreak>
<afterbreak>\s</afterbreak>
</rule>
```

especifica que una conjunt de dues o més lletres majúscules seguides d'un espai formen un punt on no s'ha de produir un tall de segments. En canvi, la segona regla:

```
<rule break="yes">
<beforebreak>\.</beforebreak>
<afterbreak>\s</afterbreak>
</rule>
```

especifica que un punt seguit d'un espai sí que formen un punt on s'ha de produir un tall de segment.

Amb el nostre conjunt de dues regles ara la segmentació de la nostra frase es faria de la següent manera:

El Sr.

Martínez vindrà amb l'A.V.E. de les 15.30 h.

Després es reunirà amb el Dr.

Pérez en el nostre despatx de l'Av.

Diagonal.

Per poder segmentar correctament el nostre text, hauríem d'incorporar algunes regles. Si afegim com a primeres regles:

```
<rule break="no">
<beforebreak>\b(Sr|Dr|Av)\.</beforebreak>
<afterbreak>\s</afterbreak>
</rule>
<rule break="yes">
<beforebreak>\bh\.</beforebreak>
<afterbreak>[A-Z]+</afterbreak>
</rule>
```

Ara sí que aconseguirem un text segmentat correctament. Cal tenir en compte que l'ordre de les regles és significatiu, ja que si una regla ha produït un tall de segment, una regla posterior no el desfarà. *Ratel*, una de les eines que es distribueix amb el paquet *Okapi* (http://okapi.opentag.com/), permet editar de manera fàcil regles de segmentació i provar-les sobre fragments de text. A la figura 9 podem veure una captura de pantalla d'aquesta eina.

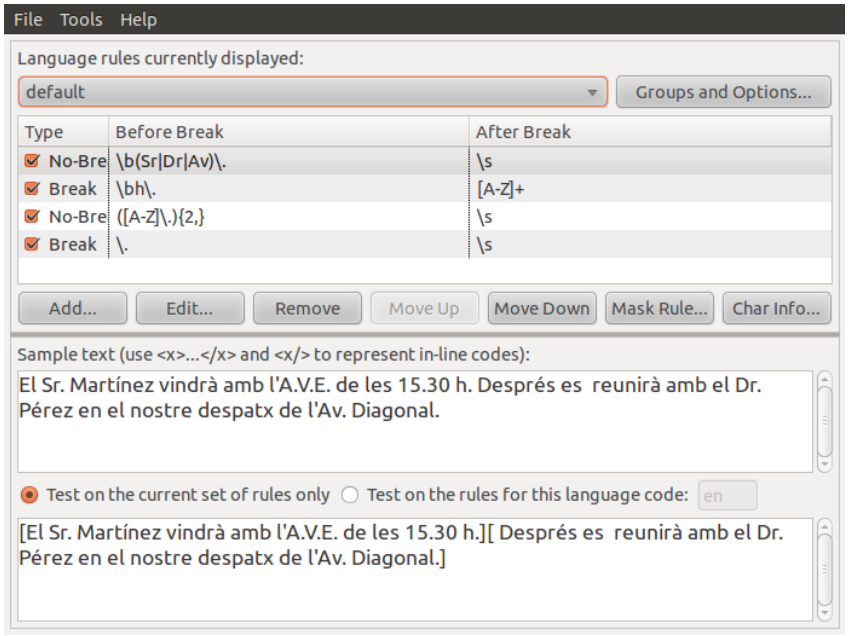


Figura 9. Ratel d'Okapi: eina per a l'edició de regles SRX

Afortunadament, ja existeixen conjunts de regles per diferents llengües i el traductor haurà d'intervenir molt poc sobre aquestes. Només en casos especials haurà d'afegir o treure alguna regla.

La importància del format SRX és que permet un intercanvi ràpid de les regles. Per un més gran aprofitament de les memòries de traducció, convé que les regles de segmentació que fem servir per crear un projecte siguin les mateixes que les regles que es van fer servir per crear el projecte que ha generat la memòria de traducció que volem fer servir. D'aquesta manera, la probabilitat de trobar segments coincidents augmenta significativament.

1.3.c. Formats de projectes de traducció: el format XLIFF

Cada eina de traducció assistida té un format propi per emmagatzemar els projectes de traducció. Algunes eines emmagatzemen els projectes com a bases de dades, on en una determinada taula guarden els segments a traduir i els segments traduïts, en altra taula la memòria de traducció ja indexada, etc. Altres formats inclouen també estructura de carpetes i subcarpetes, on en cada carpeta es guarda certa informació: en una els documents traduïts, en altra les memòries de traducció, en altra les bases de dades terminològiques, etc.

Aquesta multiplicitat de formats dificulten que un traductor amb una determinada eina pugui traduir projectes de traducció creats amb una altra eina. Hi ha un format estàndard per a l'intercanvi de projectes de traducció o localització: l'XLIFF (*XML Localisation Interchange File Format*). Aquest format també està basat en XML. A continuació podem veure un fragment d'un projecte de traducció en format XLIFF:


```
<?xml version="1.0" encoding="UTF-8"?>
<xliff version="1.2" xmlns= "urn: oasis: names: tc: xliff: document:1.2" xmlns:okp="ok
api-framework:xliff-extensions">
<body>
<trans-unit id="1" restype="x-text:p">
<source xml:lang="en">Only free resources available online have been used.</source>
<target xml:lang="es">Únicamente se han utilizado recursos libres disponibles en
Internet.</target>
</trans-unit>
</body>
</xliff>
```

Com podem observar, es tracta d'un projecte de traducció que només té un segment en anglès que ja està traduït al castellà. Si el projecte estigués sense traduir el segment corresponent a target estaria buit, o bé tindria copiat el segment original.

La majoria d'eines de traducció assistida poden traduir documents XLIFF, tot i que no totes les eines són capaces de crear aquests arxius. Es pot crear fàcilment projectes de traducció en format XLIFF amb l'eina *Rainbow* d'*Okapi Tools* (http://okapi.opentag.com/). El *Translate Toolkit* (http://translate.sourceforge.net) també proporciona eines per crear projectes de traducció en aquest format.

1.3.d. Assignació de recursos a un projecte de traducció

Un cop creat un projecte cal dotar-lo dels recursos de consulta disponibles, principalment memòries de traducció i bases de dades terminològiques. Traduir amb un sistema de traducció assistida sense recursos de poc serveix, tot i que no és del tot inútil tampoc. Normalment, quan comencem a fer servir sistemes de traducció assistida no disposem de recursos de consulta. En aquest cas el sistema només serà capaç d'ajudar-nos si dins del projecte hi ha alguna repetició, el que es coneix com a *repeticions internes*. A més, un cop acabat el projecte ja haurem generat la nostra primera memòria de traducció. Si a més aprofitem per anar recopilant la terminologia a mesura que l'anem trobant en el text, també anirem confeccionant la nostra primera base de dades terminològica que podrem fer servir en projectes futurs. A la figura 10 podem veure la pantalla que permet afegir nova terminologia a mesura que traduïm a l'eina OmegaT.

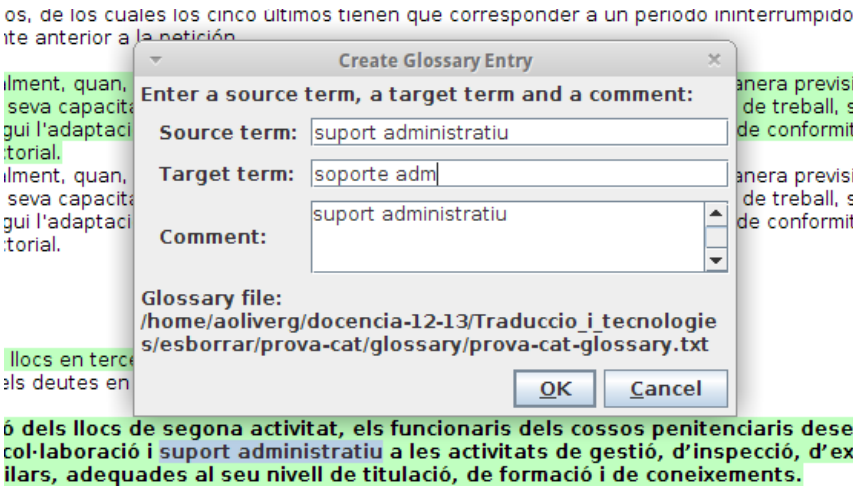


Figura 10. Pantalla d'entrada de terminologia a OmegaT

1.3.e. Traducció

Un cop creat el projecte i assignats els diferents recursos ja podem començar a traduir. Com hem comentat anteriorment, és un bon moment per ampliar les nostres bases de dades terminològiques. Les traduccions que anem fent s'emmagatzemaran automàticament en una memòria de traducció.

1.3.f. Revisió dins de l'eina de traducció assistida

Mentre anem traduint disposarem de diversos ajuts addicionals, com correctors ortogràfics i gramaticals. A la figura 11 podem veure el corrector ortogràfic d'OmegaT en acció:

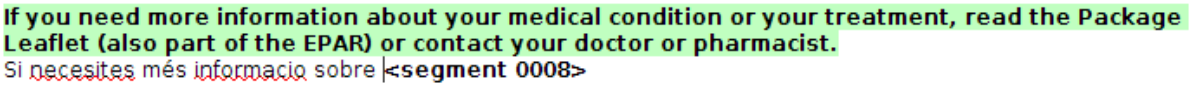

If you need more information about your medical condition or your treatment, read the Package Leaflet (also part of the EPAR) or contact your doctor or pharmacist.
Si necesites més informacio sobre <segment 0008>

Figura 11. Correcció ortogràfica a OmegaT

Un cop finalitzat el projecte hem podem fer diverses revisions:

- Una nova correcció ortogràfica i gramatical automàtica
- Una rellegida a fons de tot el projecte
- Verificacions automàtiques de consistència terminològica. Algunes eines permeten verificar si la terminologia emprada al projecte coincideix amb la de les bases de dades terminològiques assignades
- Verificació automàtica de consistència d'etiquetes, per assegurar que els documents finals apareixeran correctament en el seu format original.

L'avantatge de fer la revisió dins del sistema de traducció assistida és que tots els canvis que fem es veuran reflectits a les memòries de traducció generades al projecte. En canvi, és possible passar per alt algun aspecte relacionat amb el format final dels documents.

1.3.g. Creació dels documents traduïts

Un cop fetes les revisions dins de l'eina de traducció assistida ja podrem generar els documents finals. Aquests programes són capaços de generar documents traduïts que mantenen el format dels documents originals.

Un cas especial són els projectes que hem creat en format XLIFF amb alguna eina específica i els hem traduït amb alguna eina de traducció assistida. Per exemple, creem un projecte en format XLIFF amb Rainbow d'Okapi Tools i el traduïm amb OmegaT. En aquest cas, quan finalitzem el projecte amb OmegaT i creem el fitxer traduït obtindrem un XLIFF traduït que caldrà convertir en el format corresponent a l'original utilitzant de nou Rainbow.

1.3.h. Revisió en format final

És important fer una segona revisió amb els documents traduïts en el format final. A més de detectar alguna errada no detectada anteriorment, podrem veure si hi ha algun problema amb el format dels documents. L'inconvenient és que els canvis que fem no es veuran reflectits en les memòries de traducció generades en el projecte. Si l'errada és severa, pot ser una bona idea entrar de nou al projecte i fer també la correcció, per assegurar-nos que els canvis apareguin també a les memòries de traducció.

1.3.i. Gestió dels recursos generats

Durant la traducció s'hauran generat dos recursos importants: una memòria de traducció i una base de dades terminològica. Un cop finalitzar el projecte és important gestionar correctament aquests recursos. La idea bàsica és conservar-los en un lloc accessible i saber en tot moment quins recursos podem fer servir per a cada nou projecte. Depenent de l'eina emprada aquesta gestió dels recursos es pot fer d'una manera diferenciada. Una de les pràctiques associades a aquest llibre tracta en profunditat aquest tema.

1.4. Els sistemes de traducció assistida

Hi ha una gran quantitat d'eines de traducció assistida al mercat. Moltes d'elles, com per exemple OmegaT, són eines de programari lliure i es poden fer servir de manera lliure i gratuïta. Moltes altres, en canvi, són eines propietàries que requereixen adquirir algun tipus de llicència. Per obtenir informació actualitzada sobre les eines de traducció assistida existents recomano consultar dos enllaços:

- La plana de Wikibooks dedicada a CAT-Tools: http://en.wikibooks.org/wiki/CAT-Tools
- La plana de la Vikipèdia anglesa dedicada a la traducció assistida: http://en.wikipedia.org/wiki/Computer-assisted_translation

1.5. Conclusions

En aquest capítol hem presentat les principals funcionalitats dels sistemes de traducció assistida per ordinador i hem detallat les fases dels treballs de traducció quan es fan servir aquest tipus d'eines. En els propers capítols aprofundirem en els principals recursos associats a les eines de traducció assistida: les memòries de traducció i les bases de dades terminològiques.

En la secció *Per ampliar coneixements* d'aquest mateix capítol trobarem informació sobre eines de traducció assistida on-line i comentaris sobre els avantatges i inconvenients de l'ús d'aquest tipus d'eines. També hi ha una ampliació del tema de les expressions regulars, amb l'objectiu de dominar més l'ús del format SRX per a la creació de regles de segmentació. Per últim presentarem un corrector gramatical lliure, el LanguageTool, que es pot integrar en algunes eines de traducció assistida.

1.6. Per ampliar coneixements

1.6.1. Eines de traducció assistida on-line

Tradicionalment les eines de traducció assistida per ordinador han estat unes aplicacions informàtiques que s'instal·laven en l'ordinador del traductor i s'executaven des d'aquest mateix ordinador. En els darrers anys han aparegut algunes eines de traducció assistida que funcionen remotament en un servidor a les que accedim mitjançant un navegador web.

Aquestes eines ofereixen l'avantatge que no requereixen cap tipus d'instal·lació i que tant l'eina com els arxius de treball estan disponibles on-line des de qualsevol ordinador. Aquestes eines generalment permeten un treball col·laboratiu d'una manera molt senzilla, simplement compartint el projecte de traducció entre diverses persones. D'aquesta manera tots els traductors poden treballar sobre el mateix projecte, afegir comentaris, fer revisions de parts fetes per un altre, etc. L'únic inconvenient destacable és que per fer-les servir és imprescindible disposar d'una connexió a Internet i si per algun motiu la connexió falla no podem continuar treballant en el projecte.

Entre aquestes eines en podem destacar dues, que són d'ús gratuït:

- Google Translator Toolkit (http://translate.google.com/toolkit/)
- WordFast Anywhere (http://www.wordfast.net/?whichpage=anywhere)

L'ús d'aquestes eines és força senzill i no requereixen de cap instal·lació pel que són una bona solució en aquells casos que no disposem d'una eina instal·lada o en els casos que haguem de col·laborar amb persones que no disposen de cap eina.

Hi ha altres eines pensades per a agències de traducció que permeten fer una gestió integral del projecte via web: des de la creació del projecte, a l'assignació als traductors així com la traducció i revisió dels arxius per part dels col·laboradors. Entre aquestes eines en podem destacar dues:

- GlobalSight (http://www.globalsight.com/): es tracta d'una eina de codi obert.
- Memsource (http://www.memsource.com/): és una eina propietària de pagament que té una versió limitada d'ús gratuït per a traductors.

1.6.2. SRX i les expressions regulars

Les regles de segmentació en format SRX s'expressen mitjançant expressions regulars, cosa que permet una gran flexibilitat en la definició de les regles. A continuació podem observar un resum de la sintaxi de les expressions regulars emprades en les regles SRX (resum de la taula que s'ofereix a http://www.gala-global.org/oscarStandards/srx/srx10.html[2]):

Caràcter	**Descripció**
\A	Coincideix amb el principi de l'entrada. Difereix de `^` en el fet que `\A` no coincideix després d'una nova línia dins de l'entrada.
\b, outside of a [Set] \b, fora d'un [Conjunt]	Coincideix si la posició actual és el límit d'una paraula. Els límits tenen lloc en les transicions entre caràcters paraula (`\w`) i no-paraula (`\W`), ignorant les marques de combinació.
\b, within a [Set] \b, dins d'un [Conjunt]	Coincideix amb un RETROCÉS `\u0008`.
\B	Coincideix si la posició actual no és un límit d'una paraula.
\d	Coincideix amb qualsevol caràcter de la Categoria Nd (Número, Decimal, Dígit) d'Unicode.
\D	Coincideix amb qualsevol caràcter que no sigui un dígit decimal.
\e	Coincideix amb un ESCAPE, `\u001B`.
\f	Coincideix amb un FORM FEED, `\u000C`.
\G	Coincideix si la posició actual és el final de la coincidència anterior.
\n	Coincideix amb un LINE FEED, `\u000A`.
\r	Coincideix amb un CARRIAGE RETURN, `\u000D`.
\s	Coincideix amb un caràcter d'espai en blanc. L'espai en blanc es defineix com `[\t\n\f\r\p{Z}]`.
\S	Coincideix amb qualsevol caràcter que no sigui espai en blanc.
\t	Coincideix amb una TABULACIÓ HORITZONTAL, `\u0009`.
\uhhhh	Coincideix amb el caràcter amb el valor hexadecimal hhhh.
\Uhhhhhhhh	Coincideix amb el caràcter amb el valor hexadecimal hhhhhhhh. S'han de donar exactament 8 dígits hexadecimals, tot i que el punt de codi Unicode més llarg és `\U0010ffff`.
\w	Coincideix amb un caràcter paraula. Els caràcters paraula són `[\p{Ll}\p{Lu}\p{Lt}\p{Lo}\p{Nd}]`.
\W	Coincideix amb qualsevol caràcter no-paraula.
\x{hhhh}	Coincideix amb el caràcter amb valor hexadecimal hhhh
\xhh	Coincideix amb el caràcter amb valor de dos dígits hexadecimals hh
\Z	Coincideix si la posició actual està al final de l'entrada, però abans del darrer terminador de línia, si és que hi ha algun.
\z	Coincideix si la posició actual està al final de l'entrada.
\0nnn	Coincideix amb el caràcter amb valor octal nnn
\n	Referència. Coincideix amb l'n-enèsim grup coincident, n ha de ser >1 i < que el total de grups de l'expressió
[pattern]	Coincideix amb qualsevol dels caràcters del conjunt.

2 SRX 1.0 Specification. Copyright © The Localisation Industry Standards Association [LISA] 2004. All Rights Reserved.

.	Coincideix amb qualsevol caràcter.
^	Coincideix amb el principi d'una línia.
$	Coincideix amb el final d'una línia.
\	Cal posar la barra invertida davant de certs caràcters per referir-se a ells, ja que tenen un significat especial dins de les expressions regulars. Aquests caràcters són: `* ? + [ ( ) { } ^ $ | \ . /`

També es poden fer servir els següent operadors (és un resum de http://www.gala-global.org/oscarStandards/srx/srx10.html[3]) :

Operator	Description
`|`	Alternança. A|B coincideix amb A o B.
`*`	Coincideix 0 o més vegades. Coincideix tantes vegades com sigui possible.
`+`	Coincideix 1 o més vegades. Coincideix tantes vegades com sigui possible.
`?`	Coincideix cap o una vegada. Prefereix una vegada.
`{n}`	Coincideix exactament n vegades
`{n,}`	Coincideix com a mínim n vegades. Coincideix tantes vegades com sigui possible.
`{n,m}`	Coincideix entre n i m vegades. Coincideix tantes vegades com sigui possible, però no més d'm vegades.
`*?`	Coincideix 0 o més vegades. Coincideix les mínimes vegades possibles.
`+?`	Coincideix 1 o més vegades. Coincideix les mínimes vegades possibles.
`??`	Coincidex zero o una vegada. Prefereix zero.
`{n}?`	Coincideix exactament n vegades.
`{n,}?`	Coincideix com a minim n vegades, però no més vegades de les necessàries per coincidir amb tota l'expressió.
`{n,m}?`	Coincideix entre n i m vegades. Coincideix les mínimes vegades possibles, però no menys d'n.

3 SRX 1.0 Specification. Copyright © The Localisation Industry Standards Association [LISA] 2004. All Rights Reserved.

1.6.7. El corrector gramatical LanguageTool

Un corrector ortogràfic és capaç de detectar les paraules mal escrites en una llengua comparant-les amb un diccionari de paraules existents a la llengua (amb totes les seves formes flexionades).

Per exemple: qualsevol corrector ortogràfic del català és capaç de trobar els errors de la següent oració:

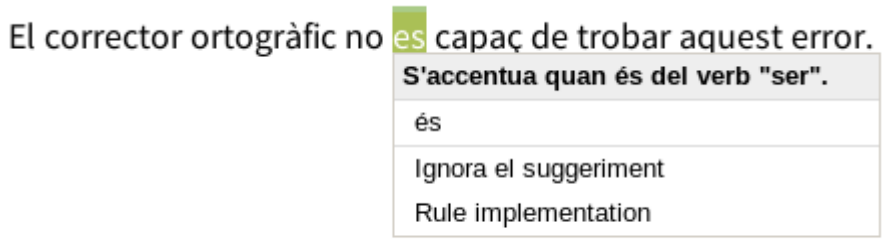
Cualsevol corector ortografic del catala és capac de trovar els errors d'aquesta oracio.

En canvi, un corrector simplement ortogràfic no trobarà el següent error:

El corrector ortogràfic no es capaç de trobar aquest error.

Ja que "es" hauria d'anar accentuat ja que és del ver "ser". Com que "es" és una paraula correcta en la llengua (però no en aquest context) el corrector ortogràfic no el marca com error.

Els correctors gramaticals van més enllà i permeten trobar alguns d'aquests error a més de construccions gramaticals incorrectes. Per exemple, veiem l'actuació del corrector LanguageTool en l'exemple anterior:



Veiem que aquest corrector ha pogut detectar que en aquest cas "es" hauria de portar accent ja que correspon a una forma del verb ser.

Veiem a continuació la regla de LanguageTool que permet fer aquesta detecció:

```
<rule>
  <pattern><!-- hi ha moltes possibles ambigüitats (es fera...), però són combinacions poc habituals
o del balear--><marker>
    <token>es<exception postag="DA0MN0|.*LOC_ADV.*|(&lt;)?NP.*"
postag_regexp="yes"/></token></marker>
   <token><exception postag="V.[^NPG].3.*|_possible_nompropi" postag_regexp="yes"/><exception
regexp="yes">['"'‘"«]|mig|ben</exception></token>
  </pattern>
  <message>S'accentua quan és del verb "ser".</message>
  <suggestion>és</suggestion>
  <short>Accent diacrític</short><!--
  <example type="incorrect" correction="és">Ell <marker>es</marker> Joan.</example> -->
  <example correction="és" type="incorrect">La bruixa <marker>es</marker> vella.</example>
  <example correction="és" type="incorrect">Que <marker>es</marker> penso això.</example>
  <example type="correct">pujar a la punta d'es Mut</example>
```

```
  <example type="correct">Que es pense això</example>
  <example type="correct">es "carrega"</example>
  <example type="correct">es ben mereixeria</example>
  <example type="correct">es mig situava</example>
  <example type="correct">És així es digui com es digui.</example>
</rule>
```

No entrarem en detall d'implementació però ens fixarem només en el fet que la regla defineix un patró (*pattern*) que en aquest cas té dues paraules (*tokens*), la pròpia *es* (amb unes excepcions donades per unes etiquetes) i una altra que respon a una etiqueta verbal (comença per V) o a una paraula d'una llista definida de possibles noms propis. Si es donen dues paraules seguides que acompleixin aquestes condicions el corrector detecta un error i mostra els missatges i les possibles correccions.
A continuació podem observar la frase de l'exemple etiquetada per l'etiquetador de català integrat a LanguageTool:

```
<S> El[el/DA0MS0,El/_GN_MS,El/_GN_MS,]
corrector[corrector/NCMS000,corrector/_GN_MS,corrector/_GN_MS,corrector/_GN_MS,]
ortogràfic[ortogràfic/AQ0MS0,ortogràfic/_GN_MS,ortogràfic/_GN_MS,ortogràfic/ignore_concordance,]
no[no/RN,] es[es/P0300000,] capaç[capaç/AQ0CS0,] de[de/SPS00,]
trobar[trobar/VMN00000,trobar/complement,] aquest[aquest/DD0MS0,aquest/_GN_MS,]
error[error/_GN_MS,error/NCMS000,].[</S>./_PUNCT,]
```

El corrector gramatical és capaç d'etiquetar el text que analitza, és a dir, donar-li a les paraules les etiquetes que expressen les seves categories gramaticals i algunes subcategoritzacions, tot i que no és capaç de desambiguar i assignar totes les etiquetes possibles.

LanguageTool (https://www.languagetool.org/) és un corrector gramatical de codi obert que està disponible per més de 20 llengües, entre elles el català, castellà i anglès. Els usuaris avançats poden modificar les regles o crear-ne de noves de manera que el corrector va millorant amb el temps.

LanguageTool pot funcionar com a aplicació independent i a més s'integra perfectament a LibreOffice/OpenOffice i també es pot instal·lar com a extensió de Firefox. També és possible instal·lar LanguageTool a l'eina de traducció assistida OmegaT.

Bibliografia

Lippmann E. O. (1971) *An approach to Computer-Aided Translation.* IEE Transactions on Engineering Writing and Speech, Vol. EWS-14, No. 1, February 1971

Llicència d'aquest document